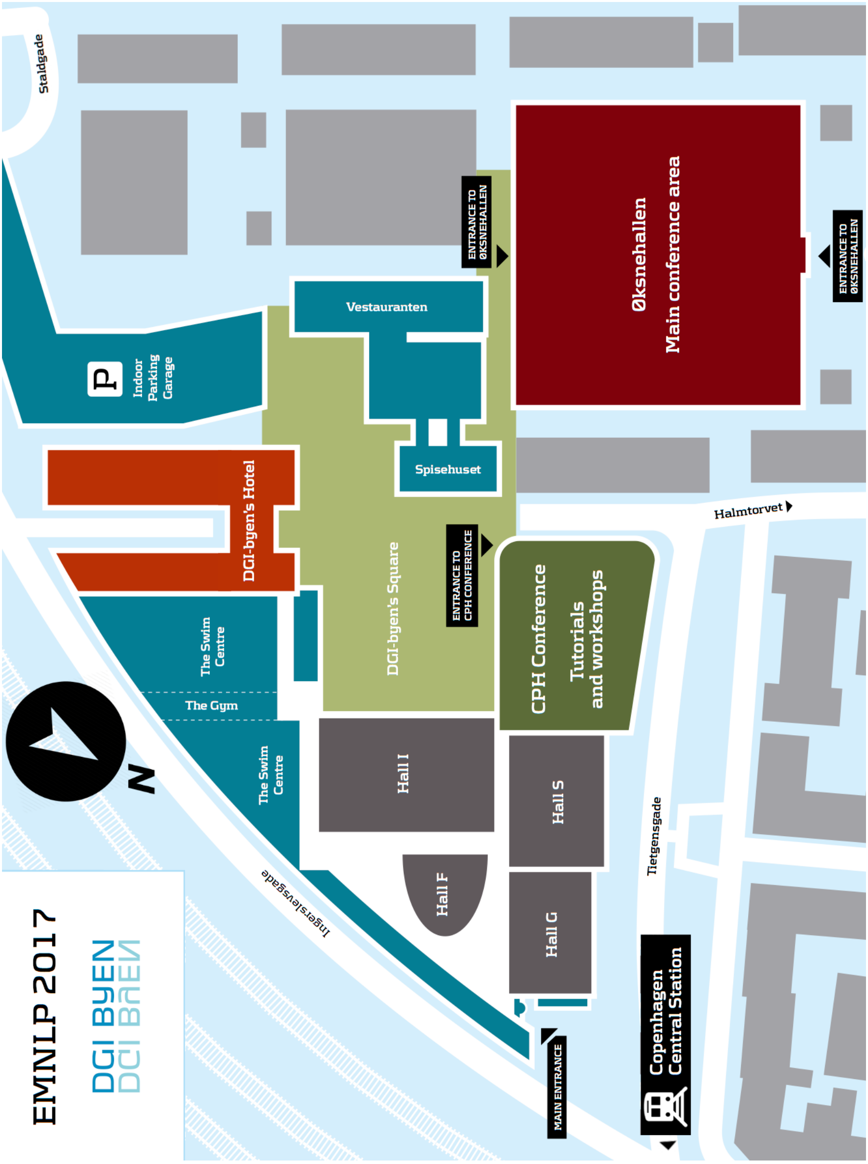
EMNLP 2018

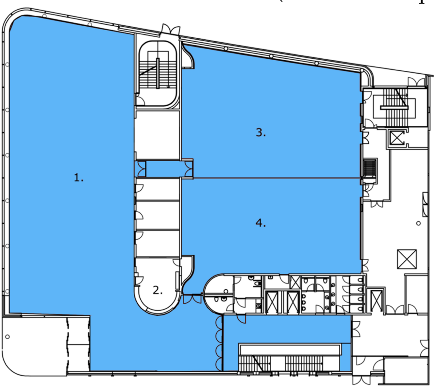
Conference Handbook

Brussels, Belgium
Oct. 31-Nov. 4

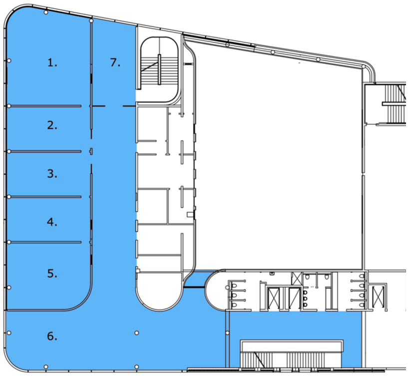
Cover design by
Cover photo by Matthew James
Deep gratitude to Joachim Bingel for invaluable advice with creating this handbook
Handbook assembled by Kai-Wei Chang and proofread by He He
Printing by Frederiksbergs Bogtrykkeri A/S


EMNLP 2017
DGI BYEN
Staldgade
P
Indoor Parking Garage
DGI-byen's Hotel
The Swim Centre
The Gym
The Swim Centre
Vestauranten
Spisehuset
DGI-byen's Square
ENTRANCE TO ØKSNEHALLEN
Øksnehallen
Main conference area
ENTRANCE TO ØKSNEHALLEN
ENTRANCE TO CPH CONFERENCE
CPH Conference
Tutorials and workshops
Halmtorvet
Hall I
Hall S
Hall F
Hall G
N
Ingerslevsgade
Tietgensgade
MAIN ENTRANCE
Copenhagen Central Station

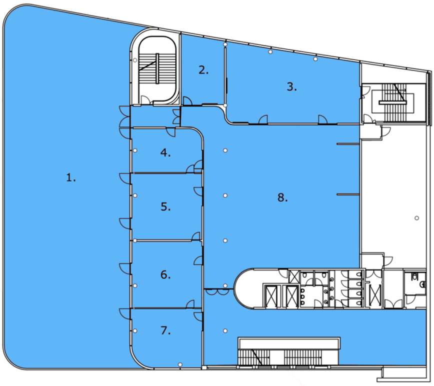
CPH Conference Floorplan
(Most workshops, three tutorials)



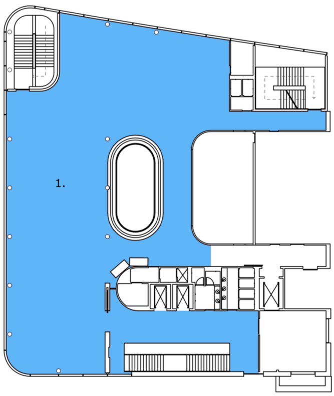
CPH CONFERENCE: GROUND FLOOR
1. Lobby
2. Reception
3. Sankt Hans Torv
4. Nørrebros Runddel



CPH CONFERENCE: 1st FLOOR
1. Kastrup Lufthavn
2. Christianshavn
3. Islands Brygge
4. Christiania
5. Amager Strandpark
6. Break/ Lounge area
7. Break/ Lounge area

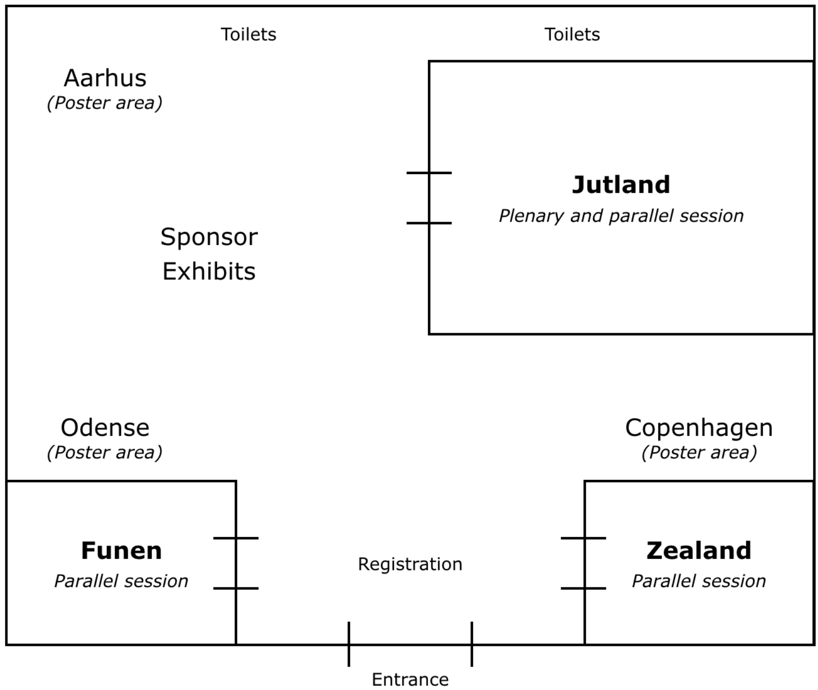


CPH CONFERENCE: 2nd FLOOR
1. Roof terrace
2. Istedgade
3. Hovedbanen
4. Tivoli
5. Vesterbro Torv
6. Enghave Plads
7. Kødbyen
8. Break / Lounge area



CPH CONFERENCE: 3rd FLOOR
1. Østerbro

Øksnehallen Floorplan
(Main Conference, WMT, Tutorials 5 and 6)



Contents

1

Conference Information

Message from the General Chair

Thank you so much for joining us in Copenhagen. Welcome to a cosmopolitan city of fantastic restaurants, lovely seascapes, rich history, and lots and lots of cyclists!

We have an exciting program lined up for you, with three invited talks, fifteen workshops, seven tutorials, nine TACL presentations, 323 reviewed papers presented as both oral talks and posters, and twenty-one demos. I am especially grateful to our Program Chairs, Rebecca Hwa and Sebastian Riedel, who did a fantastic job managing a backbreaking 1,500 paper submissions (1466 reviewed papers). This involved 51 Area chairs and 980 reviewers. We tried some new things this year (never conducive to a smooth process) including a more careful handling of the COIs that result from Area Chair submissions, and the addition of a meta-review step to encourage more thoughtful reviewing. We are soliciting feedback on the meta-review process, from both reviewers and authors. Despite the additional time involvement, many of the Area Chairs embraced this new approach, and would like to repeat it. However, there are clearly a few dissenters, since Rebecca and Sebastian ended up writing around 200 meta-reviews themselves at the last minute. We are also trying to raise the visibility and status of the poster sessions by integrating them as parallel sessions alongside oral talks, with poster session chairs. This is in response to the survey results from EMNLP 2015 that indicated a decided preference for smaller, more frequent poster sessions during the day rather than evening mega-sessions. Finally, Rebecca and Sebastian are bringing you three outstanding invited speakers, Dan Jurafsky, Sharon Goldwater, and Nando de Freitas. No program chairs ever worked harder to bring you a superb set of presentations in an attendee friendly setting.

I am also very grateful to Victoria Fossum and Karl Moritz Hermann, our Workshop Chairs, who put together a terrific slate of fifteen workshops, and paid meticulous attention to ensuring that each workshop could hold exactly the poster sessions, invited talks and special events that it required. Our tutorial chairs, Alexandra Birch and Nathan Schneider, also outdid themselves, providing seven especially tempting tutorial offerings. Matt Post deserves to be singled out, for being an Advisor to our conscientious and successful Handbook Chair, Joachim Bingel, as well as becoming a welcome last minute addition to our excellent team of Demo Chairs, Lucia Specia and Michael Paul. Thanks are due to our Website Chair, Anders Johannsen, who responded promptly and deftly to all of our requests, and to Chloé Braud, who single-handedly took on the task of creating the conference app in Conference4me. Our Student Volunteer and Student Sponsorship Chairs, Zeljko Agic and Yonatan Bisk, worked hard to bring you the helpful and energetic volunteers who keep things running smoothly, for which we are also grateful.

Last but not least, many thanks to your hosts, our Local Arrangements Chairs, Dirk Hovy and Anders Søgaard and their team. Their concern has been increasing the enjoyment of your experience, and to that end they proposed a stunning venue, put together an amazing reception and Social Event, chose your conference bags, issued all the invitation letters for visas, helped create all the signs, etc., etc., etc. Dan Hardt, our Sponsorship chair, working with Anders and Dirk, raised an unusual amount of local sponsorships, all to defray the cost of the Social Event.

As always, we are extremely indebted to our generous sponsors. Our platinum sponsors are Amazon, Apple, Baidu, Bloomberg, Facebook, Google, and Siteimprove. Gold sponsors include Deloitte, ebay, IBM Research, Maluuba, Microsoft, SAP, Recruit Institute of Technology, textkernel, and Zalando. Silver sponsors are CVTE, Duolingo, Huawei, Nuance, Oracle, Snapchat, Sogou, Unsilo and Wizkids. Grammarly, NextAI and Yandex are our Bronze sponsors.

Finally, many, many thanks to our Area Chairs, our reviewers, and our authors, whose outstanding research is being showcased here for your delectation. *Nyd det mens det varer!*

Best Regards,
Martha Palmer
EMNLP 2017 General Chair

Message from the Program Committee Co-Chairs

Welcome to the 2017 Conference on Empirical Methods in Natural Language Processing! This is an exciting year; we have received a new record-high in the number of submissions: 1,509 papers. After discounting early withdraws, duplicates, and other invalid submissions, we sent out 1,418 submissions (836 long papers, 582 short papers) to be reviewed by the program committee. Ultimately, 216 long papers (25.8% acceptance rate) and 107 short papers (18.4% acceptance rate) have been accepted for presentation, making a total of 323 papers and an overall acceptance rate of 22.8%.

This year's technical program consists of three invited talks and 113 oral presentations and 219 poster presentations for the 323 long and short accepted papers as well as nine papers accepted to the Transactions of the Association for Computational Linguistics. To accommodate all the presentations in a compressed timeframe, we opted to have plenary sessions for the invited talks and the winners of the Best Paper Awards, while allotting three parallel oral sessions and thematically related poster sessions for all other presentations. We chose to have concurrent poster and oral sessions for several reasons. First, this is the preferred model of the majority (51.6%) of participants who filled out the EMNLP 2015 post-conference survey. Second, this allows us to spread out the poster presentations across three days in smaller thematically related clusters. Finally, this maximises the number of acceptances for the high quality submissions we received; by having more poster sessions, we are able to maintain the acceptance rates at the previous year's level despite an increase in submissions by 40%.

It would not have been possible to properly handle such a large number of submissions without the generous voluntary help from all the members of the program committee, which consists of 980 reviewers overseen by 51 area chairs. We continued last year's experiment of defining twelve relatively broad topic areas and assigning multiple area chairs to facilitate consistent ranking of larger sets of papers. Most technical program decisions, from the selection of papers to the modes of presentation to the choice of outstanding papers, are primarily made in a bottom-up fashion: reviewers assessed and scored papers, made recommendations for oral vs poster decisions, and marked papers suitable for best paper awards; area chairs ensured the quality of assessments, encouraged discussions and assembled opinions into their own recommendations; finally, we construct the technical program, considering the recommendations from the area chairs while taking into account venue constraints and balance across areas. A new experimental feature of this year's EMNLP reviewing process is the "meta review," in which the area chairs briefly summarize the major discussions between the reviewers to give authors a more transparent view of the process.

Per EMNLP tradition, awards are given to outstanding papers in three categories: Best Long Paper, Best Short Paper, and Best Resource Paper. The selection process is bottom-up: based on the reviewers and area chairs' recommendations, we nominated four papers for each category; we invited expert members to form a Best Papers committee for each category; each committee reviews the candidates and select the winners. The awarded papers will be presented at a special plenary session on the last day of the conference.

We are extremely grateful that three amazing speakers have agreed to give invited talks at EMNLP. Nando de Freitas (Google Deepmind) will discuss simulated physical environments, and whether language would benefit from the development of such environments, and could contribute toward improving such environments and agents within them. Sharon Goldwater (University of Edinburgh) will describe work on developing unsupervised speech technology for those of the world's 7,000 or so languages not spoken in large rich countries. Dan Jurafsky (Stanford University) will talk about processing the language of policing to automatically measure linguistic aspects of the interaction from discourse factors like conversational structure to social factors like respect.

The conference would not have been possible without the support of various people inside and outside of the committee. In particular, we would like to thank:

- Martha Palmer, whose encouragement and advice as the general chair has been invaluable every step of the way;
- Chris Callison-Burch, who has given us excellent advice and support in his capacity as the SIGDAT Secretary;
- Priscilla Rasmussen, who always has the right answers;
- Xavier Carreras and Kevin Duh, who generously shared their experiences as the chairs of EMNLP 2016;
- Anders Johannsen, who is lightning fast with website updates;
- Our 51 area chairs: David Bamman, Mohit Bansal, Roberto Basili, Chris Biemann, Jordan Boyd-Graber, Marine Carpuat, Joyce Chai, David Chiang, Jinho Choi, Jennifer Chu-Carroll, Trevor Cohn, Cristian Danescu-Niculescu-Mizil, Dipanjan Das, Hal Daume, Mona Diab, Mark Dredze, Jacob Eisenstein, Sanja Fidler, Alona Fyshe, Dan Gildea, Ed Grefenstette, Hannaneh Hajishirzi, Julia Hockenmaier, Kentaro Inui, Jing Jiang, Philipp Koehn, Mamoru Komachi, Anna Korhonen, Tom Kwiatkowski, Gina Levow, Bing Liu, Nitin Madnani, Mausam, Rada Mihalcea, Marie-Francine Moens, Saif M. Mohammad, Mari Ostendorf, Sameer Pradhan, Alexander Rush, Anoop Sarkar, William Schuler, Hinrich Schütze, Sameer Singh, Thamar Solorio, Vivek Srikumar, Amanda Stent, Tomek Strzalkowski, Mihai Surdeanu, Andreas Vlachos, Scott Wentau Yih, Zhang Yue;
- The best papers award committee members: Chris Brew, Mike Collins, Kevin Duh, Adam Lopez, Ani Nenkova, Bonnie Webber, Luke Zettlemoyer;
- Preethi Raghavan and Siddharth Patwardhan, the publications co-chairs and Joachim Bingel, the conference handbook chair;
- Dirk Hovy and Anders Søgaard, the local arrangements co-chairs;
- Rich Gerber and Paolo Gai at SoftConf.

Finally, we'd like to thank SIGDAT for the opportunity to serve as Program Co-Chairs of EMNLP 2017. It is an honor and a rewarding learning experience. We hope you will be as inspired by the technical program as we are.

EMNLP 2017 Program Co-Chairs
Rebecca Hwa, University of Pittsburgh
Sebastian Riedel, University College London

Organizing Committee

General Chair
Ellen Riloff, University of Utah, USA

Program Committee Co-chairs
David Chiang, University of Notre Dame, USA
Julia Hockenmaier, University of Illinois Urbana-Champaign, USA
Junichi Tsujii, Artificial Intelligence Research Center, Japan

Local Arrangements Committee
Marie-Francine Moens, KU Leuven, Belgium (chair)
Dominique De Brabanter, KU Leuven, Belgium
Frieda Steurs, KU Leuven, Belgium
Hugo Van hamme, KU Leuven, Belgium

Local Sponsorship Chair
Katrien Beuls, Vrije Universiteit Brussel, Belgium

Workshop Co-chairs
Marieke van Erp, KNAW Humanities Cluster, Netherlands
Vincent Ng, University of Texas at Dallas, USA

Tutorial Co-chairs
Mausam, Indian Institute of Technology Delhi, India
Lu Wang, Northeastern University, USA

Demos Co-chairs
Eduardo Blanco, University of North Texas, USA
Wei Lu, Singapore University of Technology and Design, Singapore

Publications Sr Chair
Preethi Raghavan, IBM, USA

Publications Jr Chair
Micha Elsner, Ohio State University, USA

Publicity Chair
Mohit Iyyer, University of Massachusetts at Amherst, USA

Video Chair
Bonan Min, BBN Technologies, USA

Website and Conference App Chair
Nitin Madnani, Educational Testing Service, USA

Conference Handbook Chair
Kai-Wei Chang, University of California at Los Angeles, USA

Conference Handbook Advisor
Joachim Bingel, University of Copenhagen, Denmark

Handbook Proofreader
He He, Stanford University, USA

Student Scholarship Co-chair and Student Volunteer Coordinator
Anna Rumshisky, University of Massachusetts at Lowell, USA Hugo Van hamme, KU Leuven, Belgium

Childcare Policy and Grant Coordinator
Swapna Somasundaran, Educational Testing Service, USA
Aoife Cahill, Educational Testing Service, USA

Program Committee

Program Committee Co-chairs
David Chiang, University of Notre Dame, USA
Julia Hockenmaier, University of Illinois Urbana-Champaign, USA
Junichi Tsujii, Artificial Intelligence Research Center, Japan

Area Chairs
Information Extraction, Information Retrieval, and Question Answering
Jordan Boyd-Graber, University of Maryland, USA (senior chair)
Isabelle Augenstein, University of Copenhagen, Denmark
Ming-Wei Chang, Google, USA
Doug Downey, Northwestern University, USA
Ruihong Huang, Texas A&M University, USA
Mausam, IIT Delhi, India
Makoto Miwa, Toyota Technological Institute, Japan
William Wang, University of California at Santa Barbara, USA
Scott Yih, Allen Institute for Artificial Intelligence, USA

Text Mining and Information Retrieval
Alessandro Moschitti, University of Trento, Italy (senior chair)
Sophia Ananiadou, University of Manchester, United Kingdom
Hsin-Hsi Chen, National Taiwan University
Marius Pasca, Google, USA
Xiang Ren, University of Southern California, USA
Alan Ritter, Ohio State University, USA
David Smith, Northeastern University, USA

Social Media, Computational Social Science, and Sentiment/Opinion Analysis
Kam-Fai Wong, Chinese University of Hong Kong (senior chair)
Eiji Aramaki, Nara Institute of Science and Technology, Japan
Mona Diab, George Washington University, USA
Yulan He, Aston University, United Kingdom
Dirk Hovy, Bocconi University, Italy
Rada Mihalcea, University of Michigan, USA
Alice Oh, KAIST, Korea
Wei Xu, Ohio State University, USA

Morphology, Syntax, and Psycholinguistics
Slav Petrov, Google, USA (senior chair)
Liang Huang, Oregon State University, USA
Roger Levy, Massachusetts Institute of Technology, USA
Stephan Oepen, University of Oslo, Norway
Emily Pitler, Google, USA
Reut Tsarfaty, Open University of Isra

Semantics
Massimo Poesio, Queen Mary University of London, United Kingdom (senior chair)
Omri Abend, Hebrew University of Jerusalem, Israel
Raffaella Bernardi, University of Trento, Italy
Michael Lewis, Facebook, USA
Yulia Tsvetkov, Carnegie Mellon University, USA
Benjamin Van Durme, Johns Hopkins University, USA
Nianwen Xue, Brandeis University, USA

Discourse, Dialogue, Summarization, Generation, and Multimodal NLP Yejin Choi, University of Washington, USA (senior chair)
Mohit Bansal, University of North Carolina, USA
Michel Galley, Microsoft, USA
Grzegorz Chrupała, Tilburg University, Netherlands

Haizhou Li, National University of Singapore
Fei Liu, University of Central Florida, USA
Karen Livescu, Toyota Technological Institute - Chicago, USA
Meg Mitchell, Google, USA
Rashmi Prasad, Interactions, USA
Xiaojun Wan, Peking University, China
Zhou Yu, University of California at Davis, USA

Machine Translation and Multilinguality
Philipp Koehn, Johns Hopkins University, USA (senior chair)
Arianna Bisazza, Leiden University, Netherlands
Qun Liu, Dublin City University, Ireland
Yang Liu, Tsinghua University, China
Zhaopeng Tu, Tencent, China
Taro Watanabe, Google, Japan

Machine Learning
Xavier Carreras, dMetrics, USA (senior chair)
Shay Cohen, University of Edinburgh, United Kingdom
Kevin Gimpel, Toyota Technological Institute - Chicago, USA
Stefan Riezler, Heidelberg University, Germany
Karl Stratos, Toyota Technological Institute - Chicago, USA
Jun Suzuki, Tohoku University, Japan

Venue Info

Please note that EMNLP takes place in *four different buildings*. Almost all workshops are located in **CPH Conference**, while the WMT workshop, Tutorials 5 and 6 as well as the entire main conference take place in **Øksnehallen**.

Besides these, we will have Tutorials 2 and 4 located in **Spisehuset**, and the BEA workshop will take place in the Conference Hotel (**DGI Byen Hotel, Room 3**).

There are maps with room locations, as well as an area map, in the front matter.

Meal Info

The following meals are provided as part of your registration fee:

- During the coffee breaks in the mornings and afternoons, **snacks** are provided.
- There will also be **snacks** at the welcome reception.
- A **full dinner** with various options is provided as part of the social event.

Welcome Reception

Friday, September 8, 2017, 19:00 – 22:00

TBD

Catch up with your colleagues at the **Welcome Reception**! It will be held following the tutorials and workshops on Friday evening.

2

Tutorials: Wednesday, October 31

Overview

7:30 – 18:00	**Registration**	*Lobby*
9:00 – 12:30	**Morning Tutorials**	
	Joint models for NLP *Yue Zhang*	*Skt. Hans Torv*
	Graph Formalisms for Meaning Representations *Adam Lopez and Sorcha Gilroy*	*Spisehuset*
10:30 – 11:00	**Coffee break**	*Multiple levels*
12:30 – 14:00	**Lunch break**	
14:00 – 17:30	**Afternoon Tutorials**	
	Writing Code for NLP Research *Matt Gardner, Mark Neumann, Grus, Joel, and Nicholas Lourie*	*Skt. Hans Torv*
15:30 – 16:00	**Coffee break**	*Multiple levels*

Tutorial 1

Joint models for NLP

Yue Zhang

Wednesday, October 31, 2018 1, 9:00–12:30

Location: Skt. Hans Torv

Joint models have received much research attention in NLP, allowing relevant tasks to share common information while avoiding error propagation in multi-stage pepelines. Several main approaches have been taken by statistical joint modeling, while neural models allow parameter sharing and adversarial training. This tutorial reviews main approaches to joint modeling for both statistical and neural methods.

Yue Zhang is currently an assistant professor at Singapore University of Technology and Design. His research interests include machine learning methods for structured prediction, syntactic parsing, text generation and information extraction. Yue Zhang has served as area co-chairs of ACL 2017, 2018, EMNLP 2015, 2017, NAAACL 2015 and COLING 2014, 2018. He has given tutorials at NAACL 2010, ACL 2014 and EMNLP 2016.

Tutorial 2

Graph Formalisms for Meaning Representations

Adam Lopez and Sorcha Gilroy

Wednesday, October 31, 2018 1, 9:00–12:30

Location: Spisehuset

In this tutorial we will focus on Hyperedge Replacement Languages (HRL; Drewes et al. 1997), a context-free graph rewriting system. HRL are one of the most popular graph formalisms to be studied in NLP (Chiang et al., 2013; Peng et al., 2015; Bauer and Rambow, 2016). We will discuss HRL by formally defining them, studying several examples, discussing their properties, and providing exercises for the tutorial. While HRL have been used in NLP in the past, there is some speculation that they are more expressive than is necessary for graphs representing natural language (Drewes, 2017). Part of our own research has been exploring what restrictions of HRL could yield languages that are more useful for NLP and also those that have desirable properties for NLP models, such as being closed under intersection.

With that in mind, we also plan to discuss Regular Graph Languages (RGL; Courcelle 1991), a subfamily of HRL which are closed under intersection. The definition of RGL is relatively simple after being introduced to HRL. We do not plan on discussing any proofs of why RGL are also a subfamily of MSOL, as described in Gilroy et al. (2017b). We will briefly mention the other formalisms shown in Figure 1 such as MSOL and DAGAL but this will focus on their properties rather than any formal definitions.

Adam Lopez develops computational models of natural language learning, understanding and generation in people and machines, and his research focuses on basic scientific, mathematical, and engineering problems related to these models. He is especially interested in models that handle the diversity of morphological, syntactic, and semantic phenomena across the world?s languages. He teaches undergraduate and master?s-level courses in natural language processing at the School of Informatics at the University of Edinburgh, and previously in the Department of Computer Science at the Johns Hopkins University. These courses have ranged in size from 10 to 200 students from a variety of backgrounds, including computer science, mathematics, and linguistics. He previously taught a popular course in machine translation at ESSLLI in 2010 and NASSLLI in 2012. He has given many tutorial lectures at outreach events for high school students, and once at a poetry event for a local arts festival.

Sorcha Gilroy is currently a third year PhD student in the Edinburgh NLP group. Her research is focused on probabilistic models of graphs. She has worked closely with Hyperedge Replacement Grammars, Monadic Second-Order Logic and Directed Acyclic Graph Automata, which will all be at least touched on in this course. She has given several talks on this topic in the past at the Universities of Edinburgh, Amsterdam, Melbourne, the Information Sciences Institute, and the IT University of Copenhagen. She has been a teaching assistant in courses on formal language theory, machine translation, and machine learning at the University of Edinburgh; her responsibilities included leading tutorial sessions for students. She once gave a public talk on machine translation (using magnets) in front of the Scottish National Gallery as part of the UK?s Soapbox Science festival.

Tutorial 3

Writing Code for NLP Research

Matt Gardner, Mark Neumann, Grus, Joel, and Nicholas Lourie

Wednesday, October 31, 2018 1, 14:00–17:30

Location: Skt. Hans Torv

Doing modern NLP research requires writing code. Good code enables fast prototyping, easy debugging, controlled experiments, and accessible visualizations that help researchers understand what a model is doing. Bad code leads to research that is at best hard to reproduce and extend, and at worst simply incorrect. Indeed, there is a growing recognition of the importance of having good tools to assist good research in our field, as the upcoming workshop on open source software for NLP demonstrates. This tutorial aims to share best practices for writing code for NLP research, drawing on the instructors' experience designing the recently-released AllenNLP toolkit, a PyTorch-based library for deep learning NLP research. We will explain how a library with the right abstractions and components enables better code and better science, using models implemented in AllenNLP as examples. Participants will learn how to write research code in a way that facilitates good science and easy experimentation, regardless of what framework they use.

Matt Gardner is a research scientist at the Allen Institute for Artificial Intelligence. His research focuses on question answering and semantic parsing. He is the lead designer of the AllenNLP toolkit, and a host of the NLP Highlights podcast.

Joel Grus is a research engineer at the Allen Institute for Artificial Intelligence, where he works on problems at the intersection of engineering and machine learning. He is the author of Data Science from Scratch: First Principles with Python, the viral ?Fizz Buzz in Tensorflow? blog post, and the ?Livecoding Madness: Let?s Build a Deep Learning Library? video.

Mark Neumann is a research engineer at the Allen Institute for Artificial Intelligence, where he supports research on deep representation learning and natural language processing. He is a core developer of the AllenNLP toolkit.

Nicholas Lourie is a research engineer at the Allen Institute for Artificial Intelligence. Previously, he worked on large scale processing and information extraction from scientific documents. Currently, he works on the AllenNLP toolkit and NLP research.

3

Tutorials: Thursday, November 1

Overview

7:30 – 18:00	**Registration**	*Lobby*
14:00 – 17:30	**Afternoon Tutorials**	
	Deep Latent Variable Models of Natural Language *Alexander Rush, Yoon Kim, and Sam Wiseman*	*Zealand*
	Standardized Tests as benchmarks for Artificial Intelligence *Mrinmaya Sachan, Minjoon Seo, Hannaneh Hajishirzi, and Eric Xing*	*Jutland*
	Deep Chit-Chat: Deep Learning for ChatBots *Wei Wu and Rui Yan*	*Nørrebros Runddel*
15:30 – 16:00	**Coffee break**	*Multiple levels*
19:00 – 22:00	**Welcome Reception**	*TBD*

Tutorial 4

Deep Latent Variable Models of Natural Language

Alexander Rush, Yoon Kim, and Sam Wiseman

Thursday, November 1, 2018 2, 9:00–12:30

Location: Spisehuset

The proposed tutorial will cover deep latent variable models both in the case where exact inference over the latent variables is tractable and when it is not. The former case includes neural extensions of unsupervised tagging and parsing models. Our discussion of the latter case, where inference cannot be performed tractably, will restrict itself to continuous latent variables. In particular, we will discuss recent developments both in neural variational inference (e.g., relating to Variational Auto-encoders) and in implicit density modeling (e.g., relating to Generative Adversarial Networks). We will highlight the challenges of applying these families of methods to NLP problems, and discuss recent successes and best practices.

Alexander ?Sasha? Rush is an assistant professor at Harvard University. His research interest is in ML methods for NLP with recent focus on deep learning for text generation including applications in machine translation, data and document summarization, and diagram-to-text generation, as well as the development of the OpenNMT translation system. His past work focused on structured prediction and combinatorial optimization for NLP.

Yoon Kim is a PhD student at Harvard University. He is interested in deep learning approaches to natural language processing, and especially in deep models for text generation. He has done previous work in text classification, language modeling, knowledge distillation, and generation.

Sam Wiseman is a PhD student at Harvard University. He is interested in the intersection of deep learning, structured prediction, and natural language processing. He has done previous work in coreference resolution, structured prediction for NLP, and text generation.

Tutorial 5

Standardized Tests as benchmarks for Artificial Intelligence

Mrinmaya Sachan, Minjoon Seo, Hannaneh Hajishirzi, and Eric Xing

Thursday, November 1, 2018 2, 9:00–12:30

Location: Zealand

Standardized tests have recently been proposed as replacements to the Turing test as a driver for progress in AI (Clark, 2015). These include tests on understanding passages and stories and answering questions about them (Richardson et al., 2013; Rajpurkar et al., 2016a, inter alia), science question answering (Schoenick et al., 2016, inter alia), algebra word problems (Kushman et al., 2014, inter alia), geometry problems (Seo et al., 2015; Sachan et al., 2016), visual question answering (Antol et al., 2015), etc. Many of these tests require sophisticated understanding of the world, aiming to push the boundaries of AI.

For this tutorial, we broadly categorize these tests into two categories: open domain tests such as reading comprehensions and elementary school tests where the goal is to find the support for an answer from the student curriculum, and closed domain tests such as intermediate level math and science tests (algebra, geometry, Newtonian physics problems, etc.). Unlike open domain tests, closed domain tests require the system to have significant domain knowledge and reasoning capabilities. For example, geometry questions typically involve a number of geometry primitives (lines, quadrilaterals, circles, etc) and require students to use axioms and theorems of geometry (Pythagoras theorem, alternating angles, etc) to solve them. These closed domains often have a formal logical basis and the question can be mapped to a formal language by semantic parsing. The formal question representation can then provided as an input to an expert system to solve the question.

Mrinmaya Sachan is a final-year PhD student at Carnegie Mellon University advised by Prof. Eric Xing. His work focuses on building automated solvers for standardized tests using instructional material. He also loves problems in NLP, Structured Prediction and Statistical Machine Learning. His work on machine comprehension was one of the outstanding papers at ACL 2015. He has been awarded a number of fellowships namely the Siebel Scholarship (2013-14), IBM Fellowship (2016-17) and CMLH Fellowship (2017-18). He was also a finalist for the Facebook Fellowship in 2014-15. He regularly publishes in top NLP conferences.

Minjoon Seo is a research scientist at NAVER Clova and a Ph.D. candidate in the Allen School of Computer Science & Engineering at the University of Washington. He is advised by Hannaneh Hajishirzi, Ali Farhadi and Oren Etzioni. He is interested in question-driven learning models for extracting, accessing, and combining knowledge from natural language. He has served as a program committee member in ACL, EMNLP, ICLR, and AAAI. He recently won AI2 Key Scientific Challenges Award (2018). Previously, he received B.S. at the University of California, Berkeley.

Hannaneh Hajishirzi is an assistant research professor in the Department of Electrical Engineering and an adjunct assistant professor in the Allen School of Computer Science & Engineering at the University of Washington. Her research interests are in natural language processing, machine learning, and artificial intelligence. Her research is currently focused designing algorithms for semantic understanding, question answering, and information extraction about different types of textual and visual data such as web data, news articles, scientific articles, and conversations. Her prior research was on designing statistical relational frameworks to learn, control, and reason about complex dynamic domains.

Eric P. Xing is a professor in the School of Computer Science at Carnegie Mellon University. His principal research interests lie in the development of machine learning and statistical methodology, and large-scale computational system and architecture, for solving problems involving automated learning, reasoning, and decision-making in high-dimensional, multimodal, and dynamic possible worlds in complex systems. Professor Xing received a Ph.D. in Molecular Biology from Rutgers University, and another Ph.D. in Computer Science from UC Berkeley. His current work involves, 1) foundations of statistical learning, including theory and algorithms for estimating time/space varying-coefficient models, sparse structured input/output models, and nonparametric Bayesian models; 2) framework for parallel machine learning on big data with big model in distributed systems or in the cloud; 3) computational and statistical analysis of gene regulation, genetic variation, and disease associations; and 4) application of statistical learning in natural language, social networks, data mining, and vision. Professor Xing has published over 200 peer-reviewed papers, and is an associate editor of the Journal of the American Statistical Association, Annals of Applied Statistics, the IEEE Transactions of Pattern Analysis and Machine Intelligence, the PLoS Journal of Computational Biology, and an Action Editor of the Machine Learning journal, and the Journal of Machine Learning Research. He is a member of the DARPA Information Science and Technology (ISAT) Advisory Group, a recipient of the NSF Career Award, the Alfred P. Sloan Research Fellowship, the United States Air Force Young Investigator Award, and the IBM Open Collaborative Research Faculty Award.

Tutorial 6

Deep Chit-Chat: Deep Learning for ChatBots

Wei Wu and Rui Yan

Thursday, November 1, 2018 2, 14:00–17:30

Location: Jutland

The tutorial is based on the long-term efforts on building conversational models with deep learning approaches for chatbots. We will summarize the fundamental challenges in modeling open domain dialogues, clarify the difference from modeling goal-oriented dialogues, and give an overview of state-of-the-art methods for open domain conversation including both retrieval-based methods and generation-based methods. In addition to these, our tutorial will also cover some new trends of research of chatbots, such as how to design a reasonable evaluation system and how to "control" conversations from a chatbot with some specific information such as personas, styles, and emotions, etc.

WORKSHOPS

Wei Wu is an applied scientist in Microsoft XiaoIce team from 2018. Before that, he was a lead researcher of Microsoft Research Asia (MSRA) from 2012 to 2017. He obtained a B.S. in Applied Mathematics from Peking University in 2007 and earned Ph.D. in Applied Mathematics from Peking University in 2012. His research interests include machine learning, natural language processing (NLP), and information retrieval. His current research focus is building conversational engines for chatbots with machine learning and NLP techniques. he has been working on single-turn conversation and multi-turn conversation. He is a key technology contributor to core chat engines of Microsoft XiaoIce and Microsoft Rinna. His recent achievement with XiaoIce team is launching a fully generative chatbot in Indonesia with full dialogue generation technologies. The chatbot now has more than 1.5 million users on LINE Indonesia.

Rui Yan is a tenure-track assistant professor at Peking University, and is dual affiliated with Beijing Institute of Big Data Research. He is also an adjunct professor at central China Normal University and Central University of Finance and Economics. Before he returned to academia, he was a senior researcher at Baidu Inc. For the past 10 years, he has been working on Artificial Intelligence (AI) for Natural Language Processing (NLP). Rui Yan has a broad interest in real world problems related to natural languages, text information, social media, and web applications. Rui's research focuses on Natural Language Processing, Information Retrieval, Machine Learning and Artificial Intelligence. More specifically, he conducts research into conversational systems, natural language cognition and generation, as well as NLP-related applications (i.e., summarization, artistic writing, etc.). When he worked in Natural Language Processing Department (now AI Group) at Baidu, his team launched the conversational system product from scratch, named Duer ChatBot.

4

Workshops

Note: almost all workshops are located in **CPH Conference** – please see the map with room locations in the front matter. Exceptions are Workshop 1 (WMT) and Workshop 10 (BEA), which take place in the main building and the conference hotel, respectively.

Wednesday–Thursday

Funen	Third Conference on Machine Translation (WMT)	p.20
Nørrebros Runddel	The Conference on Computational Natural Language Learning	p.28

Wednesday

Amager Strandpark	The Ninth International Workshop on Health Text Mining and Information Analysis	p.32
Tivoli & Vesterbro Torv	The Second Workshop on Abusive Language Online	p.34
Hovedbanen	The Search-Oriented Conversational AI workshop	p.36
Enghave Plads & Kødbyen	The Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology	p.38
Kastrup Airport	The 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis	p.**??**
Tivoli & Vesterbro Torv	The 3rd Workshop on Social Media Mining for Health Applications Workshop and Shared Task	p.40

Thursday

Amager Strandpark	BioASQ: Large-scale Biomedical Semantic Indexing and Question Answering workshop	p.42
DGI Byen Hotel, Room 3	BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP workshop	p.43
Kastrup Airport	The First Workshop on Fact Extraction and VERification	p.46
Hovedbanen	the 5th International Workshop on Argument Mining	p.49
Skt. Hans Torv	The 4th Workshop on Noisy User-generated Text	p.51
Enghave Plads & Kødbyen	The Second Workshop on Universal Dependencies	p.54

Workshop 1: Third Conference on Machine Translation

Organizers: *Barry Haddow, Philipp Koehn, Ondřej Bojar, Christian Federmann, Yvette Graham, Matthias Huck, Christof Monz, and Lucia Specia*

Location: Funen

Wednesday, October 31, 2018

	Boaster Session for Research Papers Presented as Posters
10:16–10:18	Scaling Neural Machine Translation *Myle Ott, Sergey Edunov, David Grangier, and Michael Auli*
10:18–10:20	Character-level Chinese-English Translation through ASCII Encoding *Nikola Nikolov, Yuhuang Hu, Mi Xue Tan, and Richard H.R. Hahnloser*
10:20–10:22	Neural Machine Translation of Logographic Language Using Sub-character Level Information *Longtu Zhang and Mamoru Komachi*
10:22–10:24	An Analysis of Attention Mechanisms: The Case of Word Sense Disambiguation in Neural Machine Translation *Gongbo Tang, Rico Sennrich, and Joakim Nivre*
10:24–10:26	Discourse-Related Language Contrasts in English-Croatian Human and Machine Translation *Margita Šoštarić, Christian Hardmeier, and Sara Stymne*
10:26–10:28	Coreference and Coherence in Neural Machine Translation: A Study Using Oracle Experiments *Dario Stojanovski and Alexander Fraser*
10:28–10:30	A Large-Scale Test Set for the Evaluation of Context-Aware Pronoun Translation in Neural Machine Translation *Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich*
14:00–15:20	**Session 3: Research Papers on Embedding and Context**
14:00–14:20	Beyond Weight Tying: Learning Joint Input-Output Embeddings for Neural Machine Translation *Nikolaos Pappas, Lesly Miculicich, and James Henderson*
14:20–14:40	A neural interlingua for multilingual machine translation *yichao lu yichao, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun*
14:40–15:00	Improving Neural Language Models with Weight Norm Initialization and Regularization *Christian Herold, Yingbo Gao, and Hermann Ney*
15:00–15:20	Contextual Neural Model for Translating Bilingual Multi-Speaker Conversations *Sameen Maruf, André F. T. Martins, and Gholamreza Haffari*
15:30–16:00	**Coffee Break**
16:00–17:20	**Session 4: Research Papers on Analysis and Data**
16:00–16:20	Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation *Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way*

16:20–16:40 Freezing Subnetworks to Analyze Domain Adaptation in Neural Machine Translation
Brian Thompson, Huda Khayrallah, Antonios Anastasopoulos, Arya D. McCarthy, Kevin Duh, Rebecca Marvin, Paul McNamee, Jeremy Gwinnup, Tim Anderson, and Philipp Koehn

16:40–17:00 Denoising Neural Machine Translation Training with Trusted Data and Online Data Selection
Wei Wang, Taro Watanabe, Macduff Hughes, Tetsuji Nakagawa, and Ciprian Chelba

17:00–17:20 Using Monolingual Data in Neural Machine Translation: a Systematic Study
Franck Burlot and François Yvon

10:30-11:00 **Coffee Break**

8:45–9:00 **Opening Remarks**

Boaster Session for Research Papers Presented as Posters

10:14–10:16 Neural Machine Translation into Language Varieties
Surafel Melaku Lakew, Aliia Erofeeva, and Marcello Federico

10:16–10:18 Effective Parallel Corpus Mining using Bilingual Sentence Embeddings
Mandy Guo, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernandez Abrego, Keith Stevens, Noah Constant, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil

10:18–10:20 On The Alignment Problem In Multi-Head Attention-Based Neural Machine Translation
Tamer Alkhouli, Gabriel Bretschner, and Hermann Ney

10:20–10:22 A Call for Clarity in Reporting BLEU Scores
Matt Post

10:22–10:24 Exploring gap filling as a cheaper alternative to reading comprehension questionnaires when evaluating machine translation for gisting
Mikel L. Forcada, Carolina Scarton, Lucia Specia, Barry Haddow, and Alexandra Birch

10:24–10:26 Simple Fusion: Return of the Language Model
Felix Stahlberg, James Cross, and Veselin Stoyanov

10:26–10:28 Correcting Length Bias in Neural Machine Translation
Kenton Murray and David Chiang

10:28–10:30 Extracting In-domain Training Corpora for Neural Machine Translation Using Data Selection Methods
Catarina Cruz Silva, Chao-Hong Liu, Alberto Poncelas, and Andy Way

16:00–17:20 **Session 8: Research Papers on Multilingual Translation**

16:00–16:20 Massively Parallel Cross-Lingual Learning in Low-Resource Target Language Translation
Zhong Zhou, Matthias Sperber, and Alexander Waibel

16:20–16:40 Trivial Transfer Learning for Low-Resource Neural Machine Translation
Tom Kocmi and Ondřej Bojar

16:40–17:00 Input Combination Strategies for Multi-Source Transformer Decoder
Jindřich Libovický, Jindřich Helcl, and David Mareček

17:00–17:20 Parameter Sharing Methods for Multilingual Self-Attentional Translation Models
Devendra Sachan and Graham Neubig

9:00–10:30 **Session 1: Shared Tasks Overview Presentations I**

9:00–9:30 Findings of the 2018 Conference on Machine Translation (WMT18)
Ondřej Bojar, Christian Federmann, Mark Fishel, Barry Haddow, Philipp Koehn, and Christof Monz

9:30–9:50 Findings of the Third Shared Task on Multimodal Machine Translation
Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank

9:50–10:10 Findings of the WMT 2018 Biomedical Translation Shared Task: Evaluation on Medline test sets
Mariana Neves, Antonio Jimeno Yepes, Aurélie Névéol, Cristian Grozea, Amy Siu, Madeleine Kittner, and Karin Verspoor

11:00–12:30 **Session 2: Poster Session I**

11:00–12:30 **Shared Task: News Translation**

- An Empirical Study of Machine Translation for the Shared Task of WMT18
  Chao Bei, Hao Zong, Yiming Wang, Baoyong Fan, Shiqi Li, and Conghu Yuan
- Robust parfda Statistical Machine Translation Results
  Ergun Biçici
- The TALP-UPC Machine Translation Systems for WMT18 News Shared Translation Task
  noe casas noe, Carlos Escolano, Marta R. Costa-jussà, and José A. R. Fonollosa
- Phrase-based Unsupervised Machine Translation with Compositional Phrase Embeddings
  Maksym Del, Andre Tättar, and Mark Fishel
- Alibaba's Neural Machine Translation Systems for WMT18
  Yongchao DENG, Shanbo Cheng, Jun Lu, Kai Song, Jingang Wang, Shenglan Wu, Liang Yao, Guchun Zhang, Haibo Zhang, Pei Zhang, Changfeng Zhu, and Boxing Chen
- The RWTH Aachen University English-German and German-English Unsupervised Neural Machine Translation Systems for WMT 2018
  Miguel Graça, Yunsu Kim, Julian Schamper, Jiahui Geng, and Hermann Ney
- Cognate-aware morphological segmentation for multilingual neural translation
  Stig-Arne Grönroos, Sami Virpioja, and Mikko Kurimo
- The AFRL WMT18 Systems: Ensembling, Continuation and Combination
  Jeremy Gwinnup, Tim Anderson, Grant Erdmann, and Katherine Young
- The University of Edinburgh's Submissions to the WMT18 News Translation Task
  Barry Haddow, Nikolay Bogoychev, Denis Emelin, Ulrich Germann, Roman Grundkiewicz, Kenneth Heafield, Antonio Valerio Miceli Barone, and Rico Sennrich
- TencentFmRD Neural Machine Translation for WMT18
  Bojie Hu, Ambyer Han, and Shen Huang
- The MLLP-UPV German-English Machine Translation System for WMT18
  Javier Iranzo-Sánchez, Pau Baquero-Arnal, Gonçal V. Garcés Díaz-Munío, Adrià Martínez-Villaronga, Jorge Civera, and Alfons Juan
- Microsoft's Submission to the WMT2018 News Translation Task: How I Learned to Stop Worrying and Love the Data
  Marcin Junczys-Dowmunt
- CUNI Submissions in WMT18
  Tom Kocmi, Roman Sudarikov, and Ondřej Bojar
- The JHU Machine Translation Systems for WMT 2018
  Philipp Koehn, Kevin Duh, and Brian Thompson
- JUCBNMT at WMT2018 News Translation Task: Character Based Neural Machine Translation of Finnish to English
  Sainik Kumar Mahata, Dipankar Das, and Sivaji Bandyopadhyay

- NICT's Neural and Statistical Machine Translation Systems for the WMT18 News Translation Task
  Benjamin Marie, Rui Wang, Atsushi Fujita, Masao Utiyama, and Eiichiro Sumita
- PROMT Systems for WMT 2018 Shared Translation Task
  Alexander Molchanov
- NTT's Neural Machine Translation Systems for WMT 2018
  Makoto Morishita, Jun Suzuki, and Masaaki Nagata
- The Karlsruhe Institute of Technology Systems for the News Translation Task in WMT 2018
  Ngoc-Quan Pham, Jan Niehues, and Alexander Waibel
- Tilde's Machine Translation Systems for WMT 2018
  Mārcis Pinnis, Matīss Rikters, and Rihards Krišlauks
- CUNI Transformer Neural MT System for WMT18
  Martin Popel
- The University of Helsinki submissions to the WMT18 news task
  Alessandro Raganato, Yves Scherrer, Tommi Nieminen, Arvi Hurskainen, and Jörg Tiedemann
- The RWTH Aachen University Supervised Machine Translation Systems for WMT 2018
  Julian Schamper, Jan Rosendahl, Parnia Bahar, Yunsu Kim, Arne Nix, and Hermann Ney
- The University of Cambridge's Machine Translation Systems for WMT18
  Felix Stahlberg, Adrià de Gispert, and Bill Byrne
- The LMU Munich Unsupervised Machine Translation Systems
  Dario Stojanovski, Viktor Hangya, Matthias Huck, and Alexander Fraser
- Tencent Neural Machine Translation Systems for WMT18
  Mingxuan Wang, Li Gong, Wenhuan Zhu, Jun Xie, and Chao Bian
- The NiuTrans Machine Translation System for WMT18
  Qiang Wang, Bei Li, Jiqiang Liu, Bojian Jiang, Zheyang Zhang, Yinqiao Li, Ye Lin, Tong Xiao, and Jingbo Zhu
- The University of Maryland's Chinese-English Neural Machine Translation Systems at WMT18
  Weijia Xu and Marine Carpuat

11:00–12:30 **Extra Test Suites**

- EvalD Reference-Less Discourse Evaluation for WMT18
  Ondřej Bojar, Jiří Mírovský, Kateřina Rysová, and Magdaléna Rysová
- The WMT'18 Morpheval test suites for English-Czech, English-German, English-Finnish and Turkish-English
  Franck Burlot, Yves Scherrer, Vinit Ravishankar, Ondřej Bojar, Stig-Arne Grönroos, Maarit Koponen, Tommi Nieminen, and François Yvon
- Testsuite on Czech—English Grammatical Contrasts
  Silvie Cinkova and Ondřej Bojar
- A Pronoun Test Suite Evaluation of the English—German MT Systems at WMT 2018
  Liane Guillou, Christian Hardmeier, Ekaterina Lapshinova-Koltunski, and Sharid Loáiciga
- Fine-grained evaluation of German-English Machine Translation based on a Test Suite
  Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, and Hans Uszkoreit
- The Word Sense Disambiguation Test Suite at WMT18
  Annette Rios, Mathias Müller, and Rico Sennrich

11:00–12:30 **Shared Task: Multimodal Translation**

- LIUM-CVC Submissions for WMT18 Multimodal Translation Task
  Ozan Caglayan, Adrien Bardet, Fethi Bougares, Loïc Barrault, Kai Wang, Marc Masana, Luis Herranz, and Joost van de Weijer
- The MeMAD Submission to the WMT18 Multimodal Translation Task
  Stig-Arne Grönroos, Benoit Huet, Mikko Kurimo, Jorma Laaksonen, Bernard Merialdo, Phu Pham, Mats Sjöberg, Umut Sulubacak, Jörg Tiedemann, Raphael Troncy, and Raúl Vázquez
- The AFRL-Ohio State WMT18 Multimodal System: Combining Visual with Traditional
  Jeremy Gwinnup, Joshua Sandvick, Michael Hutt, Grant Erdmann, John Duselis, and James Davis
- CUNI System for the WMT18 Multimodal Translation Task
  Jindřich Helcl, Jindřich Libovický, and Dusan Varis
- Sheffield Submissions for WMT18 Multimodal Translation Shared Task
  Chiraag Lala, Pranava Swaroop Madhyastha, Carolina Scarton, and Lucia Specia
- Ensemble Sequence Level Training for Multimodal MT: OSU-Baidu WMT18 Multimodal Machine Translation System Report
  Renjie Zheng, Yilin Yang, Mingbo Ma, and Liang Huang

11:00–12:30 **Shared Task: Biomedical Translation**

- Translation of Biomedical Documents with Focus on Spanish-English
  Mirela-Stefania Duma and Wolfgang Menzel
- Ensemble of Translators with Automatic Selection of the Best Translation – the submission of FOKUS to the WMT 18 biomedical translation task –
  Cristian Grozea
- LMU Munich's Neural Machine Translation Systems at WMT 2018
  Matthias Huck, Dario Stojanovski, Viktor Hangya, and Alexander Fraser
- Hunter NMT System for WMT18 Biomedical Translation Task: Transfer Learning in Neural Machine Translation
  Abdul Khan, Subhadarshi Panda, Jia Xu, and Lampros Flokas
- UFRGS Participation on the WMT Biomedical Translation Shared Task
  Felipe Soares and Karin Becker
- Neural Machine Translation with the Transformer and Multi-Source Romance Languages for the Biomedical WMT 2018 task
  Brian Tubay and Marta R. Costa-jussà

12:30–14:00 **Lunch**

Thursday, November 1, 2018

9:00–10:30 **Session 5: Shared Tasks Overview Presentations I**

9:00–9:25 Findings of the WMT 2018 Shared Task on Quality Estimation
Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón Astudillo, and André F. T. Martins

9:25–9:50 Findings of the WMT 2018 Shared Task on Automatic Post-Editing
Rajen Chatterjee, Matteo Negri, Raphael Rubino, and Marco Turchi

9:50–10:10 Findings of the WMT 2018 Shared Task on Parallel Corpus Filtering
Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada

10:30-11:00 **Coffee Break**

11:00–12:30 **Session 6: Poster Session II**

11:00–12:30 **Shared Task: Metrics**

- Meteor++: Incorporating Copy Knowledge into Machine Translation Evaluation
Yinuo Guo, Chong Ruan, and Junfeng Hu
- ITER: Improving Translation Edit Rate through Optimizable Edit Costs
Joybrata Panja and Sudip Kumar Naskar
- RUSE: Regressor Using Sentence Embeddings for Automatic Machine Translation Evaluation
Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi

11:00–12:30 **Shared Task: Quality Estimation**

- Keep It or Not: Word Level Quality Estimation for Post-Editing
Prasenjit Basu, Santanu Pal, and Sudip Kumar Naskar
- RTM results for Predicting Translation Performance
Ergun Biçici
- Neural Machine Translation for English-Tamil
Himanshu Choudhary, Aditya Kumar Pathak, Rajiv Ratan Saha, and Ponnurangam Kumaraguru
- The Benefit of Pseudo-Reference Translations in Quality Estimation of MT Output
Melania Duma and Wolfgang Menzel
- Supervised and Unsupervised Minimalist Quality Estimators: Vicomtech's Participation in the WMT 2018 Quality Estimation Task
Thierry Etchegoyhen, Eva Martínez Garcia, and Andoni Azpeitia
- Contextual Encoding for Translation Quality Estimation
Junjie Hu, Wei-Cheng Chang, Yuexin Wu, and Graham Neubig
- Sheffield Submissions for the WMT18 Quality Estimation Shared Task
Julia Ive, Carolina Scarton, Frédéric Blain, and Lucia Specia
- UAlacant machine translation quality estimation at WMT 2018: a simple approach using phrase tables and feed-forward neural networks
Felipe Sánchez-Martínez, Miquel Esplà-Gomis, and Mikel L. Forcada
- Alibaba Submission for WMT18 Quality Estimation Task
Jiayi Wang, Kai Fan, Bo Li, Fengming Zhou, Boxing Chen, Yangbin Shi, and Luo Si
- Quality Estimation with Force-Decoded Attention and Cross-lingual Embeddings
Elizaveta Yankovskaya, Andre Tättar, and Mark Fishel

11:00–12:30 **Shared Task: Automatic Post-Editing**

- MS-UEdin Submission to the WMT2018 APE Shared Task: Dual-Source Transformer for Automatic Post-Editing
Marcin Junczys-Dowmunt and Roman Grundkiewicz
- A Transformer-Based Multi-Source Automatic Post-Editing System
Santanu Pal, Nico Herbig, Antonio Krüger, and Josef van Genabith
- DFKI-MLT System Description for the WMT18 Automatic Post-editing Task
Daria Pylypenko and Raphael Rubino
- Multi-encoder Transformer Network for Automatic Post-Editing
Jaehun Shin and Jong-Hyeok Lee

- Multi-source transformer with combined losses for automatic post editing
amirhossein tebbifakhr amirhossein, Ruchit Agrawal, Matteo Negri, and Marco Turchi

11:00–12:30 **Shared Task: Parallel Corpus Filtering**

- The Speechmatics Parallel Corpus Filtering System for WMT18
Tom Ash, Remi Francis, and Will Williams
- STACC, OOV Density and N-gram Saturation: Vicomtech's Participation in the WMT 2018 Shared Task on Parallel Corpus Filtering
Andoni Azpeitia, Thierry Etchegoyhen, and Eva Martínez garcia
- A hybrid pipeline of rules and machine learning to filter web-crawled parallel corpora
Eduard Barbu and Verginica Barbu Mititelu
- Coverage and Cynicism: The AFRL Submission to the WMT 2018 Parallel Corpus Filtering Task
Grant Erdmann and Jeremy Gwinnup
- MAJE Submission to the WMT2018 Shared Task on Parallel Corpus Filtering
Marina Fomicheva and Jesús González-Rubio
- An Unsupervised System for Parallel Corpus Filtering
Viktor Hangya and Alexander Fraser
- Dual Conditional Cross-Entropy Filtering of Noisy Parallel Corpora
Marcin Junczys-Dowmunt
- The JHU Parallel Corpus Filtering Systems for WMT 2018
Huda Khayrallah, Hainan Xu, and Philipp Koehn
- Measuring sentence parallelism using Mahalanobis distances: The NRC unsupervised submissions to the WMT18 Parallel Corpus Filtering shared task
Patrick Littell, Samuel Larkin, Darlene Stewart, Michel Simard, Cyril Goutte, and Chi-kiu Lo
- Accurate semantic textual similarity for cleaning noisy parallel corpora using semantic machine translation evaluation metric: The NRC supervised submissions to the Parallel Corpus Filtering task
Chi-kiu Lo, Michel Simard, Darlene Stewart, Samuel Larkin, Cyril Goutte, and Patrick Littell
- Alibaba Submission to the WMT18 Parallel Corpus Filtering Task
Jun Lu, Xiaoyu Lv, Yangbin Shi, and Boxing Chen
- UTFPR at WMT 2018: Minimalistic Supervised Corpora Filtering for Machine Translation
Gustavo Paetzold
- The ILSP/ARC submission to the WMT 2018 Parallel Corpus Filtering Shared Task
Vassilis Papavassiliou, Sokratis Sofianopoulos, Prokopis Prokopidis, and Stelios Piperidis
- SYSTRAN Participation to the WMT2018 Shared Task on Parallel Corpus Filtering
Minh Quang Pham, Josep Crego, and Jean Senellart
- Tilde's Parallel Corpus Filtering Methods for WMT 2018
Mārcis Pinnis
- The RWTH Aachen University Filtering System for the WMT 2018 Parallel Corpus Filtering Task
Nick Rossenbach, Jan Rosendahl, Yunsu Kim, Miguel Graça, Aman Gokrani, and Hermann Ney

- Prompsit's submission to WMT 2018 Parallel Corpus Filtering shared task
  Víctor M. Sánchez-Cartagena, Marta Bañón, Sergio Ortiz Rojas, and Gema Ramírez
- NICT's Corpus Filtering Systems for the WMT18 Parallel Corpus Filtering Task
  Rui Wang, Benjamin Marie, Masao Utiyama, and Eiichiro Sumita

12:30–14:00 **Lunch**

14:00–15:30 **Session 3: Invited Talk**

14:00–15:30 **TBD**

15:30–16:00 **Coffee Break**

Workshop 2: The Conference on Computational Natural Language Learning

Organizers: *Ivan Titov and Anna Korhonen*

Location: Nørrebros Runddel

Wednesday, October 31, 2018

08:45–09:00 **Opening remarks**

09:00–10:30 **CoNLL Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies**

10:30–11:00 **Coffee break**

11:00–12:30 **CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection**

12:30–14:00 **Lunch**

Invited Talk by Max Welling

14:00–15:00 ***Inductive Bias in Deep Learning***

Session 1

15:00–15:15 Embedded-State Latent Conditional Random Fields for Sequence Labeling
Dung Thai, Sree Harsha Ramesh, Shikhar Murty, Luke Vilnis, and Andrew McCallum

15:15–15:30 Continuous Word Embedding Fusion via Spectral Decomposition
Tianfan Fu, Cheng Zhang, and Stephan Mandt

15:30–16:00 **Coffee break**

16:00–17:30 **Poster session 1**

- Dual Latent Variable Model for Low-Resource Natural Language Generation in Dialogue Systems
  Van-Khanh Tran and Le-Minh Nguyen
- A Trio Neural Model for Dynamic Entity Relatedness Ranking
  Tu Nguyen, Tuan Tran, and Wolfgang Nejdl
- A Unified Neural Network Model for Geolocating Twitter Users
  Mohammad Ebrahimi, Elaheh ShafieiBavani, Raymond Wong, and Fang Chen
- Corpus-Driven Thematic Hierarchy Induction
  Ilia Kuznetsov and Iryna Gurevych
- Adversarially Regularising Neural NLI Models to Integrate Logical Background Knowledge
  Pasquale Minervini and Sebastian Riedel
- From Strings to Other Things: Linking the Neighborhood and Transposition Effects in Word Reading
  Stephan Tulkens, Dominiek Sandra, and Walter Daelemans
- Global Attention for Name Tagging
  Boliang Zhang, Spencer Whitehead, Lifu Huang, and Heng Ji

- Pervasive Attention: 2D Convolutional Neural Networks for Sequence-to-Sequence Prediction
  Maha Elbayad, Laurent Besacier, and Jakob Verbeek
- Comparing Attention-Based Convolutional and Recurrent Neural Networks: Success and Limitations in Machine Reading Comprehension
  Matthias Blohm, Glorianna Jagfeld, Ekta Sood, Xiang Yu, and Ngoc Thang Vu
- Uncovering Code-Mixed Challenges: A Framework for Linguistically Driven Question Generation and Neural Based Question Answering
  Deepak Gupta, Pabitra Lenka, Asif Ekbal, and Pushpak Bhattacharyya
- Learning to Embed Semantic Correspondence for Natural Language Understanding
  Sangkeun Jung, Jinsik Lee, and Jiwon Kim
- Commonsense Knowledge Base Completion and Generation
  Itsumi Saito, Kyosuke Nishida, Hisako Asano, and Junji Tomita
- Active Learning for Interactive Neural Machine Translation of Data Streams
  Álvaro Peris and Francisco Casacuberta
- Churn Intent Detection in Multilingual Chatbot Conversations and Social Media
  Christian Abbet, Meryem M'hamdi, Athanasios Giannakopoulos, Robert West, Andreea Hossmann, Michael Baeriswyl, and Claudiu Musat
- Learning Text Representations for 500K Classification Tasks on Named Entity Disambiguation
  Ander Barrena, Aitor Soroa, and Eneko Agirre
- Hierarchical Attention Based Position-Aware Network for Aspect-Level Sentiment Analysis
  Lishuang Li, Yang Liu, and AnQiao Zhou
- Bidirectional Generative Adversarial Networks for Neural Machine Translation
  Zhirui Zhang, Shujie Liu, Mu Li, Ming Zhou, and Enhong Chen
- Latent Entities Extraction: How to Extract Entities that Do Not Appear in the Text?
  Eylon Shoshan and Kira Radinsky
- Generalizing Procrustes Analysis for Better Bilingual Dictionary Induction
  Yova Kementchedjhieva, Sebastian Ruder, Ryan Cotterell, and Anders Søgaard
- Simple Unsupervised Keyphrase Extraction using Sentence Embeddings
  Kamil Bennani-Smires, Claudiu Musat, Andreea Hossmann, Michael Baeriswyl, and Martin Jaggi

Thursday, November 1, 2018

Session 2

09:00–09:15 A Temporally Sensitive Submodularity Framework for Timeline Summarization
Sebastian Martschat and Katja Markert

09:15–09:30 Chinese Poetry Generation with a Salient-Clue Mechanism
Xiaoyuan Yi, Ruoyu Li, and Maosong Sun

09:30–09:45 Multi-Modal Sequence Fusion via Recursive Attention for Emotion Recognition
Rory Beard, Ritwik Das, Raymond W. M. Ng, P. G. Keerthana Gopalakrishnan, Luka Eerens, Pawel Swietojanski, and Ondrej Miksik

09:45–10:00 Using Sparse Semantic Embeddings Learned from Multimodal Text and Image Data to Model Human Conceptual Knowledge
Steven Derby, Paul Miller, Brian Murphy, and Barry Devereux

10:00–10:15 Similarity Dependent Chinese Restaurant Process for Cognate Identification in Multilingual Wordlists
Taraka Rama

10:15–10:30 Uncovering Divergent Linguistic Information in Word Embeddings with Lessons for Intrinsic and Extrinsic Evaluation
Mikel Artetxe, Gorka Labaka, Inigo Lopez-Gazpio, and Eneko Agirre

10:30–11:00 **Coffee break**

Invited talk by Asifa Majid

11:00–12:00 **Semantic Spaces Across Diverse Languages**

Session 3

12:00–12:15 Comparing Models of Associative Meaning: An Empirical Investigation of Reference in Simple Language Games
Judy Hanwen Shen, Matthias Hofer, Bjarke Felbo, and Roger Levy

12:15–12:30 Sequence Classification with Human Attention
Maria Barrett, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard

12:30–14:00 **Lunch**

Session 4

14:00–14:15 Sentence-Level Fluency Evaluation: References Help, But Can Be Spared!
Katharina Kann, Sascha Rothe, and Katja Filippova

14:15–14:30 Predefined Sparseness in Recurrent Sequence Models
Thomas Demeester, Johannes Deleu, Fréderic Godin, and Chris Develder

14:30–14:45 Learning to Actively Learn Neural Machine Translation
Ming Liu, Wray Buntine, and Gholamreza Haffari

14:45–15:00 Upcycle Your OCR: Reusing OCRs for Post-OCR Text Correction in Romanised Sanskrit
Amrith Krishna, Bodhisattwa P. Majumder, Rajesh Bhat, and Pawan Goyal

15:00–15:15 Weakly-Supervised Neural Semantic Parsing with a Generative Ranker
Jianpeng Cheng and Mirella Lapata

15:15–15:30 Modeling Composite Labels for Neural Morphological Tagging
Alexander Tkachenko and Kairit Sirts

15:30–16:00 **Coffee break**

16:00–17:30 **Poster session 2**

- Evolutionary Data Measures: Understanding the Difficulty of Text Classification Tasks
  Edward Collins, Nikolai Rozanov, and Bingbing Zhang
- Vectorial Semantic Spaces Do Not Encode Human Judgments of Intervention Similarity
  Paola Merlo and Francesco Ackermann
- Lessons Learned in Multilingual Grounded Language Learning
  Ákos Kádár, Desmond Elliott, Marc-Alexandre Côté, Grzegorz Chrupała, and Afra Alishahi

- Unsupervised Sentence Compression using Denoising Auto-Encoders
  Thibault Fevry and Jason Phang
- Resources to Examine the Quality of Word Embedding Models Trained on n-Gram Data
  Ábel Elekes, Adrian Englhardt, Martin Schäler, and Klemens Böhm
- Linguistically-Based Deep Unstructured Question Answering
  Ahmad Aghaebrahimian
- DIMSIM: An Accurate Chinese Phonetic Similarity Algorithm Based on Learned High Dimensional Encoding
  Min Li, Marina Danilevsky, Sara Noeman, and Yunyao Li
- Challenge or Empower: Revisiting Argumentation Quality in a News Editorial Corpus
  Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein
- Bringing Order to Neural Word Embeddings with Embeddings Augmented by Random Permutations (EARP)
  Trevor Cohen and Dominic Widdows
- Aggregated Semantic Matching for Short Text Entity Linking
  Feng Nie, Shuyan Zhou, Jing Liu, Jinpeng Wang, Chin-Yew Lin, and Rong Pan
- Adversarial Over-Sensitivity and Over-Stability Strategies for Dialogue Models
  Tong Niu and Mohit Bansal
- Improving Response Selection in Multi-Turn Dialogue Systems by Incorporating Domain Knowledge
  Debanjan Chaudhuri, Agustinus Kristiadi, Jens Lehmann, and Asja Fischer
- The Lifted Matrix-Space Model for Semantic Composition
  WooJin Chung, Sheng-Fu Wang, and Samuel Bowman
- End-to-End Neural Entity Linking
  Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann
- Modelling Salient Features as Directions in Fine-Tuned Semantic Spaces
  Thomas Ager, Ondřej Kuželka, and Steven Schockaert
- Model Transfer with Explicit Knowledge of the Relation between Class Definitions
  Hiyori Yoshikawa and Tomoya Iwakura
- Aiming to Know You Better Perhaps Makes Me a More Engaging Dialogue Partner
  Yury Zemlyanskiy and Fei Sha
- Neural Maximum Subgraph Parsing for Cross-Domain Semantic Dependency Analysis
  Yufei Chen, Sheng Huang, Fang Wang, Junjie Cao, Weiwei Sun, and Xiaojun Wan
- From Random to Supervised: A Novel Dropout Mechanism Integrated with Global Information
  Hengru Xu, Shen Li, Renfen Hu, Si Li, and Sheng Gao
- Sequence to Sequence Mixture Model for Diverse Machine Translation
  Xuanli He, Gholamreza Haffari, and Mohammad Norouzi

Workshop 3: Ninth International Workshop on Health Text Mining and Information Analysis

Organizers: *Alberto Lavelli, Anne-Lyse Minard, and Fabio Rinaldi*

Location: Amager Strandpark

Wednesday, October 31, 2018

9:00–10:30 **Session 1**

9:00–9:05 **Introduction**

9:05–9:30 Detecting Diabetes Risk from Social Media Activity
Dane Bell, Egoitz Laparra, Aditya Kousik, Terron Ishihara, Mihai Surdeanu, and Stephen Kobourov

9:30–9:55 Treatment Side Effect Prediction from Online User-generated Content
Hoang Nguyen, Kazunari Sugiyama, Min-Yen Kan, and Kishaloy Halder

9:55–10:15 **Poster booster**

10:15–10:30 **Poster session**

- Revisiting neural relation classification in clinical notes with external information
  Simon Suster, Madhumita Sushil, and Walter Daelemans
- Supervised Machine Learning for Extractive Query Based Summarisation of Biomedical Data
  Mandeep Kaur and Diego Molla
- Comparing CNN and LSTM character-level embeddings in BiLSTM-CRF models for chemical and disease named entity recognition
  Zenan Zhai, Dat Quoc Nguyen, and Karin Verspoor
- Deep learning for language understanding of mental health concepts derived from Cognitive Behavioural Therapy
  Lina M. Rojas Barahona, Bo-Hsiang Tseng, Yinpei Dai, Clare Mansfield, Osman Ramadan, Stefan Ultes, Michael Crawford, and Milica Gasic
- Investigating the Challenges of Temporal Relation Extraction from Clinical Text
  Diana Galvan, Naoaki Okazaki, Koji Matsuda, and Kentaro Inui
- De-identifying Free Text of Japanese Dummy Electronic Health Records
  Kohei Kajiyama, Hiromasa Horiguchi, Takashi Okumura, Mizuki Morita, and Yoshinobu Kano
- Unsupervised Identification of Study Descriptors in Toxicology Research: An Experimental Study
  Drahomira Herrmannova, Steven Young, Robert Patton, Christopher Stahl, Nicole Kleinstreuer, and Mary Wolfe
- Identification of Parallel Sentences in Comparable Monolingual Corpora from Different Registers
  Rémi Cardon and Natalia Grabar
- Evaluation of a Prototype System that Automatically Assigns Subject Headings to Nursing Narratives Using Recurrent Neural Network
  Hans Moen, Kai Hakala, Laura-Maria Peltonen, Henry Suhonen, Petri Loukasmäki, Tapio Salakoski, Filip Ginter, and Sanna Salanterä

- Automatically Detecting the Position and Type of Psychiatric Evaluation Report Sections
  Deya Banisakher, Naphtali Rishe, and Mark A. Finlayson

10:30–11:00 **Break**

11:00–12:30 **Session 2**

11:00–11:25 Iterative development of family history annotation guidelines using a synthetic corpus of clinical text
Taraka Rama, Pål Brekke, Øystein Nytrø, and Lilja Øvrelid

11:25–11:40 CAS: French Corpus with Clinical Cases
Natalia Grabar, Vincent Claveau, and Clément Dalloux

11:40–12:05 Analysis of Risk Factor Domains in Psychosis Patient Health Records
Eben Holderness, Nicholas Miller, Kirsten Bolton, Philip Cawkwell, Marie Meteer, James Pustejovsky, and Mei Hua-Hall

12:05–12:30 Patient Risk Assessment and Warning Symptom Detection Using Deep Attention-Based Neural Networks
Ivan Girardi, Pengfei Ji, An-phi Nguyen, Nora Hollenstein, Adam Ivankay, Lorenz Kuhn, Chiara Marchiori, and Ce Zhang

12:30–14:00 **Lunch**

14:00–15:30 **Session 3**

14:00–14:50 **Invited Talk (TBA)**

14:50–15:15 Syntax-based Transfer Learning for the Task of Biomedical Relation Extraction
Joël Legrand, Yannick Toussaint, Chedy Raïssi, and Adrien Coulet

15:15–15:30 In-domain Context-aware Token Embeddings Improve Biomedical Named Entity Recognition
Golnar Sheikhshabbafghi, Inanc Birol, and Anoop Sarkar

15:30–16:00 **Break**

16:00–17:30 **Session 4**

16:00–16:25 Self-training improves Recurrent Neural Networks performance for Temporal Relation Extraction
Chen Lin, Timothy Miller, Dmitriy Dligach, Hadi Amiri, Steven Bethard, and Guergana Savova

16:25–16:40 Listwise temporal ordering of events in clinical notes
Serena Jeblee and Graeme Hirst

16:40–16:55 Time Expressions in Mental Health Records for Symptom Onset Extraction
Natalia Viani, Lucia Yin, Joyce Kam, Ayunni Alawi, André Bittar, Rina Dutta, Rashmi Patel, Robert Stewart, and Sumithra Velupillai

16:55–17:10 Evaluation of a Sequence Tagging Tool for Biomedical Texts
Julien Tourille, Matthieu Doutreligne, Olivier Ferret, Aurélie Névéol, Nicolas Paris, and Xavier Tannier

17:10–17:30 Learning to Summarize Radiology Findings
Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D. Manning, and Curtis P. Langlotz

Workshop 4: Second Workshop on Abusive Language Online

Organizers: *Zeerak Waseem, Rob Voigt, Jacque Wernimont, Vinodkumar Prabhakaran, Ruihong Huang, and Darja Fišer*

Location: Tivoli & Vesterbro Torv

Wednesday, October 31, 2018

9:00–9:15 **Opening Remarks**

9:15–10:00 **Invited Speaker (Mikki Kendall)**

10:00–10:45 **Invited Speaker (Maryant Fernandez Perez)**

10:45–11:15 **Break**

11:15–12:15 **Panel**

12:15–12:45 **Plenary Discussions**

12:45–14:00 **Lunch**

14:00–15:30 **Posters**

14:00–15:30 Social contagion in ethnic abusive swearing during periods of increased terrorist activity
Christoph Spörlein

14:00–15:30 Neural Character-based Composition Models for Abuse Detection
Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova

14:00–15:30 Hate Speech Dataset from a White Supremacy Forum
Ona de Gibert, Naiara Perez, Aitor García Pablos, and Montse Cuadros

14:00–15:30 A Review of Standard Text Classification Practices for Multi-label Toxicity Identification of Online Content
Isuru Gunasekara and Isar Nejadgholi

14:00–15:30 Predictive Embeddings for Hate Speech Detection on Twitter
Rohan Kshirsagar, Tyrus Cukuvac, Kathy McKeown, and Susan McGregor

14:00–15:30 Challenges for Toxic Comment Classification: An In-Depth Error Analysis
Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser

14:00–15:30 Aggression Detection on Social Media Text Using Deep Neural Networks
Vinay Singh, Aman Varshney, Syed Sarfaraz Akhtar, Deepanshu Vijay, and Manish Shrivastava

14:00–15:30 Creating a WhatsApp Dataset to Study Pre-teen Cyberbullying
Rachele Sprugnoli, Stefano Menini, Sara Tonelli, Filippo Oncini, and Enrico Piras

14:00–15:30 Improving Moderation of Online Discussions via Interpretable Neural Models
Andrej Švec, Matúš Pikuliak, Marian Simko, and Maria Bielikova

14:00–15:30 Aggressive language in an online hacking forum
Andrew Caines, Sergio Pastrana, Alice Hutchings, and Paula Buttery

14:00–15:30 The Effects of User Features on Twitter Hate Speech Detection
Elise Fehn Unsvåg and Björn Gambäck

14:00–15:30 Interpreting Neural Network Hate Speech Classifiers
Cindy Wang

14:00–15:30 Determining Code Words in Euphemistic Hate Speech Using Word Embedding Networks
Rijul Magu and Jiebo Luo

14:00–15:30 Comparative Studies of Detecting Abusive Language on Twitter
Younghun Lee, Seunghyun Yoon, and Kyomin Jung

14:00–15:30 Boosting Text Classification Performance on Sexist Tweets by Text Augmentation and Text Generation Using a Combination of Knowledge Graphs
sima sharifirad sima, Borna Jafarpour, and Stan Matwin

14:00–15:30 Learning Representations for Detecting Abusive Language
Magnus Sahlgren, Tim Isbister, and Fredrik Olsson

14:00–15:30 Datasets of Slovene and Croatian Moderated News Comments
Nikola Ljubešić, Tomaž Erjavec, and Darja Fišer

14:00–15:30 Cross-Domain Detection of Abusive Language Online
Mladen Karan and Jan Šnajder

14:00–15:30 Did you offend me? Classification of Offensive Tweets in Hinglish Language
Puneet Mathur, Ramit Sawhney, Meghna Ayyar, and Rajiv Shah

14:00–15:30 Decipherment for Adversarial Offensive Language Detection
Zhelun Wu, Nishant Kambhatla, and Anoop Sarkar

14:00–15:30 The Linguistic Ideologies of Deep Abusive Language Classification
Michael Castelle

15:30–16:00 **Break**

16:00–17:15 **Plenary Discussions and Future Agendas**

17:15–17:30 **Wrapup of Discussions**

17:30–17:45 **Closing Remarks**

Workshop 5: Search-Oriented Conversational AI

Organizers: *Aleksandr Chuklin, Jeff Dalton, Julia Kiseleva, Alexey Borisov, and Mikhail Burtsev*

Location: Hovedbanen

Wednesday, October 31, 2018

09:00–09:10 **Introduction**

Session 1

09:10–09:50 **Keynote 1: Antoine Bordes**

09:50–10:10 Neural Response Ranking for Social Conversation: A Data-Efficient Approach
Igor Shalyminov, Ondřej Dušek, and Oliver Lemon

10:10–10:30 Autonomous Sub-domain Modeling for Dialogue Policy with Hierarchical Deep Reinforcement Learning
Giovanni Yoko Kristianto, Huiwen Zhang, Bin Tong, Makoto Iwayama, and Yoshiyuki Kobayashi

10:30–11:00 **Break**

Session 2

11:00–11:40 **Keynote 2: Milica Gašić**

11:40–12:00 Building Dialogue Structure from Discourse Tree of a Question
Boris Galitsky and Dmitry Ilvovsky

12:00–12:20 A Methodology for Evaluating Interaction Strategies of Task-Oriented Conversational Agents
Marco Guerini, Sara Falcone, and Bernardo Magnini

12:20–12:40 A Reinforcement Learning-driven Translation Model for Search-Oriented Conversational Systems
Wafa Aissa, Laure Soulier, and Ludovic Denoyer

12:40–14:00 **Lunch**

Session 3

14:00–14:40 **Keynote 3: Yun-Nung (Vivian) Chen**

14:40–15:00 Research Challenges in Building a Voice-based Artificial Personal Shopper - Position Paper
Nut Limsopatham, Oleg Rokhlenko, and David Carmel

15:00–15:30 **Poster session**

15:30–16:00 **Break**

Session 4

16:00–16:40 **Keynote 4: Mari Ostendorf**

16:40–17:20 **Roundtable discussion**

17:20–17:30 **Closing**

- Curriculum Learning Based on Reward Sparseness for Deep Reinforcement Learning of Task Completion Dialogue Management
  Atsushi Saito
- Data Augmentation for Neural Online Chats Response Selection
  Wenchao Du and Alan Black
- A Knowledge-Grounded Multimodal Search-Based Conversational Agent
  Shubham Agarwal, Ondřej Dušek, Ioannis Konstas, and Verena Rieser
- Embedding Individual Table Columns for Resilient SQL Chatbots
  Bojan Petrovski, Ignacio Aguado, Andreea Hossmann, Michael Baeriswyl, and Claudiu Musat
- Exploring Named Entity Recognition As an Auxiliary Task for Slot Filling in Conversational Language Understanding
  Samuel Louvan and Bernardo Magnini
- Why are Sequence-to-Sequence Models So Dull? Understanding the Low-Diversity Problem of Chatbots
  Shaojie Jiang and Maarten de Rijke
- Retrieve and Refine: Improved Sequence Generation Models For Dialogue
  Jason Weston, Emily Dinan, and Alexander Miller

Workshop 6: Fifteenth SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology

Organizers: *Sandra Kübler and Garrett Nicolai*

Location: Enghave Plads & Kødbyen

Wednesday, October 31, 2018

08:50–09:00 **Session Opening: Opening Session**

09:00–10:30 **Session 1: Phonology**

09:00–09:30 Efficient Computation of Implicational Universals in Constraint-Based Phonology Through the Hyperplane Separation Theorem
Giorgio Magri

09:30–10:00 Lexical Networks in !Xung
Syed-Amad Hussain, Micha Elsner, and Amanda Miller

10:00–10:30 Acoustic Word Disambiguation with Phonogical Features in Danish ASR
Andreas Søeborg Kirkedal

10:30–11:00 **Tea Break 1**

11:00–12:30 **Session ST: CoNLL – SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection**

12:30–14:00 **Lunch**

Wednesday, October 31, 2018

14:00–15:30 **Session 2: Low-Resource Morphology**

14:00–14:30 Adaptor Grammars for the Linguist: Word Segmentation Experiments for Very Low-Resource Languages
Pierre Godard, Laurent Besacier, François Yvon, Martine Adda-Decker, Gilles Adda, Hélène Maynard, and Annie Rialland

14:30–15:00 String Transduction with Target Language Models and Insertion Handling
Garrett Nicolai, Saeed Najafi, and Grzegorz Kondrak

15:00–15:30 Complementary Strategies for Low Resourced Morphological Modeling
Alexander Erdmann and Nizar Habash

15:30–16:00 **Tea Break 2**

16:00–17:30 **Session P: Poster Session**

- Modeling Reduplication with 2-way Finite-State Transducers
  Hossep Dolatian and Jeffrey Heinz

- Automatically Tailoring Unsupervised Morphological Segmentation to the Language
  Ramy Eskander, Owen Rambow, and Smaranda Muresan
- A Comparison of Entity Matching Methods between English and Japanese Katakana
  Michiharu Yamashita, Hideki Awashima, and Hidekazu Oiwa
- Seq2Seq Models with Dropout can Learn Generalizable Reduplication
  Brandon Prickett, Aaron Traylor, and Joe Pater
- A Characterwise Windowed Approach to Hebrew Morphological Segmentation
  Amir Zeldes
- Phonetic Vector Representations for Sound Sequence Alignment
  Pavel Sofroniev and Çağrı Çöltekin
- Sounds Wilde. Phonetically Extended Embeddings for Author-Stylized Poetry Generation
  Aleksey Tikhonov and Ivan Yamshchikov
- On Hapax Legomena and Morphological Productivity
  Janet Pierrehumbert and Ramon Granell
- A Morphological Analyzer for Shipibo-Konibo
  Ronald Cardenas and Daniel Zeman
- An Arabic Morphological Analyzer and Generator with Copious Features
  Dima Taji, Salam Khalifa, Ossama Obeid, Fadhl Eryani, and Nizar Habash
- Sanskrit n-Retroflexion is Input-Output Tier-Based Strictly Local
  Thomas Graf and Connor Mayer
- Phonological Features for Morphological Inflection
  Adam Wiemerslage, Miikka Silfverberg, and Mans Hulden
- Extracting Morphophonology from Small Corpora
  Marina Ermolaeva

17:30–17:40 **Session Closing: Closing Session**

Workshop 8: Third Social Media Mining for Health Applications Workshop and Shared Task

Organizers: *Graciela Gonzalez, Davy Weissenbacher, Abeed Sarker, Masoud Rouhizadeh, Cecile Paris, Pierre Zweigenbaum, Michael J., Azadeh Nikfarjam, Ari Klein, and Karen O'Connor*

Location: Tivoli & Vesterbro Torv

Wednesday, October 31, 2018

9:00–9:10	**Introduction (Graciela Gonzalez)**
9:10–9:30	Football and Beer - a Social Media Analysis on Twitter in Context of the FIFA Football World Cup 2018 *Roland Roller, Philippe Thomas, and Sven Schmeier*
9:30–9:50	Stance-Taking in Topics Extracted from Vaccine-Related Tweets and Discussion Forum Posts *Maria Skeppstedt, Manfred Stede, and Andreas Kerren*
9:50–10:10	Identifying Depression on Reddit: The Effect of Training Data *Inna Pirina and Çağrı Çöltekin*
10:30–11:00	**Tea Break and Poster Session**
11:00–11:20	Overview of the Third Social Media Mining for Health (SMM4H) Shared Tasks at EMNLP 2018 *Davy Weissenbacher, Abeed Sarker, Michael J. Paul, and Graciela Gonzalez-Hernandez*
11:20–11:40	Changes in Psycholinguistic Attributes of Social Media Users Before, During, and After Self-Reported Influenza Symptoms *Lucie Flekova, Vasileios Lampos, and Ingemar Cox*
11:40–12:00	Thumbs Up and Down: Sentiment Analysis of Medical Online Forums *Victoria Bobicev and Marina Sokolova*
12:00–12:20	Identification of Emergency Blood Donation Request on Twitter *Puneet Mathur, Meghna Ayyar, Sahil Chopra, Simra Shahid, Laiba Mehnaz, and Rajiv Shah*
12:20–12:30	Dealing with Medication Non-Adherence Expressions in Twitter *Takeshi Onishi, Davy Weissenbacher, Ari Klein, Karen O'Connor, and Graciela Gonzalez-Hernandez*
12:30–14:00	**Lunch**
14:00–14:20	Detecting Tweets Mentioning Drug Name and Adverse Drug Reaction with Hierarchical Tweet Representation and Multi-Head Self-Attention *Chuhan Wu, Fangzhao Wu, Junxin Liu, Sixing Wu, Yongfeng Huang, and Xing Xie*
14:20–14:40	Classification of Medication-Related Tweets Using Stacked Bidirectional LSTMs with Context-Aware Attention *Orest Xherija*
14:40–15:00	Shot Or Not: Comparison of NLP Approaches for Vaccination Behaviour Detection *Aditya Joshi, Xiang Dai, Sarvnaz Karimi, Ross Sparks, Cecile Paris, and C Raina MacIntyre*
15:00–15:10	Neural DrugNet *Nishant Nikhil and Shivansh Mundra*

15:10–15:20 IRISA at SMM4H 2018: Neural Network and Bagging for Tweet Classification
Anne-Lyse Minard, Christian Raymond, and Vincent Claveau

15:20–15:30 Drug-Use Identification from Tweets with Word and Character N-Grams
Çağrı Çöltekin and Taraka Rama

15:30–16:00 **Tea Break and Poster Session**

16:00–16:10 Automatic Identification of Drugs and Adverse Drug Reaction Related Tweets
Segun Taofeek Aroyehun and Alexander Gelbukh

16:10–16:20 UZHSMM4H: System Descriptions
Tilia Ellendorff, Joseph Cornelius, Heath Gordon, Nicola Colic, and Fabio Rinaldi

16:20–16:30 Deep Learning for Social Media Health Text Classification
Santosh Tokala, Vaibhav Gambhir, and Animesh Mukherjee

16:30–16:40 Using PPM for Health Related Text Detection
Victoria Bobicev, Victoria Lazu, and Daniela Istrati

16:40–16:50 Leveraging Web Based Evidence Gathering for Drug Information Identification from Tweets
Rupsa Saha, Abir Naskar, Tirthankar Dasgupta, and Lipika Dey

16:50–17:00 CLaC at SMM4H Task 1, 2, and 4
Parsa Bagherzadeh, Nadia Sheikh, and Sabine Bergler

Workshop 9: BioASQ: Large-scale biomedical semantic indexing and question answering

Organizers: *Georgios Paliouras, Ioannis Kakadiaris, and Anastasia Krithara*

Location: Amager Strandpark

- Results of the sixth edition of the BioASQ Challenge
  Anastasios Nentidis, Anastasia Krithara, Konstantinos Bougiatiotis, Georgios Paliouras, and Ioannis Kakadiaris
- Semantic role labeling tools for biomedical question answering: a study of selected tools on the BioASQ datasets
  Fabian Eckert and Mariana Neves
- Macquarie University at BioASQ 6b: Deep learning and deep reinforcement learning for query-based summarisation
  Diego Molla
- AUEB at BioASQ 6: Document and Snippet Retrieval
  George Brokos, Polyvios Liosis, Ryan McDonald, Dimitris Pappas, and Ion Androutsopoulos
- MindLab Neural Network Approach at BioASQ 6B
  Andrés Rosso-Mateus, Fabio A. González, and Manuel Montes-y-Gómez
- AttentionMeSH: Simple, Effective and Interpretable Automatic MeSH Indexer
  Qiao Jin, Bhuwan Dhingra, William Cohen, and Xinghua Lu
- Extraction Meets Abstraction: Ideal Answer Generation for Biomedical Questions
  Yutong Li, Nicholas Gekakis, Qiuze Wu, Boyue Li, Khyathi Chandu, and Eric Nyberg
- UNCC QA: Biomedical Question Answering system
  Abhishek Bhandwaldar and Wlodek Zadrozny
- An Adaption of BIOASQ Question Answering dataset for Machine Reading systems by Manual Annotations of Answer Spans.
  Sanjay Kamath, Brigitte Grau, and Yue Ma
- Ontology-Based Retrieval & Neural Approaches for BioASQ Ideal Answer Generation
  Ashwin Naresh Kumar, Harini Kesavamoorthy, Madhura Das, Pramati Kalwad, Khyathi Chandu, Teruko Mitamura, and Eric Nyberg

Workshop 10: BlackboxNLP: Analyzing and interpreting neural networks for NLP

Organizers: *Tal Linzen, Grzegorz Chrupała, and Afra Alishahi*

Location: DGI Byen Hotel, Room 3

- When does deep multi-task learning work for loosely related document classification tasks?
  Emma Kerinec, Chloé Braud, and Anders Søgaard
- Analyzing Learned Representations of a Deep ASR Performance Prediction Model
  Zied Elloumi, Laurent Besacier, Olivier Galibert, and Benjamin Lecouteux
- Explaining non-linear Classifier Decisions within Kernel-based Deep Architectures
  Danilo Croce, Daniele Rossini, and Roberto Basili
- Nightmare at test time: How punctuation prevents parsers from generalizing
  Anders Søgaard, Miryam de Lhoneux, and Isabelle Augenstein
- Evaluating Textual Representations through Image Generation
  Graham Spinks and Marie-Francine Moens
- On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis
  Jose Camacho-Collados and Mohammad Taher Pilehvar
- Jump to better conclusions: SCAN both left and right
  Joost Bastings, Marco Baroni, Jason Weston, Kyunghyun Cho, and Douwe Kiela
- Understanding Convolutional Neural Networks for Text Classification
  Alon Jacovi, Oren Sar Shalom, and Yoav Goldberg
- Linguistic representations in multi-task neural networks for ellipsis resolution
  Ola Rønning, Daniel Hardt, and Anders Søgaard
- Unsupervised Token-wise Alignment to Improve Interpretation of Encoder-Decoder Models
  Shun Kiyono, Sho Takase, Jun Suzuki, Naoaki Okazaki, Kentaro Inui, and Masaaki Nagata
- Rule induction for global explanation of trained models
  Madhumita Sushil, Simon Suster, and Walter Daelemans
- Can LSTM Learn to Capture Agreement? The Case of Basque
  Shauli Ravfogel, Yoav Goldberg, and Francis Tyers
- Rearranging the Familiar: Testing Compositional Generalization in Recurrent Networks
  Joao Loula, Marco Baroni, and Brenden Lake
- Evaluating the Ability of LSTMs to Learn Context-Free Grammars
  Luzi Sennhauser and Robert Berwick
- Interpretable Neural Architectures for Attributing an Ad's Performance to its Writing Style
  Reid Pryzant, Sugato Basu, and Kazoo Sone
- Interpreting Neural Networks with Nearest Neighbors
  Eric Wallace, Shi Feng, and Jordan Boyd-Graber

- 'Indicatements' that character language models learn English morpho-syntactic units and regularities
  Yova Kementchedjhieva and Adam Lopez
- LISA: Explaining Recurrent Neural Network Judgments via Layer-wIse Semantic Accumulation and Example to Pattern Transformation
  Pankaj Gupta and Hinrich Schütze
- Analysing the potential of seq-to-seq models for incremental interpretation in task-oriented dialogue
  Dieuwke Hupkes, Sanne Bouwmeester, and Raquel Fernández
- An Operation Sequence Model for Explainable Neural Machine Translation
  Felix Stahlberg, Danielle Saunders, and Bill Byrne
- Introspection for convolutional automatic speech recognition
  Andreas Krug and Sebastian Stober
- Learning and Evaluating Sparse Interpretable Sentence Embeddings
  Valentin Trifonov, Octavian-Eugen Ganea, Anna Potapenko, and Thomas Hofmann
- What do RNN Language Models Learn about Filler—Gap Dependencies?
  Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell
- Do Language Models Understand Anything? On the Ability of LSTMs to Understand Negative Polarity Items
  Jaap Jumelet and Dieuwke Hupkes
- Closing Brackets with Recurrent Neural Networks
  Natalia Skachkova, Thomas Trost, and Dietrich Klakow
- Under the Hood: Using Diagnostic Classifiers to Investigate and Improve how Language Models Track Agreement Information
  Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema
- Iterative Recursive Attention Model for Interpretable Sequence Classification
  Martin Tutek and Jan Šnajder
- Interpreting Word-Level Hidden State Behaviour of Character-Level LSTM Language Models
  Avery Hiebert, Cole Peterson, Alona Fyshe, and Nishant Mehta
- Importance of Self-Attention for Sentiment Analysis
  Gaël Letarte, Frédérik Paradis, Philippe Giguère, and François Laviolette
- Firearms and Tigers are Dangerous, Kitchen Knives and Zebras are Not: Testing whether Word Embeddings Can Tell
  Pia Sommerauer and Antske Fokkens
- An Analysis of Encoder Representations in Transformer-Based Machine Translation
  Alessandro Raganato and Jörg Tiedemann
- Evaluating Grammaticality in Seq2seq Models with a Broad Coverage HPSG Grammar: A Case Study on Machine Translation
  Johnny Wei, Khiem Pham, Brendan O'Connor, and Brian Dillon
- Context-Free Transductions with Neural Stacks
  Yiding Hao, William Merrill, Dana Angluin, Robert Frank, Noah Amsel, Andrew Benz, and Simon Mendelsohn
- Learning Explanations from Language Data
  David Harbecke, Robert Schwarzenberg, and Christoph Alt
- How much should you ask? On the question structure in QA systems.
  Barbara Rychalska, Dominika Basaj, Anna Wróblewska, and Przemyslaw Biecek

- Does it care what you asked? Understanding Importance of Verbs in Deep Learning QA System
  Barbara Rychalska, Dominika Basaj, Anna Wróblewska, and Przemyslaw Biecek
- Interpretable Textual Neuron Representations for NLP
  Nina Poerner, Benjamin Roth, and Hinrich Schütze
- Language Models Learn POS First
  Naomi Saphra and Adam Lopez
- Predicting and interpreting embeddings for out of vocabulary words in downstream tasks
  Nicolas Garneau, Jean-Samuel Leboeuf, and Luc Lamontagne
- Probing sentence embeddings for structure-dependent tense
  Geoff Bacon and Terry Regier
- Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation
  Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme
- Interpretable Word Embedding Contextualization
  Kyoung-Rok Jang, Sung-Hyon Myaeng, and Sang-Bum Kim
- State Gradients for RNN Memory Analysis
  Lyan Verwimp, Hugo Van hamme, Vincent Renkens, and Patrick Wambacq
- Extracting Syntactic Trees from Transformer Encoder Self-Attentions
  David Mareček and Rudolf Rosa
- Portable, layer-wise task performance monitoring for NLP models
  Tom Lippincott
- GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding
  Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman
- Explicitly modeling case improves neural dependency parsing
  Clara Vania and Adam Lopez
- Language Modeling Teaches You More than Translation Does: Lessons Learned Through Auxiliary Syntactic Task Analysis
  Kelly Zhang and Samuel Bowman
- Representation of Word Meaning in the Intermediate Projection Layer of a Neural Language Model
  Steven Derby, Paul Miller, Brian Murphy, and Barry Devereux
- Interpretable Structure Induction via Sparse Attention
  Ben Peters, Vlad Niculae, and André F. T. Martins
- Debugging Sequence-to-Sequence Models with Seq2Seq-Vis
  Hendrik Strobelt, Sebastian Gehrmann, Michael Behrisch, Adam Perer, Hanspeter Pfister, and Alexander Rush
- Grammar Induction with Neural Language Models: An Unusual Replication
  Phu Mon Htut, Kyunghyun Cho, and Samuel Bowman
- Does Syntactic Knowledge in Multilingual Language Models Transfer Across Languages?
  Prajit Dhar and Arianna Bisazza
- Diagnosing Failures in Question Answering Tasks with Attention
  Aida Nematzadeh, Kaylee Burns, Erin Grant, and Tom Griffiths
- End-to-end Image Captioning Exploits Distributional Similarity in Multimodal Space
  Pranava Swaroop Madhyastha, Josiah Wang, and Lucia Specia
- Limitations in learning an interpreted language with recurrent models
  Denis Paperno

Workshop 11: First Workshop on Fact Extraction and Verification

Organizers: *James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal*

Location: Kastrup Airport

Thursday, November 1, 2018

09:15–10:00 **Invited Talk: Learning with Explanations (Tim Rocktäschel)**

10:00–10:30 **Research Talks 1**

10:00–10:15 The Data Challenge in Misinformation Detection: Source Reputation vs. Content Veracity
Fatemeh Torabi Asr and Maite Taboada

10:15–10:30 **Towards Automated Factchecking: Developing an Annotation Schema and Benchmark for Consistent Automated Claim Detection (Lev Konstantinovskiy, Oliver Price, Mevan Babakar and Arkaitz Zubiaga)**

10:30–11:30 **Research Posters**

- Crowdsourcing Semantic Label Propagation in Relation Classification
  Anca Dumitrache, Lora Aroyo, and Chris Welty
- Retrieve and Re-rank: A Simple and Effective IR Approach to Simple Question Answering over Knowledge Graphs
  Vishal Gupta, Manoj Chinnakotla, and Manish Shrivastava
- Information Nutrition Labels: A Plugin for Online News Evaluation
  Vincentius Kevin, Birte Högden, Claudia Schwenger, Ali Sahan, Neelu Madan, Piush Aggarwal, Anusha Bangaru, Farid Muradov, and Ahmet Aker
- Joint Modeling for Query Expansion and Information Extraction with Reinforcement Learning
  Motoki Taniguchi, Yasuhide Miura, and Tomoko Ohkuma
- Towards Automatic Fake News Detection: Cross-Level Stance Detection in News Articles
  Costanza Conforti, Mohammad Taher Pilehvar, and Nigel Collier
- Belittling the Source: Trustworthiness Indicators to Obfuscate Fake News on the Web
  Diego Esteves, Aniketh Janardhan Reddy, Piyush Chawla, and Jens Lehmann
- Automated Fact-Checking of Claims in Argumentative Parliamentary Debates
  Nona Naderi and Graeme Hirst
- Stance Detection in Fake News A Combined Feature Representation
  Bilal Ghanem, Paolo Rosso, and Francisco Rangel
- Zero-shot Relation Classification as Textual Entailment
  Abiola Obamuyide and Andreas Vlachos
- Teaching Syntax by Adversarial Distraction
  Juho Kim, Christopher Malon, and Asim Kadav
- Where is Your Evidence: Improving Fact-checking by Justification Modeling
  Tariq Alhindi, Savvas Petridis, and Smaranda Muresan

11:30–12:15 **Invited Talk: Argumentation Mining and Generation Supporting the Verification of Content (Marie-Francine Moens)**

12:15–12:30 **Research Talks 2**

12:15–12:30 Affordance Extraction and Inference based on Semantic Role Labeling
Daniel Loureiro and Alípio Jorge

14:00–14:45 **Invited Talk: Building a broad knowledge graph for products (Luna Dong)**

14:45–15:30 **Shared Task Flash Talks**

14:45–14:50 **The Fact Extraction and VERification (FEVER) Shared Task (Organizers)**

14:50–15:00 **Combining Fact Extraction and Claim Verification in an NLI Model (Yixin Nie, Haonan Chen and Mohit Bansal)**

15:00–15:10 UCL Machine Reading Group: Four Factor Framework For Fact Finding (HexaF)
Takuma Yoneda, Jeff Mitchell, Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel

15:10–15:20 Multi-Sentence Textual Entailment for Claim Verification
Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych

15:20–15:30 Team Papelo: Transformer Networks at FEVER
Christopher Malon

15:30–16:30 **Shared Task Posters**

- Uni-DUE Student Team: Tackling fact checking through decomposable attention neural network
Jan Kowollik and Ahmet Aker
- SIRIUS-LTG: An Entity Linking Approach to Fact Extraction and Verification
Farhad Nooralahzadeh and Lilja Øvrelid
- Integrating Entity Linking and Evidence Ranking for Fact Extraction and Verification
Motoki Taniguchi, Tomoki Taniguchi, Takumi Takahashi, Yasuhide Miura, and Tomoko Ohkuma
- Robust Document Retrieval and Individual Evidence Modeling for Fact Extraction and Verification.
Tuhin Chakrabarty, Tariq Alhindi, and Smaranda Muresan
- DeFactoNLP: Fact Verification using Entity Recognition, TFIDF Vector Comparison and Decomposable Attention
Aniketh Janardhan Reddy, Gil Rocha, and Diego Esteves
- An End-to-End Multi-task Learning Model for Fact Checking
sizhen li sizhen, Shuai Zhao, Bo Cheng, and Hao Yang
- Team GESIS Cologne: An all in all sentence-based approach for FEVER
Wolfgang Otto
- Joint Sentence Extraction and Fact Checking with Pointer Networks
Christopher Hidey and Mona Diab
- QED: A fact verification system for the FEVER shared task
Jackson Luken, Nanjiang Jiang, and Marie-Catherine de Marneffe
- Team UMBC-FEVER : Claim verification using Semantic Lexical Resources
Ankur Padia, Francis Ferraro, and Tim Finin

- A mostly unlexicalized model for recognizing textual entailment
  Mithun Paul, Rebecca Sharp, and Mihai Surdeanu

16:30–17:15 **Invited Talk: Call for Help: Putting Computation in Computational Fact Checking (Delip Rao)**

17:15–17:30 **Prizes + Closing Remarks (Organizers)**

Workshop 12: Fifth International Workshop on Argument Mining

Organizers: *Noam Slonim and Ranit Aharonov*

Location: Hovedbanen

Thursday, November 1, 2018

09:00–09:10 **Openings**

Session 1

09:10–10:10 **Keynote Talk: Argumentation and Human Reason (Hugo Mercier)**

10:10–10:30 Argumentative Link Prediction using Residual Networks and Multi-Objective Learning
Andrea Galassi, Marco Lippi, and Paolo Torroni

10:30–11:00 **Coffee Break**

Session 2

11:00–11:20 End-to-End Argument Mining for Discussion Threads Based on Parallel Constrained Pointer Architecture
Gaku Morio and Katsuhide Fujita

11:20–11:40 ArguminSci: A Tool for Analyzing Argumentation and Rhetorical Aspects in Scientific Writing
Anne Lauscher, Goran Glavaš, and Kai Eckert

11:40–12:00 Evidence Type Classification in Randomized Controlled Trials
Tobias Mayer, Elena Cabrio, and Serena Villata

12:00–12:20 Predicting the Usefulness of Amazon Reviews Using Off-The-Shelf Argumentation Mining
Marco Passon, Marco Lippi, Giuseppe Serra, and Carlo Tasso

12:20–14:30 **Lunch and Poster Presentations**

- An Argument-Annotated Corpus of Scientific Publications
Anne Lauscher, Goran Glavaš, and Simone Paolo Ponzetto
- Annotating Claims in the Vaccination Debate
Benedetta Torsi and Roser Morante
- Argument Component Classification for Classroom Discussions
Luca Lugini and Diane Litman
- Evidence Types, Credibility Factors, and Patterns or Soft Rules for Weighing Conflicting Evidence: Argument Mining in the Context of Legal Rules Governing Evidence Assessment
Vern R. Walker, Dina Foerster, Julia Monica Ponce, and Matthew Rosen
- Feasible Annotation Scheme for Capturing Policy Argument Reasoning using Argument Templates
Paul Reisert, Naoya Inoue, Tatsuki Kuribayashi, and Kentaro Inui
- Frame- and Entity-Based Knowledge for Common-Sense Argumentative Reasoning
Teresa Botschen, Daniil Sorokin, and Iryna Gurevych

- Incorporating Topic Aspects for Online Comment Convincingness Evaluation
  Yunfan Gu, Zhongyu Wei, Maoran Xu, Hao Fu, Yang Liu, and Xuanjing Huang
- Proposed Method for Annotation of Scientific Arguments in Terms of Semantic Relations and Argument Schemes
  Nancy Green
- Using context to identify the language of face-saving
  Nona Naderi and Graeme Hirst

Session 3

14:30–15:10 **Special Presentation: Project Debater (Noam Slonim and Ranit Aharonov)**

15:10–15:30 Dave the debater: a retrieval-based and generative argumentative dialogue agent
Dieu-Thu Le, Cam Tu Nguyen, and Kim Anh Nguyen

15:30–16:00 **Coffee Break**

Session 4

16:00–16:20 PD3: Better Low-Resource Cross-Lingual Transfer By Combining Direct Transfer and Annotation Projection
Steffen Eger, Andreas Rücklé, and Iryna Gurevych

16:20–16:40 Cross-Lingual Argumentative Relation Identification: from English to Portuguese
Gil Rocha, Christian Stab, Henrique Lopes Cardoso, and Iryna Gurevych

16:40–17:00 More or less controlled elicitation of argumentative text: Enlarging a microtext corpus via crowdsourcing
Maria Skeppstedt, Andreas Peldszus, and Manfred Stede

17:00–17:30 **Best paper announcement and closing**

Workshop 13: Forth Workshop on Noisy User-generated Text

Organizers: *Alan Ritter, Leon Derczynski, Wei Xu, and Timothy Baldwin*

Location: Skt. Hans Torv

Thursday, November 1, 2018

9:00–9:05 **Opening**

9:05–9:50 **Invited Talk: Leon Derczynski**

9:50–10:35 **Oral Session I**

9:50–10:05 Inducing a lexicon of sociolinguistic variables from code-mixed text
Philippa Shoemark, James Kirby, and Sharon Goldwater

10:05–10:20 Twitter Geolocation using Knowledge-Based Methods
Taro Miyazaki, Afshin Rahimi, Trevor Cohn, and Timothy Baldwin

10:20–10:35 Geocoding Without Geotags: A Text-based Approach for reddit
Keith Harrigian

10:35–11:00 **Tea Break**

11:00–12:30 **Oral Session II**

11:00–11:15 Assigning people to tasks identified in email: The EPA dataset for addressee tagging for detected task intent
Revanth Rameshkumar, Peter Bailey, Abhishek Jha, and Chris Quirk

11:15–11:30 How do you correct run-on sentences it's not as easy as it seems
Junchao Zheng, Courtney Napoles, and Joel Tetreault

11:30–11:45 A POS Tagging Model Adapted to Learner English
Ryo Nagata, Tomoya Mizumoto, Yuta Kikuchi, Yoshifumi Kawasaki, and Kotaro Funakoshi

11:45–12:00 Normalization of Transliterated Words in Code-Mixed Data Using Seq2Seq Model & Levenshtein Distance
Soumil Mandal and Karthick Nanmaran

12:00–12:15 Robust Word Vectors: Context-Informed Embeddings for Noisy Texts
Valentin Malykh, Varvara Logacheva, and Taras Khakhulin

12:15–12:30 Paraphrase Detection on Noisy Subtitles in Six Languages
Eetu Sjöblom, Mathias Creutz, and Mikko Aulamo

12:30–14:00 **Lunch**

14:00–14:45 **Invited Talk: Diyi Yang**

14:45–15:15 **Lightning Talks**

- Distantly Supervised Attribute Detection from Reviews
  Lisheng Fu and Pablo Barrio
- Using Wikipedia Edits in Low Resource Grammatical Error Correction
  Adriane Boyd
- Empirical Evaluation of Character-Based Model on Neural Named-Entity Recognition in Indonesian Conversational Texts
  Kemal Kurniawan and Samuel Louvan

- Orthogonal Matching Pursuit for Text Classification
  Konstantinos Skianis, Nikolaos Tziortziotis, and Michalis Vazirgiannis
- Training and Prediction Data Discrepancies: Challenges of Text Classification with Noisy, Historical Data
  R. Andrew Kreek and Emilia Apostolova
- Detecting Code-Switching between Turkish-English Language Pair
  Zeynep Yirmibeşoğlu and Gülşen Eryiğit
- Language Identification in Code-Mixed Data using Multichannel Neural Networks and Context Capture
  Soumil Mandal and Anil Kumar Singh
- Modeling Student Response Times: Towards Efficient One-on-one Tutoring Dialogues
  Luciana Benotti, Jayadev Bhaskaran, Sigtryggur Kjartansson, and David Lang
- Content Extraction and Lexical Analysis from Customer-Agent Interactions
  Sergiu Nisioi, Anca Bucur, and Liviu P. Dinu
- Preferred Answer Selection in Stack Overflow: Better Text Representations ... and Metadata, Metadata, Metadata
  Steven Xu, Andrew Bennett, Doris Hoogeveen, Jey Han Lau, and Timothy Baldwin
- Word-like character n-gram embedding
  Geewook Kim, Kazuki Fukui, and Hidetoshi Shimodaira
- Classification of Tweets about Reported Events using Neural Networks
  Kiminobu Makino, Yuka Takei, Taro Miyazaki, and Jun Goto
- Learning to Define Terms in the Software Domain
  Vidhisha Balachandran, Dheeraj Rajagopal, Rose Catherine Kanjirathinkal, and William Cohen
- FrameIt: Ontology Discovery for Noisy User-Generated Text
  Dan Iter, Alon Halevy, and Wang-Chiew Tan
- Using Author Embeddings to Improve Tweet Stance Classification
  Adrian Benton and Mark Dredze
- Low-resource named entity recognition via multi-source projection: Not quite there yet?
  Jan Vium Enghoff, Søren Harrison, and Željko Agić
- A Case Study on Learning a Unified Encoder of Relations
  Lisheng Fu, Bonan Min, Thien Huu Nguyen, and Ralph Grishman
- Convolutions Are All You Need (For Classifying Character Sequences)
  Zach Wood-Doughty, Nicholas Andrews, and Mark Dredze
- MTNT: A Testbed for Machine Translation of Noisy Text
  Paul Michel and Graham Neubig
- A Robust Adversarial Adaptation for Unsupervised Word Translation
  Kazuma Hashimoto, Ehsan Hosseini-Asl, Caiming Xiong, and Richard Socher
- A Comparative Study of Embeddings Methods for Hate Speech Detection from Tweets
  Shashank Gupta and Zeerak Waseem
- Step or Not: Discriminator for The Real Instructions in User-generated Recipes
  Shintaro Inuzuka, Takahiko Ito, and Jun Harashima
- Named Entity Recognition on Noisy Data using Images and Text
  Diego Esteves
- Handling Noise in Distributional Semantic Models for Large Scale Text Analytics and Media Monitoring
  Peter Sumbler, Nina Viereckel, Nazanin Afsarmanesh, and Jussi Karlgren

- Combining Human and Machine Transcriptions on the Zooniverse Platform
  Daniel Hanson and Andrea Simenstad
- Predicting Good Twitter Conversations
  Zach Wood-Doughty, Prabhanjan Kambadur, and Gideon Mann
- Automated opinion detection analysis of online conversations
  Yuki M Asano, Niccolo Pescetelli, and Jonas Haslbeck
- Classification of Written Customer Requests: Dealing with Noisy Text and Labels
  Viljami Laurmaa and Mostafa Ajallooeian

15:15–16:30 **Poster Session**

16:30–17:15 **Invited Talk: Daniel Preoţiuc-Pietro**

17:15–17:30 **Closing and Best Paper Awards**

Workshop 14: Second Workshop on Universal Dependencies

Organizers: *Marie-Catherine de, Teresa Lynn, Sebastian Schuster, Joakim Nivre, Filip Ginter, Yoav Goldberg, Jan Hajic, Sampo Pyysalo, Reut Tsarfaty, Francis Tyers, and Daniel Zeman*

Location: Enghave Plads & Kødbyen

Thursday, November 1, 2018

8:45–9:00 **Opening Remarks**

9:00–10:00 **Invited Talk by John Doe**

9:00–10:00 **Session 1: Important Matters Unresolved**

Session 1: Important Matters Resolved

10:00–10:30 Assessing the Impact of Incremental Error Detection and Correction. A Case Study on the Italian Universal Dependency Treebank
Chiara Alzetta, Felice Dell'Orletta, Simonetta Montemagni, Maria Simi, and Giulia Venturi

10:00–10:30 Using Universal Dependencies in cross-linguistic complexity research
Aleksandrs Berdicevskis, Çağrı Çöltekin, Katharina Ehret, Kilu von Prince, Daniel Ross, Bill Thompson, Chunxiao Yan, Vera Demberg, Gary Lupyan, Taraka Rama, and Christian Bentz

10:00–10:30 Expletives in Universal Dependency Treebanks
Gosse Bouma, Jan Hajic, Dag Haug, Joakim Nivre, Per Erik Solberg, and Lilja Øvrelid

10:00–10:30 Challenges in Converting the Index Thomisticus Treebank into Universal Dependencies
Flavio Massimiliano Cecchini, Marco Passarotti, Paola Marongiu, and Daniel Zeman

10:00–10:30 Er ... well, it matters, right? On the role of data representations in spoken language dependency parsing
Kaja Dobrovoljc and Matej Martinc

10:00–10:30 Mind the Gap: Data Enrichment in Dependency Parsing of Elliptical Constructions
Kira Droganova, Filip Ginter, Jenna Kanerva, and Daniel Zeman

10:00–10:30 Integration complexity and the order of cosisters
William Dyer

10:00–10:30 SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD
Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier

10:00–10:30 Parsing Japanese Tweets into Universal Dependencies
Hayate Iso, Kaoru Ito, Hiroyuki Nagai, Taro Okahisa, and Eiji Aramaki

10:00–10:30 Coordinate Structures in Universal Dependencies for Head-final Languages
Hiroshi Kanayama, Na-Rae Han, Masayuki Asahara, Jena D. Hwang, Yusuke Miyao, Jinho D. Choi, and Yuji Matsumoto

10:00–10:30 Investigating NP-Chunking with Universal Dependencies for English
Ophélie Lacroix

10:00–10:30 Marrying Universal Dependencies and Universal Morphology
Arya D. McCarthy, Miikka Silfverberg, Ryan Cotterell, Mans Hulden, and David Yarowsky

10:00–10:30 Enhancing Universal Dependency Treebanks: A Case Study
Joakim Nivre, Paola Marongiu, Filip Ginter, Jenna Kanerva, Simonetta Montemagni, Sebastian Schuster, and Maria Simi

10:00–10:30 Enhancing Universal Dependencies for Korean
Youngbin Noh, Jiyoon Han, Tae Hwan Oh, and Hansaem Kim

10:00–10:30 UD-Japanese BCCWJ: Universal Dependencies Annotation for the Balanced Corpus of Contemporary Written Japanese
Mai Omura and Masayuki Asahara

10:00–10:30 The First Komi-Zyrian Universal Dependencies Treebanks
Niko Partanen, Rogier Blokland, KyungTae Lim, Thierry Poibeau, and Michael Rießler

10:00–10:30 All Roads Lead to UD: Converting Stanford and Penn Parses to English Universal Dependencies with Multilayer Annotations
Siyao Peng and Amir Zeldes

10:00–10:30 Arguments and Adjuncts in Universal Dependencies
Adam Przepiórkowski and Agnieszka Patejuk

10:00–10:30 From LFG to Enhanced Universal Dependencies
Adam Przepiórkowski and Agnieszka Patejuk

10:00–10:30 The Hebrew Universal Dependency Treebank: Past Present and Future
Shoval Sade, Amit Seker, and Reut Tsarfaty

10:00–10:30 Multi-source synthetic treebank creation for improved cross-lingual dependency parsing
Francis Tyers, Mariya Sheyanova, Aleksandra Martynova, Pavel Stepachev, and Konstantin Vinogorodskiy

10:00–10:30 Toward Universal Dependencies for Shipibo-Konibo
Alonso Vásquez, Renzo Ego Aguirre, Candy Angulo, John Miller, Claudia Villanueva, Željko Agić, Roberto Zariquiey, and Arturo Oncevay

10:00–10:30 Transition-based Parsing with Lighter Feed-Forward Networks
David Vilares and Carlos Gómez-Rodríguez

10:00–10:30 Extended and Enhanced Polish Dependency Bank in Universal Dependencies Format
Alina Wróblewska

10:00–10:30 Approximate Dynamic Oracle for Dependency Parsing with Reinforcement Learning
Xiang Yu, Ngoc Thang Vu, and Jonas Kuhn

10:00–10:30 The Coptic Universal Dependency Treebank
Amir Zeldes and Mitchell Abrams

5

Main Conference: Friday, November 2

Overview

09:00 – 09:30 **Opening remarks (Gold Hall)** *Gold Hall*

09:30 – 10:30 **Keynote I: Julia Hirschberg "Truth or Lie? Spoken Indicators of Deception in Speech" (Gold Hall)** *Gold Hall*

10:30 – 11:00 **Coffee Break** *Multiple levels*

Session 1

11:00 – 12:30				
	Social Applications I *Gold Hall*	Semantics I *Copper Hall*	Vision *Silver Hall*	Entities and Coreference *Hall 100*

11:00 – 12:30 **[Posters and Demos]: Machine Translation and Multilingual Methods** *Grand Hall 2*

12:30 – 13:45 **Lunch**

Session 2

13:45 – 14:45				
	Question Answering I *Gold Hall*	Semantics II *Copper Hall*	Multilingual Methods I *Silver Hall*	Social Media *Hall 100*

13:45 – 14:45 **[Posters and Demos]: Short Posters I** *Grand Hall 2*

14:45 – 15:00 **Mini-Break** *Multiple levels*

Session 3

15:00 – 16:00				
	Machine Translation I *Gold Hall*	Machine Learning I *Copper Hall*	Semantic Parsing / Generation *Silver Hall*	Vision / Discourse *Hall 100*

15:00 – 16:00 **[Posters and Demos]: Short Posters II** *Grand Hall 2*

16:00 – 16:30 **Coffee Break** *Multiple levels*

Session 4

16:30 – 18:00				
	Language Models *Gold Hall*	Information Extraction *Copper Hall*	Syntactic Parsing *Silver Hall*	Visual QA *Hall 100*

16:30 – 18:00 **[Posters and Demos]: Semantics III** *Grand Hall 2*

Invited Talk: Johan Bos

The Moment of Meaning and the Future of Computational Semantics

Friday, November 2, 2018, 9:00–10:00

Gold Hall

Abstract: There are many recent advances in semantic parsing: we see a rising number of semantically annotated corpora and there is exciting technology (such as neural networks) to be explored. In this talk I will discuss what role computational semantics could play in future natural language processing applications (including fact checking and machine translation). I will argue that we should not just look at semantic parsing, but that things can get really interesting when we can use language-neutral meaning representations to draw (transparent) inferences. The main ideas will be exemplified by the parallel meaning bank, a new corpus comprising texts annotated with formal meaning representations for English, Dutch, German and Italian.

Biography: Johan Bos is Professor of Computational Semantics at the University of Groningen (Netherlands). He received his doctorate from the Computational Linguistics Department at the University of the Saarland (Germany) and held post-doc positions at the University of Edinburgh (UK) and the La Sapienza University in Rome (Italy). In 2010, he moved to his current position in Groningen, leading the computational semantics group. Bos is the developer of Boxer, a state-of-the-art wide-coverage semantic parser for English, initiator of the Groningen Meaning Bank, a large semantically-annotated corpus of texts, and inventor of Wordrobe, a game with a purpose for semantic annotation. Bos received a $1.5-million Vici grant from NWO (Netherlands Organisation for Scientific Research) in 2015 to investigate the role of meaning in human and machine translation. A concrete outcome of this project is the Parallel Meaning Bank containing detailed meaning representations for English, German, Dutch and Italian sentences.

Session 1 Overview – Friday, November 2, 2018

Oral tracks

Track A *Social Applications I* Gold Hall	Track B *Semantics I* Copper Hall	Track C *Vision* Silver Hall	Track D *Entities and Coreference* Hall 100	
Privacy-preserving Neural Representations of Text *Coavoux, Narayan, and Cohen*	Process Paragraph Comprehension using World Knowledge *Tandon, Dalvi, Grus, Yih, Bosselut, and Clark*	Associative Multichannel Autoencoder for Multimodal Word Representation *Wang, Zhang, and Zong*	PreCo: A Large-scale Dataset in Preschool Vocabulary for Coreference Resolution *Chen, Fan, Lu, Yuille, and Rong*	11:00
Adversarial Removal of Demographic Attributes from Text Data *Elazar and Goldberg*	Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation *Poliak, Haldar, Rudinger, Hu, Pavlick, White, and Van Durme*	Modeling Game-Based Video-Context Dialogue *Pasunuru and Bansal*	Adversarial Transfer Learning for Chinese Named Entity Recognition with Self-Attention Mechanism *Cao, Chen, Liu, Zhao, and Liu*	11:18
DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning *Popat, Mukherjee, Yates, and Weikum*	Textual Analogy Parsing: What's Shared and What's Compared among Analogous Facts *Lamm, Chaganty, Manning, Jurafsky, and Liang*	simNet: Stepwise Image-Topic Merging Network for Generating Detailed and Comprehensive Image Captions *Liu, Ren, LIU, WANG, and SUN*	Using Linguistic Features to Improve Generalization in Neural Coreference Resolvers *Moosavi and Strube*	11:36
It's going to be okay: Measuring Access to Support in Online Communities *Wang and Jurgens*	SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference *Zellers, Bisk, Schwartz, and Choi*	Multimodal Language Analysis with Recurrent Multistage Fusion *Liang, Liu, Bagher Zadeh, and Morency*	Neural Segmental Hypergraphs for Overlapping Mention Recognition *Wang and Lu*	11:54
Detecting Gang-Involved Escalation on Social Media Using Context *Chang, Zhong, Adams, Lee, Varia, Patton, Frey, Kedzie, and McKeown*	TwoWingOS: Claim Entailment with Precise Evidence via a Two-Wing Optimization Strategy *Yin and Roth*	Temporally Grounding Natural Sentence in Video *Chen, Chen, Ma, Jie, and Chua*	Variational Sequential Labelers for Semi-Supervised Sequence Labeling *Chen, Tang, Livescu, and Gimpel*	12:12

Poster tracks

11:00–12:30

Track D: *Machine Translation and Multilingual Methods* Hall 100

Parallel Session 1

Session 1A: Social Applications I

Gold Hall Chair: *chairname*

Privacy-preserving Neural Representations of Text
Maximin Coavoux, Shashi Narayan, and Shay B. Cohen 11:00–11:18

This article deals with adversarial attacks towards deep learning systems for Natural Language Processing (NLP), in the context of privacy protection. We study a specific type of attack: an attacker eavesdrops on the hidden representations of a neural text classifier and tries to recover information about the input text. Such scenario may arise in situations when the computation of a neural network is shared across multiple devices, e.g. some hidden representation is computed by a user's device and sent to a cloud-based model. We measure the privacy of a hidden representation by the ability of an attacker to predict accurately specific private information from it and characterize the tradeoff between the privacy and the utility of neural representations. Finally, we propose several defense methods based on modified training objectives and show that they improve the privacy of neural representations.

Adversarial Removal of Demographic Attributes from Text Data
Yanai Elazar and Yoav Goldberg 11:18–11:36

Recent advances in Representation Learning and Adversarial Training seem to succeed in removing unwanted features from the learned representation. We show that demographic information of authors is encoded in—and can be recovered from—the intermediate representations learned by text-based neural classifiers. The implication is that decisions of classifiers trained on textual data are not agnostic to—and likely condition on—demographic attributes. When attempting to remove such demographic information using adversarial training, we find that while the adversarial component achieves chance-level development-set accuracy during training, a post-hoc classifier, trained on the encoded sentences from the first part, still manages to reach substantially higher classification accuracies on the same data. This behavior is consistent across several tasks, demographic properties and datasets. We explore several techniques to improve the effectiveness of the adversarial component. Our main conclusion is a cautionary one: do not rely on the adversarial training to achieve invariant representation to sensitive features.

DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning
Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum 11:36–11:54

Misinformation such as fake news is one of the big challenges of our society. Research on automated fact-checking has proposed methods based on supervised learning, but these approaches do not consider external evidence apart from labeled training instances. Recent approaches counter this deficit by considering external sources related to a claim. However, these methods require substantial feature modeling and rich lexicons. This paper overcomes these limitations of prior work with an end-to-end model for evidence-aware credibility assessment of arbitrary textual claims, without any human intervention. It presents a neural network model that judiciously aggregates signals from external evidence articles, the language of these articles and the trustworthiness of their sources. It also derives informative features for generating user-comprehensible explanations that makes the neural network predictions transparent to the end-user. Experiments with four datasets and ablation studies show the strength of our method.

It's going to be okay: Measuring Access to Support in Online Communities
Zijian Wang and David Jurgens 11:54–12:12

People use online platforms to seek out support for their informational and emotional needs. Here, we ask what effect does revealing one's gender have on receiving support. To answer this, we create (i) a new dataset and method for identifying supportive replies and (ii) new methods for inferring gender from text and name. We apply these methods to create a new massive corpus of 102M online interactions with gender-labeled users, each rated by degree of supportiveness. Our analysis shows wide-spread and consistent disparity in support: identifying as a woman is associated with higher rates of support - but also higher rates of disparagement.

Detecting Gang-Involved Escalation on Social Media Using Context
Serina Chang, Ruiqi Zhong, Ethan Adams, Fei-Tzin Lee, Siddharth Varia, Desmond Patton, William Frey, Chris Kedzie, and Kathy McKeown 12:12–12:30

Gang-involved youth in cities such as Chicago have increasingly turned to social media to post about their experiences and intents online. In some situations, when they experience the loss of a loved one, their online expression of emotion may evolve into aggression towards rival gangs and ultimately into real-world violence. In this paper, we present a novel system for detecting Aggression and Loss in social media. Our system features the use of domain-specific resources automatically derived from a large unlabeled corpus, and contextual representations of the emotional and semantic content of the user's recent tweets as well as their interactions with other users. Incorporating context in our Convolutional Neural Network (CNN) leads to a significant improvement.

Session 1B: Semantics I

Copper Hall Chair: *chairname*

Process Paragraph Comprehension using World Knowledge

Niket Tandon, Bhavana Dalvi, Joel Grus, Wen-tau Yih, Antoine Bosselut, and Peter Clark 11:00–11:18

Comprehending procedural text, e.g., a paragraph describing photosynthesis, requires modeling actions and the state changes they produce, so that questions about entities at different timepoints can be answered. Although several recent systems have shown impressive progress in this task, their predictions can be globally inconsistent or highly improbable. In this paper, we show how the predicted effects of actions in the context of a paragraph can be improved in two ways: (1) by incorporating global, commonsense constraints (e.g., a non-existent entity cannot be destroyed), and (2) by biasing reading with preferences from large-scale corpora (e.g., trees rarely move). Unlike earlier methods, we treat the problem as a neural structured prediction task, allowing hard and soft constraints to steer the model away from unlikely predictions. We show that the new model significantly outperforms earlier systems on a benchmark dataset for procedural text comprehension (+8% relative gain), and that it also avoids some of the nonsensical predictions that earlier systems make.

Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme 11:18–11:36

We present a large-scale collection of diverse natural language inference (NLI) datasets that help provide insight into how well a sentence representation captures distinct types of reasoning. The collection results from recasting 13 existing datasets from 7 semantic phenomena into a common NLI structure, resulting in over half a million labeled context-hypothesis pairs in total. We refer to our collection as the DNC: Diverse Natural Language Inference Collection. The DNC is available online at https://www.decomp.net, and will grow over time as additional resources are recast and added from novel sources.

Textual Analogy Parsing: What's Shared and What's Compared among Analogous Facts

Matthew Lamm, Arun Chaganty, Christopher D. Manning, Dan Jurafsky, and Percy Liang 11:36–11:54

To understand a sentence like “whereas only 10% of White Americans live at or below the poverty line, 28% of African Americans do” it is important not only to identify individual facts, e.g., poverty rates of distinct demographic groups, but also the higher-order relations between them, e.g., the disparity between them. In this paper, we propose the task of Textual Analogy Parsing (TAP) to model this higher-order meaning. Given a sentence such as the one above, TAP outputs a frame-style meaning representation which explicitly specifies what is shared (e.g., poverty rates) and what is compared (e.g., White Americans vs. African Americans, 10% vs. 28%) between its component facts. Such a meaning representation can enable new applications that rely on discourse understanding such as automated chart generation from quantitative text. We present a new dataset for TAP, baselines, and a model that successfully uses an ILP to enforce the structural constraints of the problem.

SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi 11:54–12:12

Given a partial description like “she opened the hood of the car,” humans can reason about the situation and anticipate what might come next (”then, she examined the engine”). In this paper, we introduce the task of grounded commonsense inference, unifying natural language inference and commonsense reasoning. We present SWAG, a new dataset with 113k multiple choice questions about a rich spectrum of grounded situations. To address the recurring challenges of the annotation artifacts and human biases found in many existing datasets, we propose Adversarial Filtering (AF), a novel procedure that constructs a de-biased dataset by iteratively training an ensemble of stylistic classifiers, and using them to filter the data. To account for the aggressive adversarial filtering, we use state-of-the-art language models to massively oversample a diverse set of potential counterfactuals. Empirical results demonstrate that while humans can solve the resulting inference problems with high accuracy (88%), various competitive models struggle on our task. We provide comprehensive analysis that indicates significant opportunities for future research.

TwoWingOS: Claim Entailment with Precise Evidence via a Two-Wing Optimization Strategy

Wenpeng Yin and Dan Roth 12:12–12:30

Determining whether a given claim is supported by evidence is a fundamental NLP problem that is best modeled as Textual Entailment. However, given a large collection of text, finding evidence that could support or refute a given claim is a challenge in itself, amplified by the fact that different evidence might be needed to support or refute a claim. Nevertheless, most prior work decouples evidence finding from determining the truth value of the claim given the evidence. We propose to consider these two aspects jointly. We develop TwoWingOS (two-wing optimization strategy), a system that, while identifying appropriate evidence for a claim, also determines whether or not the claim is supported by the evidence. Given the claim, TwoWingOS attempts to identify a subset of the evidence candidates; given the predicted evidence, it then attempts to determine the truth value of the corresponding claim entailment problem. We treat this problem as coupled optimization problems, training a joint model for it. TwoWingOS offers two advantages: (i) Unlike pipeline systems it facilitates flexible-size evidence set, and (ii) Joint training improves both the claim entailment and the evidence identification. Experiments on a benchmark dataset show state-of-the-art performance.

Session 1C: Vision

Silver Hall Chair: *chairname*

Associative Multichannel Autoencoder for Multimodal Word Representation
Shaonan Wang, Jiajun Zhang, and Chengqing Zong 11:00–11:18

In this paper we address the problem of learning multimodal word representations by integrating textual, visual and auditory inputs. Inspired by the re-constructive and associative nature of human memory, we propose a novel associative multichannel autoencoder (AMA). Our model first learns the associations between textual and perceptual modalities, so as to predict the missing perceptual information of concepts. Then the textual and predicted perceptual representations are fused through reconstructing their original and associated embeddings. Using a gating mechanism our model assigns different weights to each modality according to the different concepts. Results on six benchmark concept similarity tests show that the proposed method significantly outperforms strong unimodal baselines and state-of-the-art multi-modal models.

Modeling Game-Based Video-Context Dialogue
Ramakanth Pasunuru and Mohit Bansal 11:18–11:36

Current dialogue systems focus more on textual and speech context knowledge and are usually based on two speakers. Some recent work has investigated static image-based dialogue. However, several real-world human interactions also involve dynamic visual context (similar to videos) as well as dialogue exchanges among multiple speakers. To move closer towards such multimodal conversational skills and visually-situated applications, we introduce a new video-context, many-speaker dialogue dataset based on live-broadcast soccer game videos and chats from Twitch.tv. This challenging testbed allows us to develop visually-grounded dialogue models that should generate relevant temporal and spatial event language from the live video, while also being relevant to the chat history. For strong baselines, we also present several discriminative and generative models, e.g., based on tridirectional attention flow (TriDAF). We evaluate these models via retrieval ranking-recall, automatic phrase-matching metrics, as well as human evaluation studies. We also present dataset analyses, model ablations, and visualizations to understand the contribution of different modalities and model components.

simNet: Stepwise Image-Topic Merging Network for Generating Detailed and Comprehensive Image Captions
Fenglin Liu, Xuancheng Ren, Yuanxin LIU, Houfeng WANG, and Xu SUN 11:36–11:54

The encode-decoder framework has shown recent success in image captioning. Visual attention, which is good at detailedness, and semantic attention, which is good at comprehensiveness, have been separately proposed to ground the caption on the image. In this paper, we propose the Stepwise Image-Topic Merging Network (simNet) that makes use of the two kinds of attention at the same time. At each time step when generating the caption, the decoder adaptively merges the attentive information in the extracted topics and the image according to the generated context, so that the visual information and the semantic information can be effectively combined. The proposed approach is evaluated on two benchmark datasets and reaches the state-of-the-art performances.

Multimodal Language Analysis with Recurrent Multistage Fusion
Paul Pu Liang, Ziyin Liu, AmirAli Bagher Zadeh, and Louis-Philippe Morency 11:54–12:12

Computational modeling of human multimodal language is an emerging research area in natural language processing spanning the language, visual and acoustic modalities. Comprehending multimodal language requires modeling not only the interactions within each modality (intra-modal interactions) but more importantly the interactions between modalities (cross-modal interactions). In this paper, we propose the Recurrent Multistage Fusion Network (RMFN) which decomposes the fusion problem into multiple stages, each of them focused on a subset of multimodal signals for specialized, effective fusion. Cross-modal interactions are modeled using this multistage fusion approach which builds upon intermediate representations of previous stages. Temporal and intra-modal interactions are modeled by integrating our proposed fusion approach with a system of recurrent neural networks. The RMFN displays state-of-the-art performance in modeling human multimodal language across three public datasets relating to multimodal sentiment analysis, emotion recognition, and speaker traits recognition. We provide visualizations to show that each stage of fusion focuses on a different subset of multimodal signals, learning increasingly discriminative multimodal representations.

Temporally Grounding Natural Sentence in Video
Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua 12:12–12:30

We introduce an effective and efficient method that grounds (i.e., localizes) natural sentences in long, untrimmed video sequences. Specifically, a novel Temporal GroundNet (TGN) is proposed to temporally capture the evolving fine-grained frame-by-word interactions between video and sentence. TGN sequentially scores a set of temporal candidates ended at each frame based on the exploited frame-by-word interactions, and finally grounds the segment corresponding to the sentence. Unlike traditional methods treating the overlapping segments separately in a sliding window fashion, TGN aggregates the historical information and generates the final grounding result in one single pass. We extensively evaluate our proposed TGN on three public datasets with significant improvements over the state-of-the-arts. We further show the consistent effectiveness and efficiency of TGN through an ablation study and a runtime test.

Session 1D: Entities and Coreference

Hall 100 Chair: *chairname*

PreCo: A Large-scale Dataset in Preschool Vocabulary for Coreference Resolution
Hong Chen, Zhenhua Fan, Hao Lu, Alan Yuille, and Shu Rong 11:00–11:18

We introduce PreCo, a large-scale English dataset for coreference resolution. The dataset is designed to embody the core challenges in coreference, such as entity representation, by alleviating the challenge of low overlap between training and test sets and enabling separated analysis of mention detection and mention clustering. To strengthen the training-test overlap, we collect a large corpus of 38K documents and 12.5M words which are mostly from the vocabulary of English-speaking preschoolers. Experiments show that with higher training-test overlap, error analysis on PreCo is more efficient than the one on OntoNotes, a popular existing dataset. Furthermore, we annotate singleton mentions making it possible for the first time to quantify the influence that a mention detector makes on coreference resolution performance. The dataset is freely available at https://preschool-lab.github.io/PreCo/.

Adversarial Transfer Learning for Chinese Named Entity Recognition with Self-Attention Mechanism
Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, and Shengping Liu 11:18–11:36

Named entity recognition (NER) is an important task in natural language processing area, which needs to determine entities boundaries and classify them into pre-defined categories. For Chinese NER task, there is only a very small amount of annotated data available. Chinese NER task and Chinese word segmentation (CWS) task have many similar word boundaries. There are also specificities in each task. However, existing methods for Chinese NER either do not exploit word boundary information from CWS or cannot filter the specific information of CWS. In this paper, we propose a novel adversarial transfer learning framework to make full use of task-shared boundaries information and prevent the task-specific features of CWS. Besides, since arbitrary character can provide important cues when predicting entity type, we exploit self-attention to explicitly capture long range dependencies between two tokens. Experimental results on two different widely used datasets show that our proposed model significantly and consistently outperforms other state-of-the-art methods.

Using Linguistic Features to Improve Generalization in Neural Coreference Resolvers
Nafise Sadat Moosavi and Michael Strube 11:36–11:54

Coreference resolution is an intermediate step for text understanding. It is used in tasks and domains for which we do not necessarily have coreference annotated corpora. Therefore, generalization is of special importance for coreference resolution. However, while recent coreference resolvers have notable improvements on the CoNLL dataset, they struggle to generalize properly to new domains or datasets. In this paper, we investigate the role of linguistic features in building more generalizable coreference resolvers. We show that generalization improves only slightly by merely using a set of additional linguistic features. However, employing features and subsets of their values that are informative for coreference resolution, considerably improves generalization. Thanks to better generalization, our system achieves state-of-the-art results in out-of-domain evaluations, e.g., on WikiCoref, our system, which is trained on CoNLL, achieves on-par performance with a system designed for this dataset.

Neural Segmental Hypergraphs for Overlapping Mention Recognition
Bailin Wang and Wei Lu 11:54–12:12

In this work, we propose a novel segmental hypergraph representation to model overlapping entity mentions that are prevalent in many practical datasets. We show that our model built on top of such a new representation is able to capture features and interactions that cannot be captured by previous models while maintaining a low time complexity for inference. We also present a theoretical analysis to formally assess how our representation is better than alternative representations reported in the literature in terms of representational power. Coupled with neural networks for feature learning, our model achieves the state-of-the-art performance in three benchmark datasets annotated with overlapping mentions.

Variational Sequential Labelers for Semi-Supervised Sequence Labeling
Mingda Chen, Qingming Tang, Karen Livescu, and Kevin Gimpel 12:12–12:30

We introduce a family of multitask variational methods for semi-supervised sequence labeling. Our model family consists of a latent-variable generative model and a discriminative labeler. The generative models use latent variables to define the conditional probability of a word given its context, drawing inspiration from word prediction objectives commonly used in learning word embeddings. The labeler helps inject discriminative information into the latent space. We explore several latent variable configurations, including ones with hierarchical structure, which enables the model to account for both label-specific and word-specific information. Our models consistently outperform standard sequential baselines on 8 sequence labeling datasets, and improve further with unlabeled data.

Session 1E: Machine Translation and Multilingual Methods 11:00–12:30

Hall 100 Chair: *chairname*

Joint Representation Learning of Cross-lingual Words and Entities via Attentive Distant Supervision

Yixin Cao, Lei Hou, Juanzi Li, Zhiyuan Liu, Chengjiang Li, Xu Chen, and Tiansi Dong

Jointly representation learning of words and entities benefits many NLP tasks, but has not been well explored in cross-lingual settings. In this paper, we propose a novel method for joint representation learning of cross-lingual words and entities. It captures mutually complementary knowledge, and enables cross-lingual inferences among knowledge bases and texts. Our method does not require parallel corpus, and automatically generates comparable data via distant supervision using multi-lingual knowledge bases. We utilize two types of regularizers to align cross-lingual words and entities, and design knowledge attention and cross-lingual attention to further reduce noises. We conducted a series of experiments on three tasks: word translation, entity relatedness, and cross-lingual entity linking. The results, both qualitative and quantitative, demonstrate the significance of our method.

Deep Pivot-Based Modeling for Cross-language Cross-domain Transfer with Minimal Guidance

Yftah Ziser and Roi Reichart

While cross-domain and cross-language transfer have long been prominent topics in NLP research, their combination has hardly been explored. In this work we consider this problem, and propose a framework that builds on pivot-based learning, structure-aware Deep Neural Networks (particularly LSTMs and CNNs) and bilingual word embeddings, with the goal of training a model on labeled data from one (language, domain) pair so that it can be effectively applied to another (language, domain) pair. We consider two setups, differing with respect to the unlabeled data available for model training. In the full setup the model has access to unlabeled data from both pairs, while in the lazy setup, which is more realistic for truly resource-poor languages, unlabeled data is available for both domains but only for the source language. We design our model for the lazy setup so that for a given target domain, it can train once on the source language and then be applied to any target language without re-training. In experiments with nine English-German and nine English-French domain pairs our best model substantially outperforms previous models even when it is trained in the lazy setup and previous models are trained in the full setup.

Multi-lingual Common Semantic Space Construction via Cluster-Consistent Word Embedding

Lifu Huang, Kyunghyun Cho, Boliang Zhang, Heng Ji, and Kevin Knight

We construct a multilingual common semantic space based on distributional semantics, where words from multiple languages are projected into a shared space via which all available resources and knowledge can be shared across multiple languages. Beyond word alignment, we introduce multiple cluster-level alignments and enforce the word clusters to be consistently distributed across multiple languages. We exploit three signals for clustering: (1) neighbor words in the monolingual word embedding space; (2) character-level information; and (3) linguistic properties (e.g., apposition, locative suffix) derived from linguistic structure knowledge bases available for thousands of languages. We introduce a new cluster-consistent correlational neural network to construct the common semantic space by aligning words as well as clusters. Intrinsic evaluation on monolingual and multilingual QVEC tasks shows our approach achieves significantly higher correlation with linguistic features which are extracted from manually crafted lexical resources than state-of-the-art multi-lingual embedding learning methods do. Using low-resource language name tagging as a case study for extrinsic evaluation, our approach achieves up to 14.6% absolute F-score gain over the state of the art on cross-lingual direct transfer. Our approach is also shown to be robust even when the size of bilingual dictionary is small.

Unsupervised Multilingual Word Embeddings

Xilun Chen and Claire Cardie

Multilingual Word Embeddings (MWEs) represent words from multiple languages in a single distributional vector space. Unsupervised MWE (UMWE) methods acquire multilingual embeddings without cross-lingual supervision, which is a significant advantage over traditional supervised approaches and opens many new possibilities for low-resource languages. Prior art for learning UMWEs, however, merely relies on a number of independently trained Unsupervised Bilingual Word Embeddings (UBWEs) to obtain multilingual embeddings. These methods fail to leverage the interdependencies that exist among many languages. To address this shortcoming, we propose a fully unsupervised framework for learning MWEs that directly exploits the relations between all language pairs. Our model substantially outperforms previous approaches in the experiments on multilingual word translation and cross-lingual word similarity. In addition, our model even beats supervised approaches trained with cross-lingual resources.

Crosslingual Sense Embeddings in One Space
Ta Chung Chi and Yun-Nung Chen

This paper proposes a modularized sense induction and representation learning model that jointly learns bilingual sense embeddings that align well in the vector space, where the cross-lingual signal in the English-Chinese parallel corpus is exploited to capture the collocation and distributed characteristics in the language pair. The model is evaluated on the Stanford Contextual Word Similarity (SCWS) dataset to ensure the quality of monolingual sense embeddings. In addition, we introduce Bilingual Contextual Word Similarity (BCWS), a large and high-quality dataset for evaluating cross-lingual sense embeddings, which is the first attempt of measuring whether the learned embeddings are indeed aligned well in the vector space. The proposed approach shows the superior quality of sense embeddings evaluated in both monolingual and bilingual spaces.

Adversarial Propagation and Zero-Shot Cross-Lingual Transfer of Word Vector Specialization
Edoardo Maria Ponti, Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen

Semantic specialization is a process of fine-tuning pre-trained distributional word vectors using external lexical knowledge (e.g., WordNet) to accentuate a particular semantic relation in the specialized vector space. While post-processing specialization methods are applicable to arbitrary distributional vectors, they are limited to updating only the vectors of words occurring in external lexicons (i.e., seen words), leaving the vectors of all other words unchanged. We propose a novel approach to specializing the full distributional vocabulary. Our adversarial post-specialization method propagates the external lexical knowledge to the full distributional space. We exploit words seen in the resources as training examples for learning a global specialization function. This function is learned by combining a standard L2-distance loss with a adversarial loss: the adversarial component produces more realistic output vectors. We show the effectiveness and robustness of the proposed method across three languages and on three tasks: word similarity, dialog state tracking, and lexical simplification. We report consistent improvements over distributional word vectors and vectors specialized by other state-of-the-art specialization frameworks. Finally, we also propose a cross-lingual transfer method for zero-shot specialization which successfully specializes a full target distributional space without any lexical knowledge in the target language and without any bilingual data.

Improving Cross-Lingual Word Embeddings by Meeting in the Middle
Yerai Doval, Jose Camacho-Collados, Luis Espinosa Anke, and Steven Schockaert

Cross-lingual word embeddings are becoming increasingly important in multilingual NLP. Recently, it has been shown that these embeddings can be effectively learned by aligning two disjoint monolingual vector spaces through linear transformations, using no more than a small bilingual dictionary as supervision. In this work, we propose to apply an additional transformation after the initial alignment step, which moves cross-lingual synonyms towards a middle point between them. By applying this transformation our aim is to obtain a better cross-lingual integration of the vector spaces. In addition, and perhaps surprisingly, the monolingual spaces also improve by this transformation. This is in contrast to the original alignment, which is typically learned such that the structure of the monolingual spaces is preserved. Our experiments confirm that the resulting cross-lingual embeddings outperform state-of-the-art models in both monolingual and cross-lingual evaluation tasks.

AtomicWikiEdits: A Multilingual Corpus of Wikipedia Edits for Modeling Language and Discourse
Manaal Faruqui, Ellie Pavlick, Ian Tenney, and Dipanjan Das

We release a corpus of 43 million atomic edits across 8 languages. These edits are mined from Wikipedia edit history and consist of instances in which a human editor has inserted a single contiguous phrase into, or deleted a single contiguous phrase from, an existing sentence. We use the collected data to show that the language generated during editing differs from the language that we observe in standard corpora, and that models trained on edits encode different aspects of semantics and discourse than models trained on raw text. We release the full corpus as a resource to aid ongoing research in semantics, discourse, and representation learning.

On the Relation between Linguistic Typology and (Limitations of) Multilingual Language Modeling

Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen

A key challenge in cross-lingual NLP is developing general language-independent architectures that are equally applicable to any language. However, this ambition is largely hampered by the variation in structural and semantic properties, i.e. the typological profiles of the world's languages. In this work, we analyse the implications of this variation on the language modeling (LM) task. We present a large-scale study of state-of-the art n-gram based and neural language models on 50 typologically diverse languages covering a wide variety of morphological systems. Operating in the full vocabulary LM setup focused on word-level prediction, we demonstrate that a coarse typology of morphological systems is predictive of absolute LM performance. Moreover, fine-grained typological features such as exponence, flexivity, fusion, and inflectional synthesis are borne out to be responsible for the proliferation of low-frequency phenomena which are organically difficult to model by statistical architectures, or for the meaning ambiguity of character n-grams. Our study strongly suggests that these features have to be taken into consideration during the construction of next-level language-agnostic LM architectures, capable of handling morphologically complex languages such as Tamil or Korean.

A Fast, Compact, Accurate Model for Language Identification of Codemixed Text

Yuan Zhang, Jason Riesa, Daniel Gillick, Anton Bakalov, Jason Baldridge, and David Weiss

We address fine-grained multilingual language identification: providing a language code for every token in a sentence, including codemixed text containing multiple languages. Such text is prevalent online, in documents, social media, and message boards. We show that a feed-forward network with a simple globally constrained decoder can accurately and rapidly label both codemixed and monolingual text in 100 languages and 100 language pairs. This model outperforms previously published multilingual approaches in terms of both accuracy and speed, yielding an 800x speed-up and a 19.5% averaged absolute gain on three codemixed datasets. It furthermore outperforms several benchmark systems on monolingual language identification.

Personalized Microblog Sentiment Classification via Adversarial Cross-lingual learning

Weichao Wang, Shi Feng, Wei Gao, Daling Wang, and Yifei Zhang

Sentiment expression in microblog posts can be affected by user's personal character, opinion bias, political stance and so on. Most of existing personalized microblog sentiment classification methods suffer from the insufficiency of discriminative tweets for personalization learning. We observed that microblog users have consistent individuality and opinion bias in different languages. Based on this observation, in this paper we propose a novel user-attention-based Convolutional Neural Network (CNN) model with adversarial cross-lingual learning framework. The user attention mechanism is leveraged in CNN model to capture user's language-specific individuality from the posts. Then the attention-based CNN model is incorporated into a novel adversarial cross-lingual learning framework, in which with the help of user properties as bridge between languages, we can extract the language-specific features and language-independent features to enrich the user post representation so as to alleviate the data insufficiency problem. Results on English and Chinese microblog datasets confirm that our method outperforms state-of-the-art baseline algorithms with large margins.

Cross-lingual Knowledge Graph Alignment via Graph Convolutional Networks

Zhichun Wang, Qingsong Lv, Xiaohan Lan, and Yu Zhang

Multilingual knowledge graphs (KGs) such as DBpedia and YAGO contain structured knowledge of entities in several distinct languages, and they are useful resources for cross-lingual AI and NLP applications. Cross-lingual KG alignment is the task of matching entities with their counterparts in different languages, which is an important way to enrich the cross-lingual links in multilingual KGs. In this paper, we propose a novel approach for cross-lingual KG alignment via graph convolutional networks (GCNs). Given a set of pre-aligned entities, our approach trains GCNs to embed entities of each language into a unified vector space. Entity alignments are discovered based on the distances between entities in the embedding space. Embeddings can be learned from both the structural and attribute information of entities, and the results of structure embedding and attribute embedding are combined to get accurate alignments. In the experiments on aligning real multilingual KGs, our approach gets the best performance compared with other embedding-based KG alignment approaches.

Cross-lingual Lexical Sememe Prediction
Fanchao Qi, Yankai Lin, Maosong Sun, Hao Zhu, Ruobing Xie, and Zhiyuan Liu

Sememes are defined as the minimum semantic units of human languages. As important knowledge sources, sememe-based linguistic knowledge bases have been widely used in many NLP tasks. However, most languages still do not have sememe-based linguistic knowledge bases. Thus we present a task of cross-lingual lexical sememe prediction, aiming to automatically predict sememes for words in other languages. We propose a novel framework to model correlations between sememes and multi-lingual words in low-dimensional semantic space for sememe prediction. Experimental results on real-world datasets show that our proposed model achieves consistent and significant improvements as compared to baseline methods in cross-lingual sememe prediction. The codes and data of this paper are available at https://github.com/thunlp/CL-SP.

Neural Cross-lingual Named Entity Recognition with Minimal Resources
Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell

For languages with no annotated resources, unsupervised transfer of natural language processing models such as named-entity recognition (NER) from resource-rich languages would be an appealing capability. However, differences in words and word order across languages make it a challenging problem. To improve mapping of lexical items across languages, we propose a method that finds translations based on bilingual word embeddings. To improve robustness to word order differences, we propose to use self-attention, which allows for a degree of flexibility with respect to word order. We demonstrate that these methods achieve state-of-the-art or competitive NER performance on commonly tested languages under a cross-lingual setting, with much lower resource requirements than past approaches. We also evaluate the challenges of applying these methods to Uyghur, a low-resource language.

A Stable and Effective Learning Strategy for Trainable Greedy Decoding
Yun Chen, Victor O.K. Li, Kyunghyun Cho, and Samuel Bowman

Beam search is a widely used approximate search strategy for neural network decoders, and it generally outperforms simple greedy decoding on tasks like machine translation. However, this improvement comes at substantial computational cost. In this paper, we propose a flexible new method that allows us to reap nearly the full benefits of beam search with nearly no additional computational cost. The method revolves around a small neural network actor that is trained to observe and manipulate the hidden state of a previously-trained decoder. To train this actor network, we introduce the use of a pseudo-parallel corpus built using the output of beam search on a base model, ranked by a target quality metric like BLEU. Our method is inspired by earlier work on this problem, but requires no reinforcement learning, and can be trained reliably on a range of models. Experiments on three parallel corpora and three architectures show that the method yields substantial improvements in translation quality and speed over each base system.

Addressing Troublesome Words in Neural Machine Translation
Yang Zhao, Jiajun Zhang, Zhongjun He, Chengqing Zong, and Hua Wu

One of the weaknesses of Neural Machine Translation (NMT) is in handling lowfrequency and ambiguous words, which we refer as troublesome words. To address this problem, we propose a novel memoryenhanced NMT method. First, we investigate different strategies to define and detect the troublesome words. Then, a contextual memory is constructed to memorize which target words should be produced in what situations. Finally, we design a hybrid model to dynamically access the contextual memory so as to correctly translate the troublesome words. The extensive experiments on Chineseto- English and English-to-German translation tasks demonstrate that our method significantly outperforms the strong baseline models in translation quality, especially in handling troublesome words.

Top-down Tree Structured Decoding with Syntactic Connections for Neural Machine Translation and Parsing
Jetic Gū, Hassan S. Shavarani, and Anoop Sarkar

The addition of syntax-aware decoding in Neural Machine Translation (NMT) systems requires an effective tree-structured neural network, a syntax-aware attention model and a language generation model that is sensitive to sentence structure. Recent approaches resort to sequential decoding by adding additional neural network units to capture bottom-up structural information, or serialising structured data into sequence. We exploit a top-down tree-structured model called DRNN (Doubly-Recurrent Neural Networks) first proposed by Alvarez-Melis and Jaakola (2017) to create an NMT model called Seq2DRNN that combines a sequential encoder with tree-structured decoding augmented with a syntax-aware attention model. Unlike previous approaches to syntax-based NMT which use dependency parsing models our method uses constituency parsing which we argue provides useful information for translation. In addition, we use the syntactic structure of the sentence to add new connections to the tree-structured decoder neural network (Seq2DRNN+SynC). We compare our NMT model with sequential and state of the art syntax-based NMT models and show that our model produces more fluent translations with better reordering. Since our model is capable of doing translation and constituency parsing at the same time we also compare our parsing accuracy against other neural parsing models.

XL-NBT: A Cross-lingual Neural Belief Tracking Framework
Wenhu Chen, Jianshu Chen, Yu Su, Xin Wang, Dong Yu, Xifeng Yan, and William Yang Wang

Task-oriented dialog systems are becoming pervasive, and many companies heavily rely on them to complement human agents for customer service in call centers. With globalization, the need for providing cross-lingual customer support becomes more urgent than ever. However, cross-lingual support poses great challenges—it requires a large amount of additional annotated data from native speakers. In order to bypass the expensive human annotation and achieve the first step towards the ultimate goal of building a universal dialog system, we set out to build a cross-lingual state tracking framework. Specifically, we assume that there exists a source language with dialog belief tracking annotations while the target languages have no annotated dialog data of any form. Then, we pre-train a state tracker for the source language as a teacher, which is able to exploit easy-to-access parallel data. We then distill and transfer its own knowledge to the student state tracker in target languages. We specifically discuss two types of common parallel resources: bilingual corpus and bilingual dictionary, and design different transfer learning strategies accordingly. Experimentally, we successfully use English state tracker as the teacher to transfer its knowledge to both Italian and German trackers and achieve promising results.

Contextual Parameter Generation for Universal Neural Machine Translation
Emmanouil Antonios Platanios, Mrinmaya Sachan, Graham Neubig, and Tom Mitchell

We propose a simple modification to existing neural machine translation (NMT) models that enables using a single universal model to translate between multiple languages while allowing for language specific parameterization, and that can also be used for domain adaptation. Our approach requires no changes to the model architecture of a standard NMT system, but instead introduces a new component, the contextual parameter generator (CPG), that generates the parameters of the system (e.g., weights in a neural network). This parameter generator accepts source and target language embeddings as input, and generates the parameters for the encoder and the decoder, respectively. The rest of the model remains unchanged and is shared across all languages. We show how this simple modification enables the system to use monolingual data for training and also perform zero-shot translation. We further show it is able to surpass state-of-the-art performance for both the IWSLT-15 and IWSLT-17 datasets and that the learned language embeddings are able to uncover interesting relationships between languages.

Back-Translation Sampling by Targeting Difficult Words in Neural Machine Translation
Marzieh Fadaee and Christof Monz

Neural Machine Translation has achieved state-of-the-art performance for several language pairs using a combination of parallel and synthetic data. Synthetic data is often generated by back-translating sentences randomly sampled from monolingual data using a reverse translation model. While back-translation has been shown to be very effective in many cases, it is not entirely clear why. In this work, we explore different aspects of back-translation, and show that words with high prediction loss during training benefit most from the addition of synthetic data. We introduce several variations of sampling strategies targeting difficult-to-predict words using prediction losses and frequencies of words. In addition, we also target the contexts of difficult words and sample sentences that are similar in context. Experimental results for the WMT news translation task show that our method improves translation quality by up to 1.7 and 1.2 Bleu points over back-translation using random sampling for German-English and English-German, respectively.

Multi-Domain Neural Machine Translation with Word-Level Domain Context Discrimination
Jiali Zeng, Jinsong Su, Huating Wen, Yang Liu, Jun Xie, Yongjing Yin, and Jianqiang Zhao

With great practical value, the study of Multi-domain Neural Machine Translation (NMT) mainly focuses on using mixed-domain parallel sentences to construct a unified model that allows translation to switch between different domains. Intuitively, words in a sentence are related to its domain to varying degrees, so that they will exert disparate impacts on the multi-domain NMT modeling. Based on this intuition, in this paper, we devote to distinguishing and exploiting word-level domain contexts for multi-domain NMT. To this end, we jointly model NMT with monolingual attention-based domain classification tasks and improve NMT as follows: 1) Based on the sentence representations produced by a domain classifier and an adversarial domain classifier, we generate two gating vectors and use them to construct domain-specific and domain-shared annotations, for later translation predictions via different attention models; 2) We utilize the attention weights derived from target-side domain classifier to adjust the weights of target words in the training objective, enabling domain-related words to have greater impacts during model training. Experimental results on Chinese-English and English-French multi-domain translation tasks demonstrate the effectiveness of the proposed model. Source codes of this paper are available on Github https://github.com/DeepLearnXMU/WDCNMT.

A Discriminative Latent-Variable Model for Bilingual Lexicon Induction
Sebastian Ruder, Ryan Cotterell, Yova Kementchedjhieva, and Anders Søgaard

We introduce a novel discriminative latent-variable model for the task of bilingual lexicon induction. Our model combines the bipartite matching dictionary prior of Haghighi et al. (2008) with a state-of-the-art embedding-based approach. To train the model, we derive an efficient Viterbi EM algorithm. We provide empirical improvements on six language pairs under two metrics and show that the prior theoretically and empirically helps to mitigate the hubness problem. We also demonstrate how previous work may be viewed as a similarly fashioned latent-variable model, albeit with a different prior.

Non-Adversarial Unsupervised Word Translation
Yedid Hoshen and Lior Wolf

Unsupervised word translation from non-parallel inter-lingual corpora has attracted much research interest. Very recently, neural network methods trained with adversarial loss functions achieved high accuracy on this task. Despite the impressive success of the recent techniques, they suffer from the typical drawbacks of generative adversarial models: sensitivity to hyper-parameters, long training time and lack of interpretability. In this paper, we make the observation that two sufficiently similar distributions can be aligned correctly with iterative matching methods. We present a novel method that first aligns the second moment of the word distributions of the two languages and then iteratively refines the alignment. Extensive experiments on word translation of European and Non-European languages show that our method achieves better performance than recent state-of-the-art deep adversarial approaches and is competitive with the supervised baseline. It is also efficient, easy to parallelize on CPU and interpretable.

Semi-Autoregressive Neural Machine Translation
Chunqi Wang, Ji Zhang, and Haiqing Chen

Existing approaches to neural machine translation are typically autoregressive models. While these models attain state-of-the-art translation quality, they are suffering from low parallelizability and thus slow at decod- ing long sequences. In this paper, we propose a novel model for fast sequence generation — the semi-autoregressive Transformer (SAT). The SAT keeps the autoregressive property in global but relieves in local and thus are able to produce multiple successive words in parallel at each time step. Experiments conducted on English-German and Chinese- English translation tasks show that the SAT achieves a good balance between translation quality and decoding speed. On WMT'14 English-German translation, the SAT achieves 5.58× speedup while maintaining 88% trans- lation quality, significantly better than the previous non-autoregressive methods. When produces two words at each time step, the SAT is almost lossless (only 1% degeneration in BLEU score).

Understanding Back-Translation at Scale
Sergey Edunov, Myle Ott, Michael Auli, and David Grangier

An effective method to improve neural machine translation with monolingual data is to augment the parallel training corpus with back-translations of target language sentences. This work broadens the understanding of back-translation and investigates a number of methods to generate synthetic source sentences. We find that in all but resource poor settings back-translations obtained via sampling or noised beam outputs are most effective. Our analysis shows that sampling or noisy synthetic data gives a much stronger training signal than data generated by beam or greedy search. We also compare how synthetic data compares to genuine bitext and study various domain effects. Finally, we scale to hundreds of millions of monolingual sentences and achieve a new state of the art of 35 BLEU on the WMT'14 English-German test set.

Bootstrapping Transliteration with Guided Discovery for Low-Resource Languages
Shyam Upadhyay, Jordan Kodner, and Dan Roth

Generating the English transliteration of a name written in a foreign script is an important and challenging step in multilingual knowledge acquisition and information extraction. Existing approaches to transliteration generation require a large (>5000) number of training examples. This difficulty contrasts with transliteration discovery, a somewhat easier task that involves picking a plausible transliteration from a given list. In this work, we present a bootstrapping algorithm that uses constrained discovery to improve generation, and can be used with as few as 500 training examples, which we show can be sourced from annotators in a matter of hours. This opens the task to languages for which large number of training examples are unavailable. We evaluate transliteration generation performance itself, as well the improvement it brings to cross-lingual candidate generation for entity linking, a typical downstream task. We present a comprehensive evaluation of our approach on nine languages, each written in a unique script.

NORMA: Neighborhood Sensitive Maps for Multilingual Word Embeddings
Ndapa Nakashole

Inducing multilingual word embeddings by learning a linear map between embedding spaces of different languages achieves remark- able accuracy on related languages. How- ever, accuracy drops substantially when trans- lating between distant languages. Given that languages exhibit differences in vocabu- lary, grammar, written form, or syntax, one would expect that embedding spaces of dif- ferent languages have different structures es- pecially for distant languages. With the goal of capturing such differences, we propose a method for learning neighborhood sensitive maps, NORMA. Our experiments show that NORMA outperforms current state-of-the-art methods for word translation between distant languages.

Adaptive Multi-pass Decoder for Neural Machine Translation with Reinforcement Learning
Xinwei Geng, Xiaocheng Feng, Bing Qin, and Ting Liu

Although end-to-end neural machine translation (NMT) has achieved remarkable progress in the recent years, the idea of adopting multi-pass decoding mechanism into conventional NMT is not well explored. In this paper, we propose a novel architecture called adaptive multi-pass decoder, which introduces a flexible multi-pass polishing mechanism to extend the capacity of NMT via reinforcement learning. More specifically, we adopt an extra policy network to automatically choose a suitable and effective number of decoding passes, according to the complexity of source sentences and the quality of the generated translations. Extensive experiments on Chinese-English translation demonstrate the effectiveness of our proposed adaptive multi-pass decoder upon the conventional NMT with a significant improvement about 1.55 BLEU.

Improving the Transformer Translation Model with Document-Level Context
Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu

Although the Transformer translation model (Vaswani et al., 2017) has achieved state-of-the-art performance in a variety of translation tasks, how to use document-level context to deal with discourse phenomena problematic for Transformer still remains a challenge. In this work, we extend the Transformer model with a new context encoder to represent document-level context, which is then incorporated into the original encoder and decoder. As large-scale document-level parallel corpora are usually not available, we introduce a two-step training method to take full advantage of abundant sentence-level parallel corpora and limited document-level parallel corpora. Experiments on the NIST Chinese-English datasets and the IWSLT French-English datasets show that our approach improves over Transformer significantly.

MTNT: A Testbed for Machine Translation of Noisy Text
Paul Michel and Graham Neubig

Noisy or non-standard input text can cause disastrous mistranslations in most modern Machine Translation (MT) systems, and there has been growing research interest in creating noise-robust MT systems. However, as of yet there are no publicly available parallel corpora of with naturally occurring noisy inputs and translations, and thus previous work has resorted to evaluating on synthetically created datasets. In this paper, we propose a benchmark dataset for Machine Translation of Noisy Text (MTNT), consisting of noisy comments on Reddit (www.reddit.com) and professionally sourced translations. We commissioned translations of English comments into French and Japanese, as well as French and Japanese comments into English, on the order of 7k-37k sentences per language pair. We qualitatively and quantitatively examine the types of noise included in this dataset, then demonstrate that existing MT models fail badly on a number of noise-related phenomena, even after performing adaptation on a small training set of in-domain data. This indicates that this dataset can provide an attractive testbed for methods tailored to handling noisy text in MT.

CytonMT: an Efficient Neural Machine Translation Open-source Toolkit Implemented in C++

Xiaolin Wang, Masao Utiyama, and Eiichiro Sumita

This paper presents an open-source neural machine translation toolkit named CytonMT[1]. The toolkit is built from scratch only using C++ and NVIDIA's GPU-accelerated libraries. The toolkit features training efficiency, code simplicity and translation quality. Benchmarks show that cytonMT accelerates the training speed by 64.5% to 110.8% on neural networks of various sizes, and achieves competitive translation quality.

SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing.

Taku Kudo and John Richardson

This paper describes SentencePiece, a language-independent subword tokenizer and detokenizer designed for Neural-based text processing, including Neural Machine Translation. It provides open-source C++ and Python implementations for subword units. While existing subword segmentation tools assume that the input is pre-tokenized into word sequences, SentencePiece can train subword models directly from raw sentences, which allows us to make a purely end-to-end and language independent system. We perform a validation experiment of NMT on English-Japanese machine translation, and find that it is possible to achieve comparable accuracy to direct subword training from raw sentences. We also compare the performance of subword training and segmentation with various configurations. SentencePiece is available under the Apache 2 license at https://github.com/google/sentencepiece.

[1] https://github.com/arthurxlw/cytonMt

Session 2 Overview – Friday, November 2, 2018

Oral tracks

Track A *Question Answering I* Gold Hall	Track B *Semantics II* Copper Hall	Track C *Multilingual Methods I* Silver Hall	Track D *Social Media* Hall 100	
SimpleQuestions Nearly Solved: A New Upperbound and Baseline Approach *Petrochuk and Zettlemoyer*	Why is unsupervised alignment of English embeddings from different algorithms so hard? *Hartmann, Kementchedjhieva, and Søgaard*	Distant Supervision from Disparate Sources for Low-Resource Part-of-Speech Tagging *Plank and Agić*	Joint Learning for Emotion Classification and Emotion Cause Detection *Chen, Hou, Cheng, and Li*	13:45
Phrase-Indexed Question Answering: A New Challenge for Scalable Document Comprehension *Seo, Kwiatkowski, Parikh, Farhadi, and Hajishirzi*	Quantifying Context Overlap for Training Word Embeddings *Zhuang, Xie, Zheng, and Zhu*	Unsupervised Bilingual Lexicon Induction via Latent Variable Models *Dou, Zhou, and Huang*	Exploring Optimism and Pessimism in Twitter Using Deep Learning *Caragea, Dinu, and Dumitru*	13:57
Ranking Paragraphs for Improving Answer Recall in Open-Domain Question Answering *Lee, Yun, Kim, Ko, and Kang*	Neural Latent Relational Analysis to Capture Lexical Semantic Relation *Washio and Kato*	Learning Unsupervised Word Translations Without Adversaries *Mukherjee, Yamada, and Hospedales*	Predicting News Headline Popularity with Syntactic and Semantic Knowledge Using Multi-Task Learning *Lamprinidis, Hardt, and Hovy*	14:09
Cut to the Chase: A Context Zoom-in Network for Reading Comprehension *Indurthi, Yu, Back, and Cuayahuitl*	Generalizing Word Embeddings using Bag of Subwords *Zhao, Mudgal, and Liang*	Adversarial Training for Multi-task and Multi-lingual Joint Modeling of Utterance Intent Classification *Masumura, Shinohara, Higashinaka, and Aono*	Hybrid Neural Attention for Agreement/Disagreement Inference in Online Debates *Chen, Du, Bing, and Xu*	14:21
Adaptive Document Retrieval for Deep Question Answering *Kratzwald and Feuerriegel*	Neural Metaphor Detection in Context *Gao, Choi, Choi, and Zettlemoyer*	Surprisingly Easy Hard-Attention for Sequence to Sequence Learning *Shankar, Garg, and Sarawagi*	Improving Author Attribute Prediction by Retrofitting Linguistic Representations with Homophily *Hovy and Fornaciari*	14:33

Poster tracks

13:45–14:45

Track D: *Short Posters I* Hall 100

Parallel Session 2

Session 2A: Question Answering I

Gold Hall Chair: *chairname*

SimpleQuestions Nearly Solved: A New Upperbound and Baseline Approach
Michael Petrochuk and Luke Zettlemoyer 13:45–13:57

The SimpleQuestions dataset is one of the most commonly used benchmarks for studying single-relation factoid questions. In this paper, we present new evidence that this benchmark can be nearly solved by standard methods. First, we show that ambiguity in the data bounds performance at 83.4%; many questions have more than one equally plausible interpretation. Second, we introduce a baseline that sets a new state-of-the-art performance level at 78.1% accuracy, despite using standard methods. Finally, we report an empirical analysis showing that the upperbound is loose; roughly a quarter of the remaining errors are also not resolvable from the linguistic signal. Together, these results suggest that the SimpleQuestions dataset is nearly solved.

Phrase-Indexed Question Answering: A New Challenge for Scalable Document Comprehension
Minjoon Seo, Tom Kwiatkowski, Ankur Parikh, Ali Farhadi, and Hannaneh Hajishirzi 13:57–14:09

We formalize a new modular variant of current question answering tasks by enforcing complete independence of the document encoder from the question encoder. This formulation addresses a key challenge in machine comprehension by building a standalone representation of the document discourse. It additionally leads to a significant scalability advantage since the encoding of the answer candidate phrases in the document can be pre-computed and indexed offline for efficient retrieval. We experiment with baseline models for the new task, which achieve a reasonable accuracy but significantly underperform unconstrained QA models. We invite the QA research community to engage in Phrase-Indexed Question Answering (PIQA, pika) for closing the gap. The leaderboard is at: nlp.cs.washington.edu/piqa

Ranking Paragraphs for Improving Answer Recall in Open-Domain Question Answering
Jinhyuk Lee, Seongjun Yun, Hyunjae Kim, Miyoung Ko, and Jaewoo Kang 14:09–14:21

Recently, open-domain question answering (QA) has been combined with machine comprehension models to find answers in a large knowledge source. As open-domain QA requires retrieving relevant documents from text corpora to answer questions, its performance largely depends on the performance of document retrievers. However, since traditional information retrieval systems are not effective in obtaining documents with a high probability of containing answers, they lower the performance of QA systems. Simply extracting more documents increases the number of irrelevant documents, which also degrades the performance of QA systems. In this paper, we introduce Paragraph Ranker which ranks paragraphs of retrieved documents for a higher answer recall with less noise. We show that ranking paragraphs and aggregating answers using Paragraph Ranker improves performance of open-domain QA pipeline on the four open-domain QA datasets by 7.8% on average.

Cut to the Chase: A Context Zoom-in Network for Reading Comprehension
Sathish Reddy Indurthi, Seunghak Yu, Seohyun Back, and Heriberto Cuayahuitl 14:21–14:33

In recent years many deep neural networks have been proposed to solve Reading Comprehension (RC) tasks. Most of these models suffer from reasoning over long documents and do not trivially generalize to cases where the answer is not present as a span in a given document. We present a novel neural-based architecture that is capable of extracting relevant regions based on a given question-document pair and generating a well-formed answer. To show the effectiveness of our architecture, we conducted several experiments on the recently proposed and challenging RC dataset 'NarrativeQA'. The proposed architecture outperforms state-of-the-art results by 12.62% (ROUGE-L) relative improvement.

Adaptive Document Retrieval for Deep Question Answering
Bernhard Kratzwald and Stefan Feuerriegel 14:33–14:45

State-of-the-art systems in deep question answering proceed as follows: (1)an initial document retrieval selects relevant documents, which (2) are then processed by a neural network in order to extract the final answer. Yet the exact interplay between both components is poorly understood, especially concerning the number of candidate documents that should be retrieved. We show that choosing a static number of documents - as used in prior research - suffers from a noise-information trade-off and yields suboptimal results. As a remedy, we propose an adaptive document retrieval model. This learns the optimal candidate number for document retrieval, conditional on the size of the corpus and the query. We report extensive experimental results showing that our adaptive approach outperforms state-of-the-art methods on multiple benchmark datasets, as well as in the context of corpora with variable sizes.

Session 2B: Semantics II

Copper Hall Chair: *chairname*

Why is unsupervised alignment of English embeddings from different algorithms so hard?
Mareike Hartmann, Yova Kementchedjhieva, and Anders Søgaard 13:45–13:57

This paper presents a challenge to the community: Generative adversarial networks (GANs) can perfectly align independent English word embeddings induced using the same algorithm, based on distributional information alone; but fails to do so, for two different embeddings algorithms. Why is that? We believe understanding why, is key to understand both modern word embedding algorithms and the limitations and instability dynamics of GANs. This paper shows that (a) in all these cases, where alignment fails, there exists a linear transform between the two embeddings (so algorithm biases do not lead to non-linear differences), and (b) similar effects can not easily be obtained by varying hyper-parameters. One plausible suggestion based on our initial experiments is that the differences in the inductive biases of the embedding algorithms lead to an optimization landscape that is riddled with local optima, leading to a very small basin of convergence, but we present this more as a challenge paper than a technical contribution.

Quantifying Context Overlap for Training Word Embeddings
Yimeng Zhuang, Jinghui Xie, Yinhe Zheng, and Xuan Zhu 13:57–14:09

Most models for learning word embeddings are trained based on the context information of words, more precisely first order co-occurrence relations. In this paper, a metric is designed to estimate second order co-occurrence relations based on context overlap. The estimated values are further used as the augmented data to enhance the learning of word embeddings by joint training with existing neural word embedding models. Experimental results show that better word vectors can be obtained for word similarity tasks and some downstream NLP tasks by the enhanced approach.

Neural Latent Relational Analysis to Capture Lexical Semantic Relation
Koki Washio and Tsuneaki Kato 14:09–14:21

Capturing the semantic relations of words in a vector space contributes to many natural language processing tasks. One promising approach exploits lexico-syntactic patterns as features of word pairs. In this paper, we propose a novel model of this pattern-based approach, neural latent relational analysis (NLRA). NLRA can generalize co-occurrences of word pairs and lexico-syntactic patterns, and obtain embeddings of the word pairs that do not co-occur. This overcomes the critical data sparseness problem encountered in previous pattern-based models. Our experimental results on measuring relational similarity demonstrate that NLRA outperforms the previous pattern-based models. In addition, when combined with a vector offset model, NLRA achieves a performance comparable to that of the state-of-the-art model that exploits additional semantic relational data.

Generalizing Word Embeddings using Bag of Subwords
Jinman Zhao, Sidharth Mudgal, and Yingyu Liang 14:21–14:33

We approach the problem of generalizing pre-trained word embeddings beyond fixed-size vocabularies without using additional contextual information. We propose a subword-level word vector generation model that views words as bags of character n-grams. The model is simple, fast to train and provides good vectors for rare or unseen words. Experiments show that our model achieves state-of-the-art performances in English word similarity task and in joint prediction of part-of-speech tag and morphosyntactic attributes in 23 languages, suggesting our model's ability in capturing the relationship between words' textual representations and their embeddings.

Neural Metaphor Detection in Context
Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer 14:33–14:45

We present end-to-end neural models for detecting metaphorical word use in context. We show that relatively standard BiLSTM models which operate on complete sentences work well in this setting, in comparison to previous work that used more restricted forms of linguistic context. These models establish a new state-of-the-art on existing verb metaphor detection benchmarks, and show strong performance on jointly predicting the metaphoricity of all words in a running text.

Session 2C: Multilingual Methods I

Silver Hall Chair: *chairname*

Distant Supervision from Disparate Sources for Low-Resource Part-of-Speech Tagging
Barbara Plank and Željko Agić 13:45–13:57

a cross-lingual neural part-of-speech tagger that learns from disparate sources of distant supervision, and realistically scales to hundreds of low-resource languages. The model exploits annotation projection, instance selection, tag dictionaries, morphological lexicons, and distributed representations, all in a uniform framework. The approach is simple, yet surprisingly effective, resulting in a new state of the art without access to any gold annotated data.

Unsupervised Bilingual Lexicon Induction via Latent Variable Models
Zi-Yi Dou, Zhi-Hao Zhou, and Shujian Huang 13:57–14:09

Bilingual lexicon extraction has been studied for decades and most previous methods have relied on parallel corpora or bilingual dictionaries. Recent studies have shown that it is possible to build a bilingual dictionary by aligning monolingual word embedding spaces in an unsupervised way. With the recent advances in generative models, we propose a novel approach which builds cross-lingual dictionaries via latent variable models and adversarial training with no parallel corpora. To demonstrate the effectiveness of our approach, we evaluate our approach on several language pairs and the experimental results show that our model could achieve competitive and even superior performance compared with several state-of-the-art models.

Learning Unsupervised Word Translations Without Adversaries
Tanmoy Mukherjee, Makoto Yamada, and Timothy Hospedales 14:09–14:21

Word translation, or bilingual dictionary induction, is an important capability that impacts many multilingual language processing tasks. Recent research has shown that word translation can be achieved in an unsupervised manner, without parallel seed dictionaries or aligned corpora. However, state of the art methods unsupervised bilingual dictionary induction are based on generative adversarial models, and as such suffer from their well known problems of instability and hyper-parameter sensitivity. We present a statistical dependency-based approach to bilingual dictionary induction that is unsupervised – no seed dictionary or parallel corpora required; and introduces no adversary – therefore being much easier to train. Our method performs comparably to adversarial alternatives and outperforms prior non-adversarial methods.

Adversarial Training for Multi-task and Multi-lingual Joint Modeling of Utterance Intent Classification
Ryo Masumura, Yusuke Shinohara, Ryuichiro Higashinaka, and Yushi Aono 14:21–14:33

This paper proposes an adversarial training method for the multi-task and multi-lingual joint modeling needed for utterance intent classification. In joint modeling, common knowledge can be efficiently utilized among multiple tasks or multiple languages. This is achieved by introducing both language-specific networks shared among different tasks and task-specific networks shared among different languages. However, the shared networks are often specialized in majority tasks or languages, so performance degradation must be expected for some minor data sets. In order to improve the invariance of shared networks, the proposed method introduces both language-specific task adversarial networks and task-specific language adversarial networks; both are leveraged for purging the task or language dependencies of the shared networks. The effectiveness of the adversarial training proposal is demonstrated using Japanese and English data sets for three different utterance intent classification tasks.

Surprisingly Easy Hard-Attention for Sequence to Sequence Learning
Shiv Shankar, Siddhant Garg, and Sunita Sarawagi 14:33–14:45

In this paper we show that a simple beam approximation of the joint distribution between attention and output is an easy, accurate, and efficient attention mechanism for sequence to sequence learning. The method combines the advantage of sharp focus in hard attention and the implementation ease of soft attention. On five translation tasks we show effortless and consistent gains in BLEU compared to existing attention mechanisms.

Session 2D: Social Media

Hall 100 Chair: *chairname*

Joint Learning for Emotion Classification and Emotion Cause Detection
Ying Chen, Wenjun Hou, Xiyao Cheng, and Shoushan Li 13:45–13:57

We present a neural network-based joint approach for emotion classification and emotion cause detection, which attempts to capture mutual benefits across the two sub-tasks of emotion analysis. Considering that emotion classification and emotion cause detection need different kinds of features (affective and event-based separately), we propose a joint encoder which uses a unified framework to extract features for both sub-tasks and a joint model trainer which simultaneously learns two models for the two sub-tasks separately. Our experiments on Chinese microblogs show that the joint approach is very promising.

Exploring Optimism and Pessimism in Twitter Using Deep Learning
Cornelia Caragea, Liviu P. Dinu, and Bogdan Dumitru 13:57–14:09

Identifying optimistic and pessimistic viewpoints and users from Twitter is useful for providing better social support to those who need such support, and for minimizing the negative influence among users and maximizing the spread of positive attitudes and ideas. In this paper, we explore a range of deep learning models to predict optimism and pessimism in Twitter at both tweet and user level and show that these models substantially outperform traditional machine learning classifiers used in prior work. In addition, we show evidence that a sentiment classifier would not be sufficient for accurately predicting optimism and pessimism in Twitter. Last, we study the verb tense usage as well as the presence of polarity words in optimistic and pessimistic tweets.

Predicting News Headline Popularity with Syntactic and Semantic Knowledge Using Multi-Task Learning
Sotiris Lamprinidis, Daniel Hardt, and Dirk Hovy 14:09–14:21

Newspapers need to attract readers with headlines, anticipating their readers' preferences. These preferences rely on topical, structural, and lexical factors. We model each of these factors in a multi-task GRU network to predict headline popularity. We find that pre-trained word embeddings provide significant improvements over untrained embeddings, as do the combination of two auxiliary tasks, news-section prediction and part-of-speech tagging. However, we also find that performance is very similar to that of a simple Logistic Regression model over character n-grams. Feature analysis reveals structural patterns of headline popularity, including the use of forward-looking deictic expressions and second person pronouns.

Hybrid Neural Attention for Agreement/Disagreement Inference in Online Debates
Di Chen, Jiachen Du, Lidong Bing, and Ruifeng Xu 14:21–14:33

Inferring the agreement/disagreement relation in debates, especially in online debates, is one of the fundamental tasks in argumentation mining. The expressions of agreement/disagreement usually rely on argumentative expressions in text as well as interactions between participants in debates. Previous works usually lack the capability of jointly modeling these two factors. To alleviate this problem, this paper proposes a hybrid neural attention model which combines self and cross attention mechanism to locate salient part from textual context and interaction between users. Experimental results on three (dis)agreement inference datasets show that our model outperforms the state-of-the-art models.

Improving Author Attribute Prediction by Retrofitting Linguistic Representations with Homophily
Dirk Hovy and Tommaso Fornaciari 14:33–14:45

Most text-classification approaches represent the input based on textual features, either feature-based or continuous. However, this ignores strong non-linguistic similarities like homophily: people within a demographic group use language more similar to each other than to non-group members. We use homophily cues to retrofit text-based author representations with non-linguistic information, and introduce a trade-off parameter. This approach increases in-class similarity between authors, and improves classification performance by making classes more linearly separable. We evaluate the effect of our method on two author-attribute prediction tasks with various training-set sizes and parameter settings. We find that our method can significantly improve classification performance, especially when the number of labels is large and limited labeled data is available. It is potentially applicable as preprocessing step to any text-classification task.

Session 2E: Short Posters I

13:45–14:45

Hall 100

Chair: *chairname*

A Syntactic Constraint Based Bidirectional-Asynchronous Approach for Emotional Conversation Generation

Jingyuan Li and Xiao Sun

Traditional neural language models tend to generate generic replies with poor logic and no emotion. In this paper, a syntactically constrained bidirectional-asynchronous approach for emotional conversation generation (E-SCBA) is proposed to address this issue. In our model, pre-generated emotion keywords and topic keywords are asynchronously introduced into the process of decoding. It is much different from most existing methods which generate replies from the first word to the last. Through experiments, the results indicate that our approach not only improves the diversity of replies, but gains a boost on both logic and emotion compared with baselines.

Auto-Dialabel: Labeling Dialogue Data with Unsupervised Learning

Chen Shi, Qi Chen, Lei Sha, Sujian Li, Xu Sun, Houfeng WANG, and Lintao Zhang

The lack of labeled data is one of the main challenges when building a task-oriented dialogue system. Existing dialogue datasets usually rely on human labeling, which is expensive, limited in size, and in low coverage. In this paper, we instead propose our framework auto-dialabel to automatically cluster the dialogue intents and slots. In this framework, we collect a set of context features, leverage an autoencoder for feature assembly, and adapt a dynamic hierarchical clustering method for intent and slot labeling. Experimental results show that our framework can promote human labeling cost to a great extent, achieve good intent clustering accuracy (84.1%), and provide reasonable and instructive slot labeling results.

Extending Neural Generative Conversational Model using External Knowledge Sources

Prasanna Parthasarathi and Joelle Pineau

The use of connectionist approaches in conversational agents has been progressing rapidly due to the availability of large corpora. However current generative dialogue models often lack coherence and are content poor. This work proposes an architecture to incorporate unstructured knowledge sources to enhance the next utterance prediction in chit-chat type of generative dialogue models. We focus on Sequence-to-Sequence (Seq2Seq) conversational agents trained with the Reddit News dataset, and consider incorporating external knowledge from Wikipedia summaries as well as from the NELL knowledge base. Our experiments show faster training time and improved perplexity when leveraging external knowledge.

Modeling Temporality of Human Intentions by Domain Adaptation

Xiaolei Huang, Lixing Liu, Kate Carey, Joshua Woolley, Stefan Scherer, and Brian Borsari

Categorizing patient's intentions in conversational assessment can help decision making in clinical treatments. Many conversation corpora span broaden a series of time stages. However, it is not clear that how the themes shift in the conversation impact on the performance of human intention categorization (eg., patients might show different behaviors during the beginning versus the end). This paper proposes a method that models the temporal factor by using domain adaptation on clinical dialogue corpora, Motivational Interviewing (MI). We deploy Bi-LSTM and topic model jointly to learn language usage change across different time sessions. We conduct experiments on the MI corpora to show the promising improvement after considering temporality in the classification task.

An Auto-Encoder Matching Model for Learning Utterance-Level Semantic Dependency in Dialogue Generation

Liangchen Luo, Jingjing Xu, Junyang Lin, Qi Zeng, and Xu SUN

Generating semantically coherent responses is still a major challenge in dialogue generation. Different from conventional text generation tasks, the mapping between inputs and responses in conversations is more complicated, which highly demands the understanding of utterance-level semantic dependency, a relation between the whole meanings of inputs and outputs. To address this problem, we propose an Auto-Encoder Matching (AEM) model to learn such dependency. The model contains two auto-encoders and one mapping module. The auto-encoders learn the semantic representations of inputs and responses, and the mapping module learns to connect the utterance-level representations. Experimental results from automatic and human evaluations demonstrate that our model is capable of generating responses of high coherence and fluency compared to baseline models.

A Dataset for Document Grounded Conversations
Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black

This paper introduces a document grounded dataset for conversations. We define "Document Grounded Conversations" as conversations that are about the contents of a specified document. In this dataset the specified documents were Wikipedia articles about popular movies. The dataset contains 4112 conversations with an average of 21.43 turns per conversation. This positions this dataset to not only provide a relevant chat history while generating responses but also provide a source of information that the models could use. We describe two neural architectures that provide benchmark performance on the task of generating the next response. We also evaluate our models for engagement and fluency, and find that the information from the document helps in generating more engaging and fluent responses.

Out-of-domain Detection based on Generative Adversarial Network
Seonghan Ryu, Sangjun Koo, Hwanjo Yu, and Gary Geunbae Lee

The main goal of this paper is to develop out-of-domain (OOD) detection for dialog systems. We propose to use only in-domain sentences to build a generative adversarial network (GAN) of which the discriminator generates low scores for OOD sentences. To improve basic GANs, we apply feature matching loss in the discriminator, use domain-category analysis as an additional task in the discriminator, and remove the biases in the generator. Thereby, we reduce the huge effort of collecting OOD sentences for training OOD detection. For evaluation, we experimented OOD detection on a multi-domain dialog system. The experimental results showed the proposed method was most accurate compared to the existing methods.

Listening Comprehension over Argumentative Content
Shachar Mirkin, Guy Moshkowich, Matan Orbach, Lili Kotlerman, Yoav Kantor, Tamar Lavee, Michal Jacovi, Yonatan Bilu, Ranit Aharonov, and Noam Slonim

This paper presents a task for machine listening comprehension in the argumentation domain and a corresponding dataset in English. We recorded 200 spontaneous speeches arguing for or against 50 controversial topics. For each speech, we formulated a question, aimed at confirming or rejecting the occurrence of potential arguments in the speech. Labels were collected by listening to the speech and marking which arguments were mentioned by the speaker. We applied baseline methods addressing the task, to be used as a benchmark for future work over this dataset. All data used in this work is freely available for research.

Using Active Learning to Expand Training Data for Implicit Discourse Relation Recognition
Yang Xu, Yu Hong, Huibin Ruan, Jianmin Yao, Min Zhang, and Guodong Zhou

We tackle discourse-level relation recognition, a problem of determining semantic relations between text spans. Implicit relation recognition is challenging due to the lack of explicit relational clues. The increasingly popular neural network techniques have been proven effective for semantic encoding, whereby widely employed to boost semantic relation discrimination. However, learning to predict semantic relations at a deep level heavily relies on a great deal of training data, but the scale of the publicly available data in this field is limited. In this paper, we follow Rutherford and Xue (2015) to expand the training data set using the corpus of explicitly-related arguments, by arbitrarily dropping the overtly presented discourse connectives. On the basis, we carry out an experiment of sampling, in which a simple active learning approach is used, so as to take the informative instances for data expansion. The goal is to verify whether the selective use of external data not only reduces the time consumption of retraining but also ensures a better system performance. Using the expanded training data, we retrain a convolutional neural network (CNN) based classifer which is a simplified version of Qin et al. (2016)'s stacking gated relation recognizer. Experimental results show that expanding the training set with small-scale carefully-selected external data yields substantial performance gain, with the improvements of about 4% for accuracy and 3.6% for F-score. This allows a weak classifier to achieve a comparable performance against the state-of-the-art systems.

Learning To Split and Rephrase From Wikipedia Edit History
Jan A. Botha, Manaal Faruqui, John Alex, Jason Baldridge, and Dipanjan Das

Split and rephrase is the task of breaking down a sentence into shorter ones that together convey the same meaning. We extract a rich new dataset for this task by mining Wikipedia's edit history: WikiSplit contains one million naturally occurring sentence rewrites, providing sixty times more distinct split examples and a ninety times larger vocabulary than the WebSplit corpus introduced by Narayan et al. (2017) as a benchmark for this task. Incorporating WikiSplit as training data produces a model with qualitatively better predictions that score 32 BLEU points above the prior best result on the WebSplit benchmark.

BLEU is Not Suitable for the Evaluation of Text Simplification

Elior Sulem, Omri Abend, and Ari Rappoport

BLEU is widely considered to be an informative metric for text-to-text generation, including Text Simplification (TS). TS includes both lexical and structural aspects. In this paper we show that BLEU is not suitable for the evaluation of sentence splitting, the major structural simplification operation. We manually compiled a sentence splitting gold standard corpus containing multiple structural paraphrases, and performed a correlation analysis with human judgments. We find low or no correlation between BLEU and the grammaticality and meaning preservation parameters where sentence splitting is involved. Moreover, BLEU often negatively correlates with simplicity, essentially penalizing simpler sentences.

S2SPMN:A Simple and Effective Framework for Response Generation with Relevant Information

Jiaxin Pei and Chenliang Li

How to generate relevant and informative responses is one of the core topics in response generation area. Following the task formulation of machine translation, previous works mainly consider response generation task as a mapping from a source sentence to a target sentence. To realize this mapping, existing works tend to design intuitive but complex models. However, the relevant information existed in large dialogue corpus is mainly overlooked. In this paper, we propose Sequence to Sequence with Prototype Memory Network (S2SPMN) to exploit the relevant information provided by the large dialogue corpus to enhance response generation. Specifically, we devise two simple approaches in S2SPMN to select the relevant information (named prototypes) from the dialogue corpus. These prototypes are then saved into prototype memory network (PMN). Furthermore, a hierarchical attention mechanism is devised to extract the semantic information from the PMN to assist the response generation process. Empirical studies reveal the advantage of our model over several classical and strong baselines.

Improving Reinforcement Learning Based Image Captioning with Natural Language Prior

Tszhang Guo, Shiyu Chang, Mo Yu, and Kun Bai

Recently, Reinforcement Learning (RL) approaches have demonstrated advanced performance in image captioning by directly optimizing the metric used for testing. However, this shaped reward introduces learning biases, which reduces the readability of generated text. In addition, the large sample space makes training unstable and slow.To alleviate these issues, we propose a simple coherent solution that constrains the action space using an n-gram language prior. Quantitative and qualitative evaluations on benchmarks show that RL with the simple add-on module performs favorably against its counterpart in terms of both readability and speed of convergence. Human evaluation results show that our model is more human readable and graceful. The implementation will become publicly available upon the acceptance of the paper.

Training for Diversity in Image Paragraph Captioning

Luke Melas-Kyriazi, Alexander Rush, and George Han

Image paragraph captioning models aim to produce detailed descriptions of a source image. These models use similar techniques as standard image captioning models, but they have encountered issues in text generation, notably a lack of diversity between sentences, that have limited their effectiveness. In this work, we consider applying sequence-level training for this task. We find that standard self-critical training produces poor results, but when combined with an integrated penalty on trigram repetition produces much more diverse paragraphs. This simple training approach improves on the best result on the Visual Genome paragraph captioning dataset from 16.9 to 30.6 CIDEr, with gains on METEOR and BLEU as well, without requiring any architectural changes.

A Graph-Theoretic Summary Evaluation for Rouge

Elaheh ShafieiBavani, Mohammad Ebrahimi, Raymond Wong, and Fang Chen

ROUGE is one of the first and most widely used evaluation metrics for text summarization. However, its assessment merely relies on surface similarities between peer and model summaries. Consequently, ROUGE is unable to fairly evaluate summaries including lexical variations and paraphrasing. We propose a graph-based approach adopted into ROUGE to evaluate summaries based on both lexical and semantic similarities. Experiment results over TAC AESOP datasets show that exploiting the lexico-semantic similarity of the words used in summaries would significantly help ROUGE correlate better with human judgments.

Guided Neural Language Generation for Abstractive Summarization using Abstract Meaning Representation
Hardy Hardy and Andreas Vlachos

Recent work on abstractive summarization has made progress with neural encoder-decoder architectures. However, such models are often challenged due to their lack of explicit semantic modeling of the source document and its summary. In this paper, we extend previous work on abstractive summarization using Abstract Meaning Representation (AMR) with a neural language generation stage which we guide using the source document. We demonstrate that this guidance improves summarization results by 7.4 and 10.5 points in ROUGE-2 using gold standard AMR parses and parses obtained from an off-the-shelf parser respectively. We also find that the summarization performance on later parses is 2 ROUGE-2 points higher than that of a well-established neural encoder-decoder approach trained on a larger dataset.

Evaluating Multiple System Summary Lengths: A Case Study
Ori Shapira, David Gabay, Hadar Ronen, Judit Bar-Ilan, Yael Amsterdamer, Ani Nenkova, and Ido Dagan

Practical summarization systems are expected to produce summaries of varying lengths, per user needs. While a couple of early summarization benchmarks tested systems across multiple summary lengths, this practice was mostly abandoned due to the assumed cost of producing reference summaries of multiple lengths. In this paper, we raise the research question of whether reference summaries of a single length can be used to reliably evaluate system summaries of multiple lengths. For that, we have analyzed a couple of datasets as a case study, using several variants of the ROUGE metric that are standard in summarization evaluation. Our findings indicate that the evaluation protocol in question is indeed competitive. This result paves the way to practically evaluating varying-length summaries with simple, possibly existing, summarization benchmarks.

Neural Latent Extractive Document Summarization
Xingxing Zhang, Mirella Lapata, Furu Wei, and Ming Zhou

Extractive summarization models need sentence level labels, which are usually created with rule-based methods since most summarization datasets only have document summary pairs. These labels might be suboptimal. We propose a latent variable extractive model, where sentences are viewed as latent variables and sentences with activated variables are used to infer gold summaries. During training, the loss can come directly from gold summaries. Experiments on CNN/Dailymail dataset show our latent extractive model outperforms a strong extractive baseline trained on rule-based labels and also performs competitively with several recent models.

On the Abstractiveness of Neural Document Summarization
Fangfang Zhang, Jin-ge Yao, and Rui Yan

Many modern neural document summarization systems based on encoder-decoder networks are designed to produce abstractive summaries. We attempted to verify the degree of abstractiveness of modern neural abstractive summarization systems by calculating overlaps in terms of various types of units. Upon the observation that many abstractive systems tend to be near-extractive in practice, we also implemented a pure copy system, which achieved comparable results as abstractive summarizers while being far more computationally efficient. These findings suggest the possibility for future efforts towards more efficient systems that could better utilize the vocabulary in the original document.

A Reinforcement Learning Framework for Automatic Essay Scoring Incorporating Rating Schema
Yucheng Wang, Zhongyu Wei, Yaqian Zhou, and Xuanjing Huang

Automatic essay scoring (AES) is the task of assigning grades to essays without human interference. Existing systems for AES are typically trained to predict the score of each single essay at a time without considering the rating schema. In order to address this issue, we propose a reinforcement learning framework for essay scoring that incorporates quadratic weighted kappa as guidance to optimize the scoring system. Experiment results on benchmark datasets show the effectiveness of our framework.

Identifying Well-formed Natural Language Questions
Manaal Faruqui and Dipanjan Das

Understanding search queries is a hard problem as it involves dealing with "word salad" text ubiquitously issued by users. However, if a query resembles a well-formed question, a natural language processing pipeline is able to perform more accurate interpretation, thus reducing downstream compounding errors.

Hence, identifying whether or not a query is well formed can enhance query understanding. Here, we introduce a new task of identifying a well-formed natural language question. We construct and release a dataset of 25,100 publicly available questions classified into well-formed and non-wellformed categories and report an accuracy of 70.7% on the test set. We also show that our classifier can be used to improve the performance of neural sequence-to-sequence models for generating questions for reading comprehension.

Self-Governing Neural Networks for On-Device Short Text Classification
Sujith Ravi and Zornitsa Kozareva

Deep neural networks reach state-of-the-art performance for wide range of natural language processing, computer vision and speech applications. Yet, one of the biggest challenges is running these complex networks on devices such as mobile phones or smart watches with tiny memory footprint and low computational capacity. We propose on-device Self-Governing Neural Networks (SGNNs), which learn compact projection vectors with local sensitive hashing. The key advantage of SGNNs over existing work is that they surmount the need for pre-trained word embeddings and complex networks with huge parameters. We conduct extensive evaluation on dialog act classification and show significant improvement over state-of-the-art results. Our findings show that SGNNs are effective at capturing low-dimensional semantic text representations, while maintaining high accuracy.

HFT-CNN: Learning Hierarchical Category Structure for Multi-label Short Text Categorization
Kazuya Shimura, Jiyi Li, and Fumiyo Fukumoto

We focus on the multi-label categorization task for short texts and explore the use of a hierarchical structure (HS) of categories. In contrast to the existing work using non-hierarchical flat model, the method leverages the hierarchical relations between the pre-defined categories to tackle the data sparsity problem. The lower the HS level, the less the categorization performance. Because the number of training data per category in a lower level is much smaller than that in an upper level. We propose an approach which can effectively utilize the data in the upper levels to contribute the categorization in the lower levels by applying the Convolutional Neural Network (CNN) with a fine-tuning technique. The results using two benchmark datasets show that proposed method, Hierarchical Fine-Tuning based CNN (HFT-CNN) is competitive with the state-of-the-art CNN based methods.

A Hierarchical Neural Attention-based Text Classifier
Koustuv Sinha, Yue Dong, Jackie Chi Kit Cheung, and Derek Ruths

Deep neural networks have been displaying superior performance over traditional supervised classifiers in text classification. They learn to extract useful features automatically when sufficient amount of data is presented. However, along with the growth in the number of documents comes the increase in the number of categories, which often results in poor performance of the multiclass classifiers. In this work, we use external knowledge in the form of topic category taxonomies to aide the classification by introducing a deep hierarchical neural attention-based classifier. Our model performs better than or comparable to state-of-the-art hierarchical models at significantly lower computational cost while maintaining high interpretability.

Labeled Anchors and a Scalable, Transparent, and Interactive Classifier
Jeffrey Lund, Stephen Cowley, Wilson Fearn, Emily Hales, and Kevin Seppi

We propose Labeled Anchors, an interactive and supervised topic model based on the anchor words algorithm (Arora et al., 2013). Labeled Anchors is similar to Supervised Anchors (Nguyen et al., 2014) in that it extends the vector-space representation of words to include document labels. However, our formulation also admits a classifier which requires no training beyond inferring topics, which means our approach is also fast enough to be interactive. We run a small user study that demonstrates that untrained users can interactively update topics in order to improve classification accuracy.

Coherence-Aware Neural Topic Modeling
Ran Ding, Ramesh Nallapati, and Bing Xiang

Topic models are evaluated based on their ability to describe documents well (i.e. low perplexity) and to produce topics that carry coherent semantic meaning. In topic modeling so far, perplexity is a direct optimization target. However, topic coherence, owing to its challenging computation, is not optimized for and is only evaluated after training. In this work, under a neural variational inference framework, we propose methods to incorporate a topic coherence objective into the training process. We demonstrate that such a coherence-aware topic model exhibits a similar level of perplexity as baseline models but achieves substantially higher topic coherence.

Utilizing Character and Word Embeddings for Text Normalization with Sequence-to-Sequence Models
Daniel Watson, Nasser Zalmout, and Nizar Habash

Text normalization is an important enabling technology for several NLP tasks. Recently, neural-network-based approaches have outperformed well-established models in this task. However, in languages other than English, there has been little exploration in this direction. Both the scarcity of annotated data and the complexity of the language increase the difficulty of the problem. To address these challenges, we use a sequence-to-sequence model with character-based attention, which in addition to its self-learned character embeddings, uses word embeddings pre-trained with an approach that also models subword information. This provides the neural model with access to more linguistic information especially suitable for text normalization, without large parallel corpora. We show that providing the model with word-level features bridges the gap for the neural network approach to achieve a state-of-the-art F1 score on a standard Arabic language correction shared task dataset.

Topic Intrusion for Automatic Topic Model Evaluation
Shraey Bhatia, Jey Han Lau, and Timothy Baldwin

Topic coherence is increasingly being used to evaluate topic models and filter topics for end-user applications. Topic coherence measures how well topic words relate to each other, but offers little insight on the utility of the topics in describing the documents. In this paper, we explore the topic intrusion task — the task of guessing an outlier topic given a document and a few topics — and propose a method to automate it. We improve upon the state-of-the-art substantially, demonstrating its viability as an alternative method for topic model evaluation.

Supervised and Unsupervised Methods for Robust Separation of Section Titles and Prose Text in Web Documents
Abhijith Athreya Mysore Gopinath, Shomir Wilson, and Norman Sadeh

The text in many web documents is organized into a hierarchy of section titles and corresponding prose content, a structure which provides potentially exploitable information on discourse structure and topicality. However, this organization is generally discarded during text collection, and collecting it is not straightforward: the same visual organization can be implemented in a myriad of different ways in the underlying HTML. To remedy this, we present a flexible system for automatically extracting the hierarchical section titles and prose organization of web documents irrespective of differences in HTML representation. This system uses features from syntax, semantics, discourse and markup to build two models which classify HTML text into section titles and prose text. When tested on three different domains of web text, our domain-independent system achieves an overall precision of 0.82 and a recall of 0.98. The domain-dependent variation produces very high precision (0.99) at the expense of recall (0.75). These results exhibit a robust level of accuracy suitable for enhancing question answering, information extraction, and summarization.

Session 3 Overview – Friday, November 2, 2018

Oral tracks

	Track A *Machine Translation I* Gold Hall	**Track B** *Machine Learning I* Copper Hall	**Track C** *Semantic Parsing / Generation* Silver Hall	**Track D** *Vision / Discourse* Hall 100
15:00	SwitchOut: an Efficient Data Augmentation Algorithm for Neural Machine Translation *Wang, Pham, Dai, and Neubig*	Self-Governing Neural Networks for On-Device Short Text Classification *Ravi and Kozareva*	Exploiting Rich Syntactic Information for Semantic Parsing with Graph-to-Sequence Model *Xu, Wu, Wang, Yu, Chen, and Sheinin*	Conversational Decision Making Model for Predicting King's Decision in the Annals of the Joseon Dynasty *Bak and Oh*
15:12	Improving Unsupervised Word-by-Word Translation with Language Model and Denoising Autoencoder *Kim, Geng, and Ney*	Supervised Domain Enablement Attention for Personalized Domain Classification *Kim and Kim*	Retrieval-Based Neural Code Generation *Hayati, Olivier, Avvaru, Yin, Tomasic, and Neubig*	Toward Fast and Accurate Neural Discourse Segmentation *Wang, Li, and Yang*
15:24	Decipherment of Substitution Ciphers with Neural Language Models *Kambhatla, Mansouri Bigvand, and Sarkar*	A Deep Neural Network Sentence Level Classification Method with Context Information *Song, Petrak, and Roberts*	SQL-to-Text Generation with Graph-to-Sequence Model *Xu, Wu, Wang, Feng, and Sheinin*	A Dataset for Telling the Stories of Social Media Videos *Gella, Lewis, and Rohrbach*
15:36	Rapid Adaptation of Neural Machine Translation to New Languages *Neubig and Hu*	Towards Dynamic Computation Graphs via Sparse Latent Structure *Niculae, Martins, and Cardie*	Generating Syntactic Paraphrases *Colin and Gardent*	Cascaded Mutual Modulation for Visual Reasoning *Yao, Xu, Wang, and Xu*
15:48	Compact Personalized Models for Neural Machine Translation *Wuebker, Simianer, and DeNero*	Convolutional Neural Networks with Recurrent Neural Filters *Yang*	Neural Davidsonian Semantic Proto-role Labeling *Rudinger, Teichert, Culkin, Zhang, and Van Durme*	How agents see things: On visual representations in an emergent language game *Bouchacourt and Baroni*

Poster tracks

15:00–16:00

Track D: *Short Posters II* — Hall 100

Parallel Session 3

Session 3A: Machine Translation I

Gold Hall Chair: *chairname*

SwitchOut: an Efficient Data Augmentation Algorithm for Neural Machine Translation
Xinyi Wang, Hieu Pham, Zihang Dai, and Graham Neubig 15:00–15:12

In this work, we examine methods for data augmentation for text-based tasks such as neural machine translation (NMT). We formulate the design of a data augmentation policy with desirable properties as an optimization problem, and derive a generic analytic solution. This solution not only subsumes some existing augmentation schemes, but also leads to an extremely simple data augmentation strategy for NMT: randomly replacing words in both the source sentence and the target sentence with other random words from their corresponding vocabularies. We name this method SwitchOut. Experiments on three translation datasets of different scales show that SwitchOut yields consistent improvements of about 0.5 BLEU, achieving better or comparable performances to strong alternatives such as word dropout (Sennrich et al., 2016a). Code to implement this method is included in the appendix.

Improving Unsupervised Word-by-Word Translation with Language Model and Denoising Autoencoder
Yunsu Kim, Jiahui Geng, and Hermann Ney 15:12–15:24

Unsupervised learning of cross-lingual word embedding offers elegant matching of words across languages, but has fundamental limitations in translating sentences. In this paper, we propose simple yet effective methods to improve word-by-word translation of cross-lingual embeddings, using only monolingual corpora but without any back-translation. We integrate a language model for context-aware search, and use a novel denoising autoencoder to handle reordering. Our system surpasses state-of-the-art unsupervised translation systems without costly iterative training. We also analyze the effect of vocabulary size and denoising type on the translation performance, which provides better understanding of learning the cross-lingual word embedding and its usage in translation.

Decipherment of Substitution Ciphers with Neural Language Models
Nishant Kambhatla, Anahita Mansouri Bigvand, and Anoop Sarkar 15:24–15:36

Decipherment of homophonic substitution ciphers using language models is a well-studied task in NLP. Previous work in this topic scores short local spans of possible plaintext decipherments using n-gram language models. The most widely used technique is the use of beam search with n-gram language models proposed by Nuhn et al.(2013). We propose a beam search algorithm that scores the entire candidate plaintext at each step of the decipherment using a neural language model. We augment beam search with a novel rest cost estimation that exploits the prediction power of a neural language model. We compare against the state of the art n-gram based methods on many different decipherment tasks. On challenging ciphers such as the Beale cipher we provide significantly better error rates with much smaller beam sizes.

Rapid Adaptation of Neural Machine Translation to New Languages
Graham Neubig and Junjie Hu 15:36–15:48

This paper examines the problem of adapting neural machine translation systems to new, low-resourced languages (LRLs) as effectively and rapidly as possible. We propose methods based on starting with massively multilingual “seed models”, which can be trained ahead-of-time, and then continuing training on data related to the LRL. We contrast a number of strategies, leading to a novel, simple, yet effective method of “similar-language regularization”, where we jointly train on both a LRL of interest and a similar high-resourced language to prevent over-fitting to small LRL data. Experiments demonstrate that massively multilingual models, even without any explicit adaptation, are surprisingly effective, achieving BLEU scores of up to 15.5 with no data from the LRL, and that the proposed similar-language regularization method improves over other adaptation methods by 1.7 BLEU points average over 4 LRL settings.

Compact Personalized Models for Neural Machine Translation
Joern Wuebker, Patrick Simianer, and John DeNero 15:48–16:00

We propose and compare methods for gradient-based domain adaptation of self-attentive neural machine translation models. We demonstrate that a large proportion of model parameters can be frozen during adaptation with minimal or no reduction in translation quality by encouraging structured sparsity in the set of offset tensors during learning via group lasso regularization. We evaluate this technique for both batch and incremental adaptation across multiple data sets and language pairs. Our system architecture–

combining a state-of-the-art self-attentive model with compact domain adaptation–provides high quality personalized machine translation that is both space and time efficient.

Session 3B: Machine Learning I

Copper Hall Chair: *chairname*

Self-Governing Neural Networks for On-Device Short Text Classification
Sujith Ravi and Zornitsa Kozareva 15:00–15:12

Deep neural networks reach state-of-the-art performance for wide range of natural language processing, computer vision and speech applications. Yet, one of the biggest challenges is running these complex networks on devices such as mobile phones or smart watches with tiny memory footprint and low computational capacity. We propose on-device Self-Governing Neural Networks (SGNNs), which learn compact projection vectors with local sensitive hashing. The key advantage of SGNNs over existing work is that they surmount the need for pre-trained word embeddings and complex networks with huge parameters. We conduct extensive evaluation on dialog act classification and show significant improvement over state-of-the-art results. Our findings show that SGNNs are effective at capturing low-dimensional semantic text representations, while maintaining high accuracy.

Supervised Domain Enablement Attention for Personalized Domain Classification
Joo-Kyung Kim and Young-Bum Kim 15:12–15:24

In large-scale domain classification for natural language understanding, leveraging each user's domain enablement information, which refers to the preferred or authenticated domains by the user, with attention mechanism has been shown to improve the overall domain classification performance. In this paper, we propose a supervised enablement attention mechanism, which utilizes sigmoid activation for the attention weighting so that the attention can be computed with more expressive power without the weight sum constraint of softmax attention. The attention weights are explicitly encouraged to be similar to the corresponding elements of the output one-hot vector, and self-distillation is used to leverage the attention information of the other enabled domains. By evaluating on the actual utterances from a large-scale IPDA, we show that our approach significantly improves domain classification performance

A Deep Neural Network Sentence Level Classification Method with Context Information
Xingyi Song, Johann Petrak, and Angus Roberts 15:24–15:36

In the sentence classification task, context formed from sentences adjacent to the sentence being classified can provide important information for classification. This context is, however, often ignored. Where methods do make use of context, only small amounts are considered, making it difficult to scale. We present a new method for sentence classification, Context-LSTM-CNN, that makes use of potentially large contexts. The method also utilizes long-range dependencies within the sentence being classified, using an LSTM, and short-span features, using a stacked CNN. Our experiments demonstrate that this approach consistently improves over previous methods on two different datasets.

Towards Dynamic Computation Graphs via Sparse Latent Structure
Vlad Niculae, André F. T. Martins, and Claire Cardie 15:36–15:48

Deep NLP models benefit from underlying structures in the data—e.g., parse trees—typically extracted using off-the-shelf parsers. Recent attempts to jointly learn the latent structure encounter a tradeoff: either make factorization assumptions that limit expressiveness, or sacrifice end-to-end differentiability. Using the recently proposed SparseMAP inference, which retrieves a sparse distribution over latent structures, we propose a novel approach for end-to-end learning of latent structure predictors jointly with a downstream predictor. To the best of our knowledge, our method is the first to enable unrestricted dynamic computation graph construction from the global latent structure, while maintaining differentiability.

Convolutional Neural Networks with Recurrent Neural Filters
Yi Yang 15:48–16:00

We introduce a class of convolutional neural networks (CNNs) that utilize recurrent neural networks (RNNs) as convolution filters. A convolution filter is typically implemented as a linear affine transformation followed by a non-linear function, which fails to account for language compositionality. As a result, it limits the use of high-order filters that are often warranted for natural language processing tasks. In this work, we model convolution filters with RNNs that naturally capture compositionality and long-term dependencies in language. We show that simple CNN architectures equipped with recurrent neural filters (RNFs) achieve results that are on par with the best published ones on the Stanford Sentiment Treebank and two answer sentence selection datasets.

Session 3C: Semantic Parsing / Generation

Silver Hall Chair: *chairname*

Exploiting Rich Syntactic Information for Semantic Parsing with Graph-to-Sequence Model

Kun Xu, Lingfei Wu, Zhiguo Wang, Mo Yu, Liwei Chen, and Vadim Sheinin 15:00–15:12

Existing neural semantic parsers mainly utilize a sequence encoder, i.e., a sequential LSTM, to extract word order features while neglecting other valuable syntactic information such as dependency or constituent trees. In this paper, we first propose to use the syntactic graph to represent three types of syntactic information, i.e., word order, dependency and constituency features; then employ a graph-to-sequence model to encode the syntactic graph and decode a logical form. Experimental results on benchmark datasets show that our model is comparable to the state-of-the-art on Jobs640, ATIS, and Geo880. Experimental results on adversarial examples demonstrate the robustness of the model is also improved by encoding more syntactic information.

Retrieval-Based Neural Code Generation

Shirley Anugrah Hayati, Raphael Olivier, Pravalika Avvaru, Pengcheng Yin, Anthony Tomasic, and Graham Neubig 15:12–15:24

In models to generate program source code from natural language, representing this code in a tree structure has been a common approach. However, existing methods often fail to generate complex code correctly due to a lack of ability to memorize large and complex structures. We introduce RECODE, a method based on subtree retrieval that makes it possible to explicitly reference existing code examples within a neural code generation model. First, we retrieve sentences that are similar to input sentences using a dynamic- programming-based sentence similarity scoring method. Next, we extract n-grams of action sequences that build the associated abstract syntax tree. Finally, we increase the probability of actions that cause the retrieved n-gram action subtree to be in the predicted code. We show that our approach improves the performance on two code generation tasks by up to +2.6 BLEU.

SQL-to-Text Generation with Graph-to-Sequence Model

Kun Xu, Lingfei Wu, Zhiguo Wang, Yansong Feng, and Vadim Sheinin 15:24–15:36

Previous work approaches the SQL-to-text generation task using vanilla Seq2Seq models, which may not fully capture the inherent graph-structured information in SQL query. In this paper, we propose a graph-to-sequence model to encode the global structure information into node embeddings. This model can effectively learn the correlation between the SQL query pattern and its interpretation. Experimental results on the WikiSQL dataset and Stackoverflow dataset show that our model outperforms the Seq2Seq and Tree2Seq baselines, achieving the state-of-the-art performance.

Generating Syntactic Paraphrases

Emilie Colin and Claire Gardent 15:36–15:48

We study the automatic generation of syntactic paraphrases using four different models for generation: data-to-text generation, text-to-text generation, text reduction and text expansion, We derive training data for each of these tasks from the WebNLG dataset and we show (i) that conditioning generation on syntactic constraints effectively permits the generation of syntactically distinct paraphrases for the same input and (ii) that exploiting different types of input (data, text or data+text) further increases the number of distinct paraphrases that can be generated for a given input.

Neural Davidsonian Semantic Proto-role Labeling

Rachel Rudinger, Adam Teichert, Ryan Culkin, Sheng Zhang, and Benjamin Van Durme 15:48–16:00

We present a model for semantic proto-role labeling (SPRL) using an adapted bidirectional LSTM encoding strategy that we call NeuralDavidsonian: predicate-argument structure is represented as pairs of hidden states corresponding to predicate and argument head tokens of the input sequence. We demonstrate: (1) state-of-the-art results in SPRL, and (2) that our network naturally shares parameters between attributes, allowing for learning new attribute types with limited added supervision.

Session 3D: Vision / Discourse

Hall 100 Chair: *chairname*

Conversational Decision Making Model for Predicting King's Decision in the Annals of the Joseon Dynasty
JinYeong Bak and Alice Oh 15:00–15:12

Styles of leaders when they make decisions in groups vary, and the different styles affect the performance of the group. To understand the key words and speakers associated with decisions, we initially formalize the problem as one of predicting leaders' decisions from discussion with group members. As a dataset, we introduce conversational meeting records from a historical corpus, and develop a hierarchical RNN structure with attention and pre-trained speaker embedding in the form of a, Conversational Decision Making Model (CDMM). The CDMM outperforms other baselines to predict leaders' final decisions from the data. We explain why CDMM works better than other methods by showing the key words and speakers discovered from the attentions as evidence.

Toward Fast and Accurate Neural Discourse Segmentation
Yizhong Wang, Sujian Li, and Jingfeng Yang 15:12–15:24

Discourse segmentation, which segments texts into Elementary Discourse Units, is a fundamental step in discourse analysis. Previous discourse segmenters rely on complicated hand-crafted features and are not practical in actual use. In this paper, we propose an end-to-end neural segmenter based on BiLSTM-CRF framework. To improve its accuracy, we address the problem of data insufficiency by transferring a word representation model that is trained on a large corpus. We also propose a restricted self-attention mechanism in order to capture useful information within a neighborhood. Experiments on the RST-DT corpus show that our model is significantly faster than previous methods, while achieving new state-of-the-art performance.

A Dataset for Telling the Stories of Social Media Videos
Spandana Gella, Mike Lewis, and Marcus Rohrbach 15:24–15:36

Video content on social media platforms constitutes a major part of the communication between people, as it allows everyone to share their stories. However, if someone is unable to consume video, either due to a disability or network bandwidth, this severely limits their participation and communication. Automatically telling the stories using multi-sentence descriptions of videos would allow bridging this gap. To learn and evaluate such models, we introduce VideoStory a new large-scale dataset for video description as a new challenge for multi-sentence video description. Our VideoStory captions dataset is complementary to prior work and contains 20k videos posted publicly on a social media platform amounting to 396 hours of video with 123k sentences, temporally aligned to the video.

Cascaded Mutual Modulation for Visual Reasoning
Yiqun Yao, Jiaming Xu, Feng Wang, and Bo Xu 15:36–15:48

Visual reasoning is a special visual question answering problem that is multi-step and compositional by nature, and also requires intensive text-vision interactions. We propose CMM: Cascaded Mutual Modulation as a novel end-to-end visual reasoning model. CMM includes a multi-step comprehension process for both question and image. In each step, we use a Feature-wise Linear Modulation (FiLM) technique to enable textual/visual pipeline to mutually control each other. Experiments show that CMM significantly outperforms most related models, and reach state-of-the-arts on two visual reasoning benchmarks: CLEVR and NLVR, collected from both synthetic and natural languages. Ablation studies confirm the effectiveness of CMM to comprehend natural language logics under the guidence of images. Our code is available at https://github.com/FlamingHorizon/CMM-VR.

How agents see things: On visual representations in an emergent language game
Diane Bouchacourt and Marco Baroni 15:48–16:00

There is growing interest in the language developed by agents interacting in emergent-communication settings. Earlier studies have focused on the agents' symbol usage, rather than on their representation of visual input. In this paper, we consider the referential games of Lazaridou et al. (2017), and investigate the representations the agents develop during their evolving interaction. We find that the agents establish successful communication by inducing visual representations that almost perfectly align with each other, but, surprisingly, do not capture the conceptual properties of the objects depicted in the input images. We conclude that, if we care about developing language-like communication systems, we must pay more attention to the visual semantics agents associate to the symbols they use.

Session 3E: Short Posters II

15:00–16:00

Hall 100

Chair: *chairname*

Attention-Based Capsule Network with Dynamic Routing for Relation Extraction
Ningyu Zhang, Shumin Deng, Zhanling Sun, Xi Chen, Wei Zhang, and Huajun Chen

A capsule is a group of neurons, whose activity vector represents the instantiation parameters of a specific type of entity. In this paper, we explore the capsule networks used for relation extraction in a multi-instance multi-label learning framework and propose a novel neural approach based on capsule networks with attention mechanisms. We evaluate our method with different benchmarks, and it is demonstrated that our method improves the precision of the predicted relations. Particularly, we show that capsule networks improve multiple entity pairs relation extraction.

Put It Back: Entity Typing with Language Model Enhancement
Ji Xin, Hao Zhu, Xu Han, Zhiyuan Liu, and Maosong Sun

Entity typing aims to classify semantic types of an entity mention in a specific context. Most existing models obtain training data using distant supervision, and inevitably suffer from the problem of noisy labels. To address this issue, we propose entity typing with language model enhancement. It utilizes a language model to measure the compatibility between context sentences and labels, and thereby automatically focuses more on context-dependent labels. Experiments on benchmark datasets demonstrate that our method is capable of enhancing the entity typing model with information from the language model, and significantly outperforms the state-of-the-art baseline. Code and data for this paper can be found from https://github.com/thunlp/LME.

Event Detection with Neural Networks: A Rigorous Empirical Evaluation
Walker Orr, Prasad Tadepalli, and Xiaoli Fern

Detecting events and classifying them into predefined types is an important step in knowledge extraction from natural language texts. While the neural network models have generally led the state-of-the-art, the differences in performance between different architectures have not been rigorously studied. In this paper we present a novel GRU-based model that combines syntactic information along with temporal structure through an attention mechanism. We show that it is competitive with other neural network architectures through empirical evaluations under different random initializations and training-validation-test splits of ACE2005 dataset.

PubSE: A Hierarchical Model for Publication Extraction from Academic Homepages
Yiqing Zhang, Jianzhong Qi, Rui Zhang, and Chuandong Yin

Publication information in a researcher's academic homepage provides insights about the researcher's expertise, research interests, and collaboration networks. We aim to extract all the publication strings from a given academic homepage. This is a challenging task because the publication strings in different academic homepages may be located at different positions with different structures. To capture the positional and structural diversity, we propose an end-to-end hierarchical model named PubSE based on Bi-LSTM-CRF. We further propose an alternating training method for training the model. Experiments on real data show that PubSE outperforms the state-of-the-art models by up to 11.8% in F1-score.

Neural Transition-based Model for Nested Mention Recognition
Bailin Wang, Wei Lu, Yu Wang, and Hongxia Jin

It is common that entity mentions can contain other mentions recursively. This paper introduces a scalable transition-based method to model the nested structure of mentions. We first map a sentence with nested mentions to a designated forest where each mention corresponds to a constituent of the forest. Our shift-reduce based system then learns to construct the forest structure in a bottom-up manner through an action sequence whose maximal length is guaranteed to be three times of the sentence length. Based on Stack-LSTM which is employed to efficiently and effectively represent the states of the system in a continuous space, our system is further incorporated with a character-based component to capture letter-level patterns. Our model gets the state-of-the-art performances in ACE datasets, showing its effectiveness in detecting nested mentions.

Genre Separation Network with Adversarial Training for Cross-genre Relation Extraction
Ge Shi, Chong Feng, Lifu Huang, Boliang Zhang, Heng Ji, Lejian Liao, and Heyan Huang

Relation Extraction suffers from dramatical performance decrease when training a model on one genre and directly applying it to a new genre, due to the distinct feature distributions. Previous studies address this problem by discovering a shared space across genres using manually crafted features, which requires great human effort. To effectively automate this process, we design a genre-separation network, which applies two encoders, one genre-independent and one genre-shared, to explicitly extract genre-specific and genre-agnostic features. Then we train a relation classifier using the genre-agnostic features on the source genre and directly apply to the target genre. Experiment results on three distinct genres of the ACE dataset show that our approach achieves up to 6.1% absolute F1-score gain compared to previous methods. By incorporating a set of external linguistic features, our approach outperforms the state-of-the-art by 1.7% absolute F1 gain. We make all programs of our model publicly available for research purpose

Effective Use of Context in Noisy Entity Linking
David Mueller and Greg Durrett

To disambiguate between closely related concepts, entity linking systems need to effectively distill cues from their context, which may be quite noisy. We investigate several techniques for using these cues in the context of noisy entity linking on short texts. Our starting point is a state-of-the-art attention-based model from prior work; while this model's attention typically identifies context that is topically relevant, it fails to identify some of the most indicative surface strings, especially those exhibiting lexical overlap with the true title. Augmenting the model with convolutional networks over characters still leaves it largely unable to pick up on these cues compared to sparse features that target them directly, indicating that automatically learning how to identify relevant character-level context features is a hard problem. Our final system outperforms past work on the WikilinksNED test set by 2.8% absolute.

Exploiting Contextual Information via Dynamic Memory Network for Event Detection
Shaobo Liu, Rui Cheng, Xiaoming Yu, and Xueqi Cheng

The task of event detection involves identifying and categorizing event triggers. Contextual information has been shown effective on the task. However, existing methods which utilize contextual information only process the context once. We argue that the context can be better exploited by processing the context multiple times, allowing the model to perform complex reasoning and to generate better context representation, thus improving the overall performance. Meanwhile, dynamic memory network (DMN) has demonstrated promising capability in capturing contextual information and has been applied successfully to various tasks. In light of the multi-hop mechanism of the DMN to model the context, we propose the trigger detection dynamic memory network (TD-DMN) to tackle the event detection problem. We performed a five-fold cross-validation on the ACE-2005 dataset and experimental results show that the multi-hop mechanism does improve the performance and the proposed model achieves best F1 score compared to the state-of-the-art methods.

Do explanation modalities make VQA models more predictable to a human?
Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Devi Parikh

A rich line of research attempts to make deep neural networks more transparent by generating human-interpretable 'explanations' of their decision process, especially for interactive tasks like Visual Question Answering (VQA). In this work, we analyze if existing explanations indeed make a VQA model — its responses as well as failures — more predictable to a human. Surprisingly, we find that they do not. On the other hand, we find that human-in-the-loop approaches that treat the model as a black-box do.

Facts That Matter
Marco Ponza, Luciano Del Corro, and Gerhard Weikum

This work introduces fact salience: The task of generating a machine-readable representation of the most prominent information in a text document as a set of facts. We also present SalIE, the first fact salience system. SalIE is unsupervised and knowledge agnostic, based on open information extraction to detect facts in natural language text, PageRank to determine their relevance, and clustering to promote diversity. We compare SalIE with several baselines (including positional, standard for saliency tasks), and in an extrinsic evaluation, with state-of-the-art automatic text summarizers. SalIE outperforms baselines and text summarizers showing that facts are an effective way to compress information.

Entity Tracking Improves Cloze-style Reading Comprehension
Luong Hoang, Sam Wiseman, and Alexander Rush

Recent work has improved on modeling for reading comprehension tasks with simple approaches such as the Attention Sum-Reader; however, automatic systems still significantly trail human performance. Analysis suggests that many of the remaining hard instances are related to the inability to track entity-references throughout documents. This work focuses on these hard entity tracking cases with two extensions: (1) additional entity features, and (2) training with a multi-task tracking objective. We show that these simple modifications improve performance both independently and in combination, and we outperform the previous state of the art on the LAMBADA dataset by 8 pts, particularly on difficult entity examples. We also effectively match the performance of more complicated models on the named entity portion of the CBT dataset.

Adversarial Domain Adaptation for Duplicate Question Detection
Darsh Shah, Tao Lei, Alessandro Moschitti, Salvatore Romeo, and Preslav Nakov

We address the problem of detecting duplicate questions in forums, which is an important step towards automating the process of answering new questions. As finding and annotating such potential duplicates manually is very tedious and costly, automatic methods based on machine learning are a viable alternative. However, many forums do not have annotated data, i.e., questions labeled by experts as duplicates, and thus a promising solution is to use domain adaptation from another forum that has such annotations. Here we focus on adversarial domain adaptation, deriving important findings about when it performs well and what properties of the domains are important in this regard. Our experiments with StackExchange data show an average improvement of 5.6% over the best baseline across multiple pairs of domains.

Translating Math Word Problem to Expression Tree
Lei Wang, Yan Wang, Deng Cai, Dongxiang Zhang, and Xiaojiang Liu

Sequence-to-sequence (SEQ2SEQ) models have been successfully applied to automatic math word problem solving. Despite its simplicity, a drawback still remains: a math word problem can be correctly solved by more than one equations. This non-deterministic transduction harms the performance of maximum likelihood estimation. In this paper, by considering the uniqueness of expression tree, we propose an equation normalization method to normalize the duplicated equations. Moreover, we analyze the performance of three popular SEQ2SEQ models on the math word problem solving. We find that each model has its own specialty in solving problems, consequently an ensemble model is then proposed to combine their advantages. Experiments on dataset Math23K show that the ensemble model with equation normalization significantly outperforms the previous state-of-the-art methods.

Semantic Linking in Convolutional Neural Networks for Answer Sentence Selection
Massimo Nicosia and Alessandro Moschitti

State-of-the-art networks that model relations between two pieces of text often use complex architectures and attention. In this paper, instead of focusing on architecture engineering, we take advantage of small amounts of labelled data that model semantic phenomena in text to encode matching features directly in the word representations. This greatly boosts the accuracy of our reference network, while keeping the model simple and fast to train. Our approach also beats a tree kernel model that uses similar input encodings, and neural models which use advanced attention and compare-aggregate mechanisms.

A dataset and baselines for sequential open-domain question answering
Ahmed Elgohary, Chen Zhao, and Jordan Boyd-Graber

Previous work on question-answering systems mainly focuses on answering individual questions, assuming they are independent and devoid of context. Instead, we investigate sequential question answering, asking multiple related questions. We present QBLink, a new dataset of fully human-authored questions. We extend existing strong question answering frameworks to include previous questions to improve the overall question-answering accuracy in open-domain question answering. The dataset is publicly available at `http://sequential.qanta.org`.

Improving the results of string kernels by adapting them to your test set
Radu Tudor Ionescu and Andrei M. Butnaru

Recently, string kernels have obtained state-of-the-art results in various text classification tasks such as Arabic dialect identification or native language identification. In this paper, we apply two simple yet effective transductive learning approaches to further improve the results of string kernels. The first approach is based on interpreting the pairwise string kernel similarities between samples in the training set and samples in the test set as features. Our second approach is a simple self-training method based on two learning iterations. In the first iteration, a classifier is trained on the training set and tested on the test set, as usual. In the second iteration, a number of test samples (to which the classifier associated higher confidence scores) are added to the training set for another round of training. However, the ground-truth labels of the added test samples are not necessary. Instead, we use the labels predicted by the classifier in the first training iteration. By adapting string kernels to the test set, we report significantly better accuracy rates in English polarity classification and Arabic dialect identification.

Parameterized Convolutional Neural Networks for Aspect Level Sentiment Classification
Binxuan Huang and Kathleen Carley

We introduce a novel parameterized convolutional neural network for aspect level sentiment classification. Using parameterized filters and parameterized gates, we incorporate aspect information into convolutional neural networks (CNN). Experiments demonstrate that our parameterized filters and parameterized gates effectively capture the aspect-specific features, and our CNN-based models achieve excellent results on SemEval 2014 datasets.

Improving Multi-label Emotion Classification via Sentiment Classification with Dual Attention Transfer Network
Jianfei Yu, Luis Marujo, Jing Jiang, Pradeep Karuturi, and William Brendel

In this paper, we target at improving the performance of multi-label emotion classification with the help of sentiment classification. Specifically, we propose a new transfer learning architecture to divide the sentence representation into two different feature spaces, which are expected to respectively capture the general sentiment words and the other important emotion-specific words via a dual attention mechanism. Experimental results on two benchmark datasets demonstrate the effectiveness of our proposed method.

Learning Sentiment Memories for Sentiment Modification without Parallel Data
Yi Zhang, Jingjing Xu, Pengcheng Yang, and Xu SUN

The task of sentiment modification requires reversing the sentiment of the input and preserving the sentiment-independent content. However, aligned sentences with the same content but different sentiments are usually unavailable. Due to the lack of such parallel data, it is hard to extract sentiment independent content and reverse the sentiment in an unsupervised way. Previous work usually can not reconcile sentiment transformation and content preservation. In this paper, motivated by the fact the non-emotional context (e.g., “staff”) provides strong cues for the occurrence of emotional words (e.g., “friendly”), we propose a novel method that automatically extracts appropriate sentiment information from learned sentiment memories according to the specific context. Experiments show that our method substantially improves the content preservation degree and achieves the state-of-the-art performance.

Joint Aspect and Polarity Classification for Aspect-based Sentiment Analysis with End-to-End Neural Networks
Martin Schmitt, Simon Steinheber, Konrad Schreiber, and Benjamin Roth

In this work, we propose a new model for aspect-based sentiment analysis. In contrast to previous approaches, we jointly model the detection of aspects and the classification of their polarity in an end-to-end trainable neural network. We conduct experiments with different neural architectures and word representations on the recent GermEval 2017 dataset. We were able to show considerable performance gains by using the joint modeling approach in all settings compared to pipeline approaches. The combination of a convolutional neural network and fasttext embeddings outperformed the best submission of the shared task in 2017, establishing a new state of the art.

Representing Social Media Users for Sarcasm Detection
Y. Alex Kolchinski and Christopher Potts

We explore two methods for representing authors in the context of textual sarcasm detection: a Bayesian approach that directly represents authors’ propensities to be sarcastic, and a dense embedding approach that can learn interactions between the author and the text. Using the SARC dataset of Reddit comments, we show that augmenting a bidirectional RNN with these representations improves performance; the Bayesian approach suffices in homogeneous contexts, whereas the added power of the dense embeddings proves valuable in more diverse ones.

Syntactical Analysis of the Weaknesses of Sentiment Analyzers
Rohil Verma, Samuel Kim, and David Walter

We carry out a syntactic analysis of two state-of-the-art sentiment analyzers, Google Cloud Natural Language and Stanford CoreNLP, to assess their classification accuracy on sentences with negative polarity items. We were motivated by the absence of studies investigating sentiment analyzer performance on sentences with polarity items, a common construct in human language. Our analysis focuses on two sentential structures: downward entailment and non-monotone quantifiers; and demonstrates weaknesses of Google Natural Language and CoreNLP in capturing polarity item information. We describe the particular syntactic phenomenon that these analyzers fail to understand that any ideal sentiment analyzer must. We also provide a set of 150 test sentences that any ideal sentiment analyzer must be able to understand.

Is Nike female? Predicting brand name gender across product categories
Sridhar Moorthy, Ruth Pogacar, Samin Khan, and Yang Xu

Are brand names such as Nike female or male? Previous research suggests that the sound of a person's first name is associated with the person's gender, but no research has tried to use this knowledge to assess the gender of brand names. We present a simple computational approach that uses sound symbolism to address this open issue. Consistent with previous research, a model trained on various linguistic features of name endings predicts human gender with high accuracy. Applying this model to a data set of over a thousand commercially-traded brands in 17 product categories, our results reveal an overall bias toward male names, cutting across both male-oriented product categories as well as female-oriented categories. In addition, we find variation within categories, suggesting that firms might be seeking to imbue their brands with differentiating characteristics as part of their competitive strategy.

Improving Large-Scale Fact-Checking using Decomposable Attention Models and Lexical Tagging
Nayeon Lee, Chien-Sheng Wu, and Pascale Fung

Fact-checking of textual sources needs to effectively extract relevant information from large knowledge bases. In this paper, we extend an existing pipeline approach to better tackle this problem. We propose a neural ranker using a decomposable attention model that dynamically selects sentences to achieve promising improvement in evidence retrieval F1 by 38.80%, with (x65) speedup compared to a TF-IDF method. Moreover, we incorporate lexical tagging methods into our pipeline framework to simplify the tasks and render the model more generalizable. As a result, our framework achieves promising performance on a large-scale fact extraction and verification dataset with speedup.

Harnessing Popularity in Social Media for Extractive Summarization of Online Conversations
Ryuji Kano, Yasuhide Miura, Motoki Taniguchi, Yan-Ying Chen, Francine Chen, and Tomoko Ohkuma

We leverage a popularity measure in social media as a distant label for extractive summarization of online conversations. In social media, users can vote, share, or bookmark a post they prefer. The number of these actions is regarded as a measure of popularity. However, popularity is not determined solely by content of a post, e.g., a text or an image it contains, but is highly based on its contexts, e.g., timing, and authority. We propose Disjunctive model that computes the contribution of content and context separately. For evaluation, we build a dataset where the informativeness of comments is annotated. We evaluate the results with ranking metrics, and show that our model outperforms the baseline models which directly use popularity as a measure of informativeness.

Identifying Control in Social Media from Crowd Annotations
Masoud Rouhizadeh, Kokil Jaidka, Laura Smith, H. Andrew Schwartz, Anneke Buffone, and Lyle Ungar

Individuals express their locus of control, or "control", in their language when they identify whether or not they are in control of their circumstances. Although control is a core concept underlying rhetorical style, it is not clear whether control is expressed by how or by what authors write. We explore the roles of syntax and semantics in expressing users' sense of control –i.e. being "controlled by" or "in control of" their circumstances– in a corpus of annotated Facebook posts. We present rich insights into these linguistic aspects and find that while the language signaling control is easy to identify, it is more challenging to label it is internally or externally controlled, with lexical features outperforming syntactic features at the task. Our findings could have important implications for studying self-expression in social media.

Somm: Into the Model
Shengli Hu

To what extent could the sommelier profession, or wine stewardship, be displaced by machine leaning algorithms? There are at least three essential skills that make a qualified sommelier: wine theory, blind tasting, and beverage service, as exemplified in the rigorous certification processes of certified sommeliers and above (advanced and master) with the most authoritative body in the industry, the Court of Master Sommelier (hereafter CMS). We propose and train corresponding machine learning models that match these skills, and compare algorithmic results with real data collected from a large group of wine professionals. We find that our machine learning models outperform human sommeliers on most tasks — most notably in the section of blind tasting, where hierarchically supervised Latent Dirichlet Allocation outperforms sommeliers' judgment calls by over 6% in terms of F1-score; and in the section of beverage service, especially wine and food pairing, a modified Siamese neural network based on BiLSTM achieves better results than sommeliers by 2%. This demonstrates, contrary to popular opinion in the industry, that the sommelier profession is at least to some extent automatable, barring economic (Kleinberg et al., 2017) and psychological (Dietvorst et al., 2015) complications.

Fine-Grained Emotion Detection in Health-Related Online Posts
Hamed Khanpour and Cornelia Caragea

Detecting fine-grained emotions in online health communities provides insightful information about patients' emotional states. However, current computational approaches to emotion detection from health-related posts focus only on identifying messages that contain emotions, with no emphasis on the emotion type, using a set of handcrafted features. In this paper, we take a step further and propose to detect fine-grained emotion types from health-related posts and show how high-level and abstract features derived from deep neural networks combined with lexicon-based features can be employed to detect emotions.

The Remarkable Benefit of User-Level Aggregation for Lexical-based Population-Level Predictions
Salvatore Giorgi, Daniel Preoţiuc-Pietro, Anneke Buffone, Daniel Rieman, Lyle Ungar, and H. Andrew Schwartz

Nowcasting based on social media text promises to provide unobtrusive and near real-time predictions of community-level outcomes. These outcomes are typically regarding people, but the data is often aggregated without regard to users in the Twitter populations of each community. This paper describes a simple yet effective method for building community-level models using Twitter language aggregated by user. Results on four different U.S. county-level tasks, spanning demographic, health, and psychological outcomes show large and consistent improvements in prediction accuracies (e.g. from Pearson r=.73 to .82 for median income prediction or r=.37 to .47 for life satisfaction prediction) over the standard approach of aggregating all tweets. We make our aggregated and anonymized community-level data, derived from 37 billion tweets – over 1 billion of which were mapped to counties, available for research.

Session 4 Overview – Friday, November 2, 2018

Oral tracks

	Track A *Language Models* Gold Hall	**Track B** *Information Extraction* Copper Hall	**Track C** *Syntactic Parsing* Silver Hall	**Track D** *Visual QA* Hall 100
16:30	Deterministic Non-Autoregressive Neural Sequence Modeling by Iterative Refinement *Lee, Mansimov, and Cho*	Automatic Event Salience Identification *Liu, Xiong, Mitamura, and Hovy*	Valency-Augmented Dependency Parsing *Shi and Lee*	Tell-and-Answer: Towards Explainable Visual Question Answering using Attributes and Captions *Li, Fu, Yu, Mei, and Luo*
16:48	Large Margin Neural Language Model *Huang, Li, Ping, and Huang*	Temporal Information Extraction by Predicting Relative Time-lines *Leeuwenberg and Moens*	Unsupervised Learning of Syntactic Structure with Invertible Neural Projections *He, Neubig, and Berg-Kirkpatrick*	Learning a Policy for Opportunistic Active Learning *Padmakumar, Stone, and Mooney*
17:06	Targeted Syntactic Evaluation of Language Models *Marvin and Linzen*	Jointly Multiple Events Extraction via Attention-based Graph Information Aggregation *Liu, Luo, and Huang*	Dynamic Oracles for Top-Down and In-Order Shift-Reduce Constituent Parsing *Fernández-González and Gómez-Rodríguez*	RecipeQA: A Challenge Dataset for Multimodal Comprehension of Cooking Recipes *Yagcioglu, Erdem, Erdem, and Ikizler-Cinbis*
17:24	Rational Recurrences *Peng, Schwartz, Thomson, and Smith*	RESIDE: Improving Distantly-Supervised Neural Relation Extraction using Side Information *Vashishth, Joshi, Prayaga, Bhattacharyya, and Talukdar*	Constituent Parsing as Sequence Labeling *Gómez-Rodríguez and Vilares*	TVQA: Localized, Compositional Video Question Answering *Lei, Yu, Bansal, and Berg*
17:42	Efficient Contextualized Representation: Language Model Pruning for Sequence Labeling *Liu, Ren, Shang, Gu, Peng, and Han*	Collective Event Detection via a Hierarchical and Bias Tagging Networks with Gated Multi-level Attention Mechanisms *Chen, Yang, Liu, Zhao, and Jia*	Synthetic Data Made to Order: The Case of Parsing *Wang and Eisner*	Localizing Moments in Video with Temporal Language *Hendricks, Wang, Shechtman, Sivic, Darrell, and Russell*

Poster tracks

16:30–18:00

Track D: *Semantics III*

Hall 100

Parallel Session 4

Session 4A: Language Models

Gold Hall Chair: *chairname*

Deterministic Non-Autoregressive Neural Sequence Modeling by Iterative Refinement
Jason Lee, Elman Mansimov, and Kyunghyun Cho 16:30–16:48

We propose a conditional non-autoregressive neural sequence model based on iterative refinement. The proposed model is designed based on the principles of latent variable models and denoising autoencoders, and is generally applicable to any sequence generation task. We extensively evaluate the proposed model on machine translation (En-De and En-Ro) and image caption generation, and observe that it significantly speeds up decoding while maintaining the generation quality comparable to the autoregressive counterpart.

Large Margin Neural Language Model
Jiaji Huang, Yi Li, Wei Ping, and Liang Huang 16:48–17:06

We propose a large margin criterion for training neural language models. Conventionally, neural language models are trained by minimizing perplexity (PPL) on grammatical sentences. However, we demonstrate that PPL may not be the best metric to optimize in some tasks, and further propose a large margin formulation. The proposed method aims to enlarge the margin between the "good" and "bad" sentences in a task-specific sense. It is trained end-to-end and can be widely applied to tasks that involve re-scoring of generated text. Compared with minimum-PPL training, our method gains up to 1.1 WER reduction for speech recognition and 1.0 BLEU increase for machine translation.

Targeted Syntactic Evaluation of Language Models
Rebecca Marvin and Tal Linzen 17:06–17:24

We present a data set for evaluating the grammaticality of the predictions of a language model. We automatically construct a large number of minimally different pairs of English sentences, each consisting of a grammatical and an ungrammatical sentence. The sentence pairs represent different variations of structure-sensitive phenomena: subject-verb agreement, reflexive anaphora and negative polarity items. We expect a language model to assign a higher probability to the grammatical sentence than the ungrammatical one. In an experiment using this data set, an LSTM language model performed poorly on many of the constructions. Multi-task training with a syntactic objective (CCG supertagging) improved the LSTM's accuracy, but a large gap remained between its performance and the accuracy of human participants recruited online. This suggests that there is considerable room for improvement over LSTMs in capturing syntax in a language model.

Rational Recurrences
Hao Peng, Roy Schwartz, Sam Thomson, and Noah A. Smith 17:24–17:42

Despite the tremendous empirical success of neural models in natural language processing, many of them lack the strong intuitions that accompany classical machine learning approaches. Recently, connections have been shown between convolutional neural networks (CNNs) and weighted finite state automata (WFSAs), leading to new interpretations and insights. In this work, we show that some recurrent neural networks also share this connection to WFSAs. We characterize this connection formally, defining rational recurrences to be recurrent hidden state update functions that can be written as the Forward calculation of a finite set of WFSAs. We show that several recent neural models use rational recurrences. Our analysis provides a fresh view of these models and facilitates devising new neural architectures that draw inspiration from WFSAs. We present one such model, which performs better than two recent baselines on language modeling and text classification. Our results demonstrate that transferring intuitions from classical models like WFSAs can be an effective approach to designing and understanding neural models.

Efficient Contextualized Representation: Language Model Pruning for Sequence Labeling
Liyuan Liu, Xiang Ren, Jingbo Shang, Xiaotao Gu, Jian Peng, and Jiawei Han 17:42–18:00

Many efforts have been made to facilitate natural language processing tasks with pre-trained language models (LMs), and brought significant improvements to various applications. To fully leverage the nearly unlimited corpora and capture linguistic information of multifarious levels, large-size LMs are required; but for a specific task, only parts of these information are useful. Such large-sized LMs, even in the inference stage, may cause heavy computation workloads, making them too time-consuming for large-scale applications. Here we propose to compress bulky LMs while preserving useful information with regard to a specific task. As different layers of the model keep different information, we develop a layer selection method for model pruning using sparsity-inducing regularization. By introducing the dense connectivity, we can detach any layer without affecting others, and stretch shallow and wide LMs to be deep and narrow. In model training, LMs are learned with layer-wise dropouts for better robustness. Experiments on two benchmark datasets demonstrate the effectiveness of our method.

Session 4B: Information Extraction

Copper Hall Chair: *chairname*

Automatic Event Salience Identification
Zhengzhong Liu, Chenyan Xiong, Teruko Mitamura, and Eduard Hovy 16:30–16:48

Identifying the salience (i.e. importance) of discourse units is an important task in language understanding. While events play important roles in text documents, little research exists on analyzing their saliency status. This paper empirically studies Event Salience and proposes two salience detection models based on discourse relations. The first is a feature based salience model that incorporates cohesion among discourse units. The second is a neural model that captures more complex interactions between discourse units. In our new large-scale event salience corpus, both methods significantly outperform the strong frequency baseline, while our neural model further improves the feature based one by a large margin. Our analyses demonstrate that our neural model captures interesting connections between salience and discourse unit relations (e.g., scripts and frame structures).

Temporal Information Extraction by Predicting Relative Time-lines
Artuur Leeuwenberg and Marie-Francine Moens 16:48–17:06

The current leading paradigm for temporal information extraction from text consists of three phases: (1) recognition of events and temporal expressions, (2) recognition of temporal relations among them, and (3) time-line construction from the temporal relations. In contrast to the first two phases, the last phase, time-line construction, received little attention and is the focus of this work. In this paper, we propose a new method to construct a linear time-line from a set of (extracted) temporal relations. But more importantly, we propose a novel paradigm in which we directly predict start and end-points for events from the text, constituting a time-line without going through the intermediate step of prediction of temporal relations as in earlier work. Within this paradigm, we propose two models that predict in linear complexity, and a new training loss using TimeML-style annotations, yielding promising results.

Jointly Multiple Events Extraction via Attention-based Graph Information Aggregation
Xiao Liu, Zhunchen Luo, and Heyan Huang 17:06–17:24

Event extraction is of practical utility in natural language processing. In the real world, it is a common phenomenon that multiple events existing in the same sentence, where extracting them are more difficult than extracting a single event. Previous works on modeling the associations between events by sequential modeling methods suffer a lot from the low efficiency in capturing very long-range dependencies. In this paper, we propose a novel Jointly Multiple Events Extraction (JMEE) framework to jointly extract multiple event triggers and arguments by introducing syntactic shortcut arcs to enhance information flow and attention-based graph convolution networks to model graph information. The experiment results demonstrate that our proposed framework achieves competitive results compared with state-of-the-art methods.

RESIDE: Improving Distantly-Supervised Neural Relation Extraction using Side Information
Shikhar Vashishth, Rishabh Joshi, Sai Suman Prayaga, Chiranjib Bhattacharyya, and Partha Talukdar
17:24–17:42

Distantly-supervised Relation Extraction (RE) methods train an extractor by automatically aligning relation instances in a Knowledge Base (KB) with unstructured text. In addition to relation instances, KBs often contain other relevant side information, such as aliases of relations (e.g., founded and co-founded are aliases for the relation founderOfCompany). RE models usually ignore such readily available side information. In this paper, we propose RESIDE, a distantly-supervised neural relation extraction method which utilizes additional side information from KBs for improved relation extraction. It uses entity type and relation alias information for imposing soft constraints while predicting relations. RESIDE employs Graph Convolution Networks (GCN) to encode syntactic information from text and improves performance even when limited side information is available. Through extensive experiments on benchmark datasets, we demonstrate RESIDE's effectiveness. We have made RESIDE's source code available to encourage reproducible research.

Collective Event Detection via a Hierarchical and Bias Tagging Networks with Gated Multi-level Attention Mechanisms
Yubo Chen, Hang Yang, Kang Liu, Jun Zhao, and Yantao Jia 17:42–18:00

Traditional approaches to the task of ACE event detection primarily regard multiple events in one sentence as independent ones and recognize them separately by using sentence-level information. However, events in one sentence are usually interdependent and sentence-level information is often insufficient to resolve ambiguities for some types of events. This paper proposes a novel framework dubbed as Hierarchical and Bias Tagging Networks with Gated Multi-level Attention Mechanisms (HBTNGMA) to solve the two problems simultaneously. Firstly, we propose a hierachical and bias tagging networks to detect multiple events in one sentence collectively. Then, we devise a gated multi-level attention to automatically extract and dynamically fuse the sentence-level and document-level information. The experimental results on the widely used ACE 2005 dataset show that our approach significantly outperforms other state-of-the-art methods.

Session 4C: Syntactic Parsing

Silver Hall Chair: *chairname*

Valency-Augmented Dependency Parsing
Tianze Shi and Lillian Lee 16:30–16:48

We present a complete, automated, and efficient approach for utilizing valency analysis in making dependency parsing decisions. It includes extraction of valency patterns, a probabilistic model for tagging these patterns, and a joint decoding process that explicitly considers the number and types of each token's syntactic dependents. On 53 treebanks representing 41 languages in the Universal Dependencies data, we find that incorporating valency information yields higher precision and F1 scores on the core arguments (subjects and complements) and functional relations (e.g., auxiliaries) that we employ for valency analysis. Precision on core arguments improves from 80.87 to 85.43. We further show that our approach can be applied to an ostensibly different formalism and dataset, Tree Adjoining Grammar as extracted from the Penn Treebank; there, we outperform the previous state-of-the-art labeled attachment score by 0.7. Finally, we explore the potential of extending valency patterns beyond their traditional domain by confirming their helpfulness in improving PP attachment decisions.

Unsupervised Learning of Syntactic Structure with Invertible Neural Projections
Junxian He, Graham Neubig, and Taylor Berg-Kirkpatrick 16:48–17:06

Unsupervised learning of syntactic structure is typically performed using generative models with discrete latent variables and multinomial parameters. In most cases, these models have not leveraged continuous word representations. In this work, we propose a novel generative model that jointly learns discrete syntactic structure and continuous word representations in an unsupervised fashion by cascading an invertible neural network with a structured generative prior. We show that the invertibility condition allows for efficient exact inference and marginal likelihood computation in our model so long as the prior is well-behaved. In experiments we instantiate our approach with both Markov and tree-structured priors, evaluating on two tasks: part-of-speech (POS) induction, and unsupervised dependency parsing without gold POS annotation. On the Penn Treebank, our Markov-structured model surpasses state-of-the-art results on POS induction. Similarly, we find that our tree-structured model achieves state-of-the-art performance on unsupervised dependency parsing for the difficult training condition where neither gold POS annotation nor punctuation-based constraints are available.

Dynamic Oracles for Top-Down and In-Order Shift-Reduce Constituent Parsing
Daniel Fernández-González and Carlos Gómez-Rodríguez 17:06–17:24

We introduce novel dynamic oracles for training two of the most accurate known shift-reduce algorithms for constituent parsing: the top-down and in-order transition-based parsers. In both cases, the dynamic oracles manage to notably increase their accuracy, in comparison to that obtained by performing classic static training. In addition, by improving the performance of the state-of-the-art in-order shift-reduce parser, we achieve the best accuracy to date (92.0 F1) obtained by a fully-supervised single-model greedy shift-reduce constituent parser on the WSJ benchmark.

Constituent Parsing as Sequence Labeling
Carlos Gómez-Rodríguez and David Vilares 17:24–17:42

We introduce a method to reduce constituent parsing to sequence labeling. For each word w_t, it generates a label that encodes: (1) the number of ancestors in the tree that the words w_t and w_t+1 have in common, and (2) the nonterminal symbol at the lowest common ancestor. We first prove that the proposed encoding function is injective for any tree without unary branches. In practice, the approach is made extensible to all constituency trees by collapsing unary branches. We then use the PTB and CTB treebanks as testbeds and propose a set of fast baselines. We achieve 90% F-score on the PTB test set, outperforming the Vinyals et al. (2015) sequence-to-sequence parser. In addition, sacrificing some accuracy, our approach achieves the fastest constituent parsing speeds reported to date on PTB by a wide margin.

Synthetic Data Made to Order: The Case of Parsing
Dingquan Wang and Jason Eisner 17:42–18:00

To approximately parse an unfamiliar language, it helps to have a treebank of a similar language. But what if the closest available treebank still has the wrong word order? We show how to (stochastically) permute the constituents of an existing dependency treebank so that its surface part-of-speech statistics approximately match those of the target language. The parameters of the permutation model can be evaluated for quality by dynamic programming and tuned by gradient descent (up to a local optimum).

This optimization procedure yields trees for a new artificial language that resembles the target language. We show that delexicalized parsers for the target language can be successfully trained using such "made to order" artificial languages.

Session 4D: Visual QA

Hall 100 Chair: *chairname*

Tell-and-Answer: Towards Explainable Visual Question Answering using Attributes and Captions

Qing Li, Jianlong Fu, Dongfei Yu, Tao Mei, and Jiebo Luo 16:30–16:48

In Visual Question Answering, most existing approaches adopt the pipeline of representing an image via pre-trained CNNs, and then using the uninterpretable CNN features in conjunction with the question to predict the answer. Although such end-to-end models might report promising performance, they rarely provide any insight, apart from the answer, into the VQA process. In this work, we propose to break up the end-to-end VQA into two steps: explaining and reasoning, in an attempt towards a more explainable VQA by shedding light on the intermediate results between these two steps. To that end, we first extract attributes and generate descriptions as explanations for an image. Next, a reasoning module utilizes these explanations in place of the image to infer an answer. The advantages of such a breakdown include: (1) the attributes and captions can reflect what the system extracts from the image, thus can provide some insights for the predicted answer; (2) these intermediate results can help identify the inabilities of the image understanding or the answer inference part when the predicted answer is wrong. We conduct extensive experiments on a popular VQA dataset and our system achieves comparable performance with the baselines, yet with added benefits of explanability and the inherent ability to further improve with higher quality explanations.

Learning a Policy for Opportunistic Active Learning

Aishwarya Padmakumar, Peter Stone, and Raymond Mooney 16:48–17:06

Active learning identifies data points to label that are expected to be the most useful in improving a supervised model. Opportunistic active learning incorporates active learning into interactive tasks that constrain possible queries during interactions. Prior work has shown that opportunistic active learning can be used to improve grounding of natural language descriptions in an interactive object retrieval task. In this work, we use reinforcement learning for such an object retrieval task, to learn a policy that effectively trades off task completion with model improvement that would benefit future tasks.

RecipeQA: A Challenge Dataset for Multimodal Comprehension of Cooking Recipes

Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis 17:06–17:24

Understanding and reasoning about cooking recipes is a fruitful research direction towards enabling machines to interpret procedural text. In this work, we introduce RecipeQA, a dataset for multimodal comprehension of cooking recipes. It comprises of approximately 20K instructional recipes with multiple modalities such as titles, descriptions and aligned set of images. With over 36K automatically generated question-answer pairs, we design a set of comprehension and reasoning tasks that require joint understanding of images and text, capturing the temporal flow of events and making sense of procedural knowledge. Our preliminary results indicate that RecipeQA will serve as a challenging test bed and an ideal benchmark for evaluating machine comprehension systems. The data and leaderboard are available at http://hucvl.github.io/recipeqa.

TVQA: Localized, Compositional Video Question Answering

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg 17:24–17:42

Recent years have witnessed an increasing interest in image-based question-answering (QA) tasks. However, due to data limitations, there has been much less work on video-based QA. In this paper, we present TVQA, a large-scale video QA dataset based on 6 popular TV shows. TVQA consists of 152,545 QA pairs from 21,793 clips, spanning over 460 hours of video. Questions are designed to be compositional in nature, requiring systems to jointly localize relevant moments within a clip, comprehend subtitle-based dialogue, and recognize relevant visual concepts. We provide analyses of this new dataset as well as several baselines and a multi-stream end-to-end trainable neural network framework for the TVQA task. The dataset is publicly available at http://tvqa.cs.unc.edu.

Localizing Moments in Video with Temporal Language

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell 17:42–18:00

Localizing moments in a longer video via natural language queries is a new, challenging task at the intersection of language and video understanding. Though moment localization with natural language is similar to other language and vision tasks like natural language object retrieval in images, moment localization offers an interesting opportunity to model temporal dependencies and reasoning in text. We propose a new model that explicitly reasons about different temporal segments in a video, and shows that temporal context is important for localizing phrases which include temporal language. To benchmark whether our model, and other recent video localization models, can effectively reason about temporal language, we collect the novel TEMPOral reasoning in video and language (TEMPO) dataset. Our dataset consists of two parts: a dataset with real videos and template sentences (TEMPO - Template Language) which allows for controlled studies on temporal language, and a human language dataset which consists of temporal sentences annotated by humans (TEMPO - Human Language).

Session 4E: Semantics III

16:30–18:00

Hall 100

Chair: *chairname*

Card-660: A Reliable Evaluation Framework for Rare Word Representation Models

Mohammad Taher Pilehvar, Dimitri Kartsaklis, Victor Prokhorov, and Nigel Collier

Rare word representation has recently enjoyed a surge of interest, owing to the crucial role that effective handling of infrequent words can play in accurate semantic understanding. However, there is a paucity of reliable benchmarks for evaluation and comparison of these techniques. We show in this paper that the only existing benchmark (the Stanford Rare Word dataset) suffers from low-confidence annotations and limited vocabulary; hence, it does not constitute a solid comparison framework. In order to fill this evaluation gap, we propose Cambridge Rare word Dataset (Card-660), an expert-annotated word similarity dataset which provides a highly reliable, yet challenging, benchmark for rare word representation techniques. Through a set of experiments we show that even the best mainstream word embeddings, with millions of words in their vocabularies, are unable to achieve performances higher than 0.43 (Pearson correlation) on the dataset, compared to a human-level upperbound of 0.90. We release the dataset and the annotation materials at https://pilehvar.github.io/card-660/.

Leveraging Gloss Knowledge in Neural Word Sense Disambiguation by Hierarchical Co-Attention

Fuli Luo, Tianyu Liu, Zexue He, Qiaolin Xia, Zhifang Sui, and Baobao Chang

The goal of Word Sense Disambiguation (WSD) is to identify the correct meaning of a word in the particular context. Traditional supervised methods only use labeled data (context), while missing rich lexical knowledge such as the gloss which defines the meaning of a word sense. Recent studies have shown that incorporating glosses into neural networks for WSD has made significant improvement. However, the previous models usually build the context representation and gloss representation separately. In this paper, we find that the learning for the context and gloss representation can benefit from each other. Gloss can help to highlight the important words in the context, thus building a better context representation. Context can also help to locate the key words in the gloss of the correct word sense. Therefore, we introduce a co-attention mechanism to generate co-dependent representations for the context and gloss. Furthermore, in order to capture both word-level and sentence-level information, we extend the attention mechanism in a hierarchical fashion. Experimental results show that our model achieves the state-of-the-art results on several standard English all-words WSD test datasets.

Weeding out Conventionalized Metaphors: A Corpus of Novel Metaphor Annotations

Erik-Lân Do Dinh, Hannah Wieland, and Iryna Gurevych

We encounter metaphors every day, but only a few jump out on us and make us stumble. However, little effort has been devoted to investigating more novel metaphors in comparison to general metaphor detection efforts. We attribute this gap primarily to the lack of larger datasets that distinguish between conventionalized, i.e., very common, and novel metaphors. The goal of this paper is to alleviate this situation by introducing a crowdsourced novel metaphor annotation layer for an existing metaphor corpus. Further, we analyze our corpus and investigate correlations between novelty and features that are typically used in metaphor detection, such as concreteness ratings and more semantic features like the Potential for Metaphoricity. Finally, we present a baseline approach to assess novelty in metaphors based on our annotations.

Streaming word similarity mining on the cheap

Olof Görnerup and Daniel Gillblad

Accurately and efficiently estimating word similarities from text is fundamental in natural language processing. In this paper, we propose a fast and lightweight method for estimating similarities from streams by explicitly counting second-order co-occurrences. The method rests on the observation that words that are highly correlated with respect to such counts are also highly similar with respect to first-order co-occurrences. Using buffers of co-occurred words per word to count second-order co-occurrences, we can then estimate similarities in a single pass over data without having to do prohibitively expensive similarity calculations. We demonstrate that this approach is scalable, converges rapidly, behaves robustly under parameter changes, and that it captures word similarities on par with those given by state-of-the-art word embeddings.

Memory, Show the Way: Memory Based Few Shot Word Representation Learning

Jingyuan Sun, Shaonan Wang, and Chengqing Zong

Distributional semantic models (DSMs) generally require sufficient examples for a word to learn a high quality representation. This is in stark contrast with human who can guess the meaning of a word from one or a few referents only. In this paper, we propose Mem2Vec, a memory based embedding learning method capable of acquiring high quality word representations from fairly limited context. Our method directly adapts the representations produced by a DSM with a longterm memory to guide its guess of a novel word. Based on a pre-trained embedding space, the proposed method delivers impressive performance on two challenging few-shot word similarity tasks. Embeddings learned with our method also lead to considerable improvements over strong baselines on NER and sentiment classification.

Disambiguated skip-gram model
Karol Grzegorczyk and Marcin Kurdziel

We present disambiguated skip-gram: a neural-probabilistic model for learning multi-sense distributed representations of words. Disambiguated skip-gram jointly estimates a skip-gram-like context word prediction model and a word sense disambiguation model. Unlike previous probabilistic models for learning multi-sense word embeddings, disambiguated skip-gram is end-to-end differentiable and can be interpreted as a simple feed-forward neural network. We also introduce an effective pruning strategy for the embeddings learned by disambiguated skip-gram. This allows us to control the granularity of representations learned by our model. In experimental evaluation disambiguated skip-gram improves state-of-the are results in several word sense induction benchmarks.

Picking Apart Story Salads
Su Wang, Eric Holgate, Greg Durrett, and Katrin Erk

During natural disasters and conflicts, information about what happened is often confusing and messy, and distributed across many sources. We would like to be able to automatically identify relevant information and assemble it into coherent narratives of what happened. To make this task accessible to neural models, we introduce *Story Salads*, mixtures of multiple documents that can be generated at scale. By exploiting the Wikipedia hierarchy, we can generate salads that exhibit challenging inference problems. Story salads give rise to a novel, challenging clustering task, where the objective is to group sentences from the same narratives. We demonstrate that simple bag-of-words similarity clustering falls short on this task, and that it is necessary to take into account global context and coherence.

Context-Attentive Embeddings for Improved Sentence Representations
Douwe Kiela, Changhan Wang, and Kyunghyun Cho

While one of the first steps in many NLP systems is selecting what pre-trained word embeddings to use, we argue that such a step is better left for neural networks to figure out by themselves. To that end, we introduce dynamic meta-embeddings, a simple yet effective method for the supervised learning of embedding ensembles, which leads to state-of-the-art performance within the same model class on a variety of tasks. We subsequently show how the technique can be used to shed new light on the usage of word embeddings in NLP systems.

A Probabilistic Model for Joint Learning of Word Embeddings from Texts and Images
Melissa Ailem, Bowen Zhang, Aurélien Bellet, Pascal Denis, and Fei Sha

Several recent studies have shown the benefits of combining language and perception to infer word embeddings. These multimodal approaches either simply combine pre-trained textual and visual representations (e.g. features extracted from convolutional neural networks), or use the latter to bias the learning of textual word embeddings. In this work, we propose a novel probabilistic model to formalize how linguistic and perceptual inputs can work in concert to explain the observed word-context pairs in a text corpus. Our approach learns textual and visual representations jointly: latent visual factors couple together a skip-gram model for co-occurrence in linguistic data and a generative latent variable model for visual data. Extensive experimental studies validate the proposed model. Concretely, on the tasks of assessing pairwise word similarity and image/caption retrieval, our approach attains equally competitive or stronger results when compared to other state-of-the-art multimodal models.

Transfer and Multi-Task Learning for Noun—Noun Compound Interpretation
Murhaf Fares, Stephan Oepen, and Erik Velldal

In this paper, we empirically evaluate the utility of transfer and multi-task learning on a challenging semantic classification task: semantic interpretation of noun–noun compounds. Through a comprehensive series of experiments and in-depth error analysis, we show that transfer learning via parameter initialization and multi-task learning via parameter sharing can help a neural classification model generalize over a highly skewed distribution of relations. Further, we demonstrate how dual annotation with two distinct sets of relations over the same set of compounds can be exploited to improve the overall accuracy of a neural classifier and its F1 scores on the less frequent, but more difficult relations.

Dissecting Contextual Word Embeddings: Architecture and Representation
Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih

Contextual word representations derived from pre-trained bidirectional language models (biLMs) have recently been shown to provide significant improvements to the state of the art for a wide range of NLP tasks. How- ever, many questions remain as to how and why these models are so effective. In this paper, we present a detailed empirical study of how the choice of neural architecture (e.g. LSTM, CNN, or self attention) influences both end task accuracy and qualitative properties of the representations that are learned. We show there is a tradeoff between speed and accuracy, but all architectures learn high quality con- textual representations that outperform word embeddings for four challenging NLP tasks. Additionally, all architectures learn representations that vary with network depth, from exclusively morphological based at the word em- bedding layer through local syntax based in the lower contextual layers to longer range semantics such coreference at the upper layers. Together, these results suggest that unsupervised biLMs, independent of architecture, are learn- ing much more about the structure of language than previously appreciated.

Preposition Sense Disambiguation and Representation
Hongyu Gong, Jiaqi Mu, Suma Bhat, and pramod viswanath pramod

Prepositions are highly polysemous, and their variegated senses encode significant semantic information. In this paper we match each preposition's left- and right context, and their interplay to the geometry of the word vectors to the left and right of the preposition. Extracting these features from a large corpus and using them with machine learning models makes for an efficient preposition sense disambiguation (PSD) algorithm, which is comparable to and better than state-of-the-art on two benchmark datasets. Our reliance on no linguistic tool allows us to scale the PSD algorithm to a large corpus and learn sense-specific preposition representations. The crucial abstraction of preposition senses as word representations permits their use in downstream applications–phrasal verb paraphrasing and preposition selection–with new state-of-the-art results.

Auto-Encoding Dictionary Definitions into Consistent Word Embeddings
Tom Bosc and Pascal Vincent

Monolingual dictionaries are widespread and semantically rich resources. This paper presents a simple model that learns to compute word embeddings by processing dictionary definitions and trying to reconstruct them. It exploits the inherent recursivity of dictionaries by encouraging consistency between the representations it uses as inputs and the representations it produces as outputs. The resulting embeddings are shown to capture semantic similarity better than regular distributional methods and other dictionary-based methods. In addition, our method shows strong performance when trained exclusively on dictionary data and generalizes in one shot.

Spot the Odd Man Out: Exploring the Associative Power of Lexical Resources
Gabriel Stanovsky and Mark Hopkins

We propose Odd-Man-Out, a novel task which aims to test different properties of word representations. An Odd-Man-Out puzzle is composed of 5 (or more) words, and requires the system to choose the one which does not belong with the others. We show that this simple setup is capable of teasing out various properties of different popular lexical resources (like WordNet and pre-trained word embeddings), while being intuitive enough to annotate on a large scale. In addition, we propose a novel technique for training multi-prototype word representations, based on unsupervised clustering of ELMo embeddings, and show that it surpasses all other representations on all Odd-Man-Out collections.

TACL-1346
TACL-1346

Word embeddings are ubiquitous in NLP and information retrieval, but it is unclear what they represent when the word is polysemous. Here it is shown that multiple word senses reside in linear superposition within the word embedding and simple sparse coding can recover vectors that approximately capture the senses. The success of our approach, which applies to several embedding methods, is mathematically explained using a variant of the random walk on discourses model (Arora et al., 2016). A novel aspect of our technique is that each extracted word sense is accompanied by one of about 2000 "discourse atoms" that gives a succinct description of which other words co-occur with that word sense. Discourse atoms can be of independent interest, and make the method potentially more useful. Empirical tests are used to verify and support the theory.

Neural Multitask Learning for Simile Recognition
Lizhen Liu, Xiao Hu, Wei Song, Ruiji Fu, Ting Liu, and Guoping Hu

Simile is a special type of metaphor, where comparators such as like and as are used to compare two objects. Simile recognition is to recognize simile sentences and extract simile components, i.e., the tenor and the vehicle. This paper presents a study of simile recognition in Chinese. We construct an annotated corpus for this research, which consists of 11.3k sentences that contain a comparator. We propose a neural network framework for jointly optimizing three tasks: simile sentence classification, simile component extraction and language modeling. The experimental results show that the neural network based approaches can outperform all rule-based and feature-based baselines. Both simile sentence classification and simile component extraction can benefit from multitask learning. The former can be solved very well, while the latter is more difficult.

Structured Alignment Networks for Matching Sentences
Yang Liu, Matt Gardner, and Mirella Lapata

Many tasks in natural language processing involve comparing two sentences to compute some notion of relevance, entailment, or similarity. Typically this comparison is done either at the word level or at the sentence level, with no attempt to leverage the inherent structure of the sentence. When sentence structure is used for comparison, it is obtained during a non-differentiable pre-processing step, leading to propagation of errors. We introduce a model of structured alignments between sentences, showing how to compare two sentences by matching their latent structures. Using a structured attention mechanism, our model matches candidate spans in the first sentence to candidate spans in the second sentence, simultaneously discovering the tree structure of each sentence. Our model is fully differentiable and trained only on the matching objective. We evaluate this model on two tasks, natural entailment detection and answer sentence selection, and find that modeling latent tree structures results in superior performance. Analysis of the learned sentence structures shows they can reflect some syntactic phenomena.

Compare, Compress and Propagate: Enhancing Neural Architectures with Alignment Factorization for Natural Language Inference
Yi Tay, Anh Tuan Luu, and Siu Cheung Hui

This paper presents a new deep learning architecture for Natural Language Inference (NLI). Firstly, we introduce a new architecture where alignment pairs are compared, compressed and then propagated to upper layers for enhanced representation learning. Secondly, we adopt factorization layers for efficient and expressive compression of alignment vectors into scalar features, which are then used to augment the base word representations. The design of our approach is aimed to be conceptually simple, compact and yet powerful. We conduct experiments on three popular benchmarks, SNLI, MultiNLI and SciTail, achieving competitive performance on all. A lightweight parameterization of our model also enjoys a 3 times reduction in parameter size compared to the existing state-of-the-art models, e.g., ESIM and DIIN, while maintaining competitive performance. Additionally, visual analysis shows that our propagated features are highly interpretable.

Convolutional Interaction Network for Natural Language Inference
Jingjing Gong, Xipeng Qiu, Xinchi Chen, Dong Liang, and Xuanjing Huang

Attention-based neural models have achieved great success in natural language inference (NLI). In this paper, we propose the Convolutional Interaction Network (CIN), a general model to capture the interaction between two sentences, which can be an alternative to the attention mechanism for NLI. Specifically, CIN encodes one sentence with the filters dynamically generated based on another sentence. Since the filters may be designed to have various numbers and sizes, CIN can capture more complicated interaction patterns. Experiments on three large datasets demonstrate CIN's efficacy.

Lessons from Natural Language Inference in the Clinical Domain
Alexey Romanov and Chaitanya Shivade

State of the art models using deep neural networks have become very good in learning an accurate mapping from inputs to outputs. However, they still lack generalization capabilities in conditions that differ from the ones encountered during training. This is even more challenging in specialized, and knowledge intensive domains, where training data is limited. To address this gap, we introduce MedNLI - a dataset annotated by doctors, performing a natural language inference task (NLI), grounded in the medical history of patients. We present strategies to: 1) leverage transfer learning using datasets from the open domain, (e.g. SNLI) and 2) incorporate domain knowledge from external data and lexical sources (e.g. medical terminologies). Our results demonstrate performance gains using both strategies.

Question Generation from SQL Queries Improves Neural Semantic Parsing
Daya Guo, Yibo Sun, Duyu Tang, Nan Duan, Jian Yin, Hong Chi, James Cao, Peng Chen, and Ming Zhou
In this paper, we study how to learn a semantic parser of state-of-the-art accuracy with less supervised training data. We conduct our study on WikiSQL, the largest hand-annotated semantic parsing dataset to date. First, we demonstrate that question generation is an effective method that empowers us to learn a state-of-the-art neural network based semantic parser with thirty percent of the supervised training data. Second, we show that applying question generation to the full supervised training data further improves the state-of-the-art model. In addition, we observe that there is a logarithmic relationship between the accuracy of a semantic parser and the amount of training data.

SemRegex: A Semantics-Based Approach for Generating Regular Expressions from Natural Language Specifications
Zexuan Zhong, Jiaqi Guo, Wei Yang, Jian Peng, Tao Xie, Jian-Guang Lou, Ting Liu, and Dongmei Zhang
Recent research proposes syntax-based approaches to address the problem of generating programs from natural language specifications. These approaches typically train a sequence-to-sequence learning model using a syntax-based objective: maximum likelihood estimation (MLE). Such syntax-based approaches do not effectively address the goal of generating semantically correct programs, because these approaches fail to handle Program Aliasing, i.e., semantically equivalent programs may have many syntactically different forms. To address this issue, in this paper, we propose a semantics-based approach named SemRegex. SemRegex provides solutions for a subtask of the program-synthesis problem: generating regular expressions from natural language. Different from the existing syntax-based approaches, SemRegex trains the model by maximizing the expected semantic correctness of the generated regular expressions. The semantic correctness is measured using the DFA-equivalence oracle, random test cases, and distinguishing test cases. The experiments on three public datasets demonstrate the superiority of SemRegex over the existing state-of-the-art approaches.

Decoupling Structure and Lexicon for Zero-Shot Semantic Parsing
Jonathan Herzig and Jonathan Berant
Building a semantic parser quickly in a new domain is a fundamental challenge for conversational interfaces, as current semantic parsers require expensive supervision and lack the ability to generalize to new domains. In this paper, we introduce a zero-shot approach to semantic parsing that can parse utterances in unseen domains while only being trained on examples in other source domains. First, we map an utterance to an abstract, domain independent, logical form that represents the structure of the logical form, but contains slots instead of KB constants. Then, we replace slots with KB constants via lexical alignment scores and global inference. Our model reaches an average accuracy of 53.4% on 7 domains in the OVERNIGHT dataset, substantially better than other zero-shot baselines, and performs as good as a parser trained on over 30% of the target domain examples.

A Span Selection Model for Semantic Role Labeling
Hiroki Ouchi, Hiroyuki Shindo, and Yuji Matsumoto
We present a simple and accurate span-based model for semantic role labeling (SRL). Our model directly takes into account all possible argument spans and scores them for each label. At decoding time, we greedily select higher scoring labeled spans. One advantage of our model is to allow us to design and use span-level features, that are difficult to use in token-based BIO tagging approaches. Experimental results demonstrate that our ensemble model achieves the state-of-the-art results, 87.4 F1 and 87.0 F1 on the CoNLL-2005 and 2012 datasets, respectively.

Mapping Language to Code in Programmatic Context
Srinivasan Iyer, Ioannis Konstas, Alvin Cheung, and Luke Zettlemoyer

Source code is rarely written in isolation. It depends significantly on the programmatic context, such as the class that the code would reside in. To study this phenomenon, we introduce the task of generating class member functions given English documentation and the programmatic context provided by the rest of the class. This task is challenging because the desired code can vary greatly depending on the functionality the class provides (e.g., a sort function may or may not be available when we are asked to "return the smallest element" in a particular member variable list). We introduce CONCODE, a new large dataset with over 100,000 examples consisting of Java classes from online code repositories, and develop a new encoder-decoder architecture that models the interaction between the method documentation and the class environment. We also present a detailed error analysis suggesting that there is significant room for future work on this task.

SyntaxSQLNet: Syntax Tree Networks for Complex and Cross Domain Text-to-SQL Task
Tao Yu, Michihiro Yasunaga, Kai Yang, Rui Zhang, Dongxu Wang, Zifan Li, and Dragomir Radev

Most existing studies in text-to-SQL tasks do not require generating complex SQL queries with multiple clauses or sub-queries, and generalizing to new, unseen databases. In this paper we propose SyntaxSQLNet, a syntax tree network to address the complex and cross-domain text-to-SQL generation task. SyntaxSQLNet employs a SQL specific syntax tree-based decoder with SQL generation path history and table-aware column attention encoders. We evaluate SyntaxSQLNet on a new large-scale text-to-SQL corpus containing databases with multiple tables and complex SQL queries containing multiple SQL clauses and nested queries. We use a database split setting where databases in the test set are unseen during training. Experimental results show that SyntaxSQLNet can handle a significantly greater number of complex SQL examples than prior work, outperforming the previous state-of-the-art model by 9.5% in exact matching accuracy. To our knowledge, we are the first to study this complex text-to-SQL task. Our task and models with the latest updates are available at `https://yale-lily.github.io/seq2sql/spider`.

Cross-lingual Decompositional Semantic Parsing
Sheng Zhang, Xutai Ma, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme

We introduce the task of cross-lingual decompositional semantic parsing: mapping content provided in a source language into a decompositional semantic analysis based on a target language. We present: (1) a form of decompositional semantic analysis designed to allow systems to target varying levels of structural complexity (shallow to deep analysis), (2) an evaluation metric to measure the similarity between system output and reference semantic analysis, (3) an end-to-end model with a novel annotating mechanism that supports intra-sentential coreference, and (4) an evaluation dataset on which our model outperforms strong baselines by at least 1.75 F1 score.

Learning Semantic Parsers from Natural Language Supervision
Igor Labutov, Bishan Yang, and Tom Mitchell

As humans, we often rely on language to learn language. For example, when corrected in a conversation, we may learn from that correction, over time improving our language fluency. Inspired by this observation, we propose a learning algorithm for training semantic parsers from supervision (feedback) expressed in natural language. Our algorithm learns a semantic parser from users' corrections such as "no, what I really meant was before his job, not after", by also simultaneously learning to parse this natural language feedback in order to leverage it as a form of supervision. Unlike supervision with gold-standard logical forms, our method does not require the user to be familiar with the underlying logical formalism, and unlike supervision from denotation, it does not require the user to know the correct answer to their query. This makes our learning algorithm naturally scalable in settings where existing conversational logs are available and can be leveraged as training data. We construct a novel dataset of natural language feedback in a conversational setting, and show that our method is effective at learning a semantic parser from such natural language supervision.

DeepCx: A transition-based approach for shallow semantic parsing with complex constructional triggers
Jesse Dunietz, Jaime Carbonell, and Lori Levin

This paper introduces the surface construction labeling (SCL) task, which expands the coverage of Shallow Semantic Parsing (SSP) to include frames triggered by complex constructions. We present DeepCx, a neural, transition-based system for SCL. As a test case for the approach, we apply DeepCx to the task of tagging causal language in English, which relies on a wider variety of constructions than are typically addressed in SSP. We report substantial improvements over previous tagging efforts on a causal language dataset. We also propose ways DeepCx could be extended to still more difficult constructions and to other semantic domains once appropriate datasets become available.

What It Takes to Achieve 100% Condition Accuracy on WikiSQL
Semih Yavuz, Izzeddin Gur, Yu Su, and Xifeng Yan

WikiSQL is a newly released dataset for studying the natural language sequence to SQL translation problem. The SQL queries in WikiSQL are simple: Each involves one relation and does not have any join operation. Despite of its simplicity, none of the publicly reported structured query generation models can achieve an accuracy beyond 62%, which is still far from enough for practical use. In this paper, we ask two questions, "Why is the accuracy still low for such simple queries?" and "What does it take to achieve 100% accuracy on WikiSQL?" To limit the scope of our study, we focus on the WHERE clause in SQL. The answers will help us gain insights about the directions we should explore in order to further improve the translation accuracy. We will then investigate alternative solutions to realize the potential ceiling performance on WikiSQL. Our proposed solution can reach up to 88.6% condition accuracy on the WikiSQL dataset.

Better Transition-Based AMR Parsing with Refined Search Space
Zhijiang Guo and Wei Lu

This paper introduces a simple yet effective transition-based system for Abstract Meaning Representation (AMR) parsing. We argue that a well-defined search space involved in a transition system is crucial for building an effective parser. We propose to conduct the search in a refined search space based on a new compact AMR graph and an improved oracle. Our end-to-end parser achieves the state-of-the-art performance on various datasets with minimal additional information.

TRANX: A Transition-based Neural Abstract Syntax Parser for Semantic Parsing and Code Generation
Pengcheng Yin and Graham Neubig

We present TRANX, a transition-based neural semantic parser that maps natural language (NL) utterances into formal meaning representations (MRs). TRANX uses a transition system based on the abstract syntax description language for the target MR, which gives it two major advantages: (1) it is highly accurate, using information from the syntax of the target MR to constrain the output space and model the information flow, and (2) it is highly generalizable, and can easily be applied to new types of MR by just writing a new abstract syntax description corresponding to the allowable structures in the MR. Experiments on four different semantic parsing and code generation tasks show that our system is generalizable, extensible, and effective, registering strong results compared to existing neural semantic parsers.

Visual Interrogation of Attention-Based Models for Natural Language Inference and Machine Comprehension
Shusen Liu, Tao Li, Zhimin Liu, Vivek Srikumar, Valerio Pascucci, and Peer-Timo Bremer

Neural networks models have gained unprecedented popularity in natural language processing due to their state-of-the-art performance and the flexible end-to-end training scheme. Despite their advantages, the lack of interpretability hinders the deployment and refinement of the models. In this work, we present a flexible visualization library for creating customized visual analytic environments, in which the user can investigate and interrogate the relationships among the input, the model internals (i.e., attention), and the output predictions, which in turn shed light on the model decision-making process.

Magnitude: A Fast, Efficient Universal Vector Embedding Utility Package
Ajay Patel, Alexander Sands, Chris Callison-Burch, and Marianna Apidianaki

Vector space embedding models like word2vec, GloVe, and fastText are extremely popular representations in natural language processing (NLP) applications. We present Magnitude, a fast, lightweight tool for utilizing and processing embeddings. Magnitude is an open source Python package with a compact vector storage file format that allows for efficient manipulation of huge numbers of embeddings. Magnitude performs common operations up to 60 to 6,000 times faster than Gensim. Magnitude introduces several novel features for improved robustness like out-of-vocabulary lookups.

Universal Sentence Encoder for English
Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil

We present easy-to-use TensorFlow Hub sentence embedding models having good task transfer performance. Model variants allow for trade-offs between accuracy and compute resources. We report the relationship between model complexity, resources, and transfer performance. Comparisons are made with baselines without transfer learning and to baselines that incorporate word-level transfer. Transfer learning using sentence-level embeddings is shown to outperform models without transfer learning and often those that use only word-level transfer. We show good transfer task performance with minimal training data and obtain encouraging results on word embedding association tests (WEAT) of model bias.

6

Main Conference: Saturday, November 3

Overview

	Session 5			
09:00 – 10:30	Semantics IV *Gold Hall*	Summarization *Copper Hall*	IR / Text Mining *Silver Hall*	Machine Learning II *Hall 100*
09:00 – 10:30	**[Posters and Demos]: Information Extraction, Question Answering** *Grand Hall 2*			
10:30 – 11:00	**Coffee Break**			*Multiple levels*
	Session 6			
11:00 – 12:30	Dialogue I *Gold Hall*	Question Answering II *Copper Hall*	Semantics V *Silver Hall*	Multilingual Methods II *Hall 100*
11:00 – 12:30	**[Posters and Demos]: Syntax, Morphology, Vision and Language I** *Grand Hall 2*			
12:30 – 13:45	**Lunch**			
	Session 7			
13:45 – 14:45	Dialogue II *Gold Hall*	Social Applications II *Copper Hall*	NER *Silver Hall*	Morphology / Parsing *Hall 100*
13:45 – 14:45	**[Posters and Demos]: Short Posters III**			*Grand Hall 2*
14:45 – 15:00	**Mini-Break**			*Multiple levels*
15:00 – 16:00	**Keynote II: Gideon Mann "Understanding the News that Moves Markets" (Gold Hall)**			*Gold Hall*
16:00 – 16:30	**Coffee Break**			*Multiple levels*
	Session 8			
16:30 – 18:00	Text Categorization *Gold Hall*	Generation *Copper Hall*	Knowledge Graphs *Silver Hall*	Morphology / Phonology *Hall 100*
16:30 – 18:00	**[Posters and Demos]: Sentiment, Social Applications, Multimodal Semantics, Discourse**			*Grand Hall 2*

19:00 – 22:00 **Social Event (Royal Museums of Fine Arts of Belgium)**
Royal Museums of Fine Arts of Belgium

Invited Talk: Julia Hirschberg

Truth or Lie? Spoken Indicators of Deception in Speech

Saturday, November 3, 2018, 9:00–10:00

Gold Hall

Abstract: Detecting deception from various forms of human behavior is a longstanding research goal which is of considerable interest to the military, law enforcement, corporate security, social services and mental health workers. However, both humans and polygraphs are very poor at this task. We describe more accurate methods we have developed to detect deception automatically from spoken language. Our classifiers are trained on the largest cleanly recorded corpus of within-subject deceptive and non-deceptive speech that has been collected. To distinguish truth from lie we make use of acoustic-prosodic, lexical, demographic, and personality features. We further examine differences in deceptive behavior based upon gender, personality, and native language (Mandarin Chinese vs. English), comparing our systems to human performance. We extend our studies to identify cues in trusted speech vs. mistrusted speech and how these features differ by speaker and by listener. Why does a listener believe a lie?

Biography: Julia Hirschberg is Percy K. and Vida L. W. Hudson Professor and Chair of Computer Science at Columbia University. She previously worked at Bell Laboratories and AT&T Labs where she created the HCI Research Department. She has been editor of Computational Linguistics and Speech Communication, is a fellow of AAAI, ISCA, ACL, ACM, and IEEE, and a member of the National Academy of Engineering. She received the IEEE James L. Flanagan Speech and Audio Processing Award and the ISCA Medal for Scientific Achievement. She currently serves on the IEEE Speech and Language Processing Technical Committee, is co-chair of the CRA-W Board, and has worked for diversity for many years at AT&T and Columbia. She works on spoken language processing and NLP, studying text-to-speech synthesis, spoken dialogue systems, entrainment in conversation, detection of deceptive and emotional speech, hedging behavior, and linguistic code-switching (language mixing).

Session 5 Overview – Saturday, November 3, 2018

Oral tracks

	Track A *Semantics IV* Gold Hall	Track B *Summarization* Copper Hall	Track C *IR / Text Mining* Silver Hall	Track D *Machine Learning II* Hall 100
09:00	Heuristically Informed Unsupervised Idiom Usage Recognition *Liu and Hwa*	Neural Related Work Summarization with a Joint Context-driven Attention Mechanism *Wang, Liu, and Gao*	Improved Semantic-Aware Network Embedding with Fine-Grained Word Alignment *Shen, Zhang, Henao, and Carin*	Gromov-Wasserstein Alignment of Word Embedding Spaces *Alvarez-Melis and Jaakkola*
09:18	Coming to Your Senses: on Controls and Evaluation Sets in Polysemy Research *Dubossarsky, Grossman, and Weinshall*	Improving Neural Abstractive Document Summarization with Explicit Information Selection Modeling *Li, Xiao, Lyu, and Wang*	Learning Context-Aware Convolutional Filters for Text Processing *Shen, Min, Li, and Carin*	Deep Probabilistic Logic: A Unifying Framework for Indirect Supervision *Wang and Poon*
09:36	Predicting Semantic Relations using Global Graph Properties *Pinter and Eisenstein*	Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization *Narayan, Cohen, and Lapata*	Deep Relevance Ranking using Enhanced Document-Query Interactions *McDonald, Brokos, and Androutsopoulos*	Deriving Machine Attention from Human Rationales *Bao, Chang, Yu, and Barzilay*
09:54	Learning Scalar Adjective Intensity from Paraphrases *Cocos, Wharton, Pavlick, Apidianaki, and Callison-Burch*	Improving Abstraction in Text Summarization *Kryściński, Paulus, Xiong, and Socher*	Learning Neural Embeddings for CLIR with Adversarial Framework *Li and Cheng*	Semi-Supervised Sequence Modeling with Cross-View Training *Clark, Luong, Manning, and Le*
10:12	Pointwise HSIC: A Linear-Time Kernelized Co-occurrence Norm for Sparse Linguistic Expressions *Yokoi, Kobayashi, Fukumizu, Suzuki, and Inui*	Content Selection in Deep Learning Models of Summarization *Kedzie, McKeown, and Daume III*	AD3: Attentive Deep Document Dater *Ray, DASGUPTA, and Talukdar*	**TACL-1430** **TACL-1430**

Poster tracks

09:00–10:30

Track D: *Information Extraction, Question Answering* Hall 100

Parallel Session 5

Session 5A: Semantics IV

Gold Hall Chair: *chairname*

Heuristically Informed Unsupervised Idiom Usage Recognition
Changsheng Liu and Rebecca Hwa 09:00–09:18

Many idiomatic expressions can be interpreted figuratively or literally depending on their contexts. This paper proposes an unsupervised learning method for recognizing the intended usages of idioms. We treat the usages as a latent variable in probabilistic models and train them in a linguistically motivated feature space. Crucially, we show that distributional semantics is a helpful heuristic for distinguishing the literal usage of idioms, giving us a way to formulate a literal usage metric to estimate the likelihood that the idiom is intended literally. This information then serves as a form of distant supervision to guide the unsupervised training process for the probabilistic models. Experiments show that our overall model performs competitively against supervised methods.

Coming to Your Senses: on Controls and Evaluation Sets in Polysemy Research
Haim Dubossarsky, Eitan Grossman, and Daphna Weinshall 09:18–09:36

The point of departure of this article is the claim that sense-specific vectors provide an advantage over normal vectors due to the polysemy that they presumably represent. This claim is based on performance gains observed in gold standard evaluation tests such as word similarity tasks. We demonstrate that this claim, at least as it is instantiated in prior art, is unfounded in two ways. Furthermore, we provide empirical data and an analytic discussion that may account for the previously reported improved performance. First, we show that ground-truth polysemy degrades performance in word similarity tasks. Therefore word similarity tasks are not suitable as an evaluation test for polysemy representation. Second, random assignment of words to senses is shown to improve performance in the same task. This and additional results point to the conclusion that performance gains as reported in previous work may be an artifact of random sense assignment, which is equivalent to sub-sampling and multiple estimation of word vector representations. Theoretical analysis shows that this may on its own be beneficial for the estimation of word similarity, by reducing the bias in the estimation of the cosine distance.

Predicting Semantic Relations using Global Graph Properties
Yuval Pinter and Jacob Eisenstein 09:36–09:54

Semantic graphs, such as WordNet, are resources which curate natural language on two distinguishable layers. On the local level, individual relations between synsets (semantic building blocks) such as hypernymy and meronymy enhance our understanding of the words used to express their meanings. Globally, analysis of graph-theoretic properties of the entire net sheds light on the structure of human language as a whole. In this paper, we combine global and local properties of semantic graphs through the framework of Max-Margin Markov Graph Models (M3GM), a novel extension of Exponential Random Graph Model (ERGM) that scales to large multi-relational graphs. We demonstrate how such global modeling improves performance on the local task of predicting semantic relations between synsets, yielding new state-of-the-art results on the WN18RR dataset, a challenging version of WordNet link prediction in which "easy" reciprocal cases are removed. In addition, the M3GM model identifies multirelational motifs that are characteristic of well-formed lexical semantic ontologies.

Learning Scalar Adjective Intensity from Paraphrases
Anne Cocos, Veronica Wharton, Ellie Pavlick, Marianna Apidianaki, and Chris Callison-Burch 09:54–10:12

Adjectives like "warm", "hot", and "scalding" all describe temperature but differ in intensity. Understanding these differences between adjectives is a necessary part of reasoning about natural language. We propose a new paraphrase-based method to automatically learn the relative intensity relation that holds between a pair of scalar adjectives. Our approach analyzes over 36k adjectival pairs from the Paraphrase Database under the assumption that, for example, paraphrase pair "really hot" <–> "scalding" suggests that "hot" < "scalding". We show that combining this paraphrase evidence with existing, complementary pattern- and lexicon-based approaches improves the quality of systems for automatically ordering sets of scalar adjectives and inferring the polarity of indirect answers to "yes/no" questions.

Pointwise HSIC: A Linear-Time Kernelized Co-occurrence Norm for Sparse Linguistic Expressions

Sho Yokoi, Sosuke Kobayashi, Kenji Fukumizu, Jun Suzuki, and Kentaro Inui 10:12–10:30

In this paper, we propose a new kernel-based co-occurrence measure that can be applied to sparse linguistic expressions (e.g., sentences) with a very short learning time, as an alternative to pointwise mutual information (PMI). As well as deriving PMI from mutual information, we derive this new measure from the Hilbert–Schmidt independence criterion (HSIC); thus, we call the new measure the pointwise HSIC (PHSIC). PHSIC can be interpreted as a smoothed variant of PMI that allows various similarity metrics (e.g., sentence embeddings) to be plugged in as kernels. Moreover, PHSIC can be estimated by simple and fast (linear in the size of the data) matrix calculations regardless of whether we use linear or nonlinear kernels. Empirically, in a dialogue response selection task, PHSIC is learned thousands of times faster than an RNN-based PMI while outperforming PMI in accuracy. In addition, we also demonstrate that PHSIC is beneficial as a criterion of a data selection task for machine translation owing to its ability to give high (low) scores to a consistent (inconsistent) pair with other pairs.

Session 5B: Summarization

Copper Hall Chair: *chairname*

Neural Related Work Summarization with a Joint Context-driven Attention Mechanism
Yongzhen Wang, Xiaozhong Liu, and Zheng Gao 09:00–09:18

Conventional solutions to automatic related work summarization rely heavily on human-engineered features. In this paper, we develop a neural data-driven summarizer by leveraging the seq2seq paradigm, in which a joint context-driven attention mechanism is proposed to measure the contextual relevance within full texts and a heterogeneous bibliography graph simultaneously. Our motivation is to maintain the topic coherency between a related work section and its target document, where both the textual and graphic contexts play a big role in characterizing the relationship among scientific publications accurately. Experimental results on a large dataset show that our approach achieves a considerable improvement over a typical seq2seq summarizer and five classical summarization baselines.

Improving Neural Abstractive Document Summarization with Explicit Information Selection Modeling
Wei Li, Xinyan Xiao, Yajuan Lyu, and Yuanzhuo Wang 09:18–09:36

Information selection is the most important component in document summarization task. In this paper, we propose to extend the basic neural encoding-decoding framework with an information selection layer to explicitly model and optimize the information selection process in abstractive document summarization. Specifically, our information selection layer consists of two parts: gated global information filtering and local sentence selection. Unnecessary information in the original document is first globally filtered, then salient sentences are selected locally while generating each summary sentence sequentially. To optimize the information selection process directly, distantly-supervised training guided by the golden summary is also imported. Experimental results demonstrate that the explicit modeling and optimizing of the information selection process improves document summarization performance significantly, which enables our model to generate more informative and concise summaries, and thus significantly outperform state-of-the-art neural abstractive methods.

Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization
Shashi Narayan, Shay B. Cohen, and Mirella Lapata 09:36–09:54

We introduce "extreme summarization", a new single-document summarization task which does not favor extractive strategies and calls for an abstractive modeling approach. The idea is to create a short, one-sentence news summary answering the question "What is the article about?". We collect a real-world, large-scale dataset for this task by harvesting online articles from the British Broadcasting Corporation (BBC). We propose a novel abstractive model which is conditioned on the article's topics and based entirely on convolutional neural networks. We demonstrate experimentally that this architecture captures long-range dependencies in a document and recognizes pertinent content, outperforming an oracle extractive system and state-of-the-art abstractive approaches when evaluated automatically and by humans.

Improving Abstraction in Text Summarization
Wojciech Kryściński, Romain Paulus, Caiming Xiong, and Richard Socher 09:54–10:12

Abstractive text summarization aims to shorten long text documents into a human readable form that contains the most important facts from the original document. However, the level of actual abstraction as measured by novel phrases that do not appear in the source document remains low in existing approaches. We propose two techniques to improve the level of abstraction of generated summaries. First, we decompose the decoder into a contextual network that retrieves relevant parts of the source document, and a pretrained language model that incorporates prior knowledge about language generation. Second, we propose a novelty metric that is optimized directly through policy learning to encourage the generation of novel phrases. Our model achieves results comparable to state-of-the-art models, as determined by ROUGE scores and human evaluations, while achieving a significantly higher level of abstraction as measured by n-gram overlap with the source document.

Content Selection in Deep Learning Models of Summarization
Chris Kedzie, Kathleen McKeown, and Hal Daume III 10:12–10:30

We carry out experiments with deep learning models of summarization across the domains of news, personal stories, meetings, and medical articles in order to understand how content selection is performed. We find that many sophisticated features of state of the art extractive summarizers do not improve performance over simpler models. These results suggest that it is easier to create a summarizer for a new domain than previous work suggests and bring into question the benefit of deep learning models for summarization for those domains that do have massive datasets (i.e., news). At the same time, they suggest important questions for new research in summarization; namely, new forms of sentence representations or external knowledge sources are needed that are better suited to the sumarization task.

Session 5C: IR / Text Mining

Silver Hall Chair: *chairname*

Improved Semantic-Aware Network Embedding with Fine-Grained Word Alignment
Dinghan Shen, Xinyuan Zhang, Ricardo Henao, and Lawrence Carin 09:00–09:18

Network embeddings, which learns low-dimensional representations for each vertex in a large-scale network, have received considerable attention in recent years. For a wide range of applications, vertices in a network are typically accompanied by rich textual information such as user profiles, paper abstracts, etc. In this paper, we propose to incorporate semantic features into network embeddings by matching important words between text sequences for all pairs of vertices. We introduce an word-by-word alignment framework that measures the compatibility of embeddings between word pairs, and then adaptively accumulates these alignment features with a simple yet effective aggregation function. In experiments, we evaluate the proposed framework on three real-world benchmarks for downstream tasks, including link prediction and multi-label vertex classification. The experimental results demonstrate that our model outperforms state-of-the-art network embedding methods by a large margin.

Learning Context-Aware Convolutional Filters for Text Processing
Dinghan Shen, Martin Renqiang Min, Yitong Li, and Lawrence Carin 09:18–09:36

Convolutional neural networks (CNNs) have recently emerged as a popular building block for natural language processing (NLP). Despite their success, most existing CNN models employed in NLP share the same learned (and static) set of filters for all input sentences. In this paper, we consider an approach of using a small meta network to learn context-sensitive convolutional filters for text processing. The role of meta network is to abstract the contextual information of a sentence or document into a set of input-sensitive filters. We further generalize this framework to model sentence pairs, where a bidirectional filter generation mechanism is introduced to encapsulate co-dependent sentence representations. In our benchmarks on four different tasks, including ontology classification, sentiment analysis, answer sentence selection, and paraphrase identification, our proposed model, a modified CNN with context-sensitive filters, consistently outperforms the standard CNN and attention-based CNN baselines. By visualizing the learned context-sensitive filters, we further validate and rationalize the effectiveness of proposed framework.

Deep Relevance Ranking using Enhanced Document-Query Interactions
Ryan McDonald, George Brokos, and Ion Androutsopoulos 09:36–09:54

We explore several new models for document relevance ranking, building upon the Deep Relevance Matching Model (DRMM) of Guo et al. (2016). Unlike DRMM, which uses context-insensitive encodings of terms and query-document term interactions, we inject rich context-sensitive encodings throughout our models, inspired by PACRR's (Hui et al., 2017) convolutional n-gram matching features, but extended in several ways including multiple views of query and document inputs. We test our models on datasets from the BIOASQ question answering challenge (Tsatsaronis et al., 2015) and TREC ROBUST 2004 (Voorhees, 2005), showing they outperform BM25-based baselines, DRMM, and PACRR.

Learning Neural Embeddings for CLIR with Adversarial Framework
Bo Li and Ping Cheng 09:54–10:12

The existing studies in cross-language information retrieval (CLIR) mostly rely on general text representation models (e.g., vector space model or latent semantic analysis). These models are not optimized for the target retrieval task. In this paper, we follow the success of neural representation in natural language processing (NLP) and develop a novel text representation model based on adversarial learning, which seeks a task-specific embedding space for CLIR. Adversarial learning is implemented as an interplay between the generator process and the discriminator process. In order to adapt adversarial learning to CLIR, we design three constraints to direct representation learning, which are (1) a matching constraint capturing essential characteristics of cross-language ranking, (2) a translation constraint bridging language gaps, and (3) an adversarial constraint forcing both language and media invariant to be reached more efficiently and effectively. Through the joint exploitation of these constraints in an adversarial manner, the underlying cross-language semantics relevant to retrieval tasks are better preserved in the embedding space. Standard CLIR experiments show that our model significantly outperforms state-of-the-art continuous space models and is better than the strong machine translation baseline.

AD3: Attentive Deep Document Dater
Swayambhu Nath Ray, SHIB SANKAR DASGUPTA, and Partha Talukdar 10:12–10:30

Knowledge of the creation date of documents facilitates several tasks such as summarization, event extraction, temporally focused information extraction etc. Unfortunately, for most of the documents on the Web, the time-stamp metadata is either missing or can't be trusted. Thus, predicting creation time from document content itself is an important task. In this paper, we propose Attentive Deep Document Dater (AD3), an attention-based neural document dating system which utilizes both context and temporal information in documents in a flexible and principled manner. We perform extensive experimentation on multiple real-world datasets to demonstrate the effectiveness of AD3 over neural and non-neural baselines.

Session 5D: Machine Learning II

Hall 100 Chair: *chairname*

Gromov-Wasserstein Alignment of Word Embedding Spaces
David Alvarez-Melis and Tommi Jaakkola 09:00–09:18

Cross-lingual or cross-domain correspondences play key roles in tasks ranging from machine translation to transfer learning. Recently, purely unsupervised methods operating on monolingual embeddings have become effective alignment tools. Current state-of-the-art methods, however, involve multiple steps, including heuristic post-hoc refinement strategies. In this paper, we cast the correspondence problem directly as an optimal transport (OT) problem, building on the idea that word embeddings arise from metric recovery algorithms. Indeed, we exploit the Gromov-Wasserstein distance that measures how similarities between pairs of words relate across languages. We show that our OT objective can be estimated efficiently, requires little or no tuning, and results in performance comparable with the state-of-the-art in various unsupervised word translation tasks.

Deep Probabilistic Logic: A Unifying Framework for Indirect Supervision
Hai Wang and Hoifung Poon 09:18–09:36

Deep learning has emerged as a versatile tool for a wide range of NLP tasks, due to its superior capacity in representation learning. But its applicability is limited by the reliance on annotated examples, which are difficult to produce at scale. Indirect supervision has emerged as a promising direction to address this bottleneck, either by introducing labeling functions to automatically generate noisy examples from unlabeled text, or by imposing constraints over interdependent label decisions. A plethora of methods have been proposed, each with respective strengths and limitations. Probabilistic logic offers a unifying language to represent indirect supervision, but end-to-end modeling with probabilistic logic is often infeasible due to intractable inference and learning. In this paper, we propose deep probabilistic logic (DPL) as a general framework for indirect supervision, by composing probabilistic logic with deep learning. DPL models label decisions as latent variables, represents prior knowledge on their relations using weighted first-order logical formulas, and alternates between learning a deep neural network for the end task and refining uncertain formula weights for indirect supervision, using variational EM. This framework subsumes prior indirect supervision methods as special cases, and enables novel combination via infusion of rich domain and linguistic knowledge. Experiments on biomedical machine reading demonstrate the promise of this approach.

Deriving Machine Attention from Human Rationales
Yujia Bao, Shiyu Chang, Mo Yu, and Regina Barzilay 09:36–09:54

Attention-based models are successful when trained on large amounts of data. In this paper, we demonstrate that even in the low-resource scenario, attention can be learned effectively. To this end, we start with discrete human-annotated rationales and map them into continuous attention. Our central hypothesis is that this mapping is general across domains, and thus can be transferred from resource-rich domains to low-resource ones. Our model jointly learns a domain-invariant representation and induces the desired mapping between rationales and attention. Our empirical results validate this hypothesis and show that our approach delivers significant gains over state-of-the-art baselines, yielding over 15% average error reduction on benchmark datasets.

Semi-Supervised Sequence Modeling with Cross-View Training
Kevin Clark, Minh-Thang Luong, Christopher D. Manning, and Quoc Le 09:54–10:12

Unsupervised representation learning algorithms such as word2vec and ELMo improve the accuracy of many supervised NLP models, mainly because they can take advantage of large amounts of unlabeled text. However, the supervised models only learn from task-specific labeled data during the main training phase. We therefore propose Cross-View Training (CVT), a semi-supervised learning algorithm that improves the representations of a Bi-LSTM sentence encoder using a mix of labeled and unlabeled data. On labeled examples, standard supervised learning is used. On unlabeled examples, CVT teaches auxiliary prediction modules that see restricted views of the input (e.g., only part of a sentence) to match the predictions of the full model seeing the whole input. Since the auxiliary modules and the full model share intermediate representations, this in turn improves the full model. Moreover, we show that CVT is particularly effective when combined with multi-task learning. We evaluate CVT on five sequence tagging tasks, machine translation, and dependency parsing, achieving state-of-the-art results.

TACL-1430
TACL-1430 10:12–10:30

The analysis of crowdsourced annotations in NLP is concerned with identifying 1) gold standard labels, 2) annotator accuracies and biases, and 3) item difficulties and error patterns. Traditionally, majority voting was used for 1), and coefficients of agreement for 2) and 3). Lately, model-based analysis of corpus annotation have proven better at all three tasks. But there has been relatively little work comparing them on the same datasets. This paper aims to fill this gap by analyzing six models of annotation, covering different approaches to annotator ability, item difficulty, and parameter pooling (tying) across annotators and items. We evaluate these models along four aspects: comparison to gold labels, predictive accuracy for new annotations, annotator characterization, and item difficulty, using four datasets with varying degrees of noise in the form of random (spammy) annotators. We conclude with guidelines for model selection, application, and implementation.

Session 5E: Information Extraction, Question Answering 09:00–10:30

Hall 100 Chair: *chairname*

A Probabilistic Annotation Model for Crowdsourcing Coreference

Silviu Paun, Jon Chamberlain, Udo Kruschwitz, Juntao Yu, and Massimo Poesio

The availability of large scale annotated corpora for coreference is essential to the development of the field. However, creating resources at the required scale via expert annotation would be too expensive. Crowdsourcing has been proposed as an alternative; but this approach has not been widely used for coreference. This paper addresses one crucial hurdle on the way to make this possible, by introducing a new model of annotation for aggregating crowdsourced anaphoric annotations. The model is evaluated along three dimensions: the accuracy of the inferred mention pairs, the quality of the post-hoc constructed silver chains, and the viability of using the silver chains as an alternative to the expert-annotated chains in training a state of the art coreference system. The results suggest that our model can extract from crowdsourced annotations coreference chains of comparable quality to those obtained with expert annotation.

A Deterministic Algorithm for Bridging Anaphora Resolution

Yufang Hou

Previous work on bridging anaphora resolution (Poesio et al., 2004; Hou et al., 2013) use syntactic preposition patterns to calculate word relatedness. However, such patterns only consider NPs' head nouns and hence do not fully capture the semantics of NPs. Recently, Hou (2018) created word embeddings (embeddings_PP) to capture associative similarity (i.e., relatedness) between nouns by exploring the syntactic structure of noun phrases. But embeddings_PP only contains word representations for nouns. In this paper, we create new word vectors by combining embeddings_PP with GloVe. This new word embeddings (embeddings_bridging) are a more general lexical knowledge resource for bridging and allow us to represent the meaning of an NP beyond its head easily. We therefore develop a deterministic approach for bridging anaphora resolution, which represents the semantics of an NP based on its head noun and modifications. We show that this simple approach achieves the competitive results compared to the best system in Hou et al. (2013) which explores Markov Logic Networks to model the problem. Additionally, we further improve the results for bridging anaphora resolution reported in Hou (2018) by combining our simple deterministic approach with Hou et al. (2013)'s best system MLN II.

A Knowledge Hunting Framework for Common Sense Reasoning

Ali Emami, Noelia De La Cruz, Adam Trischler, Kaheer Suleman, and Jackie Chi Kit Cheung

We introduce an automatic system that achieves state-of-the-art results on the Winograd Schema Challenge (WSC), a common sense reasoning task that requires diverse, complex forms of inference and knowledge. Our method uses a knowledge hunting module to gather text from the web, which serves as evidence for candidate problem resolutions. Given an input problem, our system generates relevant queries to send to a search engine, then extracts and classifies knowledge from the returned results and weighs them to make a resolution. Our approach improves F1 performance on the full WSC by 0.21 over the previous best and represents the first system to exceed 0.5 F1. We further demonstrate that the approach is competitive on the Choice of Plausible Alternatives (COPA) task, which suggests that it is generally applicable.

Mapping Text to Knowledge Graph Entities by Vector Space Transformation

Dimitri Kartsaklis, Mohammad Taher Pilehvar, and Nigel Collier

This paper addresses the problem of mapping natural language text to knowledge base entities. The mapping process is approached as a composition of a phrase or a sentence into a point in a multi-dimensional entity space obtained from a knowledge graph. The compositional model is an LSTM equipped with a dynamic disambiguation mechanism on the input word embeddings (a Multi-Sense LSTM), addressing polysemy issues. Further, the knowledge base space is prepared by collecting random walks from a graph enhanced with textual features, which act as a set of semantic bridges between text and knowledge base entities. The ideas of this work are demonstrated on large-scale text-to-entity mapping and entity classification tasks, with state of the art results.

Differentiating Concepts and Instances for Knowledge Graph Embedding

Xin Lv, Lei Hou, Juanzi Li, and Zhiyuan Liu

Concepts, which represent a group of different instances sharing common properties, are essential information in knowledge representation. Most conventional knowledge embedding methods encode both entities (concepts and instances) and relations as vectors in a low dimensional semantic space equally, ignoring the difference between concepts and instances. In this paper, we propose a novel knowledge graph embedding model named TransC by differentiating concepts and instances. Specifically, TransC encodes each concept in knowledge graph as a sphere and each instance as a vector in the same semantic space. We use the relative positions to model the relations between concepts and instances (i.e.,instanceOf), and the relations between concepts and sub-concepts (i.e., subClassOf). We evaluate our model on both link prediction and triple classification tasks on the dataset based on YAGO. Experimental results show that TransC outperforms state-of-the-art methods, and captures the semantic transitivity for instanceOf and subClassOf relation. Our codes and datasets can be obtained from https:// github.com/davidlvxin/TransC.

One-Shot Relational Learning for Knowledge Graphs
Wenhan Xiong, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang

Knowledge graphs (KG) are the key components of various natural language processing applications. To further expand KGs' coverage, previous studies on knowledge graph completion usually require a large number of positive examples for each relation. However, we observe long-tail relations are actually more common in KGs and those newly added relations often do not have many known triples for training. In this work, we aim at predicting new facts under a challenging setting where only one training instance is available. We propose a one-shot relational learning framework, which utilizes the knowledge distilled by embedding models and learns a matching metric by considering both the learned embeddings and one-hop graph structures. Empirically, our model yields considerable performance improvements over existing embedding models, and also eliminates the need of re-training the embedding models when dealing with newly added relations.

Regular Expression Guided Entity Mention Mining from Noisy Web Data
Shanshan Zhang, Lihong He, Slobodan Vucetic, and Eduard Dragut

Many important entity types in web documents, such as dates, times, email addresses, and course numbers, follow or closely resemble patterns that can be described by Regular Expressions (REs). Due to a vast diversity of web documents and ways in which they are being generated, even seemingly straightforward tasks such as identifying mentions of date in a document become very challenging. It is reasonable to claim that it is impossible to create a RE that is capable of identifying such entities from web documents with perfect precision and recall. Rather than abandoning REs as a go-to approach for entity detection, this paper explores ways to combine the expressive power of REs, ability of deep learning to learn from large data, and human-in-the loop approach into a new integrated framework for entity identification from web data. The framework starts by creating or collecting the existing REs for a particular type of an entity. Those REs are then used over a large document corpus to collect weak labels for the entity mentions and a neural network is trained to predict those RE-generated weak labels. Finally, a human expert is asked to label a small set of documents and the neural network is fine tuned on those documents. The experimental evaluation on several entity identification problems shows that the proposed framework achieves impressive accuracy, while requiring very modest human effort.

HyTE: Hyperplane-based Temporally aware Knowledge Graph Embedding
SHIB SANKAR DASGUPTA, Swayambhu Nath Ray, and Partha Talukdar

Knowledge Graph (KG) embedding has emerged as an active area of research resulting in the development of several KG embedding methods. Relational facts in KG often show temporal dynamics, e.g., the fact (Cristiano_Ronaldo, playsFor, Manchester_United) is valid only from 2003 to 2009. Most of the existing KG embedding methods ignore this temporal dimension while learning embeddings of the KG elements. In this paper, we propose HyTE, a temporally aware KG embedding method which explicitly incorporates time in the entity-relation space by associating each timestamp with a corresponding hyperplane. HyTE not only performs KG inference using temporal guidance, but also predicts temporal scopes for relational facts with missing time annotations. Through extensive experimentation on temporal datasets extracted from real-world KGs, we demonstrate the effectiveness of our model over both traditional as well as temporal KG embedding methods.

Neural Adaptation Layers for Cross-domain Named Entity Recognition
Bill Yuchen Lin and Wei Lu

Recent research efforts have shown that neural architectures can be effective in conventional information extraction tasks such as named entity recognition, yielding state-of-the-art results on standard newswire datasets. However, despite significant resources required for training such models, the performance of a model trained on one domain typically degrades dramatically when applied to a different domain, yet extracting entities from new emerging domains such as social media can be of significant interest. In this paper, we empirically investigate effective methods for conveniently adapting an existing, well-trained neural NER model for a new domain. Unlike existing approaches, we propose lightweight yet effective methods for performing domain adaptation for neural models. Specifically, we introduce adaptation layers on top of existing neural architectures, where no re-training using the source domain data is required. We conduct extensive empirical studies and show that our approach significantly outperforms state-of-the-art methods.

Entity Linking within a Social Media Platform: A Case Study on Yelp
Hongliang Dai, Yangqiu Song, Liwei Qiu, and Rijia Liu

In this paper, we study a new entity linking problem where both the entity mentions and the target entities are within a same social media platform. Compared with traditional entity linking problems that link mentions to a knowledge base, this new problem have less information about the target entities. However, if we can successfully link mentions to entities within a social media platform, we can improve a lot of applications such as comparative study in business intelligence and opinion leader finding. To study this problem, we constructed a dataset called Yelp-EL, where the business mentions in Yelp reviews are linked to their corresponding businesses on the platform. We conducted comprehensive experiments and analysis on this dataset with a learning to rank model that takes different types of features as input, as well as a few state-of-the-art entity linking approaches. Our experimental results show that two types of features that are not available in traditional entity linking: social features and location features, can be very helpful for this task.

Annotation of Large Clinical Entity Corpus
Pinal Patel, Disha Davey, Vishal Panchal, and Parth Pathak

Having an entity annotated corpus of the clinical domain is one of the basic requirements for detection of clinical entities using machine learning (ML) approaches. Past researches have shown the superiority of statistical/ML approaches over the rule based approaches. But in order to take full advantage of the ML approaches, an accurately annotated corpus becomes an essential requirement. Though there are a few annotated corpora available either on a small data set, or covering a narrower domain (like cancer patients records, lab reports), annotation of a large data set representing the entire clinical domain has not been created yet. In this paper, we have described in detail the annotation guidelines, annotation process and our approaches in creating a CER (clinical entity recognition) corpus of 5,160 clinical documents from forty different clinical specialities. The clinical entities range across various types such as diseases, procedures, medications, medical devices and so on. We have classified them into eleven categories for annotation. Our annotation also reflects the relations among the group of entities that constitute larger concepts altogether.

Visual Supervision in Bootstrapped Information Extraction
Matthew Berger, Ajay Nagesh, Joshua Levine, Mihai Surdeanu, and Helen Zhang

We challenge a common assumption in active learning, that a list-based interface populated by informative samples provides for efficient and effective data annotation. We show how a 2D scatterplot populated with diverse and representative samples can yield improved models given the same time budget. We consider this for bootstrapping-based information extraction, in particular named entity classification, where human and machine jointly label data. To enable effective data annotation in a scatterplot, we have developed an embedding-based bootstrapping model that learns the distributional similarity of entities through the patterns that match them in a large data corpus, while being discriminative with respect to human-labeled and machine-promoted entities. We conducted a user study to assess the effectiveness of these different interfaces, and analyze bootstrapping performance in terms of human labeling accuracy, label quantity, and labeling consensus across multiple users. Our results suggest that supervision acquired from the scatterplot interface, despite being noisier, yields improvements in classification performance compared with the list interface, due to a larger quantity of supervision acquired.

Learning Named Entity Tagger using Domain-Specific Dictionary
Jingbo Shang, Liyuan Liu, Xiaotao Gu, Xiang Ren, Teng Ren, and Jiawei Han

Recent advances in deep neural models allow us to build reliable named entity recognition (NER) systems without handcrafting features. However, such methods require large amounts of manually-labeled training data. There have been efforts on replacing human annotations with distant supervision (in conjunction with external dictionaries), but the generated noisy labels pose significant challenges on learning effective neural models. Here we propose two neural models to suit noisy distant supervision from the dictionary. First, under the traditional sequence labeling framework, we propose a revised fuzzy CRF layer to handle tokens with multiple possible labels. After identifying the nature of noisy labels in distant supervision, we go beyond the traditional framework and propose a novel, more effective neural model AutoNER with a new Tie or Break scheme. In addition, we discuss how to refine distant supervision for better NER performance. Extensive experiments on three benchmark datasets demonstrate that AutoNER achieves the best performance when only using dictionaries with no additional human effort, and delivers competitive results with state-of-the-art supervised benchmarks.

Zero-Shot Open Entity Typing as Type-Compatible Grounding
Ben Zhou, Daniel Khashabi, Chen-Tse Tsai, and Dan Roth

The problem of entity-typing has been studied predominantly as a supervised learning problems, mostly with task-specific annotations (for coarse types) and sometimes with distant supervision (for fine types). While such approaches have strong performance within datasets they often lack the flexibility to transfer across text genres and to generalize to new type taxonomies. In this work we propose a zero-shot entity typing approach that requires no annotated data and can flexibly identify newly defined types. Given a type taxonomy, the entries of which we define as Boolean functions of freebase "types," we ground a given mention to a set of *type-compatible* Wikipedia entries, and then infer the target mention's type using an inference algorithm that makes use of the types of these entries. We evaluate our system on a broad range of datasets, including standard fine-grained and coarse-grained entity typing datasets, and on a dataset in the biological domain. Our system is shown to be competitive with state-of-the-art supervised NER systems, and to outperform them on out-of-training datasets. We also show that our system significantly outperforms other zero-shot fine typing systems.

Attention-Guided Answer Distillation for Machine Reading Comprehension
Minghao Hu, Yuxing Peng, Furu Wei, Zhen Huang, Dongsheng Li, Nan Yang, and Ming Zhou

Despite that current reading comprehension systems have achieved significant advancements, their promising performances are often obtained at the cost of making an ensemble of numerous models. Besides, existing approaches are also vulnerable to adversarial attacks. This paper tackles these problems by leveraging knowledge distillation, which aims to transfer knowledge from an ensemble model to a single model. We first demonstrate that vanilla knowledge distillation applied to answer span prediction is effective for reading comprehension systems. We then propose two novel approaches that not only penalize the prediction on confusing answers but also guide the training with alignment information distilled from the ensemble. Experiments show that our best student model has only a slight drop of 0.4% F1 on the SQuAD test set compared to the ensemble teacher, while running 12x faster during inference. It even outperforms the teacher on adversarial SQuAD datasets and NarrativeQA benchmark.

Interpretation of Natural Language Rules in Conversational Machine Reading
Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel

Most work in machine reading focuses on question answering problems where the answer is directly expressed in the text to read. However, many real-world question answering problems require the reading of text not because it contains the literal answer, but because it contains a recipe to derive an answer together with the reader's background knowledge. One example is the task of interpreting regulations to answer "Can I...?" or "Do I have to...?" questions such as "I am working in Canada. Do I have to carry on paying UK National Insurance?" after reading a UK government website about this topic. This task requires both the interpretation of rules and the application of background knowledge. It is further complicated due to the fact that, in practice, most questions are underspecified, and a human assistant will regularly have to ask clarification questions such as "How long have you been working abroad?" when the answer cannot be directly derived from the question and text. In this paper, we formalise this task and develop a crowd-sourcing strategy to collect 37k task instances based on real-world rules and crowd-generated questions and scenarios. We analyse the challenges of this task and assess its difficulty by evaluating the performance of rule-based and machine-learning baselines. We observe promising results when no background knowledge is necessary, and substantial room for improvement whenever background knowledge is needed.

A State-transition Framework to Answer Complex Questions over Knowledge Base
Sen Hu, Lei Zou, and Xinbo Zhang

Although natural language question answering over knowledge graphs have been studied in the literature, existing methods have some limitations in answering complex questions. To address that, in this paper, we propose a State Transition-based approach to translate a complex natural language question N to a semantic query graph (SQG), which is used to match the underlying knowledge graph to find the answers to question N. In order to generate SQG, we propose four primitive operations (expand, fold, connect and merge) and a learning-based state transition approach. Extensive experiments on several benchmarks (such as QALD, WebQuestions and ComplexQuestions) with two knowledge bases (DBpedia and Freebase) confirm the superiority of our approach compared with state-of-the-arts.

A Multi-answer Multi-task Framework for Real-world Machine Reading Comprehension
Jiahua Liu, Wan Wei, Maosong Sun, Hao Chen, Yantao Du, and Dekang Lin

The task of machine reading comprehension (MRC) has evolved from answering simple questions from well-edited text to answering real questions from users out of web data. In the real-world setting, full-body text from multiple relevant documents in the top search results are provided as context for questions from user queries, including not only questions with a single, short, and factual answer, but also questions about reasons, procedures, and opinions. In this case, multiple answers could be equally valid for a single question and each answer may occur multiple times in the context, which should be taken into consideration when we build MRC system. We propose a multi-answer multi-task framework, in which different loss functions are used for multiple reference answers. Minimum Risk Training is applied to solve the multi-occurrence problem of a single answer. Combined with a simple heuristic passage extraction strategy for overlong documents, our model increases the ROUGE-L score on the DuReader dataset from 44.18, the previous state-of-the-art, to 51.09.

Logician and Orator: Learning from the Duality between Language and Knowledge in Open Domain
Mingming Sun, Xu Li, and Ping Li

We propose the task of Open-Domain Information Narration (OIN) as the reverse task of Open Information Extraction (OIE), to implement the dual structure between language and knowledge in the open domain. Then, we develop an agent, called Orator, to accomplish the OIN task, and assemble the Orator and the recently proposed OIE agent — Logician~[**Sun2018**] into a dual system to utilize the duality structure with a reinforcement learning paradigm. Experimental results reveal the dual structure between OIE and OIN tasks helps to build better both OIE agents and OIN agents.

Deep and Wide Reader: Effective Memory Augmenting Method for Large-Scale Reading Comprehension
Seohyun Back, Seunghak Yu, Sathish Reddy Indurthi, Jihie Kim, and Jaegul Choo

Machine reading comprehension helps machines learn to utilize most of the human knowledge written in the form of text. Existing approaches made a significant progress comparable to human-level performance, but they are still limited in understanding, up to a few paragraphs, failing to properly comprehend lengthy document. In this paper, we propose a novel deep neural network architecture to handle a long-range dependency in RC tasks. In detail, our method has two novel aspects: (1) an advanced memory-augmented architecture and (2) an expanded gated recurrent unit with dense connections that mitigate potential information distortion occurring in the memory. Our proposed architecture is widely applicable to other models. We have performed extensive experiments with well-known benchmark datasets such as TriviaQA, QUASAR-T, and SQuAD. The experimental results demonstrate that the proposed method outperforms existing methods, especially for lengthy documents.

Multi-Granular Sequence Encoding via Dilated Compositional Units for Reading Comprehension
Yi Tay, Anh Tuan Luu, and Siu Cheung Hui

Sequence encoders are crucial components in many neural architectures for learning to read and comprehend. This paper presents a new compositional encoder for reading comprehension (RC). Our proposed encoder is not only aimed at being fast but also expressive. Specifically, the key novelty behind our encoder is that it explicitly models across multiple granularities using a new dilated composition mechanism. In our approach, gating functions are learned by modeling relationships and reasoning over multi-granular sequence information, enabling compositional learning that is aware of both long and short term information. We conduct experiments on three RC datasets, showing that our proposed encoder demonstrates very promising results both as a standalone encoder as well as a complementary building block. Empirical results show that simple Bi-Attentive architectures augmented with our proposed encoder not only achieves state-of-the-art / highly competitive results but is also considerably faster than other published works.

Neural Compositional Denotational Semantics for Question Answering
Nitish Gupta and Mike Lewis

Answering compositional questions requiring multi-step reasoning is challenging. We introduce an end-to-end differentiable model for interpreting questions about a knowledge graph (KG), which is inspired by formal approaches to semantics. Each span of text is represented by a denotation in a KG and a vector that captures ungrounded aspects of meaning. Learned composition modules recursively combine constituent spans, culminating in a grounding for the complete sentence which answers the question. For example, to interpret "not green", the model represents "green" as a set of KG entities and "not" as a trainable ungrounded vector—and then uses this vector to parameterize a composition function that performs a complement operation. For each sentence, we build a parse chart subsuming all possible parses, allowing the model to jointly learn both the composition operators and output structure by gradient descent from end-task supervision. The model learns a variety of challenging semantic operators, such as quantifiers, disjunctions and composed relations, and infers latent syntactic structure. It also generalizes well to longer questions than seen in its training data, in contrast to RNN, its tree-based variants, and semantic parsing baselines.

Cross-Pair Text Representations for Answer Sentence Selection
Kateryna Tymoshenko and Alessandro Moschitti

High-level semantics tasks, e.g., paraphrasing, textual entailment or question answering, involve modeling of text pairs. Before the emergence of neural networks, this has been mostly performed using intra-pair features, which incorporate similarity scores or rewrite rules computed between the members within the same pair. In this paper, we compute scalar products between vectors representing similarity between members of different pairs, in place of simply using a single vector for each pair. This allows us to obtain a representation specific to any pair of pairs, which delivers the state of the art in answer sentence selection. Most importantly, our approach can outperform much more complex algorithms based on neural networks.

QuAC: Question Answering in Context
Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer

We present QuAC, a dataset for Question Answering in Context that contains 14K information-seeking QA dialogs (100K questions in total). The dialogs involve two crowd workers: (1) a student who poses a sequence of freeform questions to learn as much as possible about a hidden Wikipedia text, and (2) a teacher who answers the questions by providing short excerpts from the text. QuAC introduces challenges not found in existing machine comprehension datasets: its questions are often more open-ended, unanswerable, or only meaningful within the dialog context, as we show in a detailed qualitative evaluation. We also report results for a number of reference models, including a recently state-of-the-art reading comprehension architecture extended to model dialog context. Our best model underperforms humans by 20 F1, suggesting that there is significant room for future work on this data. Dataset, baseline, and leaderboard available at http://quac.ai.

Knowledge Base Question Answering via Encoding of Complex Query Graphs
Kangqi Luo, Fengli Lin, Xusheng Luo, and Kenny Zhu

Answering complex questions that involve multiple entities and multiple relations using a standard knowledge base is an open and challenging task. Most existing KBQA approaches focus on simpler questions and do not work very well on complex questions because they were not able to simultaneously represent the question and the corresponding complex query structure. In this work, we encode such complex query structure into a uniform vector representation, and thus successfully capture the interactions between individual semantic components within a complex question. This approach consistently outperforms existing methods on complex questions while staying competitive on simple questions.

Neural Relation Extraction via Inner-Sentence Noise Reduction and Transfer Learning
Tianyi Liu, Xinsong Zhang, Wanhao Zhou, and Weijia Jia

Extracting relations is critical for knowledge base completion and construction in which distant supervised methods are widely used to extract relational facts automatically with the existing knowledge bases.

However, the automatically constructed datasets comprise amounts of low-quality sentences containing noisy words, which is neglected by current distant supervised methods resulting in unacceptable precisions. To mitigate this problem, we propose a novel word-level distant supervised approach for relation extraction. We first build Sub-Tree Parse(STP) to remove noisy words that are irrelevant to relations. Then we construct a neural network inputting the sub-tree while applying the entity-wise attention to identify the important semantic features of relational words in each instance. To make our model more robust against noisy words, we initialize our network with a priori knowledge learned from the relevant task of entity classification by transfer learning. We conduct extensive experiments using the corpora of New York Times(NYT) and Freebase. Experiments show that our approach is effective and improves the area of Precision/Recall(PR) from 0.35 to 0.39 over the state-of-the-art work.

Graph Convolution over Pruned Dependency Trees Improves Relation Extraction
Yuhao Zhang, Peng Qi, and Christopher D. Manning

Dependency trees help relation extraction models capture long-range relations between words. However, existing dependency-based models either neglect crucial information (e.g., negation) by pruning the dependency trees too aggressively, or are computationally inefficient because it is difficult to parallelize over different tree structures. We propose an extension of graph convolutional networks that is tailored for relation extraction, which pools information over arbitrary dependency structures efficiently in parallel. To incorporate relevant information while maximally removing irrelevant content, we further apply a novel pruning strategy to the input trees by keeping words immediately around the shortest path between the two entities among which a relation might hold. The resulting model achieves state-of-the-art performance on the large-scale TACRED dataset, outperforming existing sequence and dependency-based neural models. We also show through detailed analysis that this model has complementary strengths to sequence models, and combining them further improves the state of the art.

Multi-Level Structured Self-Attentions for Distantly Supervised Relation Extraction
Jinhua Du, Jingguang Han, Andy Way, and Dadong Wan

Attention mechanism is often used in deep neural networks for distantly supervised rela- tion extraction (DS-RE) to distinguish valid from noisy instances. However, traditional 1- D vector attention model is insufficient for learning of different contexts in the selection of valid instances to predict the relationship for an entity pair. To alleviate this issue, we propose a novel multi-level structured (2- D matrix) self-attention mechanism for DS- RE in a multi-instance learning (MIL) frame- work using bidirectional recurrent neural net- works (BiRNN). In the proposed method, a structured word-level self-attention learns a 2- D matrix where each row vector represents a weight distribution for different aspects of an instance regarding two entities. Targeting the MIL issue, the structured sentence-level attention learns a 2-D matrix where each row vector represents a weight distribution on selection of different valid instances. Experiments conducted on two publicly available DS-RE datasets show that the proposed frame- work with multi-level structured self-attention mechanism significantly outperform baselines in terms of PR curves, PN and F1 measures.

N-ary Relation Extraction using Graph State LSTM
Linfeng Song, Yue Zhang, Zhiguo Wang, and Daniel Gildea

Cross-sentence n-ary relation extraction detects relations among n entities across multiple sentences. Typical methods formulate an input as a *document graph*, integrating various intra-sentential and inter-sentential dependencies. The current state-of-the-art method splits the input graph into two DAGs, adopting a DAG-structured LSTM for each. Though being able to model rich linguistic knowledge by leveraging graph edges, important information can be lost in the splitting procedure. We propose a graph-state LSTM model, which uses a parallel state to model each word, recurrently enriching state values via message passing. Compared with DAG LSTMs, our graph LSTM keeps the original graph structure, and speeds up computation by allowing more parallelization. On a standard benchmark, our model shows the best result in the literature.

Hierarchical Relation Extraction with Coarse-to-Fine Grained Attention
Xu Han, Pengfei Yu, Zhiyuan Liu, Maosong Sun, and Peng Li

Distantly supervised relation extraction employs existing knowledge graphs to automatically collect training data. While distant supervision is effective to scale relation extraction up to large-scale corpora, it inevitably suffers from the wrong labeling problem. Many efforts have been devoted to identifying valid instances from noisy data. However, most existing methods handle each relation in isolation, regardless of rich semantic correlations located in relation hierarchies. In this paper, we aim to incorporate the hierarchical information of relations for distantly supervised relation extraction and propose a novel hierarchical attention scheme. The multiple layers of our hierarchical attention scheme provide coarse-to-fine granularity to better identify valid instances, which is especially effective for extracting those long-tail relations. The experimental results on a large-scale benchmark dataset demonstrate that our models are capable of modeling the hierarchical information of relations and significantly outperform other baselines. The source code of this paper can be obtained from https://github.com/thunlp/HNRE.

Label-Free Distant Supervision for Relation Extraction via Knowledge Graph Embedding
Guanying Wang, Wen Zhang, Ruoxu Wang, Yalin Zhou, Xi Chen, Wei Zhang, Hai Zhu, and Huajun Chen
Distant supervision is an effective method to generate large scale labeled data for relation extraction, which assumes that if a pair of entities appears in some relation of a Knowledge Graph (KG), all sentences containing those entities in a large unlabeled corpus are then labeled with that relation to train a relation classifier. However, when the pair of entities has multiple relationships in the KG, this assumption may produce noisy relation labels. This paper proposes a label-free distant supervision method, which makes no use of the relation labels under this inadequate assumption, but only uses the prior knowledge derived from the KG to supervise the learning of the classifier directly and softly. Specifically, we make use of the type information and the translation law derived from typical KG embedding model to learn embeddings for certain sentence patterns. As the supervision signal is only determined by the two aligned entities, neither hard relation labels nor extra noise-reduction model for the bag of sentences is needed in this way. The experiments show that the approach performs well in current distant supervision dataset.

Extracting Entities and Relations with Joint Minimum Risk Training
Changzhi Sun, Yuanbin Wu, Man Lan, Shiliang Sun, Wenting Wang, Kuang-Chih Lee, and Kewen Wu
We investigate the task of joint entity relation extraction. Unlike prior efforts, we propose a new lightweight joint learning paradigm based on minimum risk training (MRT). Specifically, our algorithm optimizes a global loss function which is flexible and effective to explore interactions between the entity model and the relation model. We implement a strong and simple neural network where the MRT is executed. Experiment results on the benchmark ACE05 and NYT datasets show that our model is able to achieve state-of-the-art joint extraction performances.

Large-scale Exploration of Neural Relation Classification Architectures
Hoang-Quynh Le, Duy-Cat Can, Sinh T. Vu, Thanh Hai Dang, Mohammad Taher Pilehvar, and Nigel Collier
Experimental performance on the task of relation classification has generally improved using deep neural network architectures. One major drawback of reported studies is that individual models have been evaluated on a very narrow range of datasets, raising questions about the adaptability of the architectures, while making comparisons between approaches difficult. In this work, we present a systematic large-scale analysis of neural relation classification architectures on six benchmark datasets with widely varying characteristics. We propose a novel multi-channel LSTM model combined with a CNN that takes advantage of all currently popular linguistic and architectural features. Our 'Man for All Seasons' approach achieves state-of-the-art performance on two datasets. More importantly, in our view, the model allowed us to obtain direct insights into the continued challenges faced by neural language models on this task.

Possessors Change Over Time: A Case Study with Artworks
Dhivya Chinnappa and Eduardo Blanco
This paper presents a corpus and experimental results to extract possession relations over time. We work with Wikipedia articles about artworks, and extract possession relations along with temporal information indicating when these relations are true. The annotation scheme yields many possessors over time for a given artwork, and experimental results show that an LSTM ensemble can automate the task.

CogCompTime: A Tool for Understanding Time in Natural Language
Qiang Ning, Ben Zhou, Zhili Feng, Haoruo Peng, and Dan Roth
Automatic extraction of temporal information is important for natural language understanding. It involves two basic tasks: (1) Understanding time expressions that are mentioned explicitly in text (e.g., February 27, 1998 or tomorrow), and (2) Understanding temporal information that is conveyed implicitly via relations. This paper introduces CogCompTime, a system that has these two important functionalities.

It incorporates the most recent progress, achieves state-of-the-art performance, and is publicly available at http://cogcomp.org/page/publication_view/844.

DERE: A Task and Domain-Independent Slot Filling Framework for Declarative Relation Extraction
Heike Adel, Laura Ana Maria Bostan, Sean Papay, Sebastian Padó, and Roman Klinger

Most machine learning systems for natural language processing are tailored to specific tasks. As a result, comparability of models across tasks is missing and their applicability to new tasks is limited. This affects end users without machine learning experience as well as model developers. To address these limitations, we present DERE, a novel framework for declarative specification and compilation of template-based information extraction. It uses a generic specification language for the task and for data annotations in terms of spans and frames. This formalism enables the representation of a large variety of natural language processing challenges. The backend can be instantiated by different models, following different paradigms. The clear separation of frame specification and model backend will ease the implementation of new models and the evaluation of different models across different tasks. Furthermore, it simplifies transfer learning, joint learning across tasks and/or domains as well as the assessment of model generalizability. DERE is available as open-source software.

Integrating Knowledge-Supported Search into the INCEpTION Annotation Platform
Beto Boullosa, Richard Eckart de Castilho, Naveen Kumar, Jan-Christoph Klie, and Iryna Gurevych

Annotating entity mentions and linking them to a knowledge resource are essential tasks in many domains. It disambiguates mentions, introduces cross-document coreferences, and the resources contribute extra information, e.g. taxonomic relations. Such tasks benefit from text annotation tools that integrate a search which covers the text, the annotations, as well as the knowledge resource. However, to the best of our knowledge, no current tools integrate knowledge-supported search as well as entity linking support. We address this gap by introducing knowledge-supported search functionality into the INCEpTION text annotation platform. In our approach, cross-document references are created by linking entity mentions to a knowledge base in the form of a structured hierarchical vocabulary. The resulting annotations are then indexed to enable fast and yet complex queries taking into account the text, the annotations, and the vocabulary structure.

OpenKE: An Open Toolkit for Knowledge Embedding
Xu Han, Shulin Cao, Xin Lv, Yankai Lin, Zhiyuan Liu, Maosong Sun, and Juanzi Li

We release an open toolkit for knowledge embedding (OpenKE), which provides a unified framework and various fundamental models to embed knowledge graphs into a continuous low-dimensional space. OpenKE prioritizes operational efficiency to support quick model validation and large-scale knowledge representation learning. Meanwhile, OpenKE maintains sufficient modularity and extensibility to easily incorporate new models into the framework. Besides the toolkit, the embeddings of some existing large-scale knowledge graphs pre-trained by OpenKE are also available, which can be directly applied for many applications including information retrieval, personalized recommendation and question answering. The toolkit, documentation, and pre-trained embeddings are all released on http://openke.thunlp.org/.

An Interactive Web-Interface for Visualizing the Inner Workings of the Question Answering LSTM
Ekaterina Loginova and Günter Neumann

We present a visualisation tool which aims to illuminate the inner workings of an LSTM model for question answering. It plots heatmaps of neurons' firings and allows a user to check the dependency between neurons and manual features. The system possesses an interactive web-interface and can be adapted to other models and domains. We plan to make the demo fully available online in June 2018.

An Interface for Annotating Science Questions
Michael Boratko, Harshit Padigela, Divyendra Mikkilineni, Pritish Yuvraj, Rajarshi Das, Andrew McCallum, Maria Chang, Achille Fokoue, Pavan Kapanipathi, Nicholas Mattei, Ryan Musa, Kartik Talamadupula, and Michael Witbrock

Recent work introduces the AI2 Reasoning Challenge (ARC) and the associated ARC dataset that partitions open domain, complex science questions into an Easy Set and a Challenge Set. That work includes an analysis of 100 questions with respect to the types of knowledge and reasoning required to answer them. However, it does not include clear definitions of these types, nor does it offer information about the quality of the labels or the annotation process used. In this paper, we introduce a novel interface for human annotation of science question-answer pairs with their respective knowledge and reasoning types, in order that the classification of new questions may be improved. We build on the classification schema proposed by prior work on the ARC dataset, and evaluate the effectiveness of our interface with a preliminary study involving 10 participants.

Interactive Instance-based Evaluation of Knowledge Base Question Answering
Daniil Sorokin and Iryna Gurevych

Most approaches to Knowledge Base Question Answering are based on semantic parsing. In this paper, we present a tool that aids in debugging of question answering systems that construct a structured semantic representation for the input question. Previous work has largely focused on building question answering interfaces or evaluation frameworks that unify multiple data sets. The primary objective of our system is to enable interactive debugging of model predictions on individual instances (questions) and to simplify manual error analysis. Our interactive interface helps researchers to understand the shortcomings of a particular model, qualitatively analyze the complete pipeline and compare different models. A set of sit-by sessions was used to validate our interface design.

Session 6 Overview – Saturday, November 3, 2018

Oral tracks

	Track A *Dialogue I* Gold Hall	**Track B** *Question Answering II* Copper Hall	**Track C** *Semantics V* Silver Hall	**Track D** *Multilingual Methods II* Hall 100
11:00	Using lexical alignment and referring ability to address data sparsity in situated dialog reference resolution *Shore and Skantze*	Large-scale Cloze Test Dataset Created by Teachers *Xie, Lai, Dai, and Hovy*	A Unified Syntax-aware Framework for Semantic Role Labeling *Li, He, Cai, Zhang, Zhao, Liu, Li, and Si*	Sentence Compression for Arbitrary Languages via Multilingual Pivoting *Mallinson, Sennrich, and Lapata*
11:18	Subgoal Discovery for Hierarchical Dialogue Policy Learning *Tang, Li, Gao, Wang, Li, and Jebara*	emrQA: A Large Corpus for Question Answering on Electronic Medical Records *Pampari, Raghavan, Liang, and Peng*	Semantics as a Foreign Language *Stanovsky and Dagan*	Unsupervised Cross-lingual Transfer of Word Embedding Spaces *Xu, Yang, Otani, and Wu*
11:36	Supervised Clustering of Questions into Intents for Dialog System Applications *Haponchyk, Uva, Yu, Uryupina, and Moschitti*	HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering *Yang, Qi, Zhang, Bengio, Cohen, Salakhutdinov, and Manning*	An AMR Aligner Tuned by Transition-based Parser *Liu, Che, Zheng, Qin, and Liu*	XNLI: Cross-lingual Sentence Understanding through Inference *Conneau, Rinott, Lample, Williams, Bowman, Schwenk, and Stoyanov*
11:54	Towards Exploiting Background Knowledge for Building Conversation Systems *Moghe, Arora, Banerjee, and Khapra*	Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering *Mihaylov, Clark, Khot, and Sabharwal*	Dependency-based Hybrid Trees for Semantic Parsing *Jie and Lu*	The Importance of Joint Multilingual Supervision for Cross-lingual Entity Linking *Upadhyay, Gupta, and Roth*
12:12	Decoupling Strategy and Generation in Negotiation Dialogues *He, Chen, Balakrishnan, and Liang*	Evaluating Theory of Mind in Question Answering *Nematzadeh, Burns, Grant, Gopnik, and Griffiths*	Policy Shaping and Generalized Update Equations for Semantic Parsing from Denotations *Misra, Chang, He, and Yih*	Fine-grained Coordinated Cross-lingual Text Stream Alignment for Endless Language Knowledge Acquisition *Ge, Dou, Ji, Cui, Chang, Sui, Wei, and Zhou*

Poster tracks

11:00–12:30

Track D: *Syntax, Morphology, Vision and Language I* Hall 100

Parallel Session 6

Session 6A: Dialogue I

Gold Hall Chair: *chairname*

Using lexical alignment and referring ability to address data sparsity in situated dialog reference resolution

Todd Shore and Gabriel Skantze 11:00–11:18

Referring to entities in situated dialog is a collaborative process, whereby interlocutors often expand, repair and/or replace referring expressions in an iterative process, converging on conceptual pacts of referring language use in doing so. Nevertheless, much work on exophoric reference resolution (i.e. resolution of references to entities outside of a given text) follows a literary model, whereby individual referring expressions are interpreted as unique identifiers of their referents given the state of the dialog the referring expression is initiated. In this paper, we address this collaborative nature to improve dialogic reference resolution in two ways: First, we trained a words-as-classifiers logistic regression model of word semantics and incrementally adapt the model to idiosyncratic language between dyad partners during evaluation of the dialog. We then used these semantic models to learn the general referring ability of each word, which is independent of referent features. These methods facilitate accurate automatic reference resolution in situated dialog without annotation of referring expressions, even with little background data.

Subgoal Discovery for Hierarchical Dialogue Policy Learning

Da Tang, Xiujun Li, Jianfeng Gao, Chong Wang, Lihong Li, and Tony Jebara 11:18–11:36

Developing agents to engage in complex goal-oriented dialogues is challenging partly because the main learning signals are very sparse in long conversations. In this paper, we propose a divide-and-conquer approach that discovers and exploits the hidden structure of the task to enable efficient policy learning. First, given successful example dialogues, we propose the Subgoal Discovery Network (SDN) to divide a complex goal-oriented task into a set of simpler subgoals in an unsupervised fashion. We then use these subgoals to learn a multi-level policy by hierarchical reinforcement learning. We demonstrate our method by building a dialogue agent for the composite task of travel planning. Experiments with simulated and real users show that our approach performs competitively against a state-of-the-art method that requires human-defined subgoals. Moreover, we show that the learned subgoals are often human comprehensible.

Supervised Clustering of Questions into Intents for Dialog System Applications

Iryna Haponchyk, Antonio Uva, Seunghak Yu, Olga Uryupina, and Alessandro Moschitti 11:36–11:54

Modern automated dialog systems require complex dialog managers able to deal with user intent triggered by high-level semantic questions. In this paper, we propose a model for automatically clustering questions into user intents to help the design tasks. Since questions are short texts, uncovering their semantics to group them together can be very challenging. We approach the problem by using powerful semantic classifiers from question duplicate/matching research along with a novel idea of supervised clustering methods based on structured output. We test our approach on two intent clustering corpora, showing an impressive improvement over previous methods for two languages/domains.

Towards Exploiting Background Knowledge for Building Conversation Systems

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M. Khapra 11:54–12:12

Existing dialog datasets contain a sequence of utterances and responses without any explicit background knowledge associated with them. This has resulted in the development of models which treat conversation as a sequence-to-sequence generation task (i.e., given a sequence of utterances generate the response sequence). This is not only an overly simplistic view of conversation but it is also emphatically different from the way humans converse by heavily relying on their background knowledge about the topic (as opposed to simply relying on the previous sequence of utterances). For example, it is common for humans to (involuntarily) produce utterances which are copied or suitably modified from background articles they have read about the topic. To facilitate the development of such natural conversation models which mimic the human process of conversing, we create a new dataset containing movie chats wherein each response is explicitly generated by copying and/or modifying sentences from unstructured background knowledge such as plots, comments and reviews about the movie. We establish baseline results on this dataset (90K utterances from 9K conversations) using three different models: (i) pure generation based models which ignore the background knowledge (ii) generation based models which learn to copy information from the background knowledge when required and (iii) span prediction based models which predict the appropriate response span in the background knowledge.

Decoupling Strategy and Generation in Negotiation Dialogues
He He, Derek Chen, Anusha Balakrishnan, and Percy Liang 12:12–12:30

We consider negotiation settings in which two agents use natural language to bargain on goods. Agents need to decide on both high-level strategy (e.g., proposing $50) and the execution of that strategy (e.g., generat- ing "The bike is brand new. Selling for just $50!"). Recent work on negotiation trains neural models, but their end-to-end nature makes it hard to control their strategy, and reinforcement learning tends to lead to de- generate solutions. In this paper, we pro- pose a modular approach based on coarse di- alogue acts (e.g., propose(price=50)) that de- couples strategy and generation. We show that we can flexibly set the strategy using su- pervised learning, reinforcement learning, or domain-specific knowledge without degener- acy, while our retrieval-based generation can maintain context-awareness and produce di- verse utterances. We test our approach on the recently proposed DEALORNODEAL game, and we also collect a richer dataset based on real items on Craigslist. Human evaluation shows that our systems achieve higher task success rate and more human-like negotiation behavior than previous approaches.

Session 6B: Question Answering II

Copper Hall Chair: *chairname*

Large-scale Cloze Test Dataset Created by Teachers
Qizhe Xie, Guokun Lai, Zihang Dai, and Eduard Hovy 11:00–11:18

Cloze tests are widely adopted in language exams to evaluate students' language proficiency. In this paper, we propose the first large-scale human-created cloze test dataset CLOTH, containing questions used in middle-school and high-school language exams. With missing blanks carefully created by teachers and candidate choices purposely designed to be nuanced, CLOTH requires a deeper language understanding and a wider attention span than previously automatically-generated cloze datasets. We test the performance of dedicatedly designed baseline models including a language model trained on the One Billion Word Corpus and show humans outperform them by a significant margin. We investigate the source of the performance gap, trace model deficiencies to some distinct properties of CLOTH, and identify the limited ability of comprehending the long-term context to be the key bottleneck.

emrQA: A Large Corpus for Question Answering on Electronic Medical Records
Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng 11:18–11:36

We propose a novel methodology to generate domain-specific large-scale question answering (QA) datasets by re-purposing existing annotations for other NLP tasks. We demonstrate an instance of this methodology in generating a large-scale QA dataset for electronic medical records by leveraging existing expert annotations on clinical notes for various NLP tasks from the community shared i2b2 datasets. The resulting corpus (emrQA) has 1 million questions-logical form and 400,000+ question-answer evidence pairs. We characterize the dataset and explore its learning potential by training baseline models for question to logical form and question to answer mapping.

HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering
Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning 11:36–11:54

Existing question answering (QA) datasets fail to train QA systems to perform complex reasoning and provide explanations for answers. We introduce HotpotQA, a new dataset with 113k Wikipedia-based question-answer pairs with four key features: (1) the questions require finding and reasoning over multiple supporting documents to answer; (2) the questions are diverse and not constrained to any pre-existing knowledge bases or knowledge schemas; (3) we provide sentence-level supporting facts required for reasoning, allowing QA systems to reason with strong supervision and explain the predictions; (4) we offer a new type of factoid comparison questions to test QA systems' ability to extract relevant facts and perform necessary comparison. We show that HotpotQA is challenging for the latest QA systems, and the supporting facts enable models to improve performance and make explainable predictions.

Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering
Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal 11:54–12:12

We present a new kind of question answering dataset, OpenBookQA, modeled after open book exams for assessing human understanding of a subject. The open book that comes with our questions is a set of 1326 elementary level science facts. Roughly 6000 questions probe an understanding of these facts and their application to novel situations. This requires combining an open book fact (e.g., metals conduct electricity) with broad common knowledge (e.g., a suit of armor is made of metal) obtained from other sources. While existing QA datasets over documents or knowledge bases, being generally self-contained, focus on linguistic understanding, OpenBookQA probes a deeper understanding of both the topic—in the context of common knowledge—and the language it is expressed in. Human performance on OpenBookQA is close to 92%, but many state-of-the-art pre-trained QA methods perform surprisingly poorly, worse than several simple neural baselines we develop. Our oracle experiments designed to circumvent the knowledge retrieval bottleneck demonstrate the value of both the open book and additional facts. We leave it as a challenge to solve the retrieval problem in this multi-hop setting and to close the large gap to human performance.

Evaluating Theory of Mind in Question Answering
Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Tom Griffiths 12:12–12:30

We propose a new dataset for evaluating question answering models with respect to their capacity to reason about beliefs. Our tasks are inspired by theory-of-mind experiments that examine whether children are able to reason about the beliefs of others, in particular when those beliefs differ from reality. We evaluate a number of recent neural models with memory augmentation. We find that all fail on our tasks, which require keeping track of inconsistent states of the world; moreover, the models' accuracy decreases notably when random sentences are introduced to the tasks at test.

Session 6C: Semantics V

Silver Hall Chair: *chairname*

A Unified Syntax-aware Framework for Semantic Role Labeling
Zuchao Li, Shexia He, Jiaxun Cai, Zhuosheng Zhang, Hai Zhao, Gongshen Liu, Linlin Li, and Luo Si
11:00–11:18

Semantic role labeling (SRL) aims to recognize the predicate-argument structure of a sentence. Syntactic information has been paid a great attention over the role of enhancing SRL. However, the latest advance shows that syntax would not be so important for SRL with the emerging much smaller gap between syntax-aware and syntax-agnostic SRL. To comprehensively explore the role of syntax for SRL task, we extend existing models and propose a unified framework to investigate more effective and more diverse ways of incorporating syntax into sequential neural networks. Exploring the effect of syntactic input quality on SRL performance, we confirm that high-quality syntactic parse could still effectively enhance syntactically-driven SRL. Using empirically optimized integration strategy, we even enlarge the gap between syntax-aware and syntax-agnostic SRL. Our framework achieves state-of-the-art results on CoNLL-2009 benchmarks both for English and Chinese, substantially outperforming all previous models.

Semantics as a Foreign Language
Gabriel Stanovsky and Ido Dagan 11:18–11:36

We propose a novel approach to semantic dependency parsing (SDP) by casting the task as an instance of multi-lingual machine translation, where each semantic representation is a different foreign dialect. To that end, we first generalize syntactic linearization techniques to account for the richer semantic dependency graph structure. Following, we design a neural sequence-to-sequence framework which can effectively recover our graph linearizations, performing almost on-par with previous SDP state-of-the-art while requiring less parallel training annotations. Beyond SDP, our linearization technique opens the door to integration of graph-based semantic representations as features in neural models for downstream applications.

An AMR Aligner Tuned by Transition-based Parser
Yijia Liu, Wanxiang Che, Bo Zheng, Bing Qin, and Ting Liu 11:36–11:54

In this paper, we propose a new rich resource enhanced AMR aligner which produces multiple alignments and a new transition system for AMR parsing along with its oracle parser. Our aligner is further tuned by our oracle parser via picking the alignment that leads to the highest-scored achievable AMR graph. Experimental results show that our aligner outperforms the rule-based aligner in previous work by achieving higher alignment F1 score and consistently improving two open-sourced AMR parsers. Based on our aligner and transition system, we develop a transition-based AMR parser that parses a sentence into its AMR graph directly. An ensemble of our parsers with only words and POS tags as input leads to 68.4 Smatch F1 score, which outperforms the current state-of-the-art parser.

Dependency-based Hybrid Trees for Semantic Parsing
Zhanming Jie and Wei Lu 11:54–12:12

We propose a novel dependency-based hybrid tree model for semantic parsing, which converts natural language utterance into machine interpretable meaning representations. Unlike previous state-of-the-art models, the semantic information is interpreted as the latent dependency between the natural language words in our joint representation. Such dependency information can capture the interactions between the semantics and natural language words. We integrate a neural component into our model and propose an efficient dynamic-programming algorithm to perform tractable inference. Through extensive experiments on the standard multilingual GeoQuery dataset with eight languages, we demonstrate that our proposed approach is able to achieve state-of- the-art performance across several languages. Analysis also justifies the effectiveness of using our new dependency-based representation.

Policy Shaping and Generalized Update Equations for Semantic Parsing from Denotations
Dipendra Misra, Ming-Wei Chang, Xiaodong He, and Wen-tau Yih 12:12–12:30

Semantic parsing from denotations faces two key challenges in model training: (1) given only the denotations (e.g., answers), search for good candidate semantic parses, and (2) choose the best model update algorithm. We propose effective and general solutions to each of them. Using policy shaping, we bias the search procedure towards semantic parses that are more compatible to the text, which provide better supervision signals for training. In addition, we propose an update equation that generalizes three different families of learning algorithms, which enables fast model exploration. When experimented on a recently proposed sequential question answering dataset, our framework leads to a new state-of-the-art model that outperforms previous work by 5.0% absolute on exact match accuracy.

Session 6D: Multilingual Methods II

Hall 100 Chair: *chairname*

Sentence Compression for Arbitrary Languages via Multilingual Pivoting
Jonathan Mallinson, Rico Sennrich, and Mirella Lapata 11:00–11:18

In this paper we advocate the use of bilingual corpora which are abundantly available for training sentence compression models. Our approach borrows much of its machinery from neural machine translation and leverages bilingual pivoting: compressions are obtained by translating a source string into a foreign language and then back-translating it into the source while controlling the translation length. Our model can be trained for any language as long as a bilingual corpus is available and performs arbitrary rewrites without access to compression specific data. We release. Moss, a new parallel Multilingual Compression dataset for English, German, and French which can be used to evaluate compression models across languages and genres.

Unsupervised Cross-lingual Transfer of Word Embedding Spaces
Ruochen Xu, Yiming Yang, Naoki Otani, and Yuexin Wu 11:18–11:36

Cross-lingual transfer of word embeddings aims to establish the semantic mappings among words in different languages by learning the transformation functions over the corresponding word embedding spaces. Successfully solving this problem would benefit many downstream tasks such as to translate text classification models from resource-rich languages (e.g. English) to low-resource languages. Supervised methods for this problem rely on the availability of cross-lingual supervision, either using parallel corpora or bilingual lexicons as the labeled data for training, which may not be available for many low resource languages. This paper proposes an unsupervised learning approach that does not require any cross-lingual labeled data. Given two monolingual word embedding spaces for any language pair, our algorithm optimizes the transformation functions in both directions simultaneously based on distributional matching as well as minimizing the back-translation losses. We use a neural network implementation to calculate the Sinkhorn distance, a well-defined distributional similarity measure, and optimize our objective through back-propagation. Our evaluation on benchmark datasets for bilingual lexicon induction and cross-lingual word similarity prediction shows stronger or competitive performance of the proposed method compared to other state-of-the-art supervised and unsupervised baseline methods over many language pairs.

XNLI: Cross-lingual Sentence Understanding through Inference
Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov 11:36–11:54

State-of-the-art natural language processing systems rely on supervision in the form of annotated data to learn competent models. These models are generally trained on data in a single language (usually English), and cannot be directly used beyond that language. Since collecting data in every language is not realistic, there has been a growing interest in cross-lingual language understanding (XLU) and low-resource cross-language transfer. In this work, we construct an evaluation set for XLU by extending the development and test sets of the Multi-Genre Natural Language Inference Corpus (MultiNLI) to 14 languages, including low-resource languages such as Swahili and Urdu. We hope that our dataset, dubbed XNLI, will catalyze research in cross-lingual sentence understanding by providing an informative standard evaluation task. In addition, we provide several baselines for multilingual sentence understanding, including two based on machine translation systems, and two that use parallel data to train aligned multilingual bag-of-words and LSTM encoders. We find that XNLI represents a practical and challenging evaluation suite, and that directly translating the test data yields the best performance among available baselines.

The Importance of Joint Multilingual Supervision for Cross-lingual Entity Linking
Shyam Upadhyay, Nitish Gupta, and Dan Roth 11:54–12:12

Cross-lingual Entity Linking (XEL) aims to ground entity mentions written in any language to an English Knowledge Base (KB), such as Wikipedia. XEL for most languages is challenging, owing to limited availability of resources as supervision. We address this challenge by developing the first XEL approach that combines supervision from multiple languages jointly. This enables our approach to: (a) augment the limited supervision in the target language with additional supervision from a high-resource language (like English), and (b) train a single entity linking model for multiple languages, improving upon individually trained models for each language. Extensive evaluation on three benchmark datasets across 8 languages shows that our approach significantly improves over the current state-of-the-art. We also provide analyses in two limited resource settings: (a) zero-shot setting, when no supervision in the target language is available, and in (b) low-resource setting, when some supervision in the target language is available. Our analysis provides insights into the limitations of zero-shot XEL approaches in realistic scenarios, and shows the value of joint supervision in low-resource settings.

Fine-grained Coordinated Cross-lingual Text Stream Alignment for Endless Language Knowledge Acquisition

Tao Ge, Qing Dou, Heng Ji, Lei Cui, Baobao Chang, Zhifang Sui, Furu Wei, and Ming Zhou 12:12–12:30

This paper proposes to study fine-grained coordinated cross-lingual text stream alignment through a novel information network decipherment paradigm. We use Burst Information Networks as media to represent text streams and present a simple yet effective network decipherment algorithm with diverse clues to decipher the networks for accurate text stream alignment. Experiments on Chinese-English news streams show our approach not only outperforms previous approaches on bilingual lexicon extraction from coordinated text streams but also can harvest high-quality alignments from large amounts of streaming data for endless language knowledge mining, which makes it promising to be a new paradigm for automatic language knowledge acquisition.

Session 6E: Syntax, Morphology, Vision and Language I

11:00–12:30
Hall 100 Chair: *chairname*

WECA: WordNet-Encoded Collocation-Attention Network for Homographic Pun Recognition

Yufeng Diao, Hongfei Lin, Di Wu, Liang Yang, Kan Xu, Zhihao Yang, Jian Wang, Shaowu Zhang, Bo Xu, and Dongyu Zhang

Homographic puns have a long history in human writing, widely used in written and spoken literature, which usually occur in a certain syntactic or stylistic structure. How to recognize homographic puns is an important research. However, homographic pun recognition does not solve very well in existing work. In this work, we first use WordNet to understand and expand word embedding for settling the polysemy of homographic puns, and then propose a WordNet-Encoded Collocation-Attention network model (WECA) which combined with the context weights for recognizing the puns. Our experiments on the SemEval2017 Task7 and Pun of the Day demonstrate that the proposed model is able to distinguish between homographic pun and non-homographic pun texts. We show the effectiveness of the model to present the capability of choosing qualitatively informative words. The results show that our model achieves the state-of-the-art performance on homographic puns recognition.

A Hybrid Approach to Automatic Corpus Generation for Chinese Spelling Check

Dingmin Wang, Yan Song, Jing Li, Jialong Han, and Haisong Zhang

Chinese spelling check (CSC) is a challenging yet meaningful task, which not only serves as a preprocessing in many natural language processing(NLP) applications, but also facilitates reading and understanding of running texts in peoples' daily lives. However, to utilize data-driven approaches for CSC, there is one major limitation that annotated corpora are not enough in applying algorithms and building models. In this paper, we propose a novel approach of constructing CSC corpus with automatically generated spelling errors, which are either visually or phonologically resembled characters, corresponding to the OCR- and ASR-based methods, respectively. Upon the constructed corpus, different models are trained and evaluated for CSC with respect to three standard test sets. Experimental results demonstrate the effectiveness of the corpus, therefore confirm the validity of our approach.

Neural Quality Estimation of Grammatical Error Correction

Shamil Chollampatt and Hwee Tou Ng

Grammatical error correction (GEC) systems deployed in language learning environments are expected to accurately correct errors in learners' writing. However, in practice, they often produce spurious corrections and fail to correct many errors, thereby misleading learners. This necessitates the estimation of the quality of output sentences produced by GEC systems so that instructors can selectively intervene and re-correct the sentences which are poorly corrected by the system and ensure that learners get accurate feedback. We propose the first neural approach to automatic quality estimation of GEC output sentences that does not employ any hand-crafted features. Our system is trained in a supervised manner on learner sentences and corresponding GEC system outputs with quality score labels computed using human-annotated references. Our neural quality estimation models for GEC show significant improvements over a strong feature-based baseline. We also show that a state-of-the-art GEC system can be improved when quality scores are used as features for re-ranking the N-best candidates.

Part-of-Speech Tagging for Twitter by Different Expression Styles

Tao Gui, Qi Zhang, Jingjing Gong, Minlong Peng, di liang di, Keyu Ding, and Xuanjing Huang

Part-of-Speech (POS) tagging for Twitter has received considerable attention in recent years. Because most POS tagging methods are based on supervised models, they usually require a large amount of labeled data for training. However, the existing labeled datasets for Twitter are much smaller than those for newswire text. Hence, to help POS tagging for Twitter, most domain adaptation methods try to leverage newswire datasets by learning the shared features between the two domains. However, from a linguistic perspective, Twitter users not only tend to mimic the formal expressions of traditional media, like news, but they also appear to be developing linguistically informal styles. Therefore, POS tagging for the formal Twitter context can be learned together with the newswire dataset, while POS tagging for the informal Twitter context should be learned separately. To achieve this task, in this work, we propose a hypernetwork-based method to generate different parameters to separately model contexts with different expression styles. Experimental results on three different datasets show that our approach achieves better performance than state-of-the-art methods in most cases.

Free as in Free Word Order: An Energy Based Model for Word Segmentation and Morphological Tagging in Sanskrit
Amrith Krishna, Bishal Santra, Sasi Prasanth Bandaru, Gaurav Sahu, Vishnu Dutt Sharma, Pavankumar Satuluri, and Pawan Goyal

The configurational information in sentences of a free word order language such as Sanskrit is of limited use. Thus, the context of the entire sentence will be desirable even for basic processing tasks such as word segmentation. We propose a structured prediction framework that jointly solves the word segmentation and morphological tagging tasks in Sanskrit. We build an energy based model where we adopt approaches generally employed in graph based parsing techniques (McDonald et al., 2005a; Carreras, 2007). Our model outperforms the state of the art with an F-Score of 96.92 (percentage improvement of 7.06%) while using less than one tenth of the task-specific training data. We find that the use of a graph based approach instead of a traditional lattice-based sequential labelling approach leads to a percentage gain of 12.6% in F-Score for the segmentation task.

A Challenge Set and Methods for Noun-Verb Ambiguity
Ali Elkahky, Kellie Webster, Daniel Andor, and Emily Pitler

English part-of-speech taggers regularly make egregious errors related to noun-verb ambiguity, despite having achieved 97%+ accuracy on the WSJ Penn Treebank since 2002. These mistakes have been difficult to quantify and make taggers less useful to downstream tasks such as translation and text-to-speech synthesis. This paper creates a new dataset of over 30,000 naturally-occurring non-trivial examples of noun-verb ambiguity. Taggers within 1% of each other when measured on the WSJ have accuracies ranging from 57% to 75% accuracy on this challenge set. Enhancing the strongest existing tagger with contextual word embeddings and targeted training data improves its accuracy to 89%, a 14% absolute (52% relative) improvement. Downstream, using just this enhanced tagger yields a 28% reduction in error over the prior best learned model for homograph disambiguation for textto-speech synthesis.

What do character-level models learn about morphology? The case of dependency parsing
Clara Vania, Andreas Grivas, and Adam Lopez

When parsing morphologically-rich languages with neural models, it is beneficial to model input at the character level, and it has been claimed that this is because character-level models learn morphology. We test these claims by comparing character-level models to an oracle with access to explicit morphological analysis on twelve languages with varying morphological typologies. Our results highlight many strengths of character-level models, but also show that they are poor at disambiguating some words, particularly in the face of case syncretism. We then demonstrate that explicitly modeling morphological case improves our best model, showing that character-level models can benefit from targeted forms of explicit morphological modeling.

Learning Better Internal Structure of Words for Sequence Labeling
Yingwei Xin, Ethan Hart, Vibhuti Mahajan, and Jean David Ruvini

Character-based neural models have recently proven very useful for many NLP tasks. However, there is a gap of sophistication between methods for learning representations of sentences and words. While, most character models for learning representations of sentences are deep and complex, models for learning representations of words are shallow and simple. Also, in spite of considerable research on learning character embeddings, it is still not clear which kind of architecture is the best for capturing character-to-word representations. To address these questions, we first investigate the gaps between methods for learning word and sentence representations. We conduct detailed experiments and comparisons on different state-of-the-art convolutional models, and also investigate the advantages and disadvantages of their constituents. Furthermore, we propose IntNet, a funnel-shaped wide convolutional neural architecture with no down-sampling for learning representations of the internal structure of words by composing their characters from limited, supervised training corpora. We evaluate our proposed model on six sequence labeling datasets, including named entity recognition, part-of-speech tagging, and syntactic chunking. Our in-depth analysis shows that IntNet significantly outperforms other character embedding models and obtains new state-of-the-art performance without relying on any external knowledge or resources.

ICON: Interactive Conversational Memory Network for Multimodal Emotion Detection
Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann

Emotion recognition in conversations is crucial for building empathetic machines. Present works in this domain do not explicitly consider the inter-personal influences that thrive in the emotional dynamics of dialogues. To this end, we propose Interactive COnversational memory Network (ICON), a multimodal emotion detection framework that extracts multimodal features from conversational videos and hierarchically models the self- and inter-speaker emotional influences into global memories. Such memories generate contextual summaries which aid in predicting the emotional orientation of utterance-videos. Our model outperforms state-of-the-art networks on multiple classification and regression tasks in two benchmark datasets.

Discriminative Learning of Open-Vocabulary Object Retrieval and Localization by Negative Phrase Augmentation
Ryota Hinami and Shin'ichi Satoh

Thanks to the success of object detection technology, we can retrieve objects of the specified classes even from huge image collections. However, the current state-of-the-art object detectors (such as Faster R-CNN) can only handle pre-specified classes. In addition, large amounts of positive and negative visual samples are required for training. In this paper, we address the problem of open-vocabulary object retrieval and localization, where the target object is specified by a textual query (e.g., a word or phrase). We first propose Query-Adaptive R-CNN, a simple extension of Faster R-CNN adapted to open-vocabulary queries, by transforming the text embedding vector into an object classifier and localization regressor. Then, for discriminative training, we then propose negative phrase augmentation (NPA) to mine hard negative samples which are visually similar to the query and at the same time semantically mutually exclusive of the query. The proposed method can retrieve and localize objects specified by a textual query from one million images in only 0.5 seconds with high precision.

Grounding Semantic Roles in Images
Carina Silberer and Manfred Pinkal

We address the task of visual semantic role labeling (vSRL), the identification of the participants of a situation or event in a visual scene, and their labeling with their semantic relations to the event or situation. We render candidate participants as image regions of objects, and train a model which learns to ground roles in the regions which depict the corresponding participant. Experimental results demonstrate that we can train a vSRL model without reliance on prohibitive image-based role annotations, by utilizing noisy data which we extract automatically from image captions using a linguistic SRL system. Furthermore, our model induces frame—semantic visual representations, and their comparison to previous work on supervised visual verb sense disambiguation yields overall better results.

Commonsense Action Explanation in Human-Agent Communication
Shaohua Yang, Qiaozi Gao, sari sadiya sari, and Joyce Chai

To enable collaboration and communication between humans and agents, this paper investigates learning to acquire commonsense evidence for action justification. In particular, we have developed an approach based on the generative Conditional Variational Autoencoder(CVAE) that models object relations/attributes of the world as latent variables and jointly learns a performer that predicts actions and an explainer that gathers commonsense evidence to justify the action. Our empirical results have shown that, compared to a typical attention-based model, CVAE achieves significantly higher performance in both action prediction and justification. A human subject study further shows that the commonsense evidence gathered by CVAE can be communicated to humans to achieve a significantly higher common ground between humans and agents.

Learning Personas from Dialogue with Attentive Memory Networks
Eric Chu, Prashanth Vijayaraghavan, and Deb Roy

The ability to infer persona from dialogue can have applications in areas ranging from computational narrative analysis to personalized dialogue generation. We introduce neural models to learn persona embeddings in a supervised character trope classification task. The models encode dialogue snippets from IMDB into representations that can capture the various categories of film characters. The best-performing models use a multi-level attention mechanism over a set of utterances. We also utilize prior knowledge in the form of textual descriptions of the different tropes. We apply the learned embeddings to find similar characters across different movies, and cluster movies according to the distribution of the embeddings. The use of short conversational text as input, and the ability to learn from prior knowledge using memory, suggests these methods could be applied to other domains.

Grounding language acquisition by training semantic parsers using captioned videos
Candace Ross, Andrei Barbu, Yevgeni Berzak, Battushig Myanganbayar, and Boris Katz

We develop a semantic parser that is trained in a grounded setting using pairs of videos captioned with sentences. This setting is both data-efficient, requiring little annotation, and similar to the experience of children where they observe their environment and listen to speakers. The semantic parser recovers the meaning of English sentences despite not having access to any annotated sentences. It does so despite the ambiguity inherent in vision where a sentence may refer to any combination of objects, object properties, relations or actions taken by any agent in a video. For this task, we collected a new dataset for grounded language acquisition. Learning a grounded semantic parser — turning sentences into logical forms using captioned videos — can significantly expand the range of data that parsers can be trained on, lower the effort of training a semantic parser, and ultimately lead to a better understanding of child language acquisition.

Translating Navigation Instructions in Natural Language to a High-Level Plan for Behavioral Robot Navigation
Xiaoxue Zang, Ashwini Pokle, Marynel Vázquez, Kevin Chen, Juan Carlos Niebles, Alvaro Soto, and Silvio Savarese
We propose an end-to-end deep learning model for translating free-form natural language instructions to a high-level plan for behavioral robot navigation. We use attention models to connect information from both the user instructions and a topological representation of the environment. We evaluate our model's performance on a new dataset containing 10,050 pairs of navigation instructions. Our model significantly outperforms baseline approaches. Furthermore, our results suggest that it is possible to leverage the environment map as a relevant knowledge base to facilitate the translation of free-form navigational instruction.

Mapping Instructions to Actions in 3D Environments with Visual Goal Prediction
Dipendra Misra, Andrew Bennett, Valts Blukis, Eyvind Niklasson, Max Shatkhin, and Yoav Artzi
We propose to decompose instruction execution to goal prediction and action generation. We design a model that maps raw visual observations to goals using LINGUNET, a language-conditioned image generation network, and then generates the actions required to complete them. Our model is trained from demonstration only without external resources. To evaluate our approach, we introduce two benchmarks for instruction following: LANI, a navigation task; and CHAI, where an agent executes household instructions. Our evaluation demonstrates the advantages of our model decomposition, and illustrates the challenges posed by our new benchmarks.

Deconvolutional time series regression: A technique for modeling temporally diffuse effects
Cory Shain and William Schuler
Researchers in computational psycholinguistics frequently use linear models to study time series data generated by human subjects. However, time series may violate the assumptions of these models through temporal diffusion, where stimulus presentation has a lingering influence on the response as the rest of the experiment unfolds. This paper proposes a new statistical model that borrows from digital signal processing by recasting the predictors and response as convolutionally-related signals, using recent advances in machine learning to fit latent impulse response functions (IRFs) of arbitrary shape. A synthetic experiment shows successful recovery of true latent IRFs, and psycholinguistic experiments reveal plausible, replicable, and fine-grained estimates of latent temporal dynamics, with comparable or improved prediction quality to widely-used alternatives.

Is this Sentence Difficult? Do you Agree?
Dominique Brunato, Lorenzo De Mattei, Felice Dell'Orletta, Benedetta Iavarone, and Giulia Venturi
In this paper, we present a crowdsourcing-based approach to model the human perception of sentence complexity. We collect a large corpus of sentences rated with judgments of complexity for two typologically-different languages, Italian and English. We test our approach in two experimental scenarios aimed to investigate the contribution of a wide set of lexical, morpho-syntactic and syntactic phenomena in predicting i) the degree of agreement among annotators independently from the assigned judgment and ii) the perception of sentence complexity.

Neural Transition Based Parsing of Web Queries: An Entity Based Approach
Rivka Malca and Roi Reichart
Web queries with question intent manifest a complex syntactic structure and the processing of this structure is important for their interpretation. Pinter et al. (2016) has formalized the grammar of these queries

and proposed semi-supervised algorithms for the adaptation of parsers originally designed to parse according to the standard dependency grammar, so that they can account for the unique forest grammar of queries. However, their algorithms rely on resources typically not available outside of big web corporates. We propose a new BiLSTM query parser that: (1) Explicitly accounts for the unique grammar of web queries; and (2) Utilizes named entity (NE) information from a BiLSTM NE tagger, that can be jointly trained with the parser. In order to train our model we annotate the query treebank of Pinter et al. (2016) with NEs. When trained on 2500 annotated queries our parser achieves UAS of 83.5% and segmentation F1- score of 84.5, substantially outperforming existing state-of-the-art parsers.

An Investigation of the Interactions Between Pre-Trained Word Embeddings, Character Models and POS Tags in Dependency Parsing
Aaron Smith, Miryam de Lhoneux, Sara Stymne, and Joakim Nivre

We provide a comprehensive analysis of the interactions between pre-trained word embeddings, character models and POS tags in a transition-based dependency parser. While previous studies have shown POS information to be less important in the presence of character models, we show that in fact there are complex interactions between all three techniques. In isolation each produces large improvements over a baseline system using randomly initialised word embeddings only, but combining them quickly leads to diminishing returns. We categorise words by frequency, POS tag and language in order to systematically investigate how each of the techniques affects parsing quality. For many word categories, applying any two of the three techniques is almost as good as the full combined system. Character models tend to be more important for low-frequency open-class words, especially in morphologically rich languages, while POS tags can help disambiguate high-frequency function words. We also show that large character embedding sizes help even for languages with small character sets, especially in morphologically rich languages.

Depth-bounding is effective: Improvements and Evaluation of Unsupervised PCFG Induction
Lifeng Jin, Finale Doshi-Velez, Timothy Miller, William Schuler, and Lane Schwartz

There have been several recent attempts to improve the accuracy of grammar induction systems by bounding the recursive complexity of the induction model. Modern depth-bounded grammar inducers have been shown to be more accurate than early unbounded PCFG inducers, but this technique has never been compared against unbounded induction within the same system, in part because most previous depth-bounding models are built around sequence models, the complexity of which grows exponentially with the maximum allowed depth. The present work instead applies depth bounds within a chart-based Bayesian PCFG inducer, where bounding can be switched on and off, and then samples trees with or without bounding. Results show that depth-bounding is indeed significantly effective in limiting the search space of the inducer and thereby increasing accuracy of resulting parsing model, independent of the contribution of modern Bayesian induction techniques. Moreover, parsing results on English, Chinese and German show that this bounded model is able to produce parse trees more accurately than or competitively with state-of-the-art constituency grammar induction models.

TACL-1199
TACL-1199

Both bottom-up and top-down strategies have been used for neural transition-based constituent parsing. The parsing strategies differ in terms of the order in which they recognize productions in the derivation tree, where bottom-up strategies and top-down strategies take post-order and pre-order traversal over trees, respectively. Bottom-up parsers benefit from rich features from readily built partial parses, but lack lookahead guidance in the parsing process; top-down parsers benefit from non-local guidance for local decisions, but rely on a strong encoder over the input to predict a constituent hierarchy before its construction. To mitigate both issues, we propose a novel parsing system based on in-order traversal over syntactic trees, designing a set of transition actions to find a compromise between bottom-up constituent information and top-down lookahead information. Based on stack-LSTM, our psycholinguistically motivated constituent parsing system achieves 91.8 F1 on WSJ benchmark. Furthermore, the system achieves 93.6 F1 with supervised reranking and 94.2 F1 with semi-supervised reranking, which are the best results on the WSJ benchmark.

TACL-1425
TACL-1425

We introduce a novel framework for delexicalized dependency parsing in a new language. We show that useful features of the language can be extracted automatically from an un- parsed corpus, which consists only of gold part-of-speech (POS) sequences. Providing these features to our neural parser enables it to parse sequences of this kind. Strikingly, our system has no supervision in the target language. Rather, it is a multilingual system that is trained end-to-end on a variety of other languages, so it learns a feature extractor that works well. We show experimentally across multiple languages: (1) Features computed from the unparsed corpus improve parsing accuracy. (2) Including thousands of synthetic languages in the training achieves further improvement. (3) Despite being computed from unparsed corpora, our learned task-specific features beat previous work΄92s interpretable typological features that require parsed corpora or expert categorization. Our best method improved attachment scores on held-out test languages by an average of 5.6 percentage points over past work that does not inspect the unparsed data (McDonald et al., 2011), and by 18.8 points over past ΄93grammar induction΄94 work that does not use training languages (Naseem et al., 2010).

Incremental Computation of Infix Probabilities for Probabilistic Finite Automata
Marco Cognetta, Yo-Sub Han, and Soon Chan Kwon

In natural language processing, a common task is to compute the probability of a phrase appearing in a document or to calculate the probability of all phrases matching a given pattern. For instance, one computes affix (prefix, suffix, infix, etc.) probabilities of a string or a set of strings with respect to a probability distribution of patterns. The problem of computing infix probabilities of strings when the pattern distribution is given by a probabilistic context-free grammar or by a probabilistic finite automaton is already solved, yet it was open to compute the infix probabilities in an incremental manner. The incremental computation is crucial when a new query is built from a previous query. We tackle this problem and suggest a method that computes infix probabilities incrementally for probabilistic finite automata by representing all the probabilities of matching strings as a series of transition matrix calculations. We show that the proposed approach is theoretically faster than the previous method and, using real world data, demonstrate that our approach has vastly better performance in practice.

Syntax Encoding with Application in Authorship Attribution
Richong Zhang, Zhiyuan Hu, Hongyu GUO, and Yongyi Mao

We propose a novel strategy to encode the syntax parse tree of sentence into a learnable distributed representation. The proposed syntax encoding scheme is provably information-lossless. In specific, an embedding vector is constructed for each word in the sentence, encoding the path in the syntax tree corresponding to the word. The one-to-one correspondence between these "syntax-embedding" vectors and the words (hence their embedding vectors) in the sentence makes it easy to integrate such a representation with all word-level NLP models. We empirically show the benefits of the syntax embeddings on the Authorship Attribution domain, where our approach improves upon the prior art and achieves new performance records on five benchmarking data sets.

Sanskrit Word Segmentation Using Character-level Recurrent and Convolutional Neural Networks
Oliver Hellwig and Sebastian Nehrdich

The paper introduces end-to-end neural network models that tokenize Sanskrit by jointly splitting compounds and resolving phonetic merges (Sandhi). Tokenization of Sanskrit depends on local phonetic and distant semantic features that are incorporated using convolutional and recurrent elements. Contrary to most previous systems, our models do not require feature engineering or extern linguistic resources, but operate solely on parallel versions of raw and segmented text. The models discussed in this paper clearly improve over previous approaches to Sanskrit word segmentation. As they are language agnostic, we will demonstrate that they also outperform the state of the art for the related task of German compound splitting.

TACL-1446
TACL-1446

Word segmentation is a low-level NLP task that is non-trivial for a considerable number of languages. In this paper, we present a sequence tagging framework and apply it to word segmentation for a wide range of languages with different writing systems and typological characteristics. Additionally, we investigate the correlations between various typological factors and word segmentation accuracy. The experimental results indicate that segmentation accuracy is positively related to word boundary markers and negatively to the number of unique non-segmental terms. Based on the analysis, we design a small set of language-specific settings and extensively evaluate the segmentation system on the Universal Dependencies datasets. Our model obtains state-of-the-art accuracies on all the UD languages. á0It performs substantially better on languages that are non-trivial to segment, such as Chinese, Japanese, Arabic and Hebrew, when compared to previous work.

MorAz: an Open-source Morphological Analyzer for Azerbaijani Turkish
Berke Özenç, Razieh Ehsani, and Ercan Solak

MorAz is an open-source morphological analyzer for Azerbaijani Turkish. The analyzer is available through both as a website for interactive exploration and as a RESTful web service for integration into a natural language processing pipeline. MorAz implements the morphology of Azerbaijani Turkish in two-level using Helsinki finite-state transducer and wraps the analyzer with python scripts in a Django instance.

Juman++: A Morphological Analysis Toolkit for Scriptio Continua
Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi

We present a three-part toolkit for developing morphological analyzers for languages without natural word boundaries. The first part is a C++11/14 lattice-based morphological analysis library that uses a combination of linear and recurrent neural net language models for analysis. The other parts are a tool for exposing problems in the trained model and a partial annotation tool. Our morphological analyzer of Japanese achieves new SOTA on Jumandic-based corpora while being 250 times faster than the previous one. We also perform a small experiment and quantitive analysis and experience of using development tools. All components of the toolkit is open source and available under a permissive Apache 2 License. Annotation tool screencast: https://youtu.be/2tzW2n–axM

Session 7 Overview – Saturday, November 3, 2018

Oral tracks

Track A *Dialogue II* Gold Hall	**Track B** *Social Applications II* Copper Hall	**Track C** *NER* Silver Hall	**Track D** *Morphology / Parsing* Hall 100	
Session-level Language Modeling for Conversational Speech *Xiong, Wu, Zhang, and Stolcke*	The glass ceiling in NLP *Schluter*	Marginal Likelihood Training of BiLSTM-CRF for Biomedical Named Entity Recognition from Disjoint Label Sets *Greenberg, Bansal, Verga, and McCallum*	Improved Dependency Parsing using Implicit Word Connections Learned from Unlabeled Data *Wang, Chang, and Mansur*	13:45
Toward Less Generic Responses in Neural Conversation Models: A Statistical Reweighting Method *Liu, Bi, Gao, Liu, Yao, and Shi*	Reducing Gender Bias in Abusive Language Detection *Park, Shin, and Fung*	Adversarial training for multi-context joint entity and relation extraction *Bekoulis, Deleu, Demeester, and Develder*	A Framework for Understanding the Role of Morphology in Universal Dependency Parsing *Dehouck and Denis*	13:57
Training Millions of Personalized Dialogue Agents *Mazare, Humeau, Raison, and Bordes*	SafeCity: Understanding Diverse Forms of Sexual Harassment Personal Stories *Karlekar and Bansal*	Structured Multi-Label Biomedical Text Tagging via Attentive Neural Tree Decoding *Singh, Thomas, Marshall, Shawe-Taylor, and Wallace*	The Lazy Encoder: A Fine-Grained Analysis of the Role of Morphology in Neural Machine Translation *Bisazza and Tump*	14:09
Towards Universal Dialogue State Tracking *Ren, Xie, Chen, and Yu*	Learning multiview embeddings for assessing dementia *Pou-Prom and Rudzicz*	Deep Exhaustive Model for Nested Named Entity Recognition *Sohrab and Miwa*	Imitation Learning for Neural Morphological String Transduction *Makarov and Clematide*	14:21
Semantic Parsing for Task Oriented Dialog using Hierarchical Representations *Gupta, Shah, Mohit, Kumar, and Lewis*	WikiConv: A Corpus of the Complete Conversational History of a Large Online Collaborative Community *Hua, Danescu-Niculescu-Mizil, Taraborelli, Thain, Sorensen, and Dixon*	Natural Language Processing Not-At-All from Scratch: Evaluating The Utility of Hand-crafted Features in Deep Learning *Wu, Liu, and Cohn*	An Encoder-Decoder Approach to the Paradigm Cell Filling Problem *Silfverberg and Hulden*	14:33

Poster tracks

13:45–14:45

Track D: *Short Posters III* Hall 100

Parallel Session 7

Session 7A: Dialogue II

Gold Hall Chair: *chairname*

Session-level Language Modeling for Conversational Speech
Wayne Xiong, Lingfeng Wu, Jun Zhang, and Andreas Stolcke 13:45–13:57

We propose to generalize language models for conversational speech recognition to allow them to operate across utterance boundaries and speaker changes, thereby capturing conversation-level phenomena such as adjacency pairs, lexical entrainment, and topical coherence. The model consists of a long-short-term memory (LSTM) recurrent network that reads the entire word-level history of a conversation, as well as information about turn taking and speaker overlap, in order to predict each next word. The model is applied in a rescoring framework, where the word history prior to the current utterance is approximated with preliminary recognition results. In experiments in the conversational telephone speech domain (Switchboard) we find that such a model gives substantial perplexity reductions over a standard LSTM-LM with utterance scope, as well as improvements in word error rate.

Toward Less Generic Responses in Neural Conversation Models: A Statistical Re-weighting Method
Yahui Liu, Wei Bi, Jun Gao, Xiaojiang Liu, Jian Yao, and Shuming Shi 13:57–14:09

Sequence-to-sequence neural generation models have achieved promising performance on short text conversation tasks. However, they tend to generate generic/dull responses, leading to unsatisfying dialogue experience. We observe that in the conversation tasks, each query could have multiple responses, which forms a 1-to-n or m-to-n relationship in the view of the total corpus. The objective function used in standard sequence-to-sequence models will be dominated by loss terms with generic patterns. Inspired by this observation, we introduce a statistical re-weighting method that assigns different weights for the multiple responses of the same query, and trains the common neural generation model with the weights. Experimental results on a large Chinese dialogue corpus show that our method improves the acceptance rate of generated responses compared with several baseline models and significantly reduces the number of generated generic responses.

Training Millions of Personalized Dialogue Agents
Pierre-Emmanuel Mazare, Samuel Humeau, Martin Raison, and Antoine Bordes 14:09–14:21

Current dialogue systems fail at being engaging for users, especially when trained end-to-end without relying on proactive reengaging scripted strategies. Zhang et al. (2018) showed that the engagement level of end-to-end dialogue models increases when conditioning them on text personas providing some personalized back-story to the model. However, the dataset used in Zhang et al. (2018) is synthetic and only contains around 1k different personas. In this paper we introduce a new dataset providing 5 million personas and 700 million persona-based dialogues. Our experiments show that, at this scale, training using personas still improves the performance of end-to-end systems. In addition, we show that other tasks benefit from the wide coverage of our dataset by fine-tuning our model on the data from Zhang et al. (2018) and achieving state-of-the-art results.

Towards Universal Dialogue State Tracking
Liliang Ren, Kaige Xie, Lu Chen, and Kai Yu 14:21–14:33

Dialogue state tracker is the core part of a spoken dialogue system. It estimates the beliefs of possible user's goals at every dialogue turn. However, for most current approaches, it's difficult to scale to large dialogue domains. They have one or more of following limitations: (a) Some models don't work in the situation where slot values in ontology changes dynamically; (b) The number of model parameters is proportional to the number of slots; (c) Some models extract features based on hand-crafted lexicons. To tackle these challenges, we propose StateNet, a universal dialogue state tracker. It is independent of the number of values, shares parameters across all slots, and uses pre-trained word vectors instead of explicit semantic dictionaries. Our experiments on two datasets show that our approach not only overcomes the limitations, but also significantly outperforms the performance of state-of-the-art approaches.

Semantic Parsing for Task Oriented Dialog using Hierarchical Representations
Sonal Gupta, Rushin Shah, Mrinal Mohit, Anuj Kumar, and Mike Lewis 14:33–14:45

Task oriented dialog systems typically first parse user utterances to semantic frames comprised of intents and slots. Previous work on task oriented intent and slot-filling work has been restricted to one intent per

query and one slot label per token, and thus cannot model complex compositional requests. Alternative semantic parsing systems have represented queries as logical forms, but these are challenging to annotate and parse. We propose a hierarchical annotation scheme for semantic parsing that allows the representation of compositional queries, and can be efficiently and accurately parsed by standard constituency parsing models. We release a dataset of 44k annotated queries (http://fb.me/semanticparsingdialog), and show that parsing models outperform sequence-to-sequence approaches on this dataset.

Session 7B: Social Applications II

Copper Hall Chair: *chairname*

The glass ceiling in NLP

Natalie Schluter 13:45–13:57

In this paper, we provide empirical evidence based on a rigourously studied mathematical model for bi-populated networks, that a glass ceiling within the field of NLP has developed since the mid 2000s.

Reducing Gender Bias in Abusive Language Detection

Ji Ho Park, Jamin Shin, and Pascale Fung 13:57–14:09

Abusive language detection models tend to have a problem of being biased toward identity words of a certain group of people because of imbalanced training datasets. For example, “You are a good woman” was considered “sexist” when trained on an existing dataset. Such model bias is an obstacle for models to be robust enough for practical use. In this work, we measure them on models trained with different datasets, while analyzing the effect of different pre-trained word embeddings and model architectures. We also experiment with three mitigation methods: (1) debiased word embeddings, (2) gender swap data augmentation, and (3) fine-tuning with a larger corpus. These methods can effectively reduce model bias by 90-98% and can be extended to correct model bias in other scenarios.

SafeCity: Understanding Diverse Forms of Sexual Harassment Personal Stories

Sweta Karlekar and Mohit Bansal 14:09–14:21

With the recent rise of #MeToo, an increasing number of personal stories about sexual harassment and sexual abuse have been shared online. In order to push forward the fight against such harassment and abuse, we present the task of automatically categorizing and analyzing various forms of sexual harassment, based on stories shared on the online forum SafeCity. For the labels of groping, ogling, and commenting, our single-label CNN-RNN model achieves an accuracy of 86.5%, and our multi-label model achieves a Hamming score of 82.5%. Furthermore, we present analysis using LIME, first-derivative saliency heatmaps, activation clustering, and embedding visualization to interpret neural model predictions and demonstrate how this helps extract features that can help automatically fill out incident reports, identify unsafe areas, avoid unsafe practices, and ’pin the creeps’.

Learning multiview embeddings for assessing dementia

Chloé Pou-Prom and Frank Rudzicz 14:21–14:33

As the incidence of Alzheimer’s Disease (AD) increases, early detection becomes crucial. Unfortunately, datasets for AD assessment are often sparse and incomplete. In this work, we leverage the multiview nature of a small AD dataset, DementiaBank, to learn an embedding that captures different modes of cognitive impairment. We apply generalized canonical correlation analysis (GCCA) to our dataset and demonstrate the added benefit of using multiview embeddings in two downstream tasks: identifying AD and predicting clinical scores. By including multiview embeddings, we obtain an F1 score of 0.82 in the classification task and a mean absolute error of 3.42 in the regression task. Furthermore, we show that multiview embeddings can be obtained from other datasets as well.

WikiConv: A Corpus of the Complete Conversational History of a Large Online Collaborative Community

Yiqing Hua, Cristian Danescu-Niculescu-Mizil, Dario Taraborelli, Nithum Thain, Jeffery Sorensen, and Lucas Dixon 14:33–14:45

We present a corpus that encompasses the complete history of conversations between contributors to Wikipedia, one of the largest online collaborative communities. By recording the intermediate states of conversations - including not only comments and replies, but also their modifications, deletions and restorations - this data offers an unprecedented view of online conversation. Our framework is designed to be language agnostic, and we show that it extracts high quality data in both Chinese and English. This level of detail supports new research questions pertaining to the process (and challenges) of large-scale online collaboration. We illustrate the corpus’ potential with two case studies on English Wikipedia that highlight new perspectives on earlier work. First, we explore how a person’s conversational behavior depends on how they relate to the discussion’s venue. Second, we show that community moderation of toxic behavior happens at a higher rate than previously estimated.

Session 7C: NER

Silver Hall Chair: *chairname*

Marginal Likelihood Training of BiLSTM-CRF for Biomedical Named Entity Recognition from Disjoint Label Sets
Nathan Greenberg, Trapit Bansal, Patrick Verga, and Andrew McCallum 13:45–13:57

Extracting typed entity mentions from text is a fundamental component to language understanding and reasoning. While there exist substantial labeled text datasets for multiple subsets of biomedical entity types—such as genes and proteins, or chemicals and diseases—it is rare to find large labeled datasets containing labels for all desired entity types together. This paper presents a method for training a single CRF extractor from multiple datasets with disjoint or partially overlapping sets of entity types. Our approach employs marginal likelihood training to insist on labels that are present in the data, while filling in "missing labels". This allows us to leverage all the available data within a single model. In experimental results on the Biocreative V CDR (chemicals/diseases), Biocreative VI ChemProt (chemicals/proteins) and MedMentions (19 entity types) datasets, we show that joint training on multiple datasets improves NER F1 over training in isolation, and our methods achieve state-of-the-art results.

Adversarial training for multi-context joint entity and relation extraction
Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder 13:57–14:09

Adversarial training (AT) is a regularization method that can be used to improve the robustness of neural network methods by adding small perturbations in the training data. We show how to use AT for the tasks of entity recognition and relation extraction. In particular, we demonstrate that applying AT to a general purpose baseline model for jointly extracting entities and relations, allows improving the state-of-the-art effectiveness on several datasets in different contexts (i.e., news, biomedical, and real estate data) and for different languages (English and Dutch).

Structured Multi-Label Biomedical Text Tagging via Attentive Neural Tree Decoding
Gaurav Singh, James Thomas, Iain Marshall, John Shawe-Taylor, and Byron C. Wallace 14:09–14:21

We propose a model for tagging unstructured texts with an arbitrary number of terms drawn from a tree-structured vocabulary (i.e., an ontology). We treat this as a special case of sequence-to-sequence learning in which the decoder begins at the root node of an ontological tree and recursively elects to expand child nodes as a function of the input text, the current node, and the latent decoder state. We demonstrate that this method yields state-of-the-art results on the important task of assigning MeSH terms to biomedical abstracts.

Deep Exhaustive Model for Nested Named Entity Recognition
Mohammad Golam Sohrab and Makoto Miwa 14:21–14:33

We propose a simple deep neural model for nested named entity recognition (NER). Most NER models focused on flat entities and ignored nested entities, which failed to fully capture underlying semantic information in texts. The key idea of our model is to enumerate all possible regions or spans as potential entity mentions and classify them with deep neural networks. To reduce the computational costs and capture the information of the contexts around the regions, the model represents the regions using the outputs of shared underlying bidirectional long short-term memory. We evaluate our exhaustive model on the GENIA and JNLPBA corpora in biomedical domain, and the results show that our model outperforms state-of-the-art models on nested and flat NER, achieving 77.1% and 78.4% respectively in terms of F-score, without any external knowledge resources.

Natural Language Processing Not-At-All from Scratch: Evaluating The Utility of Hand-crafted Features in Deep Learning
Minghao Wu, Fei Liu, and Trevor Cohn 14:33–14:45

Conventional wisdom is that hand-crafted features are redundant for deep learning models, as they already learn adequate representations of text automatically from corpora. In this work, we test this claim by proposing a new method for exploiting handcrafted features as part of a novel hybrid learning approach, incorporating a feature auto-encoder loss component. We evaluate on the task of named entity recognition (NER), where we show that including manual features for part-of-speech, word shapes and gazetteers can improve the performance of a neural CRF model. We obtain a F 1 of 91.89 for the CoNLL-2003 English shared task, which significantly outperforms a collection of highly competitive baseline models. We also present an ablation study showing the importance of auto-encoding, over using features as either inputs or outputs alone, and moreover, show including the autoencoder components reduces training requirements to 60%, while retaining the same predictive accuracy.

Session 7D: Morphology / Parsing

Hall 100 Chair: *chairname*

Improved Dependency Parsing using Implicit Word Connections Learned from Unlabeled Data

Wenhui Wang, Baobao Chang, and Mairgup Mansur 13:45–13:57

Pre-trained word embeddings and language model have been shown useful in a lot of tasks. However, both of them cannot directly capture word connections in a sentence, which is important for dependency parsing given its goal is to establish dependency relations between words. In this paper, we propose to implicitly capture word connections from unlabeled data by a word ordering model with self-attention mechanism. Experiments show that these implicit word connections do improve our parsing model. Furthermore, by combining with a pre-trained language model, our model gets state-of-the-art performance on the English PTB dataset, achieving 96.35% UAS and 95.25% LAS.

A Framework for Understanding the Role of Morphology in Universal Dependency Parsing

Mathieu Dehouck and Pascal Denis 13:57–14:09

This paper presents a simple framework for characterizing morphological complexity and how it encodes syntactic information. In particular, we propose a new measure of morpho-syntactic complexity in terms of governor-dependent preferential attachment that explains parsing performance. Through experiments on dependency parsing with data from Universal Dependencies (UD), we show that representations derived from morphological attributes deliver important parsing performance improvements over standard word form embeddings when trained on the same datasets. We also show that the new morpho-syntactic complexity measure is predictive of the gains provided by using morphological attributes over plain forms on parsing scores, making it a tool to distinguish languages using morphology as a syntactic marker from others.

The Lazy Encoder: A Fine-Grained Analysis of the Role of Morphology in Neural Machine Translation

Arianna Bisazza and Clara Tump 14:09–14:21

Neural sequence-to-sequence models have proven very effective for machine translation, but at the expense of model interpretability. To shed more light into the role played by linguistic structure in the process of neural machine translation, we perform a fine-grained analysis of how various source-side morphological features are captured at different levels of the NMT encoder while varying the target language. Differently from previous work, we find no correlation between the accuracy of source morphology encoding and translation quality. We do find that morphological features are only captured in context and only to the extent that they are directly transferable to the target words.

Imitation Learning for Neural Morphological String Transduction

Peter Makarov and Simon Clematide 14:21–14:33

We employ imitation learning to train a neural transition-based string transducer for morphological tasks such as inflection generation and lemmatization. Previous approaches to training this type of model either rely on an external character aligner for the production of gold action sequences, which results in a suboptimal model due to the unwarranted dependence on a single gold action sequence despite spurious ambiguity, or require warm starting with an MLE model. Our approach only requires a simple expert policy, eliminating the need for a character aligner or warm start. It also addresses familiar MLE training biases and leads to strong and state-of-the-art performance on several benchmarks.

An Encoder-Decoder Approach to the Paradigm Cell Filling Problem

Miikka Silfverberg and Mans Hulden 14:33–14:45

The Paradigm Cell Filling Problem in morphology asks to complete word inflection tables from partial ones. We implement novel neural models for this task, evaluating them on 18 data sets in 8 languages, showing performance that is comparable with previous work with far less training data. We also publish a new dataset for this task and code implementing the system described in this paper.

Session 7E: Short Posters III

13:45–14:45

Hall 100

Chair: *chairname*

Generating Natural Language Adversarial Examples

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang

Deep neural networks (DNNs) are vulnerable to adversarial examples, perturbations to correctly classified examples which can cause the model to misclassify. In the image domain, these perturbations can often be made virtually indistinguishable to human perception, causing humans and state-of-the-art models to disagree. However, in the natural language domain, small perturbations are clearly perceptible, and the replacement of a single word can drastically alter the semantics of the document. Given these challenges, we use a black-box population-based optimization algorithm to generate semantically and syntactically similar adversarial examples that fool well-trained sentiment analysis and textual entailment models with success rates of 97% and 70%, respectively. We additionally demonstrate that 92.3% of the successful sentiment analysis adversarial examples are classified to their original label by 20 human annotators, and that the examples are perceptibly quite similar. Finally, we discuss an attempt to use adversarial training as a defense, but fail to yield improvement, demonstrating the strength and diversity of our adversarial examples. We hope our findings encourage researchers to pursue improving the robustness of DNNs in the natural language domain.

Multi-Head Attention with Disagreement Regularization

Jian Li, Zhaopeng Tu, Baosong Yang, Michael R. Lyu, and Tong Zhang

Multi-head attention is appealing for the ability to jointly attend to information from different representation subspaces at different positions. In this work, we introduce a disagreement regularization to explicitly encourage the diversity among multiple attention heads. Specifically, we propose three types of disagreement regularization, which respectively encourage the subspace, the attended positions, and the output representation associated with each attention head to be different from other heads. Experimental results on widely-used WMT14 English-German and WMT17 Chinese-English translation tasks demonstrate the effectiveness and universality of the proposed approach.

Deep Bayesian Active Learning for Natural Language Processing: Results of a Large-Scale Empirical Study

Aditya Siddhant and Zachary C. Lipton

Several recent papers investigate Active Learning (AL) for mitigating the data dependence of deep learning for natural language processing. However, the applicability of AL to real-world problems remains an open question. While in supervised learning, practitioners can try many different methods, evaluating each against a validation set before selecting a model, AL affords no such luxury. Over the course of one AL run, an agent annotates its dataset exhausting its labeling budget. Thus, given a new task, we have no opportunity to compare models and acquisition functions. This paper provides a large-scale empirical study of deep active learning, addressing multiple tasks and, for each, multiple datasets, multiple models, and a full suite of acquisition functions. We find that across all settings, Bayesian active learning by disagreement, using uncertainty estimates provided either by Dropout or Bayes-by-Backprop significantly improves over i.i.d. baselines and usually outperforms classic uncertainty sampling.

Bayesian Compression for Natural Language Processing

Nadezhda Chirkova, Ekaterina Lobacheva, and Dmitry Vetrov

In natural language processing, a lot of the tasks are successfully solved with recurrent neural networks, but such models have a huge number of parameters. The majority of these parameters are often concentrated in the embedding layer, which size grows proportionally to the vocabulary length. We propose a Bayesian sparsification technique for RNNs which allows compressing the RNN dozens or hundreds of times without time-consuming hyperparameters tuning. We also generalize the model for vocabulary sparsification to filter out unnecessary words and compress the RNN even further. We show that the choice of the kept words is interpretable.

Multimodal neural pronunciation modeling for spoken languages with logographic origin

Minh Nguyen, Gia H Ngo, and Nancy Chen

Graphemes of most languages encode pronunciation, though some are more explicit than others. Languages like Spanish have a straightforward mapping between its graphemes and phonemes, while this mapping is more convoluted for languages like English. Spoken languages such as Cantonese present even more challenges in pronunciation modeling: (1) they do not have a standard written form, (2) the closest graphemic origins are logographic Han characters, of which only a subset of these logographic characters implicitly encodes pronunciation. In this work, we propose a multimodal approach to predict the pronunciation of Cantonese logographic characters, using neural networks with a geometric representation of logographs and pronunciation of cognates in historically related languages. The proposed framework improves performance by 18.1% and 25.0% respective to unimodal and multimodal baselines.

Chinese Pinyin Aided IME, Input What You Have Not Keystroked Yet

Yafang Huang and Hai Zhao

Chinese pinyin input method engine (IME) converts pinyin into character so that Chinese characters can be conveniently inputted into computer through common keyboard. IMEs work relying on its core component, pinyin-to-character conversion (P2C). Usually Chinese IMEs simply predict a list of character sequences for user choice only according to user pinyin input at each turn. However, Chinese inputting is a multi-turn online procedure, which can be supposed to be exploited for further user experience promoting. This paper thus for the first time introduces a sequence-to-sequence model with gated-attention mechanism for the core task in IMEs. The proposed neural P2C model is learned by encoding previous input utterance as extra context to enable our IME capable of predicting character sequence with incomplete pinyin input. Our model is evaluated in different benchmark datasets showing great user experience improvement compared to traditional models, which demonstrates the first engineering practice of building Chinese aided IME.

Estimating Marginal Probabilities of n-grams for Recurrent Neural Language Models

Thanapon Noraset, Doug Downey, and Lidong Bing

Recurrent neural network language models (RNNLMs) are the current standard-bearer for statistical language modeling. However, RNNLMs only estimate probabilities for complete sequences of text, whereas some applications require context-independent phrase probabilities instead. In this paper, we study how to compute an RNNLM's em marginal probability: the probability that the model assigns to a short sequence of text when the preceding context is not known. We introduce a simple method of altering the RNNLM training to make the model more accurate at marginal estimation. Our experiments demonstrate that the technique is effective compared to baselines including the traditional RNNLM probability and an importance sampling approach. Finally, we show how we can use the marginal estimation to improve an RNNLM by training the marginals to match n-gram probabilities from a larger corpus.

How to represent a word and predict it, too: improving tied architectures for language modelling

Kristina Gulordava, Laura Aina, and Gemma Boleda

Recent state-of-the-art neural language models share the representations of words given by the input and output mappings. We propose a simple modification to these architectures that decouples the hidden state from the word embedding prediction. Our architecture leads to comparable or better results compared to previous tied models and models without tying, with a much smaller number of parameters. We also extend our proposal to word2vec models, showing that tying is appropriate for general word prediction tasks.

The Importance of Generation Order in Language Modeling

Nicolas Ford, Daniel Duckworth, Mohammad Norouzi, and George Dahl

Neural language models are a critical component of state-of-the-art systems for machine translation, summarization, audio transcription, and other tasks. These language models are almost universally autoregressive in nature, generating sentences one token at a time from left to right. This paper studies the influence of token generation order on model quality via a novel two-pass language model that produces partially-filled sentence "templates" and then fills in missing tokens. We compare various strategies for structuring these two passes and observe a surprisingly large variation in model quality. We find the most effective strategy generates function words in the first pass followed by content words in the second. We believe these experimental results justify a more extensive investigation of the generation order for neural language models.

Towards Document-Level Neural Machine Translation with Hierarchical Attention Networks

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson

Neural Machine Translation (NMT) can be improved by including document-level contextual information. For this purpose, we propose a hierarchical attention model to capture the context in a structured and dynamic manner. The model is integrated in the original NMT architecture as another level of abstraction, conditioning on the NMT model's own previous hidden states. Experiments show that hierarchical attention significantly improves the BLEU score over a strong NMT baseline with the state-of-the-art in context-aware methods, and that both the encoder and decoder benefit from context in complementary ways.

Three Strategies to Improve One-to-Many Multilingual Translation
Yining Wang, Jiajun Zhang, Feifei Zhai, Jingfang Xu, and Chengqing Zong

Due to the benefits of model compactness, multilingual translation (including many-to-one, many-to-many and one-to-many) based on a universal encoder-decoder architecture attracts more and more attention. However, previous studies show that one-to-many translation based on this framework cannot perform on par with the individually trained models. In this work, we introduce three strategies to improve one-to-many multilingual translation by balancing the shared and unique features. Within the architecture of one decoder for all target languages, we first exploit the use of unique initial states for different target languages. Then, we employ language-dependent positional embeddings. Finally and especially, we propose to divide the hidden cells of the decoder into shared and language-dependent ones. The extensive experiments demonstrate that our proposed methods can obtain remarkable improvements over the strong baselines. Moreover, our strategies can achieve comparable or even better performance than the individually trained translation models.

Multi-Source Syntactic Neural Machine Translation
Anna Currey and Kenneth Heafield

We introduce a novel multi-source technique for incorporating source syntax into neural machine translation using linearized parses. This is achieved by employing separate encoders for the sequential and parsed versions of the same source sentence; the resulting representations are then combined using a hierarchical attention mechanism. The proposed model improves over both seq2seq and parsed baselines by over 1 BLEU on the WMT17 English-German task. Further analysis shows that our multi-source syntactic model is able to translate successfully without any parsed input, unlike standard parsed methods. In addition, performance does not deteriorate as much on long sentences as for the baselines.

Fixing Translation Divergences in Parallel Corpora for Neural MT
Minh Quang Pham, Josep Crego, Jean Senellart, and François Yvon

Corpus-based approaches to machine translation rely on the availability of clean parallel corpora. Such resources are scarce, and because of the automatic processes involved in their preparation, they are often noisy. This paper describes an unsupervised method for detecting translation divergences in parallel sentences. We rely on a neural network that computes cross-lingual sentence similarity scores, which are then used to effectively filter out divergent translations. Furthermore, similarity scores predicted by the network are used to identify and fix some partial divergences, yielding additional parallel segments. We evaluate these methods for English-French and English-German machine translation tasks, and show that using filtered/corrected corpora actually improves MT performance.

Adversarial Evaluation of Multimodal Machine Translation
Desmond Elliott

The promise of combining language and vision in multimodal machine translation is that systems will produce better translations by leveraging the image data. However, the evidence surrounding whether the images are useful is unconvincing due to inconsistencies between text-similarity metrics and human judgements. We present an adversarial evaluation to directly examine the utility of the image data in this task. Our evaluation tests whether systems perform better when paired with congruent images or incongruent images. This evaluation shows that only one out of three publicly available systems is sensitive to this perturbation of the data. We recommend that multimodal translation systems should be able to pass this sanity check in the future.

Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion
Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave

Continuous word representations learned separately on distinct languages can be aligned so that their words become comparable in a common space. Existing works typically solve a quadratic problem to learn a orthogonal matrix aligning a bilingual lexicon, and use a retrieval criterion for inference. In this paper, we propose an unified formulation that directly optimizes a retrieval criterion in an end-to-end fashion. Our experiments on standard benchmarks show that our approach outperforms the state of the art on word translation, with the biggest improvements observed for distant language pairs such as English-Chinese.

Learning When to Concentrate or Divert Attention: Automatic Control of Attention Temperature for Neural Machine Translation
Junyang Lin, Xu SUN, Xuancheng Ren, Muyu Li, and Qi Su

Most of the Neural Machine Translation (NMT) models are based on the sequence-to-sequence (Seq2Seq) model with an encoder-decoder framework equipped with the attention mechanism. However, the conventional attention mechanism treats the decoding at each time step equally with the same matrix, which is problematic since the softness of the attention for different types of words (e.g. content words and function words) should differ. Therefore, we propose a new model with a mechanism called Self-Adaptive Control of Temperature (SACT) to control the softness of attention by means of an attention temperature. Experimental results on the Chinese-English translation and English-Vietnamese translation demonstrate that our model outperforms the baseline models, and the analysis and the case study show that our model can attend to the most relevant elements in the source-side contexts and generate the translation of high quality.

Accelerating Asynchronous Stochastic Gradient Descent for Neural Machine Translation
Nikolay Bogoychev, Kenneth Heafield, Alham Fikri Aji, and Marcin Junczys-Dowmunt

In order to extract the best possible performance from asynchronous stochastic gradient descent one must increase the mini-batch size and scale the learning rate accordingly. In order to achieve further speedup we introduce a technique that delays gradient updates effectively increasing the mini-batch size. Unfortunately with the increase of mini-batch size we worsen the stale gradient problem in asynchronous stochastic gradient descent (SGD) which makes the model convergence poor. We introduce local optimizers which mitigate the stale gradient problem and together with fine tuning our momentum we are able to train a shallow machine translation system 27% faster than an optimized baseline with negligible penalty in BLEU.

Learning to Jointly Translate and Predict Dropped Pronouns with a Shared Reconstruction Mechanism
Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu

Pronouns are frequently omitted in pro-drop languages, such as Chinese, generally leading to significant challenges with respect to the production of complete translations. Recently, Wang et al. (2018) proposed a novel reconstruction-based approach to alleviating dropped pronoun (DP) translation problems for neural machine translation models. In this work, we improve the original model from two perspectives. First, we employ a shared reconstructor to better exploit encoder and decoder representations. Second, we jointly learn to translate and predict DPs in an end-to-end manner, to avoid the errors propagated from an external DP prediction model. Experimental results show that our approach significantly improves both translation performance and DP prediction accuracy.

Getting Gender Right in Neural MT
Eva Vanmassenhove, Christian Hardmeier, and Andy Way

Speakers of different languages must attend to and encode strikingly different aspects of the world in order to use their language correctly (Sapir, 1921; Slobin, 1996). One such difference is related to the way gender is expressed in a language. Saying “I am happy” in English, does not encode any additional knowledge of the speaker that uttered the sentence. However, many other languages do have grammatical gender systems and so such knowledge would be encoded. In order to correctly translate such a sentence into, say, French, the inherent gender information needs to be retained/recovered. The same sentence would become either “Je suis heureux”, for a male speaker or “Je suis heureuse” for a female one. Apart from morphological agreement, demographic factors (gender, age, etc.) also influence our use of language in terms of word choices or syntactic constructions (Tannen, 1991; Pennebaker et al., 2003). We integrate gender information into NMT systems. Our contribution is two-fold: (1) the compilation of large datasets with speaker information for 20 language pairs, and (2) a simple set of experiments that incorporate gender information into NMT for multiple language pairs. Our experiments show that adding a gender feature to an NMT system significantly improves the translation quality for some language pairs.

Towards Two-Dimensional Sequence to Sequence Model in Neural Machine Translation

Parnia Bahar, Christopher Brix, and Hermann Ney

This work investigates an alternative model for neural machine translation (NMT) and proposes a novel architecture, where we employ a multi-dimensional long short-term memory (MDLSTM) for translation modelling. In the state-of-the-art methods, source and target sentences are treated as one-dimensional sequences over time, while we view translation as a two-dimensional (2D) mapping using an MDLSTM layer to define the correspondence between source and target words. We extend beyond the current sequence to sequence backbone NMT models to a 2D structure in which the source and target sentences are aligned with each other in a 2D grid. Our proposed topology shows consistent improvements over attention-based sequence to sequence model on two WMT 2017 tasks, German<->English.

End-to-End Non-Autoregressive Neural Machine Translation with Connectionist Temporal Classification

Jindřich Libovický and Jindřich Helcl

Autoregressive decoding is the only part of sequence-to-sequence models that prevents them from massive parallelization at inference time. Non-autoregressive models enable the decoder to generate all output symbols independently in parallel. We present a novel non-autoregressive architecture based on connectionist temporal classification and evaluate it on the task of neural machine translation. Unlike other non-autoregressive methods which operate in several steps, our model can be trained end-to-end. We conduct experiments on the WMT English-Romanian and English-German datasets. Our models achieve a significant speedup over the autoregressive models, keeping the translation quality comparable to other non-autoregressive models.

Prediction Improves Simultaneous Neural Machine Translation

Ashkan Alinejad, Maryam Siahbani, and Anoop Sarkar

Simultaneous speech translation aims to maintain translation quality while minimizing the delay between reading input and incrementally producing the output. We propose a new general-purpose prediction action which predicts future words in the input to improve quality and minimize delay in simultaneous translation. We train this agent using reinforcement learning with a novel reward function. Our agent with prediction has better translation quality and less delay compared to an agent-based simultaneous translation system without prediction.

Training Deeper Neural Machine Translation Models with Transparent Attention

Ankur Bapna, Mia Chen, Orhan Firat, Yuan Cao, and Yonghui Wu

While current state-of-the-art NMT models, such as RNN seq2seq and Transformers, possess a large number of parameters, they are still shallow in comparison to convolutional models used for both text and vision applications. In this work we attempt to train significantly (2-3x) deeper Transformer and Bi-RNN encoders for machine translation. We propose a simple modification to the attention mechanism that eases the optimization of deeper models, and results in consistent gains of 0.7-1.1 BLEU on the benchmark WMT'14 English-German and WMT'15 Czech-English tasks for both architectures.

Context and Copying in Neural Machine Translation

Rebecca Knowles and Philipp Koehn

Neural machine translation systems with subword vocabularies are capable of translating or copying unknown words. In this work, we show that they learn to copy words based on both the context in which the words appear as well as features of the words themselves. In contexts that are particularly copy-prone, they even copy words that they have already learned they should translate. We examine the influence of context and subword features on this and other types of copying behavior.

Encoding Gated Translation Memory into Neural Machine Translation

Qian Cao and Deyi Xiong

Translation memories (TM) facilitate human translators to reuse existing repetitive translation fragments. In this paper, we propose a novel method to combine the strengths of both TM and neural machine translation (NMT) for high-quality translation. We treat the target translation of a TM match as an additional reference input and encode it into NMT with an extra encoder. A gating mechanism is further used to balance the impact of the TM match on the NMT decoder. Experiment results on the UN corpus demonstrate that when fuzzy matches are higher than 50%, the quality of NMT translation can be significantly improved by over 10 BLEU points.

Automatic Post-Editing of Machine Translation: A Neural Programmer-Interpreter Approach
Thuy-Trang Vu and Gholamreza Haffari

Automated Post-Editing (PE) is the task of automatically correct common and repetitive errors found in machine translation (MT) output. In this paper, we present a neural programmer-interpreter approach to this task, resembling the way that human perform post-editing using discrete edit operations, wich we refer to as programs. Our model outperforms previous neural models for inducing PE programs on the WMT17 APE task for German-English up to +1 BLEU score and -0.7 TER scores.

Break the Beam Search Curse: A Study of Score-Revision Methods and Stopping Criteria for Neural Machine Translation
Yilin Yang, Liang Huang, and Mingbo Ma

Beam search is widely used in neural machine translation, and usually improves translation quality compared to greedy search. It has been widely observed that, however, beam sizes larger than 5 hurt translation quality. We explain why this happens, and propose several methods to address this problem. Furthermore, we discuss the optimal stopping criteria for these methods. Results show that our hyperparameter-free methods outperform the widely-used hyperparameter-free heuristic of length normalization by +2.0 BLEU, and achieve the best results among all methods on Chinese-to-English translation.

Multi-View Learning: Multilingual and Multi-Representation Entity Typing
Yadollah Yaghoobzadeh and Hinrich Schütze

Accurate and complete knowledge bases (KBs) are paramount in NLP. We employ mul-itiview learning for increasing the accuracy and coverage of entity type information in KBs. We rely on two metaviews: language and representation. For language, we consider high- resource and low-resource languages from Wikipedia. For representation, we consider representations based on the context distribution of the entity (i.e., on its embedding), on the entity's name (i.e., on its surface form) and on its description in Wikipedia. The two metaviews language and representation can be freely combined: each pair of language and representation (e.g., German embedding, English description, Spanish name) is a distinct view. Our experiments on entity typing with fine-grained classes demonstrate the effectiveness of multiview learning. We release MVET, a large multiview — and, in particular, multilingual — entity typing dataset we created. Mono- and multilingual fine-grained entity typing systems can be evaluated on this dataset.

Word Embeddings for Code-Mixed Language Processing
Adithya Pratapa, Monojit Choudhury, and Sunayana Sitaram

We compare three existing bilingual word embedding approaches, and a novel approach of training skip-grams on synthetic code-mixed text generated through linguistic models of code-mixing, on two tasks - sentiment analysis and POS tagging for code-mixed text. Our results show that while CVM and CCA based embeddings perform as well as the proposed embedding technique on semantic and syntactic tasks respectively, the proposed approach provides the best performance for both tasks overall. Thus, this study demonstrates that existing bilingual embedding techniques are not ideal for code-mixed text processing and there is a need for learning multilingual word embedding from the code-mixed text.

The Internal Structure of Name Tokens: A Multilingual Study
Xiaodong Yu, Stephen Mayhew, Mark Sammons, and Dan Roth

Character-level patterns have been widely used as features in English Named Entity Recognition (NER) systems. However, to date there has been no direct investigation of the inherent differences between name and nonname tokens in text, nor whether this property holds across multiple languages. This paper analyzes the capabilities of corpus-agnostic Character-level Language Models (CLMs) in the binary task of distinguishing name tokens from non-name tokens. We demonstrate that CLMs provide a simple and powerful model for capturing these differences, identifying named entity tokens in a diverse set of languages at close to the performance of full NER systems. Moreover, by adding very simple CLM-based features we can significantly improve the performance of an off-the-shelf NER system for multiple languages.

Code-switched Language Models Using Dual RNNs and Same-Source Pretraining
Saurabh Garg, Tanmay Parekh, and Preethi Jyothi

This work focuses on building language models (LMs) for code-switched text. We propose two techniques that significantly improve these LMs: 1) A novel recurrent neural network unit with dual components that focus on each language in the code-switched text separately 2) Pretraining the LM using synthetic text from a generative model estimated using the training data. We demonstrate the effectiveness of our proposed techniques by reporting perplexities on a Mandarin-English task and derive significant reductions in perplexity.

Part-of-Speech Tagging for Code-Switched, Transliterated Texts without Explicit Language Identification
Kelsey Ball and Dan Garrette

Code-switching, the use of more than one language within a single utterance, is ubiquitous in much of the world, but remains a challenge for NLP largely due to the lack of representative data for training models. In this paper, we present a novel model architecture that is trained exclusively on monolingual resources, but can be applied to unseen code-switched text at inference time. The model accomplishes this by jointly maintaining separate word representations for each of the possible languages, or scripts in the case of transliteration, allowing each to contribute to inferences without forcing the model to commit to a language. Experiments on Hindi-English part-of-speech tagging demonstrate that our approach outperforms standard models when training on monolingual text without transliteration, and testing on code-switched text with alternate scripts.

Session 8 Overview – Saturday, November 3, 2018

Oral tracks

Track A *Text Categorization* Gold Hall	**Track B** *Generation* Copper Hall	**Track C** *Knowledge Graphs* Silver Hall	**Track D** *Morphology / Phonology* Hall 100	
Zero-shot User Intent Detection via Capsule Neural Networks *Xia, Zhang, Yan, Chang, and Yu*	Automatic Poetry Generation with Mutual Reinforcement Learning *Yi, Sun, Li, and Li*	Knowledge Graph Embedding with Hierarchical Relation Structure *Zhang, Zhuang, Qu, Lin, and He*	Neural Transductive Learning and Beyond: Morphological Generation in the Minimal-Resource Setting *Kann and Schütze*	16:30
Hierarchical Neural Networks for Sequential Sentence Classification in Medical Scientific Abstracts *Jin and Szolovits*	Variational Autoregressive Decoder for Neural Response Generation *Du, Li, He, Xu, Bing, and Wang*	Embedding Multimodal Relational Data for Knowledge Base Completion *Pezeshkpour, Chen, and Singh*	Implicational Universals in Stochastic Constraint-Based Phonology *Magri*	16:48
Investigating Capsule Networks with Dynamic Routing for Text Classification *Yang, Zhao, Ye, Lei, Zhao, and Zhang*	Integrating Transformer and Paraphrase Rules for Sentence Simplification *Zhao, Meng, He, Saptono, and Parmanto*	Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction *Luan, He, Ostendorf, and Hajishirzi*	Explaining Character-Aware Neural Networks for Word-Level Prediction: Do They Discover Linguistic Rules? *Godin, Demuynck, Dambre, De Neve, and Demeester*	17:06
Topic Memory Networks for Short Text Classification *Zeng, Li, Song, Gao, Lyu, and King*	Learning Neural Templates for Text Generation *Wiseman, Shieber, and Rush*	Playing 20 Question Game with Policy-Based Reinforcement Learning *Hu, Wu, Luo, Tao, Xu, and Chen*	Adapting Word Embeddings to New Languages with Morphological and Phonological Subword Representations *Chaudhary, Zhou, Levin, Neubig, Mortensen, and Carbonell*	17:24
Few-Shot and Zero-Shot Multi-Label Learning for Structured Label Spaces *Rios and Kavuluru*	Multi-Reference Training with Pseudo-Reference Generation for Text Generation *Zheng, Ma, and Huang*	Multi-Hop Knowledge Graph Reasoning with Reward Shaping *Lin, Socher, and Xiong*	**TACL-1420** **TACL-1420**	17:42

Poster tracks

16:30–18:00

Track D: *Sentiment, Social Applications, Multimodal Semantics, Discourse* Hall 100

Parallel Session 8

Session 8A: Text Categorization

Gold Hall Chair: *chairname*

Zero-shot User Intent Detection via Capsule Neural Networks
Congying Xia, Chenwei Zhang, Xiaohui Yan, Yi Chang, and Philip Yu 16:30–16:48

User intent detection plays a critical role in question-answering and dialog systems. Most previous works treat intent detection as a classification problem where utterances are labeled with predefined intents. However, it is labor-intensive and time-consuming to label users' utterances as intents are diversely expressed and novel intents will continually be involved. Instead, we study the zero-shot intent detection problem, which aims to detect emerging user intents where no labeled utterances are currently available. We propose two capsule-based architectures: IntentCapsNet that extracts semantic features from utterances and aggregates them to discriminate existing intents, and IntentCapsNet-ZSL which gives IntentCapsNet the zero-shot learning ability to discriminate emerging intents via knowledge transfer from existing intents. Experiments on two real-world datasets show that our model not only can better discriminate diversely expressed existing intents, but is also able to discriminate emerging intents when no labeled utterances are available.

Hierarchical Neural Networks for Sequential Sentence Classification in Medical Scientific Abstracts
Di Jin and Peter Szolovits 16:48–17:06

Prevalent models based on artificial neural network (ANN) for sentence classification often classify sentences in isolation without considering the context in which sentences appear. This hampers the traditional sentence classification approaches to the problem of sequential sentence classification, where structured prediction is needed for better overall classification performance. In this work, we present a hierarchical sequential labeling network to make use of the contextual information within surrounding sentences to help classify the current sentence. Our model outperforms the state-of-the-art results by 2%-3% on two benchmarking datasets for sequential sentence classification in medical scientific abstracts.

Investigating Capsule Networks with Dynamic Routing for Text Classification
Min Yang, Wei Zhao, Jianbo Ye, Zeyang Lei, Zhou Zhao, and Soufei Zhang 17:06–17:24

In this study, we explore capsule networks with dynamic routing for text classification. We propose three strategies to stabilize the dynamic routing process to alleviate the disturbance of some noise capsules which may contain "background" information or have not been successfully trained. A series of experiments are conducted with capsule networks on six text classification benchmarks. Capsule networks achieve state of the art on 4 out of 6 datasets, which shows the effectiveness of capsule networks for text classification. We additionally show that capsule networks exhibit significant improvement when transfer single-label to multi-label text classification over strong baseline methods. To the best of our knowledge, this is the first work that capsule networks have been empirically investigated for text modeling.

Topic Memory Networks for Short Text Classification
Jichuan Zeng, Jing Li, Yan Song, Cuiyun Gao, Michael R. Lyu, and Irwin King 17:24–17:42

Many classification models work poorly on short texts due to data sparsity. To address this issue, we propose topic memory networks for short text classification with a novel topic memory mechanism to encode latent topic representations indicative of class labels. Different from most prior work that focuses on extending features with external knowledge or pre-trained topics, our model jointly explores topic inference and text classification with memory networks in an end-to-end manner. Experimental results on four benchmark datasets show that our model outperforms state-of-the-art models on short text classification, meanwhile generates coherent topics.

Few-Shot and Zero-Shot Multi-Label Learning for Structured Label Spaces
Anthony Rios and Ramakanth Kavuluru 17:42–18:00

Large multi-label datasets contain labels that occur thousands of times (frequent group), those that occur only a few times (few-shot group), and labels that never appear in the training dataset (zero-shot group). Multi-label few- and zero-shot label prediction is mostly unexplored on datasets with large label spaces, especially for text classification. In this paper, we perform a fine-grained evaluation to understand how state-of-the-art methods perform on infrequent labels. Furthermore, we develop few- and zero-shot methods for multi-label text classification when there is a known structure over the label space, and evaluate them on two publicly available medical text datasets: MIMIC II and MIMIC III. For few-shot labels we achieve improvements of 6.2% and 4.8% in R10 for MIMIC II and MIMIC III, respectively, over prior efforts; the corresponding R10 improvements for zero-shot labels are 17.3% and 19%.

Session 8B: Generation

Copper Hall Chair: *chairname*

Automatic Poetry Generation with Mutual Reinforcement Learning
Xiaoyuan Yi, Maosong Sun, Ruoyu Li, and Wenhao Li 16:30–16:48

Poetry is one of the most beautiful forms of human language art. As a crucial step towards computer creativity, automatic poetry generation has drawn researchers' attention for decades. In recent years, some neural models have made remarkable progress in this task. However, they are all based on maximum likelihood estimation, which only learns common patterns of the corpus and results in loss-evaluation mismatch. Human experts evaluate poetry in terms of some specific criteria, instead of word-level likelihood. To handle this problem, we directly model the criteria and use them as explicit rewards to guide gradient update by reinforcement learning, so as to motivate the model to pursue higher scores. Besides, inspired by writing theories, we propose a novel mutual reinforcement learning schema. We simultaneously train two learners (generators) which learn not only from the teacher (rewarder) but also from each other to further improve performance. We experiment on Chinese poetry. Based on a strong basic model, our method achieves better results and outperforms the current state-of-the-art method.

Variational Autoregressive Decoder for Neural Response Generation
Jiachen Du, Wenjie Li, Yulan He, Ruifeng Xu, Lidong Bing, and Xuan Wang 16:48–17:06

Combining the virtues of probability graphic models and neural networks, Conditional Variational Autoencoder (CVAE) has shown promising performance in applications such as response generation. However, existing CVAE-based models often generate responses from a single latent variable which may not be sufficient to model high variability in responses. To solve this problem, we propose a novel model that sequentially introduces a series of latent variables to condition the generation of each word in the response sequence. In addition, the approximate posteriors of these latent variables are augmented with a backward Recurrent Neural Network (RNN), which allows the latent variables to capture long-term dependencies of future tokens in generation. To facilitate training, we supplement our model with an auxiliary objective that predicts the subsequent bag of words. Empirical experiments conducted on Opensubtitle and Reddit datasets show that the proposed model leads to significant improvement on both relevance and diversity over state-of-the-art baselines.

Integrating Transformer and Paraphrase Rules for Sentence Simplification
Sanqiang Zhao, Rui Meng, Daqing He, Andi Saptono, and Bambang Parmanto 17:06–17:24

Sentence simplification aims to reduce the complexity of a sentence while retaining its original meaning. Current models for sentence simplification adopted ideas from machine translation studies and implicitly learned simplification mapping rules from normal-simple sentence pairs. In this paper, we explore a novel model based on a multi-layer and multi-head attention architecture and we propose two innovative approaches to integrate the Simple PPDB (A Paraphrase Database for Simplification), an external paraphrase knowledge base for simplification that covers a wide range of real-world simplification rules. The experiments show that the integration provides two major benefits: (1) the integrated model outperforms multiple state-of-the-art baseline models for sentence simplification in the literature (2) through analysis of the rule utilization, the model seeks to select more accurate simplification rules. The code and models used in the paper are available at https://github.com/Sanqiang/text_simplification.

Learning Neural Templates for Text Generation
Sam Wiseman, Stuart Shieber, and Alexander Rush 17:24–17:42

While neural, encoder-decoder models have had significant empirical success in text generation, there remain several unaddressed problems with this style of generation. Encoder-decoder models are largely (a) uninterpretable, and (b) difficult to control in terms of their phrasing or content. This work proposes a neural generation system using a hidden semi-markov model (HSMM) decoder, which learns latent, discrete templates jointly with learning to generate. We show that this model learns useful templates, and that these templates make generation both more interpretable and controllable. Furthermore, we show that this approach scales to real data sets and achieves strong performance nearing that of encoder-decoder text generation models.

Multi-Reference Training with Pseudo-Reference Generation for Text Generation
Renjie Zheng, Mingbo Ma, and Liang Huang 17:42–18:00

Neural text generation, including neural machine translation, image captioning, and summarization, has been quite successful recently. However, during training time, typically only one reference is considered for each example, even though there are often multiple references available, e.g., 4 references in NIST MT evaluations, and 5 references in image captioning data. We first investigate several different ways of utilizing multiple human references during training. But more importantly, we then propose an algorithm to generate exponentially many pseudo-references by first compressing existing human references into lattices and then traversing them to generate new pseudo-references. These approaches lead to substantial improvements over strong baselines in both machine translation (+1.5 BLEU) and image captioning (+3.1 BLEU / +11.7 CIDEr).

Session 8C: Knowledge Graphs

Silver Hall Chair: *chairname*

Knowledge Graph Embedding with Hierarchical Relation Structure

Zhao Zhang, Fuzhen Zhuang, Meng Qu, Fen Lin, and Qing He 16:30–16:48

The rapid development of knowledge graphs (KGs), such as Freebase and WordNet, has changed the paradigm for AI-related applications. However, even though these KGs are impressively large, most of them are suffering from incompleteness, which leads to performance degradation of AI applications. Most existing researches are focusing on knowledge graph embedding (KGE) models. Nevertheless, those models simply embed entities and relations into latent vectors without leveraging the rich information from the relation structure. Indeed, relations in KGs conform to a three-layer hierarchical relation structure (HRS), i.e., semantically similar relations can make up relation clusters and some relations can be further split into several fine-grained sub-relations. Relation clusters, relations and sub-relations can fit in the top, the middle and the bottom layer of three-layer HRS respectively. To this end, in this paper, we extend existing KGE models TransE, TransH and DistMult, to learn knowledge representations by leveraging the information from the HRS. Particularly, our approach is capable to extend other KGE models. Finally, the experiment results clearly validate the effectiveness of the proposed approach against baselines.

Embedding Multimodal Relational Data for Knowledge Base Completion

Pouya Pezeshkpour, Liyan Chen, and Sameer Singh 16:48–17:06

Representing entities and relations in an embedding space is a well-studied approach for machine learning on relational data. Existing approaches, however, primarily focus on simple link structure between a finite set of entities, ignoring the variety of data types that are often used in knowledge bases, such as text, images, and numerical values. In this paper, we propose multimodal knowledge base embeddings (MKBE) that use different neural encoders for this variety of observed data, and combine them with existing relational models to learn embeddings of the entities and multimodal data. Further, using these learned embedings and different neural decoders, we introduce a novel multimodal imputation model to generate missing multimodal values, like text and images, from information in the knowledge base. We enrich existing relational datasets to create two novel benchmarks that contain additional information such as textual descriptions and images of the original entities. We demonstrate that our models utilize this additional information effectively to provide more accurate link prediction, achieving state-of-the-art results with a considerable gap of 5-7$%$ over existing methods. Further, we evaluate the quality of our generated multimodal values via a user study.

Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi 17:06–17:24

We introduce a multi-task setup of identifying entities, relations, and coreference clusters in scientific articles. We create SciERC, a dataset that includes annotations for all three tasks and develop a unified framework called SciIE with shared span representations. The multi-task setup reduces cascading errors between tasks and leverages cross-sentence relations through coreference links. Experiments show that our multi-task model outperforms previous models in scientific information extraction without using any domain-specific features. We further show that the framework supports construction of a scientific knowledge graph, which we use to analyze information in scientific literature.

Playing 20 Question Game with Policy-Based Reinforcement Learning

Huang Hu, Xianchao Wu, Bingfeng Luo, Chongyang Tao, Can Xu, wei wu wei, and Zhan Chen
17:24–17:42

The 20 Questions (Q20) game is a well known game which encourages deductive reasoning and creativity. In the game, the answerer first thinks of an object such as a famous person or a kind of animal. Then the questioner tries to guess the object by asking 20 questions. In a Q20 game system, the user is considered as the answerer while the system itself acts as the questioner which requires a good strategy of question selection to figure out the correct object and win the game. However, the optimal policy of question selection is hard to be derived due to the complexity and volatility of the game environment. In this paper, we propose a novel policy-based Reinforcement Learning (RL) method, which enables the questioner agent to learn the optimal policy of question selection through continuous interactions with users. To facilitate training, we also propose to use a reward network to estimate the more informative reward. Compared to previous methods, our RL method is robust to noisy answers and does not rely on the Knowledge Base of objects. Experimental results show that our RL method clearly outperforms an entropy-based engineering system and has competitive performance in a noisy-free simulation environment.

Multi-Hop Knowledge Graph Reasoning with Reward Shaping
Xi Victoria Lin, Richard Socher, and Caiming Xiong 17:42–18:00

Multi-hop reasoning is an effective approach for query answering (QA) over incomplete knowledge graphs (KGs). The problem can be formulated in a reinforcement learning (RL) setup, where a policy-based agent sequentially extends its inference path until it reaches a target. However, in an incomplete KG environment, the agent receives low-quality rewards corrupted by false negatives in the training data, which harms generalization at test time. Furthermore, since no golden action sequence is used for training, the agent can be misled by spurious search trajectories that incidentally lead to the correct answer. We propose two modeling advances to address both issues: (1) we reduce the impact of false negative supervision by adopting a pretrained one-hop embedding model to estimate the reward of unobserved facts; (2) we counter the sensitivity to spurious paths of on-policy RL by forcing the agent to explore a diverse set of paths using randomly generated edge masks. Our approach significantly improves over existing path-based KGQA models on several benchmark datasets and is comparable or better than embedding-based models.

Session 8D: Morphology / Phonology

Hall 100 Chair: *chairname*

Neural Transductive Learning and Beyond: Morphological Generation in the Minimal-Resource Setting

Katharina Kann and Hinrich Schütze 16:30–16:48

Neural state-of-the-art sequence-to-sequence (seq2seq) models often do not perform well for small training sets. We address paradigm completion, the morphological task of, given a partial paradigm, generating all missing forms. We propose two new methods for the minimal- resource setting: (i) Paradigm transduction: Since we assume only few paradigms available for training, neural seq2seq models are able to capture relationships between paradigm cells, but are tied to the idiosyncracies of the training set. Paradigm transduction mitigates this problem by exploiting the input subset of inflected forms at test time. (ii) Source selec- tion with high precision (SHIP): Multi-source models which learn to automatically select one or multiple sources to predict a target inflection do not perform well in the minimal-resource setting. SHIP is an alternative to identify a reliable source if training data is limited. On a 52-language benchmark dataset, we outperform the previous state of the art by up to 9.71% absolute accuracy.

Implicational Universals in Stochastic Constraint-Based Phonology

Giorgio Magri 16:48–17:06

This paper focuses on the most basic implicational universals in phonological theory, called T-orders after Anttila and Andrus (2006). It shows that the T-orders predicted by stochastic (and partial order) Optimality Theory coincide with those predicted by categorical OT. Analogously, the T-orders predicted by stochastic Harmonic Grammar coincide with those predicted by categorical HG. In other words, these stochastic constraint-based frameworks do not tamper with the typological structure induced by the original categorical frameworks.

Explaining Character-Aware Neural Networks for Word-Level Prediction: Do They Discover Linguistic Rules?

Fréderic Godin, Kris Demuynck, Joni Dambre, Wesley De Neve, and Thomas Demeester 17:06–17:24

Character-level features are currently used in different neural network-based natural language processing algorithms. However, little is known about the character-level patterns those models learn. Moreover, models are often compared only quantitatively while a qualitative analysis is missing. In this paper, we investigate which character-level patterns neural networks learn and if those patterns coincide with manually-defined word segmentations and annotations. To that end, we extend the contextual decomposition technique (Murdoch et al. 2018) to convolutional neural networks which allows us to compare convolutional neural networks and bidirectional long short-term memory networks. We evaluate and compare these models for the task of morphological tagging on three morphologically different languages and show that these models implicitly discover understandable linguistic rules.

Adapting Word Embeddings to New Languages with Morphological and Phonological Subword Representations

Aditi Chaudhary, Chunting Zhou, Lori Levin, Graham Neubig, David R. Mortensen, and Jaime Carbonell 17:24–17:42

Much work in Natural Language Processing (NLP) has been for resource-rich languages, making generalization to new, less-resourced languages challenging. We present two approaches for improving generalization to low-resourced languages by adapting continuous word representations using linguistically motivated subword units: phonemes, morphemes and graphemes. Our method requires neither parallel corpora nor bilingual dictionaries and provides a significant gain in performance over previous methods relying on these resources. We demonstrate the effectiveness of our approaches on Named Entity Recognition for four languages, namely Uyghur, Turkish, Bengali and Hindi, of which Uyghur and Bengali are low resource languages, and also perform experiments on Machine Translation. Exploiting subwords with transfer learning gives us a boost of +15.2 NER F1 for Uyghur and +9.7 F1 for Bengali. We also show improvements in the monolingual setting where we achieve (avg.) +3 F1 and (avg.) +1.35 BLEU.

TACL-1420

TACL-1420 17:42–18:00

Can advances in NLP help guide linguistic theory? We examine the role of neural networks, the current state of the art in many common NLP tasks. In 1986, Rumelhart and McClelland famously introduced a neural architecture that learned to transduce English verb stems to their past tense forms. Shortly thereafter, Pinker and Prince (1988) presented a comprehensive rebuttal of many of Rumelhart and McClelland9́2s claims. Much of the force of their attack centered on the empirical inadequacy of the Rumelhart and McClelland (1986) model. Today, however, that model is severely outmoded. We show that the recurrent neural networks in modern NLP systems obviate most of Pinker and Prince9́2s criticisms. We suggest that the empirical performance of recurrent neural networks warrants a reexamination of their utility in linguistic and cognitive modeling.

Session 8E: Sentiment, Social Applications, Multimodal Semantics, Discourse

16:30–18:00
Hall 100
Chair: *chairname*

A Computational Exploration of Exaggeration

Enrica Troiano, Carlo Strapparava, Gözde Özbal, and Serra Sinem Tekiroglu

Several NLP studies address the problem of figurative language, but among non-literal phenomena, they have neglected exaggeration. This paper presents a first computational approach to this figure of speech. We explore the possibility to automatically detect exaggerated sentences. First, we introduce HYPO, a corpus containing overstatements (or hyperboles) collected on the web and validated via crowdsourcing. Then, we evaluate a number of models trained on HYPO, and bring evidence that the task of hyperbole identification can be successfully performed based on a small set of semantic features.

Building Context-aware Clause Representations for Situation Entity Type Classification

Zeyu Dai and Ruihong Huang

Capabilities to categorize a clause based on the type of situation entity (e.g., events, states and generic statements) the clause introduces to the discourse can benefit many NLP applications. Observing that the situation entity type of a clause depends on discourse functions the clause plays in a paragraph and the interpretation of discourse functions depends heavily on paragraph-wide contexts, we propose to build context-aware clause representations for predicting situation entity types of clauses. Specifically, we propose a hierarchical recurrent neural network model to read a whole paragraph at a time and jointly learn representations for all the clauses in the paragraph by extensively modeling context influences and inter-dependencies of clauses. Experimental results show that our model achieves the state-of-the-art performance for clause-level situation entity classification on the genre-rich MASC+Wiki corpus, which approaches human-level performance.

Hierarchical Dirichlet Gaussian Marked Hawkes Process for Narrative Reconstruction in Continuous Time Domain

Yeon Seonwoo, Alice Oh, and Sungjoon Park

In news and discussions, many articles and posts are provided without their related previous articles or posts. Hence, it is difficult to understand the context from which the articles and posts have occurred. In this paper, we propose the Hierarchical Dirichlet Gaussian Marked Hawkes process (HD-GMHP) for reconstructing the narratives and thread structures of news articles and discussion posts. HD-GMHP unifies three modeling strategies in previous research: temporal characteristics, triggering event relations, and meta information of text in news articles and discussion threads. To show the effectiveness of the model, we perform experiments in narrative reconstruction and thread reconstruction with real world datasets: articles from the New York Times and a corpus of Wikipedia conversations. The experimental results show that HD-GMHP outperforms the baselines of LDA, HDP, and HDHP for both tasks.

Investigating the Role of Argumentation in the Rhetorical Analysis of Scientific Publications with Neural Multi-Task Learning Models

Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Kai Eckert

Exponential growth in the number of scientific publications yields the need for effective automatic analysis of rhetorical aspects of scientific writing. Acknowledging the argumentative nature of scientific text, in this work we investigate the link between the argumentative structure of scientific publications and rhetorical aspects such as discourse categories or citation contexts. To this end, we (1) augment a corpus of scientific publications annotated with four layers of rhetoric annotations with argumentation annotations and (2) investigate neural multi-task learning architectures combining argument extraction with a set of rhetorical classification tasks. By coupling rhetorical classifiers with the extraction of argumentative components in a joint multi-task learning setting, we obtain significant performance gains for different rhetorical analysis tasks.

Neural Ranking Models for Temporal Dependency Structure Parsing

Yuchen Zhang and Nianwen Xue

We design and build the first neural temporal dependency parser. It utilizes a neural ranking model with minimal feature engineering, and parses time expressions and events in a text into a temporal dependency tree structure. We evaluate our parser on two domains: news reports and narrative stories. In a parsing-only evaluation setup where gold time expressions and events are provided, our parser reaches 0.81 and

0.70 f-score on unlabeled and labeled parsing respectively, a result that is very competitive against alternative approaches. In an end-to-end evaluation setup where time expressions and events are automatically recognized, our parser beats two strong baselines on both data domains. Our experimental results and discussions shed light on the nature of temporal dependency structures in different domains and provide insights that we believe will be valuable to future research in this area.

Causal Explanation Analysis on Social Media
Youngseo Son, Nipun Bayas, and H. Andrew Schwartz

Understanding causal explanations - reasons given for happenings in one's life - has been found to be an important psychological factor linked to physical and mental health. Causal explanations are often studied through manual identification of phrases over limited samples of personal writing. Automatic identification of causal explanations in social media, while challenging in relying on contextual and sequential cues, offers a larger-scale alternative to expensive manual ratings and opens the door for new applications (e.g. studying prevailing beliefs about causes, such as climate change). Here, we explore automating causal explanation analysis, building on discourse parsing, and presenting two novel subtasks: causality detection (determining whether a causal explanation exists at all) and causal explanation identification (identifying the specific phrase that is the explanation). We achieve strong accuracies for both tasks but find different approaches best: an SVM for causality prediction (F1 = 0.791) and a hierarchy of Bidirectional LSTMs for causal explanation identification (F1 = 0.853). Finally, we explore applications of our complete pipeline (F1 = 0.868), showing demographic differences in mentions of causal explanation and that the association between a word and sentiment can change when it is used within a causal explanation.

LRMM: Learning to Recommend with Missing Modality
Cheng Wang, Mathias Niepert, and Hui Li

Multimodal learning has shown promising performance in content-based recommendation due to the auxiliary user and item information of multiple modalities such as text and images. However, the problem of incomplete and missing modality is rarely explored and most existing methods fail in learning a recommendation model with missing or corrupted modalities. In this paper, we propose LRMM, a novel framework that mitigates not only the problem of missing modalities but also more generally the cold-start problem of recommender systems. We propose modality dropout (m-drop) and a multimodal sequential autoencoder (m-auto) to learn multimodal representations for complementing and imputing missing modalities. Extensive experiments on real-world Amazon data show that LRMM achieves state-of-the-art performance on rating prediction tasks. More importantly, LRMM is more robust to previous methods in alleviating data-sparsity and the cold-start problem.

Content Explorer: Recommending Novel Entities for a Document Writer
Michal Lukasik and Richard Zens

Background research is an essential part of document writing. Search engines are great for retrieving information once we know what to look for. However, the bigger challenge is often identifying topics for further research. Automated tools could help significantly in this discovery process and increase the productivity of the writer. In this paper, we formulate the problem of recommending topics to a writer. We consider this as a supervised learning problem and run a user study to validate this approach. We propose an evaluation metric and perform an empirical comparison of state-of-the-art models for extreme multi-label classification on a large data set. We demonstrate how a simple modification of the cross-entropy loss function leads to improved results of the deep learning models.

A Genre-Aware Attention Model to Improve Likability Prediction of Books
Suraj Maharjan, Manuel Montes, Fabio A. González, and Thamar Solorio

Likability prediction of books has many uses. Readers, writers, as well as the publishing industry, can all benefit from automatic book likability prediction systems. In order to make reliable decisions, these systems need to assimilate information from different aspects of a book in a sensible way. We propose a novel multimodal neural architecture that incorporates genre supervision to assign weights to individual feature types. Our proposed method is capable of dynamically tailoring weights given to feature types based on the characteristics of each book. Our architecture achieves competitive results and even outperforms state-of-the-art for this task.

Thread Popularity Prediction and Tracking with a Permutation-invariant Model
Hou Pong Chan and Irwin King

The task of thread popularity prediction and tracking aims to recommend a few popular comments to subscribed users when a batch of new comments arrive in a discussion thread. This task has been formulated as a reinforcement learning problem, in which the reward of the agent is the sum of positive responses received by the recommended comments. In this work, we propose a novel approach to tackle this problem. First, we propose a deep neural network architecture to model the expected cumulative reward (Q-value) of a recommendation (action). Unlike the state-of-the-art approach, which treats an action as a sequence, our model uses an attention mechanism to integrate information from a set of comments. Thus, the prediction of Q-value is invariant to the permutation of the comments, which leads to a more consistent agent behavior. Second, we employ a greedy procedure to approximate the action that maximizes the predicted Q-value from a combinatorial action space. Different from the state-of-the-art approach, this procedure does not require an additional pre-trained model to generate candidate actions. Experiments on five real-world datasets show that our approach outperforms the state-of-the-art.

Modeling Inter-Aspect Relation in Aspect-Based Sentiment Analysis
Navonil Majumder, Soujanya Poria, Alexander Gelbukh, Md Shad Akhtar, Erik Cambria, and Asif Ekbal
Sentiment analysis has immense implications in e-commerce through user feedback mining. Aspect-based sentiment analysis takes this one step further by enabling businesses to extract aspect specific sentimental information. In this paper, we present a novel approach of incorporating the neighboring aspects related information into the sentiment classification of the target aspect using memory networks. We show that our method outperforms the state of the art by 1.6% on average in two distinct domains: restaurant and laptop.

Limbic: Author-Based Sentiment Aspect Modeling Regularized with Word Embeddings and Discourse Relations
Zhe Zhang and Munindar Singh

We propose Limbic, an unsupervised probabilistic model that addresses the problem of discovering aspects and sentiments and associating them with authors of opinionated texts. Limbic combines three ideas, incorporating authors, discourse relations, and word embeddings. For discourse relations, Limbic adopts a generative process regularized by a Markov Random Field. To promote words with high semantic similarity into the same topic, Limbic captures semantic regularities from word embeddings via a generalized Pólya Urn process. We demonstrate that Limbic (1) discovers aspects associated with sentiments with high lexical diversity; (2) outperforms state-of-the-art models by a substantial margin in topic cohesion and sentiment classification.

An Interpretable Neural Network with Topical Information for Relevant Emotion Ranking
Yang Yang, Deyu ZHOU, and Yulan He

Text might express or evoke multiple emotions with varying intensities. As such, it is crucial to predict and rank multiple relevant emotions by their intensities. Moreover, as emotions might be evoked by hidden topics, it is important to unveil and incorporate such topical information to understand how the emotions are evoked. We proposed a novel interpretable neural network approach for relevant emotion ranking. Specifically, motivated by transfer learning, the neural network is initialized to make the hidden layer approximate the behavior of topic models. Moreover, a novel error function is defined to optimize the whole neural network for relevant emotion ranking. Experimental results on three real-world corpora show that the proposed approach performs remarkably better than the state-of-the-art emotion detection approaches and multi-label learning methods. Moreover, the extracted emotion-associated topic words indeed represent emotion-evoking events and are in line with our common-sense knowledge.

Multi-grained Attention Network for Aspect-Level Sentiment Classification
Feifan Fan, Yansong Feng, and Dongyan Zhao

We propose a novel multi-grained attention network (MGAN) model for aspect level sentiment classification. Existing approaches mostly adopt coarse-grained attention mechanism, which may bring information loss if the aspect has multiple words or larger context. We propose a fine-grained attention mechanism, which can capture the word-level interaction between aspect and context. And then we leverage the fine-grained and coarse-grained attention mechanisms to compose the MGAN framework. Moreover, unlike previous works which train each aspect with its context separately, we design an aspect alignment loss to depict the aspect-level interactions among the aspects that have the same context. We evaluate the proposed approach on three datasets: laptop and restaurant are from SemEval 2014, and the last one is a twitter dataset. Experimental results show that the multi-grained attention network consistently outperforms the state-of-the-art methods on all three datasets. We also conduct experiments to evaluate the effectiveness of aspect alignment loss, which indicates the aspect-level interactions can bring extra useful information and further improve the performance.

Attentive Gated Lexicon Reader with Contrastive Contextual Co-Attention for Sentiment Classification
Yi Tay, Anh Tuan Luu, Siu Cheung Hui, and Jian Su

This paper proposes a new neural architecture that exploits readily available sentiment lexicon resources. The key idea is that that incorporating a word-level prior can aid in the representation learning process, eventually improving model performance. To this end, our model employs two distinctly unique components, i.e., (1) we introduce a lexicon-driven contextual attention mechanism to imbue lexicon words with long-range contextual information and (2), we introduce a contrastive co-attention mechanism that models contrasting polarities between all positive and negative words in a sentence. Via extensive experiments, we show that our approach outperforms many other neural baselines on sentiment classification tasks on multiple benchmark datasets.

Text, Visual and Acoustic are Friends! A Multi-Modal Attention Framework for Utterance-Level Sentiment Prediction
Deepanway Ghosal, Md Shad Akhtar, Dushyant Chauhan, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya

Multi-modal sentiment analysis offers various challenges, one being the effective combination of different input modalities, namely text, visual and acoustic. In this paper, we propose a recurrent neural network based multi-modal attention framework that leverages the contextual information for utterance-level sentiment prediction. The proposed approach applies attention on multi-modal multi-utterance representations and tries to learn the contributing features amongst them. We evaluate our proposed approach on two multi-modal sentiment analysis benchmark datasets, viz. CMU Multi-modal Opinion-level Sentiment Intensity (CMU-MOSI) corpus and the recently released CMU Multi-modal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) corpus. Evaluation results show the effectiveness of our proposed approach with the accuracies of 82.31% and 79.80% for the MOSI and MOSEI datasets, respectively. These are approximately 2 and 1 points performance improvement over the state-of-the-art models for the datasets.

Adaptive Semi-supervised Learning for Cross-domain Sentiment Classification
Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier

We consider the cross-domain sentiment classification problem, where a sentiment classifier is to be learned from a source domain and to be generalized to a target domain. Our approach explicitly minimizes the distance between the source and the target instances in an embedded feature space. With the difference between source and target minimized, we then exploit additional information from the target domain by consolidating the idea of semi-supervised learning, for which, we jointly employ two regularizations — entropy minimization and self-ensemble bootstrapping — to incorporate the unlabeled target data for classifier refinement. Our experimental results demonstrate that the proposed approach can better leverage unlabeled data from the target domain and achieve substantial improvements over baseline methods in various experimental settings.

ExtRA: Extracting Prominent Review Aspects from Customer Feedback
Zhiyi Luo, Shanshan Huang, Frank F. Xu, Bill Yuchen Lin, Hanyuan Shi, and Kenny Zhu

Many existing systems for analyzing and summarizing customer reviews about products or service are based on a number of prominent review aspects. Conventionally, the prominent review aspects of a product type are determined manually. This costly approach cannot scale to large and cross-domain services such as Amazon.com, Taobao.com or Yelp.com where there are a large number of product types and new products emerge almost every day. In this paper, we propose a novel framework, for extracting the most prominent aspects of a given product type from textual reviews. The proposed framework, ExtRA, extracts K most prominent aspect terms or phrases which do not overlap semantically automatically without supervision. Extensive experiments show that ExtRA is effective and achieves the state-of-the-art performance on a dataset consisting of different product types.

Cross-Lingual Cross-Platform Rumor Verification Pivoting on Multimedia Content
Weiming Wen, Songwen Su, and Zhou Yu

With the increasing popularity of smart devices, rumors with multimedia content become more and more common on social networks. The multimedia information usually makes rumors look more convincing. Therefore, finding an automatic approach to verify rumors with multimedia content is a pressing task. Previous rumor verification research only utilizes multimedia as input features. We propose not to use the multimedia content but to find external information in other news platforms pivoting on it. We introduce a new features set, cross-lingual cross-platform features that leverage the semantic similarity between the rumors and the external information. When implemented, machine learning methods utilizing such features achieved the state-of-the-art rumor verification results.

Extractive Adversarial Networks: High-Recall Explanations for Identifying Personal Attacks in Social Media Posts
Samuel Carton, Qiaozhu Mei, and Paul Resnick

We introduce an adversarial method for producing high-recall explanations of neural text classifier decisions. Building on an existing architecture for extractive explanations via hard attention, we add an adversarial layer which scans the residual of the attention for remaining predictive signal. Motivated by the important domain of detecting personal attacks in social media comments, we additionally demonstrate the importance of manually setting a semantically appropriate "default" behavior for the model by explicitly manipulating its bias term. We develop a validation set of human-annotated personal attacks to evaluate the impact of these changes.

Automatic Detection of Vague Words and Sentences in Privacy Policies
Logan Lebanoff and Fei Liu

Website privacy policies represent the single most important source of information for users to gauge how their personal data are collected, used and shared by companies. However, privacy policies are often vague and people struggle to understand the content. Their opaqueness poses a significant challenge to both users and policy regulators. In this paper, we seek to identify vague content in privacy policies. We construct the first corpus of human-annotated vague words and sentences and present empirical studies on automatic vagueness detection. In particular, we investigate context-aware and context-agnostic models for predicting vague words, and explore auxiliary-classifier generative adversarial networks for characterizing sentence vagueness. Our experimental results demonstrate the effectiveness of proposed approaches. Finally, we provide suggestions for resolving vagueness and improving the usability of privacy policies.

Multi-view Models for Political Ideology Detection of News Articles
Vivek Kulkarni, Junting Ye, Steve Skiena, and William Yang Wang

A news article's title, content and link structure often reveal its political ideology. However, most existing works on automatic political ideology detection only leverage textual cues. Drawing inspiration from recent advances in neural inference, we propose a novel attention based multi-view model to leverage cues from all of the above views to identify the ideology evinced by a news article. Our model draws on advances in representation learning in natural language processing and network science to capture cues from both textual content and the network structure of news articles. We empirically evaluate our model against a battery of baselines and show that our model outperforms state of the art by 10 percentage points F1 score.

Predicting Factuality of Reporting and Bias of News Media Sources
Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov

We present a study on predicting the factuality of reporting and bias of news media. While previous work has focused on studying the veracity of claims or documents, here we are interested in characterizing entire news media. This is an under-studied, but arguably important research problem, both in its own right and as a prior for fact-checking systems. We experiment with a large list of news websites and with a rich set of features derived from (i) a sample of articles from the target news media, (ii) its Wikipedia page, (iii) its Twitter account, (iv) the structure of its URL, and (v) information about the Web traffic it attracts. The experimental results show sizable performance gains over the baseline, and reveal the importance of each feature type.

Legal Judgment Prediction via Topological Learning
Haoxi Zhong, Guo Zhipeng, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun

Legal Judgment Prediction (LJP) aims to predict the judgment result based on the facts of a case and becomes a promising application of artificial intelligence techniques in the legal field. In real-world scenarios, legal judgment usually consists of multiple subtasks, such as the decisions of applicable law articles, charges, fines, and the term of penalty. Moreover, there exist topological dependencies among these subtasks. While most existing works only focus on a specific subtask of judgment prediction and ignore the dependencies among subtasks, we formalize the dependencies among subtasks as a Directed Acyclic Graph (DAG) and propose a topological multi-task learning framework, TopJudge, which incorporates multiple subtasks and DAG dependencies into judgment prediction. We conduct experiments on several real-world large-scale datasets of criminal cases in the civil law system. Experimental results show that our model achieves consistent and significant improvements over baselines on all judgment prediction tasks. The source code can be obtained from https://github.com/thunlp/TopJudge.

Hierarchical CVAE for Fine-Grained Hate Speech Classification
Jing Qian, Mai ElSherief, Elizabeth Belding, and William Yang Wang

Existing work on automated hate speech detection typically focuses on binary classification or on differentiating among a small set of categories. In this paper, we propose a novel method on a fine-grained hate speech classification task, which focuses on differentiating among 40 hate groups of 13 different hate group categories. We first explore the Conditional Variational Autoencoder (CVAE) as a discriminative model and then extend it to a hierarchical architecture to utilize the additional hate category information for more accurate prediction. Experimentally, we show that incorporating the hate category information for training can significantly improve the classification performance and our proposed model outperforms commonly-used discriminative models.

Residualized Factor Adaptation for Community Social Media Prediction Tasks
Mohammadzaman Zamani, H. Andrew Schwartz, Veronica Lynn, Salvatore Giorgi, and Niranjan Balasubramanian

Predictive models over social media language have shown promise in capturing community outcomes, but approaches thus far largely neglect the socio-demographic context (e.g. age, education rates, race) of the community from which the language originates. For example, it may be inaccurate to assume people in Mobile, Alabama, where the population is relatively older, will use words the same way as those from San Francisco, where the median age is younger with a higher rate of college education. In this paper, we present residualized factor adaptation, a novel approach to community prediction tasks which both (a) effectively integrates community attributes, as well as (b) adapts linguistic features to community attributes (factors). We use eleven demographic and socioeconomic attributes, and evaluate our approach over five different community-level predictive tasks, spanning health (heart disease mortality, percent fair/poor health), psychology (life satisfaction), and economics (percent housing price increase, foreclosure rate). Our evaluation shows that residualized factor adaptation significantly improves 4 out of 5 community-level outcome predictions over prior state-of-the-art for incorporating socio-demographic contexts.

Framing and Agenda-Setting in Russian News: a Computational Analysis of Intricate Political Strategies
Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov

Amidst growing concern over media manipulation, NLP attention has focused on overt strategies like censorship and "fake news". Here, we draw on two concepts from political science literature to explore subtler strategies for government media manipulation: agenda-setting (selecting what topics to cover) and framing (deciding how topics are covered). We analyze 13 years (100K articles) of the Russian newspaper Izvestia and identify a strategy of distraction: articles mention the U.S. more frequently in the month directly following an economic downturn in Russia. We introduce embedding-based methods for cross-lingually projecting English frames to Russian, and discover that these articles emphasize U.S. moral failings and threats to the U.S. Our work offers new ways to identify subtle media manipulation strategies at the intersection of agenda-setting and framing.

Identifying the narrative styles of YouTube's vloggers
Bennett Kleinberg, Maximilian Mozes, and Isabelle van der Vegt

Vlogs provide a rich public source of data in a novel setting. This paper examined the continuous sentiment styles employed in 27,333 vlogs using a dynamic intra-textual approach to sentiment analysis. Using unsupervised clustering, we identified seven distinct continuous sentiment trajectories characterized by fluctuations of sentiment throughout a vlog's narrative time. We provide a taxonomy of these seven continuous sentiment styles and found that vlogs whose sentiment builds up towards a positive ending are the most prevalent in our sample. Gender was associated with preferences for different continuous sentiment trajectories. This paper discusses the findings with respect to previous work and concludes with an outlook towards possible uses of the corpus, method and findings of this paper for related areas of research.

Native Language Identification with User Generated Content
Gili Goldin, Ella Rabinovich, and Shuly Wintner

We address the task of native language identification in the context of social media content, where authors are highly-fluent, advanced nonnative speakers (of English). Using both linguistically-motivated features and the characteristics of the social media outlet, we obtain high accuracy on this challenging task. We provide a detailed analysis of the features that sheds light on differences between native and nonnative speakers, and among nonnative speakers with different backgrounds.

Visualization of the Topic Space of Argument Search Results in args.me
Yamen Ajjour, Henning Wachsmuth, Dora Kiesel, Patrick Riehmann, Fan Fan, Giuliano Castiglia, Rosemary Adejoh, Bernd Fröhlich, and Benno Stein

In times of fake news and alternative facts, pro and con arguments on controversial topics are of increasing importance. Recently, we presented args.me as the first search engine for arguments on the web. In its initial version, args.me ranked arguments solely by their relevance to a topic queried for, making it hard to learn about the diverse topical aspects covered by the search results. To tackle this shortcoming, we integrated a visualization interface for result exploration in args.me that provides an instant overview of the main aspects in a barycentric coordinate system. This topic space is generated ad-hoc from controversial issues on Wikipedia and argument-specific LDA models. In two case studies, we demonstrate how individual arguments can be found easily through interactions with the visualization, such as highlighting and filtering.

A Multilingual Information Extraction Pipeline for Investigative Journalism
Gregor Wiedemann, Seid Muhie Yimam, and Chris Biemann

We introduce an advanced information extraction pipeline to automatically process very large collections of unstructured textual data for the purpose of investigative journalism. The pipeline serves as a new input processor for the upcoming major release of our New/s/leak 2.0 software, which we develop in cooperation with a large German news organization. The use case is that journalists receive a large collection of files up to several Gigabytes containing unknown contents. Collections may originate either from official disclosures of documents, e.g. Freedom of Information Act requests, or unofficial data leaks. Our software prepares a visually-aided exploration of the collection to quickly learn about potential stories contained in the data. It is based on the automatic extraction of entities and their co-occurrence in documents. In contrast to comparable projects, we focus on the following three major requirements particularly serving the use case of investigative journalism in cross-border collaborations: 1) composition of multiple state-of-the-art NLP tools for entity extraction, 2) support of multi-lingual document sets up to 40 languages, 3) fast and easy-to-use extraction of full-text, metadata and entities from various file formats.

When science journalism meets artificial intelligence : An interactive demonstration
Raghuram Vadapalli, Bakhtiyar Syed, Nishant Prabhu, Balaji Vasan Srinivasan, and Vasudeva Varma

We present an online interactive tool that generates titles of blog titles and thus take the first step toward automating science journalism. Science journalism aims to transform jargon-laden scientific articles into a form that the common reader can comprehend while ensuring that the underlying meaning of the article is retained. In this work, we present a tool, which, given the title and abstract of a research paper will generate a blog title by mimicking a human science journalist. The tool makes use of a model trained on a corpus of 87,328 pairs of research papers and their corresponding blogs, built from two science news aggregators. The architecture of the model is a two-stage mechanism which generates blog titles. Evaluation using standard metrics indicate the viability of the proposed system.

SIGDAT Business Meeting

Date: Saturday, November 3, 2018
Time: 12:40–13:40
Venue: Gold Hall

Chair: *Pascale Fung*

All attendees are encouraged to participate in the business meeting.

Social Event

Saturday, November 3, 2018, 18:00–22:00

Royal Museums of Fine Arts of Belgium

On the evening of Saturday, November 3rd, the EMNLP 2018 social event will take place at the Royal Museums of Fine Arts of Belgium. Four museums, housed in a single building, will welcome the EMNLP delegates with their prestigious collection of 20,000 works of art. The Museums¡¦ collections trace the history of the visual arts – painting, sculpture and drawing – from the 15th to the 21st century.

7

Main Conference: Sunday, November 4

Overview

Session 9

09:00 – 10:30	Machine Translation II *Gold Hall*	Sentiment I *Copper Hall*	Machine Learning III *Silver Hall*	Semantics VI *Hall 100*

09:00 – 10:30 **[Posters and Demos]: Generation, Dialog, Summarization; Vision and Language II** *Grand Hall 2*

10:30 – 11:00 **Coffee Break** *Multiple levels*

Session 10

11:00 – 12:30	Question Answering III *Gold Hall*	Machine Translation III *Copper Hall*	Discourse *Silver Hall*	Evolution / Sociolinguistics *Hall 100*

11:00 – 12:30 **[Posters and Demos]: Machine Learning** *Grand Hall 2*

12:30 – 13:45 **Lunch**

13:00 – 13:45 **Business Meeting** *Gold Hall*

Session 11

13:45 – 14:45	Analyzing Models *Gold Hall*	Sentiment II *Copper Hall*	Machine Translation IV *Silver Hall*	QA / Knowledge Graphs *Hall 100*

13:45 – 14:45 **[Posters and Demos]: Short Posters IV** *Grand Hall 2*

14:45 – 15:00 **Mini-Break** *Multiple levels*

15:00 – 16:00 **Keynote III: Johan Bos "The Moment of Meaning and the Future of Computational Semantics" (Gold Hall)** *Gold Hall*

16:00 – 16:30 **Coffee Break** *Multiple levels*

16:30 – 18:00 **Best Paper Awards and Closing (Gold Hall)** *Gold Hall*

Invited Talk: Gideon Mann

Understanding the News that Moves Markets

Sunday, November 4, 2018, 9:00–10:00

Gold Hall

Abstract: Since the dawn of human civilization, finance and language technology have been connected. However, only recently have advances in statistical language understanding, and an ever-increasing thirst for market advantage, led to the widespread application of natural language technology across the global capital markets. This talk will review the ways in which language technology is enabling market participants to quickly understand and respond to major world events and breaking business news. It will outline the state of the art in applications of NLP to finance and highlight open problems that are being addressed by emerging research.

Biography: Gideon Mann is the Head of Data Science at Bloomberg L.P., where he guides the strategic direction for machine learning, natural language processing (NLP) and search across the company. He is part of the leadership team for the Office of the CTO. He served as a founding member of both the Data for Good Exchange (D4GX), an annual conference on data science applications for social good, and the Shift Commission on Work, Workers and Technology. He has also been active in academic research in fact extraction, weakly-supervised learning, and distributed optimization. Recently, he has also been interested in applications of machine learning to problems in software engineering. From 2007 to 2014, he worked at Google Research in New York City, and his team built core machine learning libraries, released the Google Prediction API, and developed Colaboratory. Mann graduated Brown University in 1999 and received a Ph.D. from The Johns Hopkins University in 2006.

Session 9 Overview – Sunday, November 4, 2018

Oral tracks

	Track A *Machine Translation II* Gold Hall	**Track B** *Sentiment I* Copper Hall	**Track C** *Machine Learning III* Silver Hall	**Track D** *Semantics VI* Hall 100
09:00	Beyond Error Propagation in Neural Machine Translation: Characteristics of Language Also Matter *Wu, Tan, He, Tian, QIN, Lai, and Liu*	Sentiment Classification towards Question-Answering with Hierarchical Matching Network *Shen, Sun, Wang, Kang, Li, Liu, Si, Zhang, and Zhou*	Noise Contrastive Estimation for Conditional Models: Consistency and Statistical Efficiency *Ma and Collins*	A Word-Complexity Lexicon and A Neural Readability Ranking Model for Lexical Simplification *Maddela and Xu*
09:18	A Study of Reinforcement Learning for Neural Machine Translation *Wu, Tian, Qin, Lai, and Liu*	Cross-topic Argument Mining from Heterogeneous Sources *Stab, Miller, Schiller, Rai, and Gurevych*	CaLcs: Continuously Approximating Longest Common Subsequence for Sequence Level Optimization *Yavuz, Chiu, Nguyen, and Wu*	Learning Latent Semantic Annotations for Grounding Natural Language to Structured Data *Qin, Yao, Wang, Wang, and Lin*
09:36	Meta-Learning for Low-Resource Neural Machine Translation *Gu, Wang, Chen, Li, and Cho*	Summarizing Opinions: Aspect Extraction Meets Sentiment Prediction and They Are Both Weakly Supervised *Angelidis and Lapata*	Pathologies of Neural Models Make Interpretation Difficult *Feng, Wallace, Grissom II, Iyyer, Rodriguez, and Boyd-Graber*	Syntactic Scaffolds for Semantic Structures *Swayamdipta, Thomson, Lee, Zettlemoyer, Dyer, and Smith*
09:54	Unsupervised Statistical Machine Translation *Artetxe, Labaka, and Agirre*	CARER: Contextualized Affect Representations for Emotion Recognition *Saravia, Liu, Huang, Wu, and Chen*	Phrase-level Self-Attention Networks for Universal Sentence Encoding *Wu, WANG, Liu, and Ma*	Hierarchical Quantized Representations for Script Generation *Weber, Shekhar, Balasubramanian, and Chambers*
10:12	A Visual Attention Grounding Neural Model for Multimodal Machine Translation *Zhou, Cheng, Lee, and Yu*	**TACL-1413** **TACL-1413**	BanditSum: Extractive Summarization as a Contextual Bandit *Dong, Shen, Crawford, Hoof, and Cheung*	Semantic Role Labeling for Learner Chinese: the Importance of Syntactic Parsing and the L2-L1 Parallel Data *Lin, Duan, Zhao, Sun, and Wan*

Poster tracks

09:00–10:30

Track D: *Generation, Dialog, Summarization; Vision and Language II* Hall 100

Track D: *Generation, Dialog, Summarization; Vision and Language II*

Parallel Session 9

Session 9A: Machine Translation II

Gold Hall Chair: *chairname*

Beyond Error Propagation in Neural Machine Translation: Characteristics of Language Also Matter

Lijun Wu, Xu Tan, Di He, Fei Tian, Tao QIN, Jianhuang Lai, and Tie-Yan Liu 09:00–09:18

Neural machine translation usually adopts autoregressive models and suffers from exposure bias as well as the consequent error propagation problem. Many previous works have discussed the relationship between error propagation and the *accuracy drop* (i.e., the left part of the translated sentence is often better than its right part in left-to-right decoding models) problem. In this paper, we conduct a series of analyses to deeply understand this problem and get several interesting findings. (1) The role of error propagation on accuracy drop is overstated in the literature, although it indeed contributes to the accuracy drop problem. (2) Characteristics of a language play a more important role in causing the accuracy drop: the left part of the translation result in a right-branching language (e.g., English) is more likely to be more accurate than its right part, while the right part is more accurate for a left-branching language (e.g., Japanese). Our discoveries are confirmed on different model structures including Transformer and RNN, and in other sequence generation tasks such as text summarization.

A Study of Reinforcement Learning for Neural Machine Translation

Lijun Wu, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu 09:18–09:36

Recent studies have shown that reinforcement learning (RL) is an effective approach for improving the performance of neural machine translation (NMT) system. However, due to its instability, successfully RL training is challenging, especially in real-world systems where deep models and large datasets are leveraged. In this paper, taking several large-scale translation tasks as testbeds, we conduct a systematic study on how to train better NMT models using reinforcement learning. We provide a comprehensive comparison of several important factors (e.g., baseline reward, reward shaping) in RL training. Furthermore, to fill in the gap that it remains unclear whether RL is still beneficial when monolingual data is used, we propose a new method to leverage RL to further boost the performance of NMT systems trained with source/target monolingual data. By integrating all our findings, we obtain competitive results on WMT14 English- German, WMT17 English-Chinese, and WMT17 Chinese-English translation tasks, especially setting a state-of-the-art performance on WMT17 Chinese-English translation task.

Meta-Learning for Low-Resource Neural Machine Translation

Jiatao Gu, Yong Wang, Yun Chen, Victor O. K. Li, and Kyunghyun Cho 09:36–09:54

In this paper, we propose to extend the recently introduced model-agnostic meta-learning algorithm (MAML, Finn, et al., 2017) for low-resource neural machine translation (NMT). We frame low-resource translation as a meta-learning problem where we learn to adapt to low-resource languages based on multilingual high-resource language tasks. We use the universal lexical representation (Gu et al., 2018b) to overcome the input-output mismatch across different languages. We evaluate the proposed meta-learning strategy using eighteen European languages (Bg, Cs, Da, De, El, Es, Et, Fr, Hu, It, Lt, Nl, Pl, Pt, Sk, Sl, Sv and Ru) as source tasks and five diverse languages (Ro,Lv, Fi, Tr and Ko) as target tasks. We show that the proposed approach significantly outperforms the multilingual, transfer learning based approach (Zoph et al., 2016) and enables us to train a competitive NMT system with only a fraction of training examples. For instance, the proposed approach can achieve as high as 22.04 BLEU on Romanian-English WMT'16 by seeing only 16,000 translated words ($\sim$600 parallel sentences)

Unsupervised Statistical Machine Translation

Mikel Artetxe, Gorka Labaka, and Eneko Agirre 09:54–10:12

While modern machine translation has relied on large parallel corpora, a recent line of work has managed to train Neural Machine Translation (NMT) systems from monolingual corpora only (Artetxe et al., 2018c; Lample et al., 2018). Despite the potential of this approach for low-resource settings, existing systems are far behind their supervised counterparts, limiting their practical interest. In this paper, we propose an alternative approach based on phrase-based Statistical Machine Translation (SMT) that significantly closes the gap with supervised systems. Our method profits from the modular architecture of SMT: we first induce a phrase table from monolingual corpora through cross-lingual embedding mappings, combine it with an n-gram language model, and fine-tune hyperparameters through an unsupervised

MERT variant. In addition, iterative backtranslation improves results further, yielding, for instance, 14.08 and 26.22 BLEU points in WMT 2014 English-German and English-French, respectively, an improvement of more than 7-10 BLEU points over previous unsupervised systems, and closing the gap with supervised SMT (Moses trained on Europarl) down to 2-5 BLEU points. Our implementation is available at https://github.com/artetxem/monoses.

A Visual Attention Grounding Neural Model for Multimodal Machine Translation
Mingyang Zhou, Runxiang Cheng, Yong Jae Lee, and Zhou Yu 10:12–10:30

We introduce a novel multimodal machine translation model that utilizes parallel visual and textual information. Our model jointly optimizes the learning of a shared visual-language embedding and a translator. The model leverages a visual attention grounding mechanism that links the visual semantics with the corresponding textual semantics. Our approach achieves competitive state-of-the-art results on the Multi30K and the Ambiguous COCO datasets. We also collected a new multilingual multimodal product description dataset to simulate a real-world international online shopping scenario. On this dataset, our visual attention grounding model outperforms other methods by a large margin.

Session 9B: Sentiment I

Copper Hall Chair: *chairname*

Sentiment Classification towards Question-Answering with Hierarchical Matching Network

Chenlin Shen, Changlong Sun, Jingjing Wang, Yangyang Kang, Shoushan Li, Xiaozhong Liu, Luo Si, Min Zhang, and Guodong Zhou 09:00–09:18

In an e-commerce environment, user-oriented question-answering (QA) text pair could carry rich sentiment information. In this study, we propose a novel task/method to address QA sentiment analysis. In particular, we create a high-quality annotated corpus with specially-designed annotation guidelines for QA-style sentiment classification. On the basis, we propose a three-stage hierarchical matching network to explore deep sentiment information in a QA text pair. First, we segment both the question and answer text into sentences and construct a number of [Q-sentence, A-sentence] units in each QA text pair. Then, by leveraging a QA bidirectional matching layer, the proposed approach can learn the matching vectors of each [Q-sentence, A-sentence] unit. Finally, we characterize the importance of the generated matching vectors via a self-matching attention layer. Experimental results, comparing with a number of state-of-the-art baselines, demonstrate the impressive effectiveness of the proposed approach for QA-style sentiment classification.

Cross-topic Argument Mining from Heterogeneous Sources

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych 09:18–09:36

Argument mining is a core technology for automating argument search in large document collections. Despite its usefulness for this task, most current approaches are designed for use only with specific text types and fall short when applied to heterogeneous texts. In this paper, we propose a new sentential annotation scheme that is reliably applicable by crowd workers to arbitrary Web texts. We source annotations for over 25,000 instances covering eight controversial topics. We show that integrating topic information into bidirectional long short-term memory networks outperforms vanilla BiLSTMs by more than 3 percentage points in F1 in two- and three-label cross-topic settings. We also show that these results can be further improved by leveraging additional data for topic relevance using multi-task learning.

Summarizing Opinions: Aspect Extraction Meets Sentiment Prediction and They Are Both Weakly Supervised

Stefanos Angelidis and Mirella Lapata 09:36–09:54

We present a neural framework for opinion summarization from online product reviews which is knowledge-lean and only requires light supervision (e.g., in the form of product domain labels and user-provided ratings). Our method combines two weakly supervised components to identify salient opinions and form extractive summaries from multiple reviews: an aspect extractor trained under a multi-task objective, and a sentiment predictor based on multiple instance learning. We introduce an opinion summarization dataset that includes a training set of product reviews from six diverse domains and human-annotated development and test sets with gold standard aspect annotations, salience labels, and opinion summaries. Automatic evaluation shows significant improvements over baselines, and a large-scale study indicates that our opinion summaries are preferred by human judges according to multiple criteria.

CARER: Contextualized Affect Representations for Emotion Recognition

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen 09:54–10:12

Emotions are expressed in nuanced ways, which varies by collective or individual experiences, knowledge, and beliefs. Therefore, to understand emotion, as conveyed through text, a robust mechanism capable of capturing and modeling different linguistic nuances and phenomena is needed. We propose a semi-supervised, graph-based algorithm to produce rich structural descriptors which serve as the building blocks for constructing contextualized affect representations from text. The pattern-based representations are further enriched with word embeddings and evaluated through several emotion recognition tasks. Our experimental results demonstrate that the proposed method outperforms state-of-the-art techniques on emotion recognition tasks.

TACL-1413

TACL-1413 10:12–10:30

In recent years great success has been achieved in sentiment classification for English, thanks in part to the availability of copious annotated resources. Unfortunately, most languages do not enjoy such an abundance of labeled data. To tackle the sentiment classification problem, in low-resource languages without adequate annotated data, we propose an Adversarial Deep Averaging Network (ADAN) to transfer the knowledge learned from labeled data on a resource-rich source language to low-resource languages where only unlabeled data exists. ADAN has two discriminative branches: a sentiment classifier and an adversarial language discriminator. Both branches take input from a shared feature extractor to learn hidden representations that are simultaneously indicative for the classification task and invariant across languages. á0Experiments on Chinese and Arabic sentiment classification demonstrate that ADAN significantly outperforms state-of-the-art systems.

Session 9C: Machine Learning III

Silver Hall Chair: *chairname*

Noise Contrastive Estimation for Conditional Models: Consistency and Statistical Efficiency

Zhuang Ma and Michael Collins 09:00–09:18

Noise Contrastive Estimation (NCE) is a powerful parameter estimation method for log-linear models, which avoids calculation of the partition function or its derivatives at each training step, a computationally demanding step in many cases. It is closely related to negative sampling methods, now widely used in NLP. This paper considers NCE-based estimation of conditional models. Conditional models are frequently encountered in practice; however there has not been a rigorous theoretical analysis of NCE in this setting, and we will argue there are subtle but important questions when generalizing NCE to the conditional case. In particular, we analyze two variants of NCE for conditional models: one based on a classification objective, the other based on a ranking objective. We show that the ranking-based variant of NCE gives consistent parameter estimates under weaker assumptions than the classification-based method; we analyze the statistical efficiency of the ranking-based and classification-based variants of NCE; finally we describe experiments on synthetic data and language modeling showing the effectiveness and tradeoffs of both methods.

CaLcs: Continuously Approximating Longest Common Subsequence for Sequence Level Optimization

Semih Yavuz, Chung-Cheng Chiu, Patrick Nguyen, and Yonghui Wu 09:18–09:36

Maximum-likelihood estimation (MLE) is one of the most widely used approaches for training structured prediction models for text-generation based natural language processing applications. However, besides exposure bias, models trained with MLE suffer from wrong objective problem where they are trained to maximize the word-level correct next step prediction, but are evaluated with respect to sequence-level discrete metrics such as ROUGE and BLEU. Several variants of policy-gradient methods address some of these problems by optimizing for final discrete evaluation metrics and showing improvements over MLE training for downstream tasks like text summarization and machine translation. However, policy-gradient methods suffers from high sample variance, making the training process very difficult and unstable. In this paper, we present an alternative direction towards mitigating this problem by introducing a new objective (CaLcs) based on a differentiable surrogate of longest common subsequence (LCS) measure that captures sequence-level structure similarity. Experimental results on abstractive summarization and machine translation validate the effectiveness of the proposed approach.

Pathologies of Neural Models Make Interpretation Difficult

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber
09:36–09:54

One way to interpret neural model predictions is to highlight the most important input features—for example, a heatmap visualization over the words in an input sentence. In existing interpretation methods for NLP, a word's importance is determined by either input perturbation—measuring the decrease in model confidence when that word is removed—or by the gradient with respect to that word. To understand the limitations of these methods, we use input reduction, which iteratively removes the least important word from the input. This exposes pathological behaviors of neural models: the remaining words appear nonsensical to humans and are not the ones determined as important by interpretation methods. As we confirm with human experiments, the reduced examples lack information to support the prediction of any label, but models still make the same predictions with high confidence. To explain these counterintuitive results, we draw connections to adversarial examples and confidence calibration: pathological behaviors reveal difficulties in interpreting neural models trained with maximum likelihood. To mitigate their deficiencies, we fine-tune the models by encouraging high entropy outputs on reduced examples. Fine-tuned models become more interpretable under input reduction, without accuracy loss on regular examples.

Phrase-level Self-Attention Networks for Universal Sentence Encoding

Wei Wu, Houfeng WANG, Tianyu Liu, and Shuming Ma 09:54–10:12

Universal sentence encoding is a hot topic in recent NLP research. Attention mechanism has been an integral part in many sentence encoding models, allowing the models to capture context dependencies regardless of the distance between the elements in the sequence. Fully attention-based models have recently attracted enormous interest due to their highly parallelizable computation and significantly less training time. However, the memory consumption of their models grows quadratically with the sentence length, and the syntactic information is neglected. To this end, we propose Phrase-level Self-Attention Networks

(PSAN) that perform self-attention across words inside a phrase to capture context dependencies at the phrase level, and use the gated memory updating mechanism to refine each word's representation hierarchically with longer-term context dependencies captured in a larger phrase. As a result, the memory consumption can be reduced because the self-attention is performed at the phrase level instead of the sentence level. At the same time, syntactic information can be easily integrated in the model. Experiment results show that PSAN can achieve the state-of-the-art performance across a plethora of NLP tasks including binary and multi-class classification, natural language inference and sentence similarity.

BanditSum: Extractive Summarization as a Contextual Bandit
Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung 10:12–10:30

In this work, we propose a novel method for training neural networks to perform single-document extractive summarization without heuristically-generated extractive labels. We call our approach BanditSum as it treats extractive summarization as a contextual bandit (CB) problem, where the model receives a document to summarize (the context), and chooses a sequence of sentences to include in the summary (the action). A policy gradient reinforcement learning algorithm is used to train the model to select sequences of sentences that maximize ROUGE score. We perform a series of experiments demonstrating that BanditSum is able to achieve ROUGE scores that are better than or comparable to the state-of-the-art for extractive summarization, and converges using significantly fewer update steps than competing approaches. In addition, we show empirically that BanditSum performs significantly better than competing approaches when good summary sentences appear late in the source document.

Session 9D: Semantics VI

Hall 100 Chair: *chairname*

A Word-Complexity Lexicon and A Neural Readability Ranking Model for Lexical Simplification

Mounica Maddela and Wei Xu 09:00–09:18

Current lexical simplification approaches rely heavily on heuristics and corpus level features that do not always align with human judgment. We create a human-rated word-complexity lexicon of 15,000 English words and propose a novel neural readability ranking model with a Gaussian-based feature vectorization layer that utilizes these human ratings to measure the complexity of any given word or phrase. Our model performs better than the state-of-the-art systems for different lexical simplification tasks and evaluation datasets. Additionally, we also produce SimplePPDB++, a lexical resource of over 10 million simplifying paraphrase rules, by applying our model to the Paraphrase Database (PPDB).

Learning Latent Semantic Annotations for Grounding Natural Language to Structured Data

Guanghui Qin, Jin-Ge Yao, Xuening Wang, Jinpeng Wang, and Chin-Yew Lin 09:18–09:36

Previous work on grounded language learning did not fully capture the semantics underlying the correspondences between structured world state representations and texts, especially those between numerical values and lexical terms. In this paper, we attempt at learning explicit latent semantic annotations from paired structured tables and texts, establishing correspondences between various types of values and texts. We model the joint probability of data fields, texts, phrasal spans, and latent annotations with an adapted semi-hidden Markov model, and impose a soft statistical constraint to further improve the performance. As a by-product, we leverage the induced annotations to extract templates for language generation. Experimental results suggest the feasibility of the setting in this study, as well as the effectiveness of our proposed framework.

Syntactic Scaffolds for Semantic Structures

Swabha Swayamdipta, Sam Thomson, Kenton Lee, Luke Zettlemoyer, Chris Dyer, and Noah A. Smith
09:36–09:54

We introduce the syntactic scaffold, an approach to incorporating syntactic information into semantic tasks. Syntactic scaffolds avoid expensive syntactic processing at runtime, only making use of a treebank during training, through a multitask objective. We improve over strong baselines on PropBank semantics, frame semantics, and coreference resolution, achieving competitive performance on all three tasks.

Hierarchical Quantized Representations for Script Generation

Noah Weber, Leena Shekhar, Niranjan Balasubramanian, and Nate Chambers 09:54–10:12

Scripts define knowledge about how everyday scenarios (such as going to a restaurant) are expected to unfold. One of the challenges to learning scripts is the hierarchical nature of the knowledge. For example, a suspect arrested might plead innocent or guilty, and a very different track of events is then expected to happen. To capture this type of information, we propose an autoencoder model with a latent space defined by a hierarchy of categorical variables. We utilize a recently proposed vector quantization based approach, which allows continuous embeddings to be associated with each latent variable value. This permits the decoder to softly decide what portions of the latent hierarchy to condition on by attending over the value embeddings for a given setting. Our model effectively encodes and generates scripts, outperforming a recent language modeling-based method on several standard tasks, and allowing the autoencoder model to achieve substantially lower perplexity scores compared to the previous language modeling-based method.

Semantic Role Labeling for Learner Chinese: the Importance of Syntactic Parsing and the L2-L1 Parallel Data

Zi Lin, Yuguang Duan, Yuanyuan Zhao, Weiwei Sun, and Xiaojun Wan 10:12–10:30

This paper studies semantic parsing for interlanguage (L2), taking semantic role labeling (SRL) as a case task and learner Chinese as a case language. We first manually annotate the semantic roles for a set of learner texts to derive a gold standard for automatic SRL. Based on the new data, we then evaluate three off-the-shelf SRL systems, i.e., the PCFGLA-parser-based, neural-parser-based and neural-syntax-agnostic systems, to gauge how successful SRL for learner Chinese can be. We find two non-obvious facts: 1) the L1-sentence-trained systems performs rather badly on the L2 data; 2) the performance drop from the L1 data to the L2 data of the two parser-based systems is much smaller, indicating the importance of syntactic parsing in SRL for interlanguages. Finally, the paper introduces a new agreement-based model to explore the semantic coherency information in the large-scale L2-L1 parallel data. We then show such information is very effective to enhance SRL for learner texts. Our model achieves an F-score of 72.06, which is a 2.02 point improvement over the best baseline.

Session 9E: Generation, Dialog, Summarization; Vision and Language II

09:00–10:30

Hall 100 Chair: *chairname*

A Teacher-Student Framework for Maintainable Dialog Manager

Weikang Wang, Jiajun Zhang, Han Zhang, Mei-Yuh Hwang, Chengqing Zong, and Zhifei Li

Reinforcement learning (RL) is an attractive solution for task-oriented dialog systems. However, extending RL-based systems to handle new intents and slots requires a system redesign. The high maintenance cost makes it difficult to apply RL methods to practical systems on a large scale. To address this issue, we propose a practical teacher-student framework to extend RL-based dialog systems without retraining from scratch. Specifically, the "student" is an extended dialog manager based on a new ontology, and the "teacher" is existing resources used for guiding the learning process of the "student". By specifying constraints held in the new dialog manager, we transfer knowledge of the "teacher" to the "student" without additional resources. Experiments show that the performance of the extended system is comparable to the system trained from scratch. More importantly, the proposed framework makes no assumption about the unsupported intents and slots, which makes it possible to improve RL-based systems incrementally.

Discriminative Deep Dyna-Q: Robust Planning for Dialogue Policy Learning

Shang-Yu Su, Xiujun Li, Jianfeng Gao, Jingjing Liu, and Yun-Nung Chen

This paper presents a Discriminative Deep Dyna-Q (D3Q) approach to improving the effectiveness and robustness of Deep Dyna-Q (DDQ), a recently proposed framework that extends the Dyna-Q algorithm to integrate planning for task-completion dialogue policy learning. To obviate DDQ's high dependency on the quality of simulated experiences, we incorporate an RNN-based discriminator in D3Q to differentiate simulated experience from real user experience in order to control the quality of training data. Experiments show that D3Q significantly outperforms DDQ by controlling the quality of simulated experience used for planning. The effectiveness and robustness of D3Q is further demonstrated in a domain extension setting, where the agent's capability of adapting to a changing environment is tested.

A Self-Attentive Model with Gate Mechanism for Spoken Language Understanding

Changliang Li, Liang Li, and Ji Qi

Spoken Language Understanding (SLU), which typically involves intent determination and slot filling, is a core component of spoken dialogue systems. Joint learning has shown to be effective in SLU given that slot tags and intents are supposed to share knowledge with each other. However, most existing joint learning methods only consider joint learning by sharing parameters on surface level rather than semantic level. In this work, we propose a novel self-attentive model with gate mechanism to fully utilize the semantic correlation between slot and intent. Our model first obtains intent-augmented embeddings based on neural network with self-attention mechanism. And then the intent semantic representation is utilized as the gate for labelling slot tags. The objectives of both tasks are optimized simultaneously via joint learning in an end-to-end way. We conduct experiment on popular benchmark ATIS. The results show that our model achieves state-of-the-art and outperforms other popular methods by a large margin in terms of both intent detection error rate and slot filling F1-score. This paper gives a new perspective for research on SLU.

Learning End-to-End Goal-Oriented Dialog with Multiple Answers

Janarthanan Rajendran, Jatin Ganhotra, Satinder Singh, and Lazaros Polymenakos

In a dialog, there could be multiple valid next utterances at any point. The present end-to-end neural methods for dialog do not take this into account. They learn with the assumption that at any time there is only one correct next utterance. In this work, we focus on this problem in the goal-oriented dialog setting where there are different paths to reach a goal. We propose a new method, that uses a combination of supervised learning and reinforcement learning approaches to address this issue. We also propose a new and more effective testbed, permuted-bAbI dialog tasks, by introducing multiple valid next utterances to the original-bAbI dialog tasks, which allows evaluation of end-to-end goal-oriented dialog systems in a more realistic setting. We show that there is a significant drop in performance of existing end-to-end neural methods from 81.5% per-dialog accuracy on original-bAbI dialog tasks to 30.3% on permuted-bAbI dialog tasks. We also show that our proposed method improves the performance and achieves 47.3% per-dialog accuracy on permuted-bAbI dialog tasks. We also release permuted-bAbI dialog tasks, our proposed testbed, to the community for evaluating dialog systems in a goal-oriented setting.

AirDialogue: An Environment for Goal-Oriented Dialogue Research
Wei Wei, Quoc Le, Andrew Dai, and Jia Li

Recent progress in dialogue generation has inspired a number of studies on dialogue systems that are capable of accomplishing tasks through natural language interactions. A promising direction among these studies is the use of reinforcement learning techniques, such as self-play, for training dialogue agents. However, current datasets are limited in size, and the environment for training agents and evaluating progress is relatively unsophisticated. We present AirDialogue, a large dataset that contains 301,427 goal-oriented conversations. To collect this dataset, we create a context-generator which provides travel and flight restrictions. We then ask human annotators to play the role of a customer or an agent and interact with the goal of successfully booking a trip given the restrictions. Key to our environment is the ease of evaluating the success of the dialogue, which is achieved by using ground-truth states (e.g., the flight being booked) generated by the restrictions. Any dialogue agent that does not generate the correct states is considered to fail. Our experimental results indicate that state-of-the-art dialogue models can only achieve a score of 0.17 while humans can reach a score of 0.91, which suggests significant opportunities for future improvement.

TACL-1424
TACL-1424

Stylistic dialogue response generation, with valuable applications in personality-based conversational agents, is a challenging task because the response needs to be fluent, contextually-relevant, as well as paralinguistically accurate. Moreover, parallel datasets for regular-to-stylistic pairs are usually unavailable. We present three weakly-supervised models that can generate diverse, polite (or rude) dialogue responses without parallel data. Our late fusion model (Fusion) merges the decoder of an encoder-attention-decoder dialogue model with a language model trained on stand-alone polite utterances. Our label-fine-tuning (LFT) model prepends to each source sequence a politeness-score scaled label (predicted by our state-of-the-art politeness classifier) during training, and at test time is able to generate polite, neutral, and rude responses by simply scaling the label embedding by the corresponding score. Our reinforcement learning model (Polite-RL) encourages politeness generation by assigning rewards proportional to the politeness classifier score of the sampled response. We also present two retrieval-based, polite dialogue model baselines. Human evaluation validates that while the Fusion and the retrieval-based models achieve politeness with poorer context-relevance, the LFT and Polite-RL models can produce significantly more polite responses without sacrificing dialogue quality.

Sentence Editing under Quantifiable Guidance
Yi Liao, Lidong Bing, Piji Li, Shuming Shi, Wai Lam, and Tong Zhang

We propose the task of Quantifiable Sequence Editing (QuaSE): editing an input sequence to generate an output sequence that satisfies a given numerical outcome value measuring a certain property of the sequence, with the requirement of keeping the main content of the input sequence. For example, an input sequence could be a word sequence, such as review sentence and advertisement text. For a review sentence, the outcome could be the review rating; for an advertisement, the outcome could be the click-through rate. The major challenge in performing QuaSE is how to perceive the outcome-related wordings, and only edit them to change the outcome. In this paper, the proposed framework contains two latent factors, namely, outcome factor and content factor, disentangled from the input sentence to allow convenient editing to change the outcome and keep the content. Our framework explores the pseudo-parallel sentences by modeling their content similarity and outcome differences to enable a better disentanglement of the latent factors, which allows generating an output to better satisfy the desired outcome and keep the content. The dual reconstruction structure further enhances the capability of generating expected output by exploiting the couplings of latent factors of pseudo-parallel sentences. For evaluation, we prepared a dataset of Yelp review sentences with the ratings as outcome. Extensive experimental results are reported and discussed to elaborate the peculiarities of our framework.

Paraphrase Generation with Deep Reinforcement Learning
Zichao Li, Xin Jiang, Lifeng Shang, and Hang Li

Automatic generation of paraphrases from a given sentence is an important yet challenging task in natural language processing (NLP). In this paper, we present a deep reinforcement learning approach to paraphrase generation. Specifically, we propose a new framework for the task, which consists of a generator and an evaluator, both of which are learned from data. The generator, built as a sequence-to-sequence learning model, can produce paraphrases given a sentence. The evaluator, constructed as a deep matching model, can judge whether two sentences are paraphrases of each other. The generator is first trained by deep learning and then further fine-tuned by reinforcement learning in which the reward is given by the evaluator. For the learning of the evaluator, we propose two methods based on supervised learning and inverse reinforcement learning respectively, depending on the type of available training data. Experimental results on two datasets demonstrate the proposed models (the generators) can produce more accurate paraphrases and outperform the state-of-the-art methods in paraphrase generation in both automatic evaluation and human evaluation.

Operation Guided Neural Networks for High Fidelity Data-To-Text Generation
Feng Nie, Jinpeng Wang, Jin-Ge Yao, Rong Pan, and Chin-Yew Lin

Recent neural models for data-to-text generation are mostly based on data-driven end-to-end training over encoder-decoder networks. Even though the generated texts are mostly fluent and informative, they often generate descriptions that are not consistent with the input structured data. This is a critical issue especially in domains that require inference or calculations over raw data. In this paper, we attempt to improve the fidelity of neural data-to-text generation by utilizing pre-executed symbolic operations. We propose a framework called Operation-guided Attention-based sequence-to-sequence network (OpAtt), with a specifically designed gating mechanism as well as a quantization module for operation results to utilize information from pre-executed operations. Experiments on two sports datasets show our proposed method clearly improves the fidelity of the generated texts to the input structured data.

Generating Classical Chinese Poems via Conditional Variational Autoencoder and Adversarial Training
Juntao Li, Yan Song, Haisong Zhang, Dongmin Chen, Shuming Shi, Dongyan Zhao, and Rui Yan

It is a challenging task to automatically compose poems with not only fluent expressions but also aesthetic wording. Although much attention has been paid to this task and promising progress is made, there exist notable gaps between automatically generated ones with those created by humans, especially on the aspects of term novelty and thematic consistency. Towards filling the gap, in this paper, we propose a conditional variational autoencoder with adversarial training for classical Chinese poem generation, where the autoencoder part generates poems with novel terms and a discriminator is applied to adversarially learn their thematic consistency with their titles. Experimental results on a large poetry corpus confirm the validity and effectiveness of our model, where its automatic and human evaluation scores outperform existing models.

Paragraph-level Neural Question Generation with Maxout Pointer and Gated Self-attention Networks
Yao Zhao, Xiaochuan Ni, Yuanyuan Ding, and Qifa Ke

Question generation, the task of automatically creating questions that can be answered by a certain span of text within a given passage, is important for question-answering and conversational systems in digital assistants such as Alexa, Cortana, Google Assistant and Siri. Recent sequence to sequence neural models have outperformed previous rule-based systems. Existing models mainly focused on using one or two sentences as the input. Long text has posed challenges for sequence to sequence neural models in question generation – worse performances were reported if using the whole paragraph (with multiple sentences) as the input. In reality, however, it often requires the whole paragraph as context in order to generate high quality questions. In this paper, we propose a maxout pointer mechanism with gated self-attention encoder to address the challenges of processing long text inputs for question generation. With sentence-level inputs, our model outperforms previous approaches with either sentence-level or paragraph-level inputs. Furthermore, our model can effectively utilize paragraphs as inputs, pushing the state-of-the-art result from 13.9 to 16.3 (BLEU_4).

Spider: A Large-Scale Complex Human-Labeled SQL Corpus for Semantic Parsing and Text-to-SQL Generation
Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev

We present *Spider*, a large-scale complex and cross-domain semantic parsing and text-to-SQL dataset annotated by 11 college students. It consists of 10,181 questions and 5,693 unique complex SQL queries on 200 databases with multiple tables covering 138 different domains. We define a new complex and cross-domain semantic parsing and text-to-SQL task so that different complicated SQL queries and databases appear in train and test sets. In this way, the task requires the model to generalize well to both new SQL queries and new database schemas. Therefore, Spider is distinct from most of the previous semantic parsing tasks because they all use a single database and have the exact same program in the train set and the test set. We experiment with various state-of-the-art models and the best model achieves only 9.7% exact matching accuracy on a database split setting. This shows that Spider presents a strong challenge for future research. Our dataset and task with the most recent updates are publicly available at `https://yale-lily.github.io/seq2sql/spider`.

Unsupervised Natural Language Generation with Denoising Autoencoders
Markus Freitag and Scott Roy

Generating text from structured data is important for various tasks such as question answering and dialog systems. We show that in at least one domain, without any supervision and only based on unlabeled text, we are able to build a Natural Language Generation (NLG) system with higher performance than supervised approaches. In our approach, we interpret the structured data as a corrupt representation of the desired output and use a denoising auto-encoder to reconstruct the sentence. We show how to introduce noise into training examples that do not contain structured data, and that the resulting denoising auto-encoder generalizes to generate correct sentences when given structured data.

Answer-focused and Position-aware Neural Question Generation
Xingwu Sun, Jing Liu, Yajuan Lyu, Wei He, Yanjun Ma, and Shi Wang

In this paper, we focus on the problem of question generation (QG). Recent neural network-based approaches employ the sequence-to-sequence model which takes an answer and its context as input and generates a relevant question as output. However, we observe two major issues with these approaches: (1) The generated interrogative words (or question words) do not match the answer type. (2) The model copies the context words that are far from and irrelevant to the answer, instead of the words that are close and relevant to the answer. To address these two issues, we propose an answer-focused and position-aware neural question generation model. (1) By answer-focused, we mean that we explicitly model question word generation by incorporating the answer embedding, which can help generate an interrogative word matching the answer type. (2) By position-aware, we mean that we model the relative distance between the context words and the answer. Hence the model can be aware of the position of the context words when copying them to generate a question. We conduct extensive experiments to examine the effectiveness of our model. The experimental results show that our model significantly improves the baseline and outperforms the state-of-the-art system.

Diversity-Promoting GAN: A Cross-Entropy Based Generative Adversarial Network for Diversified Text Generation
Jingjing Xu, Xuancheng Ren, Junyang Lin, and Xu SUN

Existing text generation methods tend to produce repeated and ”boring” expressions. To tackle this problem, we propose a new text generation model, called Diversity-Promoting Generative Adversarial Network (DP-GAN). The proposed model assigns low reward for repeatedly generated text and high reward for ”novel” and fluent text, encouraging the generator to produce diverse and informative text. Moreover, we propose a novel language-model based discriminator, which can better distinguish novel text from repeated text without the saturation problem compared with existing classifier-based discriminators. The experimental results on review generation and dialogue generation tasks demonstrate that our model can generate substantially more diverse and informative text than existing baselines.

Towards a Better Metric for Evaluating Question Generation Systems
Preksha Nema and Mitesh M. Khapra

There has always been criticism for using n-gram based similarity metrics, such as BLEU, NIST, *etc*, for evaluating the performance of NLG systems. However, these metrics continue to remain popular and are recently being used for evaluating the performance of systems which automatically generate questions from documents, knowledge graphs, images, *etc*. Given the rising interest in such automatic question generation (AQG) systems, it is important to objectively examine whether these metrics are suitable for this task. In particular, it is important to verify whether such metrics used for evaluating AQG systems focus on *answerability* of the generated question by preferring questions which contain all relevant information such as question type (Wh-types), entities, relations, *etc*. In this work, we show that current automatic evaluation metrics based on n-gram similarity do not always correlate well with human judgments about *answerability* of a question. To alleviate this problem and as a first step towards better evaluation metrics for AQG, we introduce a scoring function to capture *answerability* and show that when this scoring function is integrated with existing metrics, they correlate significantly better with human judgments. The scripts and data developed as a part of this work are made publicly available.[1]

Stylistic Chinese Poetry Generation via Unsupervised Style Disentanglement
Cheng Yang, Maosong Sun, Xiaoyuan Yi, and Wenhao Li

The ability to write diverse poems in different styles under the same poetic imagery is an important characteristic of human poetry writing. Most previous works on automatic Chinese poetry generation focused on improving the coherency among lines. Some work explored style transfer but suffered from expensive expert labeling of poem styles. In this paper, we target on stylistic poetry generation in a fully unsupervised manner for the first time. We propose a novel model which requires no supervised style labeling by incorporating mutual information, a concept in information theory, into modeling. Experimental results show that our model is able to generate stylistic poems without losing fluency and coherency.

Generating More Interesting Responses in Neural Conversation Models with Distributional Constraints
Ashutosh Baheti, Alan Ritter, Jiwei Li, and Bill Dolan

Neural conversation models tend to generate safe, generic responses for most inputs. This is due to the limitations of likelihood-based decoding objectives in generation tasks with diverse outputs, such as conversation. To address this challenge, we propose a simple yet effective approach for incorporating side information in the form of distributional constraints over the generated responses. We propose two constraints that help generate more content rich responses that are based on a model of syntax and topics (Griffiths et al., 2005) and semantic similarity (Arora et al., 2016). We evaluate our approach against a variety of competitive baselines, using both automatic metrics and human judgments, showing that our proposed approach generates responses that are much less generic without sacrificing plausibility. A working demo of our code can be found at `https://github.com/abaheti95/DC-NeuralConversation`.

Better Conversations by Modeling, Filtering, and Optimizing for Coherence and Diversity
Xinnuo Xu, Ondřej Dušek, Ioannis Konstas, and Verena Rieser

We present three enhancements to existing encoder-decoder models for open-domain conversational agents, aimed at effectively modeling coherence and promoting output diversity: (1) We introduce a measure of coherence as the GloVe embedding similarity between the dialogue context and the generated response, (2) we filter our training corpora based on the measure of coherence to obtain topically coherent and lexically diverse context-response pairs, (3) we then train a response generator using a conditional variational autoencoder model that incorporates the measure of coherence as a latent variable and uses a context gate to guarantee topical consistency with the context and promote lexical diversity. Experiments on the OpenSubtitles corpus show a substantial improvement over competitive neural models in terms of BLEU score as well as metrics of coherence and diversity.

Incorporating Background Knowledge into Video Description Generation
Spencer Whitehead, Heng Ji, Mohit Bansal, Shih-Fu Chang, and Clare Voss

Most previous efforts toward video captioning focus on generating generic descriptions, such as, "A man is talking." We collect a news video dataset to generate enriched descriptions that include important background knowledge, such as named entities and related events, which allows the user to fully understand the video content. We develop an approach that uses video meta-data to retrieve topically related news documents for a video and extracts the events and named entities from these documents. Then, given the video as well as the extracted events and entities, we generate a description using a Knowledge-aware Video Description network. The model learns to incorporate entities found in the topically related documents into the description via an entity pointer network and the generation procedure is guided by the event and entity types from the topically related documents through a knowledge gate, which is a gating mechanism added to the model's decoder that takes a one-hot vector of these types. We evaluate our approach on the new dataset of news videos we have collected, establishing the first benchmark for this dataset as well as proposing a new metric to evaluate these descriptions.

Multimodal Differential Network for Visual Question Generation
Badri Narayana Patro, Sandeep Kumar, Vinod Kumar Kurmi, and Vinay Namboodiri

Generating natural questions from an image is a semantic task that requires using visual and language modality to learn multimodal representations. Images can have multiple visual and language contexts that are relevant for generating questions namely places, captions, and tags. In this paper, we propose

[1] `https://github.com/PrekshaNema25/Answerability-Metric`

the use of exemplars for obtaining the relevant context. We obtain this by using a Multimodal Differential Network to produce natural and engaging questions. The generated questions show a remarkable similarity to the natural questions as validated by a human study. Further, we observe that the proposed approach substantially improves over state-of-the-art benchmarks on the quantitative metrics (BLEU, METEOR, ROUGE, and CIDEr).

Entity-aware Image Caption Generation
Di Lu, Spencer Whitehead, Lifu Huang, Heng Ji, and Shih-Fu Chang

Current image captioning approaches generate descriptions which lack specific information, such as named entities that are involved in the images. In this paper we propose a new task which aims to generate informative image captions, given images and hashtags as in- put. We propose a simple but effective approach to tackle this problem. We first train a convolutional neural networks - long short term memory networks (CNN-LSTM) model to generate a template caption based on the in- put image. Then we use a knowledge graph based collective inference algorithm to fill in the template with specific named entities retrieved via the hashtags. Experiments on a new benchmark dataset collected from Flickr show that our model generates news-style image descriptions with much richer information. Our model outperforms unimodal baselines significantly with various evaluation metrics.

Learning to Describe Differences Between Pairs of Similar Images
Harsh Jhamtani and Taylor Berg-Kirkpatrick

In this paper, we introduce the task of automatically generating text to describe the differences between two similar images. We collect a new dataset by crowd-sourcing difference descriptions for pairs of image frames extracted from video-surveillance footage. Annotators were asked to succinctly describe all the differences in a short paragraph. As a result, our novel dataset provides an opportunity to explore models that align language and vision, and capture visual salience. The dataset may also be a useful benchmark for coherent multi-sentence generation. We perform a first-pass visual analysis that exposes clusters of differing pixels as a proxy for object-level differences. We propose a model that captures visual salience by using a latent variable to align clusters of differing pixels with output sentences. We find that, for both single-sentence generation and as well as multi-sentence generation, the proposed model outperforms the models that use attention alone.

Object Hallucination in Image Captioning
Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko

Despite continuously improving performance, contemporary image captioning models are prone to “hallucinating” objects that are not actually in a scene. One problem is that standard metrics only measure similarity to ground truth captions and may not fully capture image relevance. In this work, we propose a new image relevance metric to evaluate current models with veridical visual labels and assess their rate of object hallucination. We analyze how captioning model architectures and learning objectives contribute to object hallucination, explore when hallucination is likely due to image misclassification or language priors, and assess how well current sentence metrics capture object hallucination. We investigate these questions on the standard image captioning benchmark, MSCOCO, using a diverse set of models. Our analysis yields several interesting findings, including that models which score best on standard sentence metrics do not always have lower hallucination and that models which hallucinate more tend to make errors driven by language priors.

Abstractive Text-Image Summarization based on Multi-Modal Attentional Hierarchical RNN
Jingqiang Chen and Hai Zhuge

Rapid growth of multi-modal documents on the Internet makes multi-modal summarization research necessary. Most previous research summarizes texts or images separately. Recent neural summarization research shows the strength of the Encoder-Decoder model in text summarization. This paper proposes an abstractive text-image summarization model using the attentional hierarchical Encoder-Decoder model to summarize a text document and its accompanying images simultaneously, and then to align the sentences and images in summaries. A multi-modal attentional mechanism is proposed to attend original sentences, images, and captions when decoding. The DailyMail dataset is extended by collecting images and captions from the Web. Experiments show our model outperforms the neural abstractive and extractive text summarization methods that do not consider images. In addition, our model can generate informative summaries of images.

Keyphrase Generation with Correlation Constraints
Jun Chen, Xiaoming Zhang, Yu Wu, Zhao Yan, and Zhoujun Li

In this paper, we study automatic keyphrase generation. Although conventional approaches to this task show promising results, they neglect correlation among keyphrases, resulting in duplication and coverage issues. To solve these problems, we propose a new sequence-to-sequence architecture for keyphrase generation named CorrRNN, which captures correlation among multiple keyphrases in two ways. First, we employ a coverage vector to indicate whether the word in the source document has been summarized by previous phrases to improve the coverage for keyphrases. Second, preceding phrases are taken into account to eliminate duplicate phrases and improve result coherence. Experiment results show that our model significantly outperforms the state-of-the-art method on benchmark datasets in terms of both accuracy and diversity.

Closed-Book Training to Improve Summarization Encoder Memory
Yichen Jiang and Mohit Bansal

A good neural sequence-to-sequence summarization model should have a strong encoder that can distill and memorize the important information from long input texts so that the decoder can generate salient summaries based on the encoder's memory. In this paper, we aim to improve the memorization capabilities of the encoder of a pointer-generator model by adding an additional 'closed-book' decoder without attention and pointer mechanisms. Such a decoder forces the encoder to be more selective in the information encoded in its memory state because the decoder can't rely on the extra information provided by the attention and possibly copy modules, and hence improves the entire model. On the CNN/Daily Mail dataset, our 2-decoder model outperforms the baseline significantly in terms of ROUGE and METEOR metrics, for both cross-entropy and reinforced setups (and on human evaluation). Moreover, our model also achieves higher scores in a test-only DUC-2002 generalizability setup. We further present a memory ability test, two saliency metrics, as well as several sanity-check ablations (based on fixed-encoder, gradient-flow cut, and model capacity) to prove that the encoder of our 2-decoder model does in fact learn stronger memory representations than the baseline encoder.

Improving Neural Abstractive Document Summarization with Structural Regularization
Wei Li, Xinyan Xiao, Yajuan Lyu, and Yuanzhuo Wang

Recent neural sequence-to-sequence models have shown significant progress on short text summarization. However, for document summarization, they fail to capture the long-term structure of both documents and multi-sentence summaries, resulting in information loss and repetitions. In this paper, we propose to leverage the structural information of both documents and multi-sentence summaries to improve the document summarization performance. Specifically, we import both structural-compression and structural-coverage regularization into the summarization process in order to capture the information compression and information coverage properties, which are the two most important structural properties of document summarization. Experimental results demonstrate that the structural regularization improves the document summarization performance significantly, which enables our model to generate more informative and concise summaries, and thus significantly outperforms state-of-the-art neural abstractive methods.

Iterative Document Representation Learning Towards Summarization with Polishing
Xiuying Chen, Shen Gao, Chongyang Tao, Yan Song, Dongyan Zhao, and Rui Yan

In this paper, we introduce Iterative Text Summarization (ITS), an iteration-based model for supervised extractive text summarization, inspired by the observation that it is often necessary for a human to read an article multiple times in order to fully understand and summarize its contents. Current summarization approaches read through a document only once to generate a document representation, resulting in a sub-optimal representation. To address this issue we introduce a model which iteratively polishes the document representation on many passes through the document. As part of our model, we also introduce a selective reading mechanism that decides more accurately the extent to which each sentence in the model should be updated. Experimental results on the CNN/DailyMail and DUC2002 datasets demonstrate that our model significantly outperforms state-of-the-art extractive systems when evaluated by machines and by humans.

Bottom-Up Abstractive Summarization
Sebastian Gehrmann, Yuntian Deng, and Alexander Rush

Neural summarization produces outputs that are fluent and readable, but which can be poor at content selection, for instance often copying full sentences from the source document. This work explores the use of data-efficient content selectors to over-determine phrases in a source document that should be part of the summary. We use this selector as a bottom-up attention step to constrain the model to likely phrases. We show that this approach improves the ability to compress text, while still generating fluent summaries. This two-step process is both simpler and higher performing than other end-to-end content selection models, leading to significant improvements on ROUGE for both the CNN-DM and NYT corpus. Furthermore, the content selector can be trained with as little as 1,000 sentences making it easy to transfer a trained summarizer to a new domain.

Controlling Length in Abstractive Summarization Using a Convolutional Neural Network
Yizhu Liu, Zhiyi Luo, and Kenny Zhu

Convolutional neural networks (CNNs) have met great success in abstractive summarization, but they cannot effectively generate summaries of desired lengths. Because generated summaries are used in difference scenarios which may have space or length constraints, the ability to control the summary length in abstractive summarization is an important problem. In this paper, we propose an approach to constrain the summary length by extending a convolutional sequence to sequence model. The results show that this approach generates high-quality summaries with user defined length, and outperforms the baselines consistently in terms of ROUGE score, length variations and semantic similarity.

APRIL: Interactively Learning to Summarise by Combining Active Preference Learning and Reinforcement Learning
Yang Gao, Christian M. Meyer, and Iryna Gurevych

We propose a method to perform automatic document summarisation without using reference summaries. Instead, our method interactively learns from users' preferences. The merit of preference-based interactive summarisation is that preferences are easier for users to provide than reference summaries. Existing preference-based interactive learning methods suffer from high sample complexity, i.e. they need to interact with the oracle for many rounds in order to converge. In this work, we propose a new objective function, which enables us to leverage active learning, preference learning and reinforcement learning techniques in order to reduce the sample complexity. Both simulation and real-user experiments suggest that our method significantly advances the state of the art. Our source code is freely available at https://github.com/UKPLab/ emnlp2018-april.

Adapting the Neural Encoder-Decoder Framework from Single to Multi-Document Summarization
Logan Lebanoff, Kaiqiang Song, and Fei Liu

Generating a text abstract from a set of documents remains a challenging task. The neural encoder-decoder framework has recently been exploited to summarize single documents, but its success can in part be attributed to the availability of large parallel data automatically acquired from the Web. In contrast, parallel data for multi-document summarization are scarce and costly to obtain. There is a pressing need to adapt an encoder-decoder model trained on single-document summarization data to work with multiple-document input. In this paper, we present an initial investigation into a novel adaptation method. It exploits the maximal marginal relevance method to select representative sentences from multi-document input, and leverages an abstractive encoder-decoder model to fuse disparate sentences to an abstractive summary. The adaptation method is robust and itself requires no training data. Our system compares favorably to state-of-the-art extractive and abstractive approaches judged by automatic metrics and human assessors.

Semi-Supervised Learning for Neural Keyphrase Generation
Hai Ye and Lu Wang

We study the problem of generating keyphrases that summarize the key points for a given document. While sequence-to-sequence (seq2seq) models have achieved remarkable performance on this task (Meng et al., 2017), model training often relies on large amounts of labeled data, which is only applicable to resource-rich domains. In this paper, we propose semi-supervised keyphrase generation methods by leveraging both labeled data and large-scale unlabeled samples for learning. Two strategies are proposed. First, unlabeled documents are first tagged with synthetic keyphrases obtained from unsupervised keyphrase extraction methods or a self-learning algorithm, and then combined with labeled samples for training. Furthermore, we investigate a multi-task learning framework to jointly learn to generate keyphrases as well as the titles of the articles. Experimental results show that our semi-supervised learning-based methods outperform a state-of-the-art model trained with labeled data only.

MSMO: Multimodal Summarization with Multimodal Output
Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong

Multimodal summarization has drawn much attention due to the rapid growth of multimedia data. The output of the current multimodal summarization systems is usually represented in texts. However, we have found through experiments that multimodal output can significantly improve user satisfaction for informativeness of summaries. In this paper, we propose a novel task, multimodal summarization with multimodal output (MSMO). To handle this task, we first collect a large-scale dataset for MSMO research. We then propose a multimodal attention model to jointly generate text and select the most relevant image from the multimodal input. Finally, to evaluate multimodal outputs, we construct a novel multimodal automatic evaluation (MMAE) method which considers both intra-modality salience and inter-modality relevance. The experimental results show the effectiveness of MMAE.

Frustratingly Easy Model Ensemble for Abstractive Summarization
Hayato Kobayashi

Ensemble methods, which combine multiple models at decoding time, are now widely known to be effective for text-generation tasks. However, they generally increase computational costs, and thus, there have been many studies on compressing or distilling ensemble models. In this paper, we propose an alternative, simple but effective unsupervised ensemble method, *post-ensemble*, that combines multiple models by selecting a majority-like output in post-processing. We theoretically prove that our method is closely related to kernel density estimation based on the von Mises-Fisher kernel. Experimental results on a news-headline-generation task show that the proposed method performs better than the current ensemble methods.

Automatic Pyramid Evaluation Exploiting EDU-based Extractive Reference Summaries
Tsutomu Hirao, Hidetaka Kamigaito, and Masaaki Nagata

This paper tackles automation of the pyramid method, a reliable manual evaluation framework. To construct a pyramid, we transform human-made reference summaries into extractive reference summaries that consist of Elementary Discourse Units (EDUs) obtained from source documents and then weight every EDU by counting the number of extractive reference summaries that contain the EDU. A summary is scored by the correspondences between EDUs in the summary and those in the pyramid. Experiments on DUC and TAC data sets show that our methods strongly correlate with various manual evaluations.

Learning to Encode Text as Human-Readable Summaries using Generative Adversarial Networks
Yaushian Wang and Hung-yi Lee

Auto-encoders compress input data into a latent-space representation and reconstruct the original data from the representation. This latent representation is not easily interpreted by humans. In this paper, we propose training an auto-encoder that encodes input text into human-readable sentences, and unpaired abstractive summarization is thereby achieved. The auto-encoder is composed of a generator and a reconstructor. The generator encodes the input text into a shorter word sequence, and the reconstructor recovers the generator input from the generator output. To make the generator output human-readable, a discriminator restricts the output of the generator to resemble human-written sentences. By taking the generator output as the summary of the input text, abstractive summarization is achieved without document-summary pairs as training data. Promising results are shown on both English and Chinese corpora.

Visualizing Group Dynamics based on Multiparty Meeting Understanding
Ni Zhang, Tongtao Zhang, Indrani Bhattacharya, Heng Ji, and Rich Radke

Group discussions are usually aimed at sharing opinions, reaching consensus and making good decisions based on group knowledge. During a discussion, participants might adjust their own opinions as well as tune their attitudes towards others' opinions, based on the unfolding interactions. In this paper, we demonstrate a framework to visualize such dynamics; at each instant of a conversation, the participants' opinions and potential influence on their counterparts is easily visualized. We use multi-party meeting opinion mining based on bipartite graphs to extract opinions and calculate mutual influential factors, using the Lunar Survival Task as a study case.

PizzaPal: Conversational Pizza Ordering using a High-Density Conversational AI Platform
Antoine Raux, Yi Ma, Paul Yang, and Felicia Wong

This paper describes PizzaPal, a voice-only agent for ordering pizza, as well as the Conversational AI architecture built at b4.ai. Based on the principles of high-density conversational AI, it supports natural and flexible interactions through neural conversational language understanding, robust dialog state tracking, and hierarchical task decomposition.

Developing Production-Level Conversational Interfaces with Shallow Semantic Parsing
Arushi Raghuvanshi, Lucien Carroll, and Karthik Raghunathan

We demonstrate an end-to-end approach for building conversational interfaces from prototype to production that has proven to work well for a number of applications across diverse verticals. Our architecture improves on the standard domain-intent-entity classification hierarchy and dialogue management architecture by leveraging shallow semantic parsing. We observe that NLU systems for industry applications often require more structured representations of entity relations than provided by the standard hierarchy, yet without requiring full semantic parses which are often inaccurate on real-world conversational data. We distinguish two kinds of semantic properties that can be provided through shallow semantic parsing: entity groups and entity roles. We also provide live demos of conversational apps built for two different use cases: food ordering and meeting control.

SyntaViz: Visualizing Voice Queries through a Syntax-Driven Hierarchical Ontology
Md Iftekhar Tanveer and Ferhan Ture

This paper describes SyntaViz, a visualization interface specifically designed for analyzing natural-language queries that were created by users of a voice-enabled product. SyntaViz provides a platform for browsing the ontology of user queries from a syntax-driven perspective, providing quick access to high-impact failure points of the existing intent understanding system and evidence for data-driven decisions in the development cycle. A case study on Xfinity X1 (a voice-enabled entertainment platform from Comcast) reveals that SyntaViz helps developers identify multiple action items in a short amount of time without any special training. SyntaViz has been open-sourced for the benefit of the community.

LIA: A Natural Language Programmable Personal Assistant
Igor Labutov, Shashank Srivastava, and Tom Mitchell

We present LIA, an intelligent personal assistant that can be programmed using natural language. Our system demonstrates multiple competencies towards learning from human-like interactions. These include the ability to be taught reusable conditional procedures, the ability to be taught new knowledge about the world (concepts in an ontology) and the ability to be taught how to ground that knowledge in a set of sensors and effectors. Building such a system highlights design questions regarding the overall architecture that such an agent should have, as well as questions about parsing and grounding language in situational contexts. We outline key properties of this architecture, and demonstrate a prototype that embodies them in the form of a personal assistant on an Android device.

Data2Text Studio: Automated Text Generation from Structured Data
Longxu Dou, Guanghui Qin, Jinpeng Wang, Jin-Ge Yao, and Chin-Yew Lin

Data2Text Studio is a platform for automated text generation from structured data. It is equipped with a Semi-HMMs model to extract high-quality templates and corresponding trigger conditions from parallel data automatically, which improves the interactivity and interpretability of the generated text. In addition, several easy-to-use tools are provided for developers to edit templates of pre-trained models, and APIs are released for developers to call the pre-trained model to generate texts in third-party applications. We conduct experiments on RotoWire datasets for template extraction and text generation. The results show that our model achieves improvements on both tasks.

Demonstrating Par4Sem - A Semantic Writing Aid with Adaptive Paraphrasing
Seid Muhie Yimam and Chris Biemann

In this paper, we present Par4Sem, a semantic writing aid tool based on adaptive paraphrasing. Unlike many annotation tools that are primarily used to collect training examples, Par4Sem is integrated into a real word application, in this case a writing aid tool, in order to collect training examples from usage data. Par4Sem is a tool, which supports an adaptive, iterative, and interactive process where the underlying machine learning models are updated for each iteration using new training examples from usage data. After motivating the use of ever-learning tools in NLP applications, we evaluate Par4Sem by adopting it to a text simplification task through mere usage.

Session 10 Overview – Sunday, November 4, 2018

Oral tracks

Track A *Question Answering III* Gold Hall	Track B *Machine Translation III* Copper Hall	Track C *Discourse* Silver Hall	Track D *Evolution / Sociolinguistics* Hall 100	
Joint Multitask Learning for Community Question Answering Using Task-Specific Embeddings *Joty, Màrquez, and Nakov*	Exploiting Deep Representations for Neural Machine Translation *Dou, Tu, Wang, Shi, and Zhang*	A Skeleton-Based Model for Promoting Coherence Among Sentences in Narrative Story Generation *Xu, Ren, Zhang, Zeng, Cai, and SUN*	Making "fetch" happen: The influence of social and linguistic context on nonstandard word growth and decline *Stewart and Eisenstein*	11:00
What Makes Reading Comprehension Questions Easier? *Sugawara, Inui, Sekine, and Aizawa*	Why Self-Attention? A Targeted Evaluation of Neural Machine Translation Architectures *Tang, Müller, Rios, and Sennrich*	Nexus Network: Connecting the Preceding and the Following in Dialogue Generation *Shen, Su, Li, and Klakow*	Analyzing Correlated Evolution of Multiple Features Using Latent Representations *Murawaki*	11:18
Commonsense for Generative Multi-Hop Question Answering Tasks *Bauer, Wang, and Bansal*	Simplifying Neural Machine Translation with Addition-Subtraction Twin-Gated Recurrent Networks *Zhang, Xiong, Lin, and Zhang*	A New Local Coherence Model for Text Quality Assessment *Mesgar and Strube*	Capturing Regional Variation with Distributed Place Representations and Geographic Retrofitting *Hovy and Purschke*	11:36
Open Domain Question Answering Using Early Fusion of Knowledge Bases and Text *Sun, Dhingra, Zaheer, Mazaitis, Salakhutdinov, and Cohen*	Cube Pruning for Neural Machine Translation *Zhang, Huang, Feng, Shen, and Liu*	Deep Attentive Sentence Ordering Network *Cui, Li, Chen, and Zhang*	Characterizing Interactions and Relationships between People *Rashid and Blanco*	11:54
A Nil-Aware Answer Extraction Framework for Question Answering *Kundu and Ng*	Revisiting Character-Based Neural Machine Translation with Capacity and Compression *Cherry, Foster, Bapna, Firat, and Macherey*	Getting to "Hearer-old": Charting Referring Expressions Across Time *Staliūnaitė, Rohde, Webber, and Louis*	Why Swear? Analyzing and Inferring the Intentions of Vulgar Expressions *Holgate, Cachola, Preoţiuc-Pietro, and Li*	12:12

Poster tracks

11:00–12:30

Track D: *Machine Learning* Hall 100

Parallel Session 10

Session 10A: Question Answering III

Gold Hall Chair: *chairname*

Joint Multitask Learning for Community Question Answering Using Task-Specific Embeddings

Shafiq Joty, Lluís Màrquez, and Preslav Nakov 11:00–11:18

We address jointly two important tasks for Question Answering in community forums: given a new question, (i) find related existing questions, and (ii) find relevant answers to this new question. We further use an auxiliary task to complement the previous two, i.e., (iii) find good answers with respect to the thread question in a question-comment thread. We use deep neural networks (DNNs) to learn meaningful task-specific embeddings, which we then incorporate into a conditional random field (CRF) model for the multitask setting, performing joint learning over a complex graph structure. While DNNs alone achieve competitive results when trained to produce the embeddings, the CRF, which makes use of the embeddings and the dependencies between the tasks, improves the results significantly and consistently across a variety of evaluation metrics, thus showing the complementarity of DNNs and structured learning.

What Makes Reading Comprehension Questions Easier?

Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa 11:18–11:36

A challenge in creating a dataset for machine reading comprehension (MRC) is to collect questions that require a sophisticated understanding of language to answer beyond using superficial cues. In this work, we investigate what makes questions easier across recent 12 MRC datasets with three question styles (answer extraction, description, and multiple choice). We propose to employ simple heuristics to split each dataset into easy and hard subsets and examine the performance of two baseline models for each of the subsets. We then manually annotate questions sampled from each subset with both validity and requisite reasoning skills to investigate which skills explain the difference between easy and hard questions. From this study, we observed that (i) the baseline performances for the hard subsets remarkably degrade compared to those of entire datasets, (ii) hard questions require knowledge inference and multiple-sentence reasoning in comparison with easy questions, and (iii) multiple-choice questions tend to require a broader range of reasoning skills than answer extraction and description questions. These results suggest that one might overestimate recent advances in MRC.

Commonsense for Generative Multi-Hop Question Answering Tasks

Lisa Bauer, Yicheng Wang, and Mohit Bansal 11:36–11:54

Reading comprehension QA tasks have seen a recent surge in popularity, yet most works have focused on fact-finding extractive QA. We instead focus on a more challenging multi-hop generative task (NarrativeQA), which requires the model to reason, gather, and synthesize disjoint pieces of information within the context to generate an answer. This type of multi-step reasoning also often requires understanding implicit relations, which humans resolve via external, background commonsense knowledge. We first present a strong generative baseline that uses a multi-attention mechanism to perform multiple hops of reasoning and a pointer-generator decoder to synthesize the answer. This model performs substantially better than previous generative models, and is competitive with current state-of-the-art span prediction models. We next introduce a novel system for selecting grounded multi-hop relational commonsense information from ConceptNet via a pointwise mutual information and term-frequency based scoring function. Finally, we effectively use this extracted commonsense information to fill in gaps of reasoning between context hops, using a selectively-gated attention mechanism. This boosts the model's performance significantly (also verified via human evaluation), establishing a new state-of-the-art for the task. We also show that our background knowledge enhancements are generalizable and improve performance on QAngaroo-WikiHop, another multi-hop reasoning dataset.

Open Domain Question Answering Using Early Fusion of Knowledge Bases and Text

Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Kathryn Mazaitis, Ruslan Salakhutdinov, and William Cohen 11:54–12:12

Open Domain Question Answering (QA) is evolving from complex pipelined systems to end-to-end deep neural networks. Specialized neural models have been developed for extracting answers from either text alone or Knowledge Bases (KBs) alone. In this paper we look at a more practical setting, namely QA over the combination of a KB and entity-linked text, which is appropriate when an incomplete KB is available with a large text corpus. Building on recent advances in graph representation learning we propose a novel model, GRAFT-Net, for extracting answers from a question-specific subgraph containing text and KB entities and relations. We construct a suite of benchmark tasks for this problem, varying the difficulty of questions, the amount of training data, and KB completeness. We show that GRAFT-Net is competitive with the state-of-the-art when tested using either KBs or text alone, and vastly outperforms existing methods in the combined setting.

A Nil-Aware Answer Extraction Framework for Question Answering
Souvik Kundu and Hwee Tou Ng 12:12–12:30

Recently, there has been a surge of interest in reading comprehension-based (RC) question answering (QA). However, current approaches suffer from an impractical assumption that every question has a valid answer in the associated passage. A practical QA system must possess the ability to determine whether a valid answer exists in a given text passage. In this paper, we focus on developing QA systems that can extract an answer for a question if and only if the associated passage contains an answer. If the associated passage does not contain any valid answer, the QA system will correctly return Nil. We propose a novel nil-aware answer span extraction framework that is capable of returning Nil or a text span from the associated passage as an answer in a single step. We show that our proposed framework can be easily integrated with several recently proposed QA models developed for reading comprehension and can be trained in an end-to-end fashion. Our proposed nil-aware answer extraction neural network decomposes pieces of evidence into relevant and irrelevant parts and then combines them to infer the existence of any answer. Experiments on the NewsQA dataset show that the integration of our proposed framework significantly outperforms several strong baseline systems that use pipeline or threshold-based approaches.

Session 10B: Machine Translation III

Copper Hall Chair: *chairname*

Exploiting Deep Representations for Neural Machine Translation

Zi-Yi Dou, Zhaopeng Tu, Xing Wang, Shuming Shi, and Tong Zhang 11:00–11:18

Advanced neural machine translation (NMT) models generally implement encoder and decoder as multiple layers, which allows systems to model complex functions and capture complicated linguistic structures. However, only the top layers of encoder and decoder are leveraged in the subsequent process, which misses the opportunity to exploit the useful information embedded in other layers. In this work, we propose to simultaneously expose all of these signals with layer aggregation and multi-layer attention mechanisms. In addition, we introduce an auxiliary regularization term to encourage different layers to capture diverse information. Experimental results on widely-used WMT14 English-German and WMT17 Chinese-English translation data demonstrate the effectiveness and universality of the proposed approach.

Why Self-Attention? A Targeted Evaluation of Neural Machine Translation Architectures

Gongbo Tang, Mathias Müller, Annette Rios, and Rico Sennrich 11:18–11:36

Recently, non-recurrent architectures (convolutional, self-attentional) have outperformed RNNs in neural machine translation. CNNs and self-attentional networks can connect distant words via shorter network paths than RNNs, and it has been speculated that this improves their ability to model long-range dependencies. However, this theoretical argument has not been tested empirically, nor have alternative explanations for their strong performance been explored in-depth. We hypothesize that the strong performance of CNNs and self-attentional networks could also be due to their ability to extract semantic features from the source text, and we evaluate RNNs, CNNs and self-attention networks on two tasks: subject-verb agreement (where capturing long-range dependencies is required) and word sense disambiguation (where semantic feature extraction is required). Our experimental results show that: 1) self-attentional networks and CNNs do not outperform RNNs in modeling subject-verb agreement over long distances; 2) self-attentional networks perform distinctly better than RNNs and CNNs on word sense disambiguation.

Simplifying Neural Machine Translation with Addition-Subtraction Twin-Gated Recurrent Networks

Biao Zhang, Deyi Xiong, jinsong su jinsong, Qian Lin, and Huiji Zhang 11:36–11:54

In this paper, we propose an additionsubtraction twin-gated recurrent network (ATR) to simplify neural machine translation. The recurrent units of ATR are heavily simplified to have the smallest number of weight matrices among units of all existing gated RNNs. With the simple addition and subtraction operation, we introduce a twin-gated mechanism to build input and forget gates which are highly correlated. Despite this simplification, the essential non-linearities and capability of modeling long-distance dependencies are preserved. Additionally, the proposed ATR is more transparent than LSTM/GRU due to the simplification. Forward self-attention can be easily established in ATR, which makes the proposed network interpretable. Experiments on WMT14 translation tasks demonstrate that ATR-based neural machine translation can yield competitive performance on English-German and English-French language pairs in terms of both translation quality and speed. Further experiments on NIST Chinese-English translation, natural language inference and Chinese word segmentation verify the generality and applicability of ATR on different natural language processing tasks.

Cube Pruning for Neural Machine Translation

Wen Zhang, Liang Huang, Yang Feng, Lei Shen, and Qun Liu 11:54–12:12

Although neural machine translation has achieved promising results, it suffers from slow translation speed. The direct consequence is that a trade-off has to be made between translation quality and speed, thus its performance can not come into full play. We apply cube pruning, a popular technique to speed up dynamic programming, into neural machine translation to speed up the translation. To construct the equivalence class, similar target hidden states are combined, leading to less RNN expansion operations on the target side and less softmax operations over the large target vocabulary. The experiments show that, at the same or even better translation quality, our method can translate faster compared with naive beam search by 3.3x on GPUs and 3.5x on CPUs.

Revisiting Character-Based Neural Machine Translation with Capacity and Compression

Colin Cherry, George Foster, Ankur Bapna, Orhan Firat, and Wolfgang Macherey 12:12–12:30

Translating characters instead of words or word-fragments has the potential to simplify the processing pipeline for neural machine translation (NMT), and improve results by eliminating hyper-parameters and manual feature engineering. However, it results in longer sequences in which each symbol contains less information, creating both modeling and computational challenges. In this paper, we show that the modeling problem can be solved by standard sequence-to-sequence architectures of sufficient depth, and that deep models operating at the character level outperform identical models operating over word fragments. This result implies that alternative architectures for handling character input are better viewed as methods for reducing computation time than as improved ways of modeling longer sequences. From this perspective, we evaluate several techniques for character-level NMT, verify that they do not match the performance of our deep character baseline model, and evaluate the performance versus computation time tradeoffs they offer. Within this framework, we also perform the first evaluation for NMT of conditional computation over time, in which the model learns which timesteps can be skipped, rather than having them be dictated by a fixed schedule specified before training begins.

Session 10C: Discourse

Silver Hall Chair: *chairname*

A Skeleton-Based Model for Promoting Coherence Among Sentences in Narrative Story Generation

Jingjing Xu, Xuancheng Ren, Yi Zhang, Qi Zeng, Xiaoyan Cai, and Xu SUN 11:00–11:18

Narrative story generation is a challenging problem because it demands the generated sentences with tight semantic connections, which has not been well studied by most existing generative models. To address this problem, we propose a skeleton-based model to promote the coherence of generated stories. Different from traditional models that generate a complete sentence at a stroke, the proposed model first generates the most critical phrases, called skeleton, and then expands the skeleton to a complete and fluent sentence. The skeleton is not manually defined, but learned by a reinforcement learning method. Compared to the state-of-the-art models, our skeleton-based model can generate significantly more coherent text according to human evaluation and automatic evaluation. The G-score is improved by 20.1% in human evaluation.

Nexus Network: Connecting the Preceding and the Following in Dialogue Generation

Xiaoyu Shen, Hui Su, Wenjie Li, and Dietrich Klakow 11:18–11:36

Sequence-to-Sequence (seq2seq) models have become overwhelmingly popular in build- ing end-to-end trainable dialogue systems. Though highly efficient in learning the back- bone of human-computer communications, they suffer from the problem of strongly fa- voring short generic responses. In this pa- per, we argue that a good response should smoothly connect both the preceding dialogue history and the following conversations. We strengthen this connection by mutual infor- mation maximization. To sidestep the non- differentiability of discrete natural language tokens, we introduce an auxiliary continuous code space and map such code space to a learn- able prior distribution for generation purpose. Experiments on two dialogue datasets validate the effectiveness of our model, where the gen- erated responses are closely related to the dia- logue context and lead to more interactive con- versations.

A New Local Coherence Model for Text Quality Assessment

Mohsen Mesgar and Michael Strube 11:36–11:54

We propose a local coherence model that captures the flow of what semantically connects adjacent sentences in a text. We represent the semantics of a sentence by a vector and capture its state at each word of the sentence. We model what relates two adjacent sentences based on the two most similar semantic states, each of which is in one of the sentences. We encode the perceived coherence of a text by a vector, which represents patterns of changes in salient information that relates adjacent sentences. Our experiments demonstrate that our approach is beneficial for two downstream tasks: Readability assessment, in which our model achieves new state-of-the-art results; and essay scoring, in which the combination of our coherence vectors and other task-dependent features significantly improves the performance of a strong essay scorer.

Deep Attentive Sentence Ordering Network

Baiyun Cui, Yingming Li, Ming Chen, and Zhongfei Zhang 11:54–12:12

In this paper, we propose a novel deep attentive sentence ordering network (referred as ATTOrderNet) which integrates self-attention mechanism with LSTMs in the encoding of input sentences. It enables us to capture global dependencies among sentences regardless of their input order and obtains a reliable representation of the sentence set. With this representation, a pointer network is exploited to generate an ordered sequence. The proposed model is evaluated on Sentence Ordering and Order Discrimination tasks. The extensive experimental results demonstrate its effectiveness and superiority to the state-of-the-art methods.

Getting to "Hearer-old": Charting Referring Expressions Across Time

Ieva Staliūnaitė, Hannah Rohde, Bonnie Webber, and Annie Louis 12:12–12:30

When a reader is first introduced to an entity, its referring expression must describe the entity. For entities that are widely known, a single word or phrase often suffices. This paper presents the first study of how expressions that refer to the same entity develop over time. We track thousands of person and organization entities over 20 years of New York Times (NYT). As entities move from hearer-new (first introduction to the NYT audience) to hearer-old (common knowledge) status, we show empirically that the referring expressions along this trajectory depend on the type of the entity, and exhibit linguistic properties related to becoming common knowledge (e.g., shorter length, less use of appositives, more definiteness). These properties can also be used to build a model to predict how long it will take for an entity to reach hearer-old status. Our results reach 10-30% absolute improvement over a majority-class baseline.

Session 10D: Evolution / Sociolinguistics

Hall 100 Chair: *chairname*

Making "fetch" happen: The influence of social and linguistic context on nonstandard word growth and decline
Ian Stewart and Jacob Eisenstein 11:00–11:18

In an online community, new words come and go: today's "haha" may be replaced by tomorrow's "lol." Changes in online writing are usually studied as a social process, with innovations diffusing through a network of individuals in a speech community. But unlike other types of innovation, language change is shaped and constrained by the grammatical system in which it takes part. To investigate the role of social and structural factors in language change, we undertake a large-scale analysis of the frequencies of non-standard words in Reddit. Dissemination across many linguistic contexts is a predictor of success: words that appear in more linguistic contexts grow faster and survive longer. Furthermore, social dissemination plays a less important role in explaining word growth and decline than previously hypothesized.

Analyzing Correlated Evolution of Multiple Features Using Latent Representations
Yugo Murawaki 11:18–11:36

Statistical phylogenetic models have allowed the quantitative analysis of the evolution of a single categorical feature and a pair of binary features, but correlated evolution involving multiple discrete features is yet to be explored. Here we propose latent representation-based analysis in which (1) a sequence of discrete surface features is projected to a sequence of independent binary variables and (2) phylogenetic inference is performed on the latent space. In the experiments, we analyze the features of linguistic typology, with a special focus on the order of subject, object and verb. Our analysis suggests that languages sharing the same word order are not necessarily a coherent group but exhibit varying degrees of diachronic stability depending on other features.

Capturing Regional Variation with Distributed Place Representations and Geographic Retrofitting
Dirk Hovy and Christoph Purschke 11:36–11:54

Dialects are one of the main drivers of language variation, a major challenge for natural language processing tools. In most languages, dialects exist along a continuum, and are commonly discretized by combining the extent of several preselected linguistic variables. However, the selection of these variables is theory-driven and itself insensitive to change. We use Doc2Vec on a corpus of 16.8M anonymous online posts in the German-speaking area to learn continuous document representations of cities. These representations capture continuous regional linguistic distinctions, and can serve as input to downstream NLP tasks sensitive to regional variation. By incorporating geographic information via retrofitting and agglomerative clustering with structure, we recover dialect areas at various levels of granularity. Evaluating these clusters against an existing dialect map, we achieve a match of up to 0.77 V-score (harmonic mean of cluster completeness and homogeneity). Our results show that representation learning with retrofitting offers a robust general method to automatically expose dialectal differences and regional variation at a finer granularity than was previously possible.

Characterizing Interactions and Relationships between People
Farzana Rashid and Eduardo Blanco 11:54–12:12

This paper presents a set of dimensions to characterize the association between two people. We distinguish between interactions (when somebody refers to somebody in a conversation) and relationships (a sequence of interactions). We work with dialogue scripts from the TV show Friends, and do not impose any restrictions on the interactions and relationships. We introduce and analyze a new corpus, and present experimental results showing that the task can be automated.

Why Swear? Analyzing and Inferring the Intentions of Vulgar Expressions
Eric Holgate, Isabel Cachola, Daniel Preoţiuc-Pietro, and Junyi Jessy Li 12:12–12:30

Vulgar words are employed in language use for several different functions, ranging from expressing aggression to signaling group identity or the informality of the communication. This versatility of usage of a restricted set of words is challenging for downstream applications and has yet to be studied quantitatively or using natural language processing techniques. We introduce a novel data set of 7,800 tweets from users with known demographic traits where all instances of vulgar words are annotated with one of the six categories of vulgar word use. Using this data set, we present the first analysis of the pragmatic aspects of vulgarity and how they relate to social factors. We build a model able to predict the category of a vulgar word based on the immediate context it appears in with 67.4 macro F1 across six classes. Finally, we demonstrate the utility of modeling the type of vulgar word use in context by using this information to achieve state-of-the-art performance in hate speech detection on a benchmark data set.

Session 10E: Machine Learning

11:00–12:30

Hall 100

Chair: *chairname*

Is it Time to Swish? Comparing Deep Learning Activation Functions Across NLP tasks
Steffen Eger, Paul Youssef, and Iryna Gurevych

Activation functions play a crucial role in neural networks because they are the nonlinearities which have been attributed to the success story of deep learning. One of the currently most popular activation functions is ReLU, but several competitors have recently been proposed or 'discovered', including LReLU functions and swish. While most works compare newly proposed activation functions on few tasks (usually from image classification) and against few competitors (usually ReLU), we perform the first largescale comparison of 21 activation functions across eight different NLP tasks. We find that a largely unknown activation function performs most stably across all tasks, the so-called penalized tanh function. We also show that it can successfully replace the sigmoid and tanh gates in LSTM cells, leading to a 2 percentage point (pp) improvement over the standard choices on a challenging NLP task.

Hard Non-Monotonic Attention for Character-Level Transduction
Shijie Wu, Pamela Shapiro, and Ryan Cotterell

Character-level string-to-string transduction is an important component of various NLP tasks. The goal is to map an input string to an output string, where the strings may be of different lengths and have characters taken from different alphabets. Recent approaches have used sequence-to-sequence models with an attention mechanism to learn which parts of the input string the model should focus on during the generation of the output string. Both soft attention and hard monotonic attention have been used, but hard non-monotonic attention has only been used in other sequence modeling tasks and has required a stochastic approximation to compute the gradient. In this work, we introduce an exact, polynomial-time algorithm for marginalizing over the exponential number of non-monotonic alignments between two strings, showing that hard attention models can be viewed as neural reparameterizations of the classical IBM Model 1. We compare soft and hard non-monotonic attention experimentally and find that the exact algorithm significantly improves performance over the stochastic approximation and outperforms soft attention.

Speed Reading: Learning to Read ForBackward via Shuttle
Tsu-Jui Fu and Wei-Yun Ma

We present LSTM-Shuttle, which applies hu- man speed reading techniques to natural lan- guage processing tasks for accurate and effi- cient comprehension. In contrast to previous work, LSTM-Shuttle not only reads shuttling forward but also goes back. Shuttling for- ward enables high efficiency, and going back- ward gives the model a chance to recover lost information, ensuring better prediction. We evaluate LSTM-Shuttle on sentiment analysis, news classification, and cloze on IMDB, Rot- ten Tomatoes, AG, and Children's Book Test datasets. We show that LSTM-Shuttle predicts both better and more quickly. To demonstrate how LSTM-Shuttle actually behaves, we also analyze the shuttling operation and present a case study.

Modeling Localness for Self-Attention Networks
Baosong Yang, Zhaopeng Tu, Derek F. Wong, Fandong Meng, Lidia S. Chao, and Tong Zhang

Self-attention networks have proven to be of profound value for its strength of capturing global dependencies. In this work, we propose to model localness for self-attention networks, which enhances the ability of capturing useful local context. We cast localness modeling as a learnable Gaussian bias, which indicates the central and scope of the local region to be paid more attention. The bias is then incorporated into the original attention distribution to form a revised distribution. To maintain the strength of capturing long distance dependencies while enhance the ability of capturing short-range dependencies, we only apply localness modeling to lower layers of self-attention networks. Quantitative and qualitative analyses on Chinese-English and English-German translation tasks demonstrate the effectiveness and universality of the proposed approach.

Chargrid: Towards Understanding 2D Documents
Anoop R Katti, Christian Reisswig, Cordula Guder, Sebastian Brarda, Steffen Bickel, Johannes Höhne, and Jean Baptiste Faddoul

We introduce a novel type of text representation that preserves the 2D layout of a document. This is achieved by encoding each document page as a two-dimensional grid of characters. Based on this representation, we present a generic document understanding pipeline for structured documents. This pipeline makes use of a fully convolutional encoder-decoder network that predicts a segmentation mask and bounding boxes. We demonstrate its capabilities on an information extraction task from invoices and show that it significantly outperforms approaches based on sequential text or document images.

Training RNNs as Fast as CNNs
Tao Lei, Yu Zhang, Sida I. Wang, Hui Dai, and Yoav Artzi

Common recurrent neural architectures scale poorly due to the intrinsic difficulty in parallelizing their state computations. In this work, we propose the Simple Recurrent Unit (SRU), a light recurrent unit that balances model capacity and scalability. SRU is designed to provide expressive recurrence, enable highly parallelized implementation, and comes with careful initialization to facilitate training of deep models. We demonstrate the effectiveness of SRU on multiple NLP tasks. SRU achieves 5—9x speed-up over cuDNN-optimized LSTM on classification and question answering datasets, and delivers stronger results than LSTM and convolutional models. We also obtain an average of 0.7 BLEU improvement over the Transformer model (Vaswani et al., 2017) on translation by incorporating SRU into the architecture.

NPRF: A Neural Pseudo Relevance Feedback Framework for Ad-hoc Information Retrieval
Canjia Li, Yingfei Sun, Ben He, Le Wang, Kai Hui, Andrew Yates, Le Sun, and Jungang Xu

Pseudo relevance feedback (PRF) is commonly used to boost the performance of traditional information retrieval (IR) models by using top-ranked documents to identify and weight new query terms, thereby reducing the effect of query-document vocabulary mismatches. While neural retrieval models have recently demonstrated strong results for ad-hoc retrieval, combining them with PRF is not straightforward due to incompatibilities between existing PRF approaches and neural architectures. To bridge this gap, we propose an end-to-end neural PRF framework that can be used with existing neural IR models by embedding different neural models as building blocks. Extensive experiments on two standard test collections confirm the effectiveness of the proposed NPRF framework in improving the performance of two state-of-the-art neural IR models.

Co-Stack Residual Affinity Networks with Multi-level Attention Refinement for Matching Text Sequences
Yi Tay, Anh Tuan Luu, and Siu Cheung Hui

Learning a matching function between two text sequences is a long standing problem in NLP research. This task enables many potential applications such as question answering and paraphrase identification. This paper proposes Co-Stack Residual Affinity Networks (CSRAN), a new and universal neural architecture for this problem. CSRAN is a deep architecture, involving stacked (multi-layered) recurrent encoders. Stacked/Deep architectures are traditionally difficult to train, due to the inherent weaknesses such as difficulty with feature propagation and vanishing gradients. CSRAN incorporates two novel components to take advantage of the stacked architecture. Firstly, it introduces a new bidirectional alignment mechanism that learns affinity weights by fusing sequence pairs across stacked hierarchies. Secondly, it leverages a multi-level attention refinement component between stacked recurrent layers. The key intuition is that, by leveraging information across all network hierarchies, we can not only improve gradient flow but also improve overall performance. We conduct extensive experiments on six well-studied text sequence matching datasets, achieving state-of-the-art performance on all.

Spherical Latent Spaces for Stable Variational Autoencoders
Jiacheng Xu and Greg Durrett

A hallmark of variational autoencoders (VAEs) for text processing is their combination of powerful encoder-decoder models, such as LSTMs, with simple latent distributions, typically multivariate Gaussians. These models pose a difficult optimization problem: there is an especially bad local optimum where the variational posterior always equals the prior and the model does not use the latent variable at all, a kind of "collapse" which is encouraged by the KL divergence term of the objective. In this work, we experiment with another choice of latent distribution, namely the von Mises-Fisher (vMF) distribution, which places mass on the surface of the unit hypersphere. With this choice of prior and posterior, the KL divergence term now only depends on the variance of the vMF distribution, giving us the ability to treat it as a fixed hyperparameter. We show that doing so not only averts the KL collapse, but consistently gives better likelihoods than Gaussians across a range of modeling conditions, including recurrent language modeling and bag-of-words document modeling. An analysis of the properties of our vMF representations shows that they learn richer and more nuanced structures in their latent representations than their Gaussian counterparts.

Learning Universal Sentence Representations with Mean-Max Attention Autoencoder

Minghua Zhang, Yunfang Wu, Weikang Li, and Wei Li

In order to learn universal sentence representations, previous methods focus on complex recurrent neural networks or supervised learning. In this paper, we propose a mean-max attention autoencoder (mean-max AAE) within the encoder-decoder framework. Our autoencoder rely entirely on the MultiHead self-attention mechanism to reconstruct the input sequence. In the encoding we propose a mean-max strategy that applies both mean and max pooling operations over the hidden vectors to capture diverse information of the input. To enable the information to steer the reconstruction process dynamically, the decoder performs attention over the mean-max representation. By training our model on a large collection of unlabelled data, we obtain high-quality representations of sentences. Experimental results on a broad range of 10 transfer tasks demonstrate that our model outperforms the state-of-the-art unsupervised single methods, including the classical skip-thoughts [**kiros2015skip**] and the advanced skip-thoughts+LN model [**ba2016layer**]. Furthermore, compared with the traditional recurrent neural network, our mean-max AAE greatly reduce the training time. [2]

Word Mover's Embedding: From Word2Vec to Document Embedding

Lingfei Wu, Ian En-Hsu Yen, Kun Xu, Fangli Xu, Avinash Balakrishnan, Pin-Yu Chen, Pradeep Ravikumar, and Michael J. Witbrock

While the celebrated Word2Vec technique yields semantically rich representations for individual words, there has been relatively less success in extending to generate unsupervised sentences or documents embeddings. Recent work has demonstrated that a distance measure between documents called Word Mover's Distance (WMD) that aligns semantically similar words, yields unprecedented KNN classification accuracy. However, WMD is expensive to compute, and it is hard to extend its use beyond a KNN classifier. In this paper, we propose the Word Mover's Embedding (WME), a novel approach to building an unsupervised document (sentence) embedding from pre-trained word embeddings. In our experiments on 9 benchmark text classification datasets and 22 textual similarity tasks, the proposed technique consistently matches or outperforms state-of-the-art techniques, with significantly higher accuracy on problems of short length.

Multilingual Clustering of Streaming News

Sebastião Miranda, Arturs Znotins, Shay B. Cohen, and Guntis Barzdins

Clustering news across languages enables efficient media monitoring by aggregating articles from multilingual sources into coherent stories. Doing so in an online setting allows scalable processing of massive news streams. To this end, we describe a novel method for clustering an incoming stream of multilingual documents into monolingual and crosslingual clusters. Unlike typical clustering approaches that report results on datasets with a small and known number of labels, we tackle the problem of discovering an ever growing number of cluster labels in an online fashion, using real news datasets in multiple languages. In our formulation, the monolingual clusters group together documents while the crosslingual clusters group together monolingual clusters, one per language that appears in the stream. Our method is simple to implement, computationally efficient and produces state-of-the-art results on datasets in German, English and Spanish.

Multi-Task Label Embedding for Text Classification

Honglun Zhang, Liqiang Xiao, Wenqing Chen, Yongkun Wang, and Yaohui Jin

Multi-task learning in text classification leverages implicit correlations among related tasks to extract common features and yield performance gains. However, a large body of previous work treats labels of each task as independent and meaningless one-hot vectors, which cause a loss of potential label information. In this paper, we propose Multi-Task Label Embedding to convert labels in text classification into semantic vectors, thereby turning the original tasks into vector matching tasks. Our model utilizes semantic correlations among tasks and makes it convenient to scale or transfer when new tasks are involved. Extensive experiments on five benchmark datasets for text classification show that our model can effectively improve the performances of related tasks with semantic representations of labels and additional information from each other.

[2] Our code is publicly available at `https://github.com/Zminghua/SentEncoding`.

Semantic-Unit-Based Dilated Convolution for Multi-Label Text Classification
Junyang Lin, Qi Su, Pengcheng Yang, Shuming Ma, and Xu SUN

We propose a novel model for multi-label text classification, which is based on sequence-to-sequence learning. The model generates higher-level semantic unit representations with multi-level dilated convolution as well as a corresponding hybrid attention mechanism that extracts both the information at the word-level and the level of the semantic unit. Our designed dilated convolution effectively reduces dimension and supports an exponential expansion of receptive fields without loss of local information, and the attention-over-attention mechanism is able to capture more summary relevant information from the source context. Results of our experiments show that the proposed model has significant advantages over the baseline models on the dataset RCV1-V2 and Ren-CECps, and our analysis demonstrates that our model is competitive to the deterministic hierarchical models and it is more robust to classifying low-frequency labels

MCapsNet: Capsule Network for Text with Multi-Task Learning
Liqiang Xiao, Honglun Zhang, Wenqing Chen, Yongkun Wang, and Yaohui Jin

Multi-task learning has an ability to share the knowledge among related tasks and implicitly increase the training data. However, it has long been frustrated by the interference among tasks. This paper investigates the performance of capsule network for text, and proposes a capsule-based multi-task learning architecture, which is unified, simple and effective. With the advantages of capsules for feature clustering, proposed task routing algorithm can cluster the features for each task in the network, which helps reduce the interference among tasks. Experiments on six text classification datasets demonstrate the effectiveness of our models and their characteristics for feature clustering.

Uncertainty-aware generative models for inferring document class prevalence
Katherine Keith and Brendan O'Connor

Prevalence estimation is the task of inferring the relative frequency of classes of unlabeled examples in a group—for example, the proportion of a document collection with positive sentiment. Previous work has focused on aggregating and adjusting discriminative individual classifiers to obtain prevalence point estimates. But imperfect classifier accuracy ought to be reflected in uncertainty over the predicted prevalence for scientifically valid inference. In this work, we present (1) a generative probabilistic modeling approach to prevalence estimation, and (2) the construction and evaluation of prevalence confidence intervals; in particular, we demonstrate that an off-the-shelf discriminative classifier can be given a generative re-interpretation, by backing out an implicit individual-level likelihood function, which can be used to conduct fast and simple group-level Bayesian inference. Empirically, we demonstrate our approach provides better confidence interval coverage than an alternative, and is dramatically more robust to shifts in the class prior between training and testing.

Challenges of Using Text Classifiers for Causal Inference
Zach Wood-Doughty, Ilya Shpitser, and Mark Dredze

Causal understanding is essential for many kinds of decision-making, but causal inference from observational data has typically only been applied to structured, low-dimensional datasets. While text classifiers produce low-dimensional outputs, their use in causal inference has not previously been studied. To facilitate causal analyses based on language data, we consider the role that text classifiers can play in causal inference through established modeling mechanisms from the causality literature on missing data and measurement error. We demonstrate how to conduct causal analyses using text classifiers on simulated and Yelp data, and discuss the opportunities and challenges of future work that uses text data in causal inference.

Direct Output Connection for a High-Rank Language Model
Sho Takase, Jun Suzuki, and Masaaki Nagata

This paper proposes a state-of-the-art recurrent neural network (RNN) language model that combines probability distributions computed not only from a final RNN layer but also middle layers. This method raises the expressive power of a language model based on the matrix factorization interpretation of language modeling introduced by Yang et al. (2018). Our proposed method improves the current state-of-the-art language model and achieves the best score on the Penn Treebank and WikiText-2, which are the standard benchmark datasets. Moreover, we indicate our proposed method contributes to application tasks: machine translation and headline generation.

Disfluency Detection using Auto-Correlational Neural Networks
Paria Jamshid Lou, Peter Anderson, and Mark Johnson

In recent years, the natural language processing community has moved away from task-specific feature engineering, i.e., researchers discovering ad-hoc feature representations for various tasks, in favor of general-purpose methods that learn the input representation by themselves. However, state-of-the-art approaches to disfluency detection in spontaneous speech transcripts currently still depend on an array of hand-crafted features, and other representations derived from the output of pre-existing systems such as language models or dependency parsers. As an alternative, this paper proposes a simple yet effective model for automatic disfluency detection, called an auto-correlational neural network (ACNN). The model uses a convolutional neural network (CNN) and augments it with a new auto-correlation operator at the lowest layer that can capture the kinds of "rough copy" dependencies that are characteristic of repair disfluencies in speech. In experiments, the ACNN model outperforms the baseline CNN on a disfluency detection task with a 5% increase in f-score, which is close to the previous best result on this task.

Pyramidal Recurrent Unit for Language Modeling
Sachin Mehta, Rik Koncel-Kedziorski, Mohammad Rastegari, and Hannaneh Hajishirzi

LSTMs are powerful tools for modeling contextual information, as evidenced by their success at the task of language modeling. However, modeling contexts in very high dimensional space can lead to poor generalizability. We introduce the Pyramidal Recurrent Unit (PRU), which enables learning representations in high dimensional space with more generalization power and fewer parameters. PRUs replace the linear transformation in LSTMs with more sophisticated interactions such as pyramidal or grouped linear transformations. This architecture gives strong results on word-level language modeling while reducing parameters significantly. In particular, PRU improves the perplexity of a recent state-of-the-art language model by up to 1.3 points while learning 15-20% fewer parameters. For similar number of model parameters, PRU outperforms all previous RNN models that exploit different gating mechanisms and transformations. We provide a detailed examination of the PRU and its behavior on the language modeling tasks. Our code is open-source and available at https://sacmehta.github.io/PRU/.

On Tree-Based Neural Sentence Modeling
Haoyue Shi, Hao Zhou, Jiaze Chen, and Lei Li

Neural networks with tree-based sentence encoders have shown better results on many downstream tasks. Most of existing tree-based encoders adopt syntactic parsing trees as the explicit structure prior. To study the effectiveness of different tree structures, we replace the parsing trees with trivial trees (i.e., binary balanced tree, left-branching tree and right-branching tree) in the encoders. Though trivial trees contain no syntactic information, those encoders get competitive or even better results on all of the ten downstream tasks we investigated. This surprising result indicates that explicit syntax guidance may not be the main contributor to the superior performances of tree-based neural sentence modeling. Further analysis show that tree modeling gives better results when crucial words are closer to the final representation. Additional experiments give more clues on how to design an effective tree-based encoder. Our code is open-source and available at https://github.com/ExplorerFreda/TreeEnc.

Language Modeling with Sparse Product of Sememe Experts
Yihong Gu, Jun Yan, Hao Zhu, Zhiyuan Liu, Ruobing Xie, Maosong Sun, Fen Lin, and Leyu Lin

Most language modeling methods rely on large-scale data to statistically learn the sequential patterns of words. In this paper, we argue that words are atomic language units but not necessarily atomic semantic units. Inspired by HowNet, we use sememes, the minimum semantic units in human languages, to represent the implicit semantics behind words for language modeling, named Sememe-Driven Language Model (SDLM). More specifically, to predict the next word, SDLM first estimates the sememe distribution given textual context. Afterwards, it regards each sememe as a distinct semantic expert, and these experts jointly identify the most probable senses and the corresponding word. In this way, SDLM enables language models to work beyond word-level manipulation to fine-grained sememe-level semantics, and offers us more powerful tools to fine-tune language models and improve the interpretability as well as the robustness of language models. Experiments on language modeling and the downstream application of headline generation demonstrate the significant effectiveness of SDLM.

TACL-1379
TACL-1379

Neural architectures are prominent in the construction of language models (LMs). However, word-level prediction is typically agnostic of subword-level information (characters and character sequences) and operates over a closed vocabulary, consisting of a limited word set. Indeed, while subword-aware models boost performance across a variety of NLP tasks, previous work did not evaluate the ability of these models to assist next-word prediction in language modeling tasks. Such subword-level informed models should be particularly effective for morphologically-rich languages (MRLs) that exhibit high type-to-token ratios. In this work, we present a large-scale LM study on 50 typologically diverse languages covering a wide variety of morphological systems, and offer new LM benchmarks to the community, while considering subword-level information. The main technical contribution of our work is a novel method for injecting subword-level information into semantic word vectors, integrated into the neural language modeling training, to facilitate word-level prediction. We conduct experiments in the LM setting where the number of infrequent words is large, and demonstrate strong perplexity gains across our 50 languages, especially for morphologically-rich languages. Our code and data sets are publicly available.

TACL-1396
TACL-1396

A context-aware language model uses location, user and/or domain metadata (context) to adapt its predictions. In neural language models, context information is typically represented as an embedding and it is given to the RNN as an additional input, which has been shown to be useful in many applications. We introduce a more powerful mechanism for using context to adapt an RNN by letting the context vector control a low-rank transformation of the recurrent layer weight matrix. Experiments show that allowing a greater fraction of the model parameters to be adjusted has benefits in terms of perplexity and classification for several different types of context.

Siamese Network-Based Supervised Topic Modeling
Minghui Huang, Yanghui Rao, Yuwei Liu, Haoran Xie, and Fu Lee Wang

Label-specific topics can be widely used for supporting personality psychology, aspect-level sentiment analysis, and cross-domain sentiment classification. To generate label-specific topics, several supervised topic models which adopt likelihood-driven objective functions have been proposed. However, it is hard for them to get a precise estimation on both topic discovery and supervised learning. In this study, we propose a supervised topic model based on the Siamese network, which can trade off label-specific word distributions with document-specific label distributions in a uniform framework. Experiments on real-world datasets validate that our model performs competitive in topic discovery quantitatively and qualitatively. Furthermore, the proposed model can effectively predict categorical or real-valued labels for new documents by generating word embeddings from a label-specific topical space.

GraphBTM: Graph Enhanced Autoencoded Variational Inference forBiterm Topic Model
Qile Zhu, Zheng Feng, and Xiaolin Li

Discovering the latent topics within texts has been a fundamental task for many applications. However, conventional topic models suffer different problems in different settings. The Latent Dirichlet Allocation (LDA) may not work well for short texts due to the data sparsity (i.e. the sparse word co-occurrence patterns in short documents). The Biterm Topic Model (BTM) learns topics by modeling the word-pairs named biterms in the whole corpus. This assumption is very strong when documents are long with rich topic information and do not exhibit the transitivity of biterms. In this paper, we propose a novel way called GraphBTM to represent biterms as graphs and design a Graph Convolutional Networks (GCNs) with residual connections to extract transitive features from biterms. To overcome the data sparsity of LDA and the strong assumption of BTM, we sample a fixed number of documents to form a mini-corpus as a sample. We also propose a dataset called All News extracted from 15 news publishers, in which documents are much longer than 20 Newsgroups. We present an amortized variational inference method for GraphBTM. Our method generates more coherent topics compared with previous approaches. Experiments show that the sampling strategy improves performance by a large margin.

Modeling Online Discourse with Coupled Distributed Topics
Akshay Srivatsan, Zachary Wojtowicz, and Taylor Berg-Kirkpatrick

In this paper, we propose a deep, globally normalized topic model that incorporates structural relationships connecting documents in socially generated corpora, such as online forums. Our model (1) captures discursive interactions along observed reply links in addition to traditional topic information, and (2) incorporates latent distributed representations arranged in a deep architecture, which enables a GPU-based mean-field inference procedure that scales efficiently to large data. We apply our model to a new social media dataset consisting of 13M comments mined from the popular internet forum Reddit, a domain that poses significant challenges to models that do not account for relationships connecting user comments. We evaluate against existing methods across multiple metrics including perplexity and metadata prediction, and qualitatively analyze the learned interaction patterns.

Learning Disentangled Representations of Texts with Application to Biomedical Abstracts
Sarthak Jain, Edward Banner, Jan-Willem van de Meent, Iain J Marshall, and Byron C. Wallace

We propose a method for learning disentangled representations of texts that code for distinct and complementary aspects, with the aim of affording efficient model transfer and interpretability. To induce disentangled embeddings, we propose an adversarial objective based on the (dis)similarity between triplets of documents with respect to specific aspects. Our motivating application is embedding biomedical abstracts describing clinical trials in a manner that disentangles the populations, interventions, and outcomes in a given trial. We show that our method learns representations that encode these clinically salient aspects, and that these can be effectively used to perform aspect-specific retrieval. We demonstrate that the approach generalizes beyond our motivating application in experiments on two multi-aspect review corpora.

Multi-Source Domain Adaptation with Mixture of Experts
Jiang Guo, Darsh Shah, and Regina Barzilay

We propose a mixture-of-experts approach for unsupervised domain adaptation from multiple sources. The key idea is to explicitly capture the relationship between a target example and different source domains. This relationship, expressed by a point-to-set metric, determines how to combine predictors trained on various domains. The metric is learned in an unsupervised fashion using meta-training. Experimental results on sentiment analysis and part-of-speech tagging demonstrate that our approach consistently outperforms multiple baselines and can robustly handle negative transfer.

Sisyphus, a Workflow Manager Designed for Machine Translation and Automatic Speech Recognition
Jan-Thorsten Peter, Eugen Beck, and Hermann Ney

Training and testing many possible parameters or model architectures of state-of-the-art machine translation or automatic speech recognition system is a cumbersome task. They usually require a long pipeline of commands reaching from pre-processing the training data to post-processing and evaluating the output. This paper introduces Sisyphus, a tool that aims at managing scientific experiments in an efficient way. After defining the workflow for a given task, Sisyphus runs all required steps and ensures that all commands finish successfully. It avoids unnecessary computations by reusing tasks that are needed for multiple parts of the workflow and saves the user time by determining the order in which the tasks are to be performed. Since the program and workflow are written in Python they can be easily extended to contain arbitrary code. This makes it possible to use the rich collection of Python tools for editing, debugging, and documentation. It only has few requirements on the underlying server or cluster, and has been successfully tested in many large scale setups and can handle thousands of tasks inside the workflow.

APLenty: annotation tool for creating high-quality datasets using active and proactive learning
Minh-Quoc Nghiem and Sophia Ananiadou

In this paper, we present APLenty, an annotation tool for creating high-quality sequence labeling datasets using active and proactive learning. A major innovation of our tool is the integration of automatic annotation with active learning and proactive learning. This makes the task of creating labeled datasets easier, less time-consuming and requiring less human effort. APLenty is highly flexible and can be adapted to various other tasks.

KT-Speech-Crawler: Automatic Dataset Construction for Speech Recognition from YouTube Videos
Egor Lakomkin, Sven Magg, Cornelius Weber, and Stefan Wermter

We describe KT-Speech-Crawler: an approach for automatic dataset construction for speech recognition by crawling YouTube videos. We outline several filtering and post-processing steps, which extract samples that can be used for training end-to-end neural speech recognition systems. In our experiments, we demonstrate that a single-core version of the crawler can obtain around 150 hours of transcribed speech within a day, containing an estimated 3.5% word error rate in the transcriptions. Automatically collected samples contain reading and spontaneous speech recorded in various conditions including background noise and music, distant microphone recordings, and a variety of accents and reverberation. When training a deep neural network on speech recognition, we observed around 40% word error rate reduction on the Wall Street Journal dataset by integrating 200 hours of the collected samples into the training set.

Term Set Expansion based NLP Architect by Intel AI Lab
Jonathan Mamou, Oren Pereg, Moshe Wasserblat, Alon Eirew, Yael Green, Shira Guskin, Peter Izsak, and Daniel Korat
We present SetExpander, a corpus-based system for expanding a seed set of terms into a more complete set of terms that belong to the same semantic class. SetExpander implements an iterative end-to-end workflow. It enables users to easily select a seed set of terms, expand it, view the expanded set, validate it, re-expand the validated set and store it, thus simplifying the extraction of domain-specific fine-grained semantic classes. SetExpander has been used successfully in real-life use cases including integration into an automated recruitment system and an issues and defects resolution system.

Session 11 Overview – Sunday, November 4, 2018

Oral tracks

	Track A *Analyzing Models* Gold Hall	**Track B** *Sentiment II* Copper Hall	**Track C** *Machine Translation IV* Silver Hall	**Track D** *QA / Knowledge Graphs* Hall 100
13:45	A Neural Model of Adaptation in Reading *Schijndel and Linzen*	Joint Learning for Targeted Sentiment Analysis *Ma, Li, and WANG*	A Tree-based Decoder for Neural Machine Translation *Wang, Pham, Yin, and Neubig*	FewRel: A Large-Scale Supervised Few-shot Relation Classification Dataset with State-of-the-Art Evaluation *Han, Zhu, Yu, Wang, Yao, Liu, and Sun*
13:57	Understanding Deep Learning Performance through an Examination of Test Set Difficulty: A Psychometric Case Study *Lalor, Wu, Munkhdalai, and Yu*	Revisiting the Importance of Encoding Logic Rules in Sentiment Classification *Krishna, Jyothi, and Iyyer*	Greedy Search with Probabilistic N-gram Matching for Neural Machine Translation *Shao, Chen, and Feng*	A strong baseline for question relevancy ranking *Gonzalez, Augenstein, and Søgaard*
14:09	Lexicosyntactic inference in neural models *White, Rudinger, Rawlins, and Van Durme*	A Co-attention Neural Network Model for Emotion Cause Analysis with Emotional Context Awareness *Li, Song, Feng, Wang, and Zhang*	Exploring Recombination for Efficient Decoding of Neural Machine Translation *Zhang, Wang, Utiyama, Sumita, and Zhao*	Learning Sequence Encoders for Temporal Knowledge Graph Completion *Garcia-Duran, Dumančić, and Niepert*
14:21	Dual Fixed-Size Ordinally Forgetting Encoding (FOFE) for Competitive Neural Language Models *Watcharawittayakul, Xu, and Jiang*	Modeling Empathy and Distress in Reaction to News Stories *Buechel, Buffone, Slaff, Ungar, and Sedoc*	Has Neural Machine Translation Achieved Human Parity? A Case for Document-level Evaluation *Läubli, Sennrich, and Volk*	Similar but not the Same - Word Sense Disambiguation Improves Event Detection via Neural Representation Matching *Lu and Nguyen*
14:33	The importance of Being Recurrent for Modeling Hierarchical Structure *Tran, Bisazza, and Monz*	Interpretable Emoji Prediction via Multi-Attention LSTMs *Barbieri, Espinosa Anke, Camacho-Collados, Schockaert, and Saggion*	Automatic Reference-Based Evaluation of Pronoun Translation Misses the Point *Guillou and Hardmeier*	Learning Word Representations with Cross-Sentence Dependencyfor End-to-End Co-reference Resolution *Luo and Glass*

Poster tracks

13:45–14:45

Track D: *Short Posters IV*

Hall 100

Parallel Session 11

Session 11A: Analyzing Models

Gold Hall Chair: *chairname*

A Neural Model of Adaptation in Reading
Marten van Schijndel and Tal Linzen 13:45–13:57

It has been argued that humans rapidly adapt their lexical and syntactic expectations to match the statistics of the current linguistic context. We provide further support to this claim by showing that the addition of a simple adaptation mechanism to a neural language model improves our predictions of human reading times compared to a non-adaptive model. We analyze the performance of the model on controlled materials from psycholinguistic experiments and show that it adapts not only to lexical items but also to abstract syntactic structures.

Understanding Deep Learning Performance through an Examination of Test Set Difficulty: A Psychometric Case Study
John Lalor, Hao Wu, Tsendsuren Munkhdalai, and Hong Yu 13:57–14:09

Interpreting the performance of deep learning models beyond test set accuracy is challenging. Characteristics of individual data points are often not considered during evaluation, and each data point is treated equally. In this work we examine the impact of a test set question's difficulty to determine if there is a relationship between difficulty and performance. We model difficulty using well-studied psychometric methods on human response patterns. Experiments on Natural Language Inference (NLI) and Sentiment Analysis (SA) show that the likelihood of answering a question correctly is impacted by the question's difficulty. In addition, as DNNs are trained on larger datasets easy questions start to have a higher probability of being answered correctly than harder questions.

Lexicosyntactic inference in neural models
Aaron Steven White, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme 14:09–14:21

We investigate neural models' ability to capture lexicosyntactic inferences: inferences triggered by the interaction of lexical and syntactic information. We take the task of event factuality prediction as a case study and build a factuality judgment dataset for all English clause-embedding verbs in various syntactic contexts. We use this dataset, which we make publicly available, to probe the behavior of current state-of-the-art neural systems, showing that these systems make certain systematic errors that are clearly visible through the lens of factuality prediction.

Dual Fixed-Size Ordinally Forgetting Encoding (FOFE) for Competitive Neural Language Models
Sedtawut Watcharawittayakul, Mingbin Xu, and Hui Jiang 14:21–14:33

In this paper, we propose a new approach to employ the fixed-size ordinally-forgetting encoding (FOFE) (Zhang et al., 2015b) in neural languages modelling, called dual-FOFE. The main idea of dual-FOFE is that it allows to use two different forgetting factors so that it can avoid the trade-off in choosing either a small or large values for the single forgetting factor. In our experiments, we have compared the dual-FOFE based neural network language models (NNLM) against the original FOFE counterparts and various traditional NNLMs. Our results on the challenging Google Billion word corpus show that both FOFE and dual FOFE yield very strong performance while significantly reducing the computational complexity over other NNLMs. Furthermore, the proposed dual-FOFE method further gives over 10% improvement in perplexity over the original FOFE model.

The importance of Being Recurrent for Modeling Hierarchical Structure
Ke Tran, Arianna Bisazza, and Christof Monz 14:33–14:45

Recent work has shown that recurrent neural networks (RNNs) can implicitly capture and exploit hierarchical information when trained to solve common natural language processing tasks (Blevins et al., 2018) such as language modeling (Linzen et al., 2016; Gulordava et al., 2018) and neural machine translation (Shi et al., 2016). In contrast, the ability to model structured data with non-recurrent neural networks has received little attention despite their success in many NLP tasks (Gehring et al., 2017; Vaswani et al., 2017). In this work, we compare the two architectures—recurrent versus non-recurrent—with respect to their ability to model hierarchical structure and find that recurrency is indeed important for this purpose. The code and data used in our experiments is available at https://anonymized

Session 11B: Sentiment II

Copper Hall Chair: *chairname*

Joint Learning for Targeted Sentiment Analysis
Dehong Ma, Sujian Li, and Houfeng WANG 13:45–13:57

Targeted sentiment analysis (TSA) aims at extracting targets and classifying their sentiment classes. Previous works only exploit word embeddings as features and do not explore more potentials of neural networks when jointly learning the two tasks. In this paper, we carefully design the hierarchical stack bidirectional gated recurrent units (HSBi-GRU) model to learn abstract features for both tasks, and we propose a HSBi-GRU based joint model which allows the target label to have influence on their sentiment label. Experimental results on two datasets show that our joint learning model can outperform other baselines and demonstrate the effectiveness of HSBi-GRU in learning abstract features.

Revisiting the Importance of Encoding Logic Rules in Sentiment Classification
Kalpesh Krishna, Preethi Jyothi, and Mohit Iyyer 13:57–14:09

We analyze the performance of different sentiment classification models on syntactically complex inputs like A-but-B sentences. The first contribution of this analysis addresses reproducible research: to meaningfully compare different models, their accuracies must be averaged over far more random seeds than what has traditionally been reported. With proper averaging in place, we notice that the distillation model described in Hu et al. (2016), which incorporates explicit logic rules for sentiment classification, is ineffective. In contrast, using contextualized ELMo embeddings (Peters et al., 2018a) instead of logic rules yields significantly better performance. Additionally, we provide analysis and visualizations that demonstrate ELMo's ability to implicitly learn logic rules. Finally, a crowdsourced analysis reveals how ELMo outperforms baseline models even on sentences with ambiguous sentiment labels.

A Co-attention Neural Network Model for Emotion Cause Analysis with Emotional Context Awareness
Xiangju Li, Kaisong Song, Shi Feng, Daling Wang, and Yifei Zhang 14:09–14:21

Emotion cause analysis has been a key topic in natural language processing. Existing methods ignore the contexts around the emotion word which can provide an emotion cause clue. Meanwhile, the clauses in a document play different roles on stimulating a certain emotion, depending on their content relevance. Therefore, we propose a co-attention neural network model for emotion cause analysis with emotional context awareness. The method encodes the clauses with a co-attention based bi-directional long short-term memory into high-level input representations, which are further fed into a convolutional layer for emotion cause analysis. Experimental results show that our approach outperforms the state-of-the-art baseline methods.

Modeling Empathy and Distress in Reaction to News Stories
Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and Joao Sedoc 14:21–14:33

Computational detection and understanding of empathy is an important factor in advancing human-computer interaction. Yet to date, text-based empathy prediction has the following major limitations: It underestimates the psychological complexity of the phenomenon, adheres to a weak notion of ground truth where empathic states are ascribed by third parties, and lacks a shared corpus. In contrast, this contribution presents the first publicly available gold standard for empathy prediction. It is constructed using a novel annotation methodology which reliably captures empathy assessments by the writer of a statement using multi-item scales. This is also the first computational work distinguishing between multiple forms of empathy, empathic concern, and personal distress, as recognized throughout psychology. Finally, we present experimental results for three different predictive models, of which a CNN performs the best.

Interpretable Emoji Prediction via Multi-Attention LSTMs
Francesco Barbieri, Luis Espinosa Anke, Jose Camacho-Collados, Steven Schockaert, and Horacio Saggion 14:33–14:45

Human language has evolved towards newer forms of communication such as social media, where emojis (i.e., ideograms bearing a visual meaning) play a key role. While there is an increasing body of work aimed at the computational modeling of emoji semantics, there is currently little understanding about what makes a computational model represent or predict a given emoji in a certain way. In this paper we propose a label-wise attention mechanism with which we attempt to better understand the nuances underlying emoji prediction. In addition to advantages in terms of interpretability, we show that our proposed architecture improves over standard baselines in emoji prediction, and does particularly well when predicting infrequent emojis.

Session 11C: Machine Translation IV

Silver Hall Chair: *chairname*

A Tree-based Decoder for Neural Machine Translation
Xinyi Wang, Hieu Pham, Pengcheng Yin, and Graham Neubig 13:45–13:57

Recent advances in Neural Machine Translation (NMT) show that adding syntactic information to NMT systems can improve the quality of their translations. Most existing work utilizes some specific types of linguistically-inspired tree structures, like constituency and dependency parse trees. This is often done via a standard RNN decoder that operates on a linearized target tree structure. However, it is an open question of what specific linguistic formalism, if any, is the best structural representation for NMT. In this paper, we (1) propose an NMT model that can naturally generate the topology of an arbitrary tree structure on the target side, and (2) experiment with various target tree structures. Our experiments show the surprising result that our model delivers the best improvements with balanced binary trees constructed without any linguistic knowledge; this model outperforms standard seq2seq models by up to 2.1 BLEU points, and other methods for incorporating target-side syntax by up to 0.7 BLEU.

Greedy Search with Probabilistic N-gram Matching for Neural Machine Translation
Chenze Shao, Xilin Chen, and Yang Feng 13:57–14:09

Neural machine translation (NMT) models are usually trained with the word-level loss using the teacher forcing algorithm, which not only evaluates the translation improperly but also suffers from exposure bias. Sequence-level training under the reinforcement framework can mitigate the problems of the word-level loss, but its performance is unstable due to the high variance of the gradient estimation. On these grounds, we present a method with a differentiable sequence-level training objective based on probabilistic n-gram matching which can avoid the reinforcement framework. In addition, this method performs greedy search in the training which uses the predicted words as context just as at inference to alleviate the problem of exposure bias. Experiment results on the NIST Chinese-to-English translation tasks show that our method significantly outperforms the reinforcement-based algorithms and achieves an improvement of 1.5 BLEU points on average over a strong baseline system.

Exploring Recombination for Efficient Decoding of Neural Machine Translation
Zhisong Zhang, Rui Wang, Masao Utiyama, Eiichiro Sumita, and Hai Zhao 14:09–14:21

In Neural Machine Translation (NMT), the decoder can capture the features of the entire prediction history with neural connections and representations. This means that partial hypotheses with different prefixes will be regarded differently no matter how similar they are. However, this might be inefficient since some partial hypotheses can contain only local differences that will not influence future predictions. In this work, we introduce recombination in NMT decoding based on the concept of the "equivalence" of partial hypotheses. Heuristically, we use a simple n-gram suffix based equivalence function and adapt it into beam search decoding. Through experiments on large-scale Chinese-to-English and English-to-Germen translation tasks, we show that the proposed method can obtain similar translation quality with a smaller beam size, making NMT decoding more efficient.

Has Neural Machine Translation Achieved Human Parity? A Case for Document-level Evaluation
Samuel Läubli, Rico Sennrich, and Martin Volk 14:21–14:33

Recent research suggests that neural machine translation achieves parity with professional human translation on the WMT Chinese–English news translation task. We empirically test this claim with alternative evaluation protocols, contrasting the evaluation of single sentences and entire documents. In a pairwise ranking experiment, human raters assessing adequacy and fluency show a stronger preference for human over machine translation when evaluating documents as compared to isolated sentences. Our findings emphasise the need to shift towards document-level evaluation as machine translation improves to the degree that errors which are hard or impossible to spot at the sentence-level become decisive in discriminating quality of different translation outputs.

Automatic Reference-Based Evaluation of Pronoun Translation Misses the Point
Liane Guillou and Christian Hardmeier 14:33–14:45

We compare the performance of the APT and AutoPRF metrics for pronoun translation against a manually annotated dataset com- prising human judgements as to the correct- ness of translations of the PROTEST test suite. Although there is some correlation with the human judgements, a range of issues limit the performance of the automated met- rics. Instead, we recommend the use of semi- automatic metrics and test suites in place of fully automatic metrics.

Session 11D: QA / Knowledge Graphs

Hall 100 Chair: *chairname*

FewRel: A Large-Scale Supervised Few-shot Relation Classification Dataset with State-of-the-Art Evaluation
Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun 13:45–13:57

We present a Few-Shot Relation Classification Dataset (dataset), consisting of 70, 000 sentences on 100 relations derived from Wikipedia and annotated by crowdworkers. The relation of each sentence is first recognized by distant supervision methods, and then filtered by crowdworkers. We adapt the most recent state-of-the-art few-shot learning methods for relation classification and conduct thorough evaluation of these methods. Empirical results show that even the most competitive few-shot learning models struggle on this task, especially as compared with humans. We also show that a range of different reasoning skills are needed to solve our task. These results indicate that few-shot relation classification remains an open problem and still requires further research. Our detailed analysis points multiple directions for future research.

A strong baseline for question relevancy ranking
Ana Gonzalez, Isabelle Augenstein, and Anders Søgaard 13:57–14:09

The best systems at the SemEval-16 and SemEval-17 community question answering shared tasks – a task that amounts to question relevancy ranking – involve complex pipelines and manual feature engineering. Despite this, many of these still fail at beating the IR baseline, i.e., the rankings provided by Google's search engine. We present a strong baseline for question relevancy ranking by training a simple multi-task feed forward network on a bag of 14 distance measures for the input question pair. This baseline model, which is fast to train and uses only language-independent features, outperforms the best shared task systems on the task of retrieving relevant previously asked questions.

Learning Sequence Encoders for Temporal Knowledge Graph Completion
Alberto Garcia-Duran, Sebastijan Dumančić, and Mathias Niepert 14:09–14:21

Research on link prediction in knowledge graphs has mainly focused on static multi-relational data. In this work we consider temporal knowledge graphs where relations between entities may only hold for a time interval or a specific point in time. In line with previous work on static knowledge graphs, we propose to address this problem by learning latent entity and relation type representations. To incorporate temporal information, we utilize recurrent neural networks to learn time-aware representations of relation types which can be used in conjunction with existing latent factorization methods. The proposed approach is shown to be robust to common challenges in real-world KGs: the sparsity and heterogeneity of temporal expressions. Experiments show the benefits of our approach on four temporal KGs. The data sets are available under a permissive BSD-3 license.

Similar but not the Same - Word Sense Disambiguation Improves Event Detection via Neural Representation Matching
Weiyi Lu and Thien Huu Nguyen 14:21–14:33

Event detection (ED) and word sense disambiguation (WSD) are two similar tasks in that they both involve identifying the classes (i.e. event types or word senses) of some word in a given sentence. It is thus possible to extract the knowledge hidden in the data for WSD, and utilize it to improve the performance on ED. In this work, we propose a method to transfer the knowledge learned on WSD to ED by matching the neural representations learned for the two tasks. Our experiments on two widely used datasets for ED demonstrate the effectiveness of the proposed method.

Learning Word Representations with Cross-Sentence Dependencyfor End-to-End Co-reference Resolution
Hongyin Luo and Jim Glass 14:33–14:45

In this work, we present a word embedding model that learns cross-sentence dependency for improving end-to-end co-reference resolution (E2E-CR). While the traditional E2E-CR model generates word representations by running long short-term memory (LSTM) recurrent neural networks on each sentence of an input article or conversation separately, we propose linear sentence linking and attentional sentence linking models to learn cross-sentence dependency. Both sentence linking strategies enable the LSTMs to make use of valuable information from context sentences while calculating the representation of the current input word. With this approach, the LSTMs learn word embeddings considering knowledge not only from the current sentence but also from the entire input document. Experiments show that learning cross-sentence dependency enriches information contained by the word representations, and improves the performance of the co-reference resolution model compared with our baseline.

Session 11E: Short Posters IV

13:45–14:45

Hall 100

Chair: *chairname*

Word Relation Autoencoder for Unseen Hypernym Extraction Using Word Embeddings
Hong-You Chen, Cheng-Syuan Lee, Keng-Te Liao, and Shou-de Lin

Lexicon relation extraction given distributional representation of words is an important topic in NLP. We observe that the state-of-the-art projection-based methods cannot be generalized to handle unseen hypernyms. We propose to analyze it in the perspective of pollution and construct the corresponding indicator to measure it. We propose a word relation autoencoder (WRAE) model to address the challenge. Experiments on several hypernym-like lexicon datasets show that our model outperforms the competitors significantly.

Refining Pretrained Word Embeddings Using Layer-wise Relevance Propagation
Akira Utsumi

In this paper, we propose a simple method for refining pretrained word embeddings using layer-wise relevance propagation. Given a target semantic representation one would like word vectors to reflect, our method first trains the mapping between the original word vectors and the target representation using a neural network. Estimated target values are then propagated backward toward word vectors, and a relevance score is computed for each dimension of word vectors. Finally, the relevance score vectors are used to refine the original word vectors so that they are projected into the subspace that reflects the information relevant to the target representation. The evaluation experiment using binary classification of word pairs demonstrates that the refined vectors by our method achieve the higher performance than the original vectors.

Learning Gender-Neutral Word Embeddings
Jieyu Zhao, YICHAO ZHOU, Zeyu Li, Wei Wang, and Kai-Wei Chang

Word embedding models have become a fundamental component in a wide range of Natural Language Processing (NLP) applications. However, embeddings trained on human-generated corpora have been demonstrated to inherit strong gender stereotypes that reflect social constructs. To address this concern, in this paper, we propose a novel training procedure for learning gender-neutral word embeddings. Our approach aims to preserve gender information in certain dimensions of word vectors while compelling other dimensions to be free of gender influence. Based on the proposed method, we generate a Gender-Neutral variant of GloVe (GN-GloVe). Quantitative and qualitative experiments demonstrate that GN-GloVe successfully isolates gender information without sacrificing the functionality of the embedding model.

Learning Concept Abstractness Using Weak Supervision
Ella Rabinovich, Benjamin Sznajder, Artem Spector, Ilya Shnayderman, Ranit Aharonov, David Konopnicki, and Noam Slonim

We introduce a weakly supervised approach for inferring the property of abstractness of words and expressions in the complete absence of labeled data. Exploiting only minimal linguistic clues and the contextual usage of a concept as manifested in textual data, we train sufficiently powerful classifiers, obtaining high correlation with human labels. The results imply the applicability of this approach to additional properties of concepts, additional languages, and resource-scarce scenarios.

Word Sense Induction with Neural biLM and Symmetric Patterns
Asaf Amrami and Yoav Goldberg

An established method for Word Sense Induction (WSI) uses a language model to predict probable substitutes for target words, and induces senses by clustering these resulting substitute vectors. We replace the ngram-based language model (LM) with a recurrent one. Beyond being more accurate, the use of the recurrent LM allows us to effectively query it in a creative way, using what we call dynamic symmetric patterns. The combination of the RNN-LM and the dynamic symmetric patterns results in strong substitute vectors for WSI, allowing to surpass the current state-of-the-art on the SemEval 2013 WSI shared task by a large margin.

Strong Baselines for Learning Generic Sentence Embeddings
Jamie Kiros and William Chan

Natural language inference has been shown to be an effective supervised task for learning generic sentence embeddings. In order to better understand the components that lead to effective representations, we propose a lightweight version of InferSent, called InferLite, that does not use any recurrent layers and operates on a collection of pre-trained word embeddings. We show that a simple instance of our model that makes no use of context, word ordering or position can still obtain competitive performance on the majority of downstream prediction tasks, with most performance gaps being filled by adding local contextual information through temporal convolutions. Our models can be trained in under 1 hour on a single GPU and allows for fast inference of new representations. Finally we describe a semantic hashing layer that allows our model to learn generic binary codes for sentences.

Similarity-Based Reconstruction Loss for Meaning Representation
Olga Kovaleva, Anna Rumshisky, and Alexey Romanov

This paper addresses the problem of representation learning. Using an autoencoder framework, we propose and evaluate several loss functions that can be used as an alternative to the commonly used cross-entropy reconstruction loss. The proposed loss functions use similarities between words in the embedding space, and can be used to train any neural model for text generation. We show that the introduced loss functions amplify semantic diversity of reconstructed sentences, while preserving the original meaning of the input. We test the derived autoencoder-generated representations on paraphrase detection and language inference tasks and demonstrate performance improvement compared to the traditional cross-entropy loss.

What can we learn from Semantic Tagging?
Mostafa Abdou, Artur Kulmizev, Vinit Ravishankar, Lasha Abzianidze, and Johan Bos

We investigate the effects of multi-task learning using the recently introduced task of semantic tagging. We employ semantic tagging as an auxiliary task for three different NLP tasks: part-of-speech tagging, Universal Dependency parsing, and Natural Language Inference. We compare full neural network sharing, partial neural network sharing, and what we term the learning what to share setting where negative transfer between tasks is less likely. Our findings show considerable improvements for all tasks, particularly in the learning what to share setting which shows consistent gains across all tasks.

Conditional Word Embedding and Hypothesis Testing via Bayes-by-Backprop
Rujun Han, Michael Gill, Arthur Spirling, and Kyunghyun Cho

Conventional word embedding models do not leverage information from document meta-data, and they do not model uncertainty. We address these concerns with a model that incorporates document covariates to estimate conditional word embedding distributions. Our model allows for (a) hypothesis tests about the meanings of terms, (b) assessments as to whether a word is near or far from another conditioned on different covariate values, and (c) assessments as to whether estimated differences are statistically significant.

Classifying Referential and Non-referential It Using Gaze
Victoria Yaneva, Le An Ha, Richard Evans, and Ruslan Mitkov

When processing a text, humans and machines must disambiguate between different uses of the pronoun it, including non-referential, nominal anaphoric or clause anaphoric ones. In this paper we use eye-tracking data to learn how humans perform this disambiguation and use this knowledge to improve the automatic classification of it. We show that by using gaze data and a POS-tagger we are able to significantly outperform a common baseline and classify between three categories of it with an accuracy comparable to that of linguistic-based approaches. In addition, the discriminatory power of specific gaze features informs the way humans process the pronoun, which, to the best of our knowledge, has not been explored using data from a natural reading task.

Bi-LSTMs Are State-of-the-art for Chinese Word Segmentation
Ji Ma, Kuzman Ganchev, and David Weiss

A wide variety of neural-network architectures have been proposed for the task of Chinese word segmentation. Surprisingly, we find that a bidirectional LSTM model, when combined with standard deep learning techniques and best practices, can achieve better accuracy on many of the popular datasets as compared to models based on more complex neuralnetwork architectures. Furthermore, our error analysis shows that out-of-vocabulary words remain challenging for neural-network models, and many of the remaining errors are unlikely to be fixed through architecture changes. Instead, more effort should be made on exploring resources for further improvement.

Sanskrit Sandhi Splitting using seq2(seq)2
Rahul Aralikatte, Neelamadhav Gantayat, Naveen Panwar, Anush Sankaran, and Senthil Mani

In Sanskrit, small words (morphemes) are combined to form compound words through a process known as Sandhi. Sandhi splitting is the process of splitting a given compound word into its constituent morphemes. Although rules governing word splitting exists in the language, it is highly challenging to identify the location of the splits in a compound word. Though existing Sandhi splitting systems incorporate these pre-defined splitting rules, they have a low accuracy as the same compound word might be broken down in multiple ways to provide syntactically correct splits. In this research, we propose a novel deep learning architecture called Double Decoder RNN (DD-RNN), which (i) predicts the location of the split(s) with 95% accuracy, and (ii) predicts the constituent words (learning the Sandhi splitting rules) with 79.5% accuracy, outperforming the state-of-art by 20%. Additionally, we show the generalization capability of our deep learning model, by showing competitive results in the problem of Chinese word segmentation, as well.

Unsupervised Neural Word Segmentation for Chinese via Segmental Language Modeling
Zhiqing Sun and Zhi-Hong Deng

Previous traditional approaches to unsupervised Chinese word segmentation (CWS) can be roughly classified into discriminative and generative models. The former uses the carefully designed goodness measures for candidate segmentation, while the latter focuses on finding the optimal segmentation of the highest generative probability. However, while there exists a trivial way to extend the discriminative models into neural version by using neural language models, those of generative ones are non-trivial. In this paper, we propose the segmental language models (SLMs) for CWS. Our approach explicitly focuses on the segmental nature of Chinese, as well as preserves several properties of language models. In SLMs, a context encoder encodes the previous context and a segment decoder generates each segment incrementally. As far as we know, we are the first to propose a neural model for unsupervised CWS and achieve competitive performance to the state-of-the-art statistical models on four different datasets from SIGHAN 2005 bakeoff.

LemmaTag: Jointly Tagging and Lemmatizing for Morphologically Rich Languages with BRNNs
Daniel Kondratyuk, Tomáš Gavenčiak, Milan Straka, and Jan Hajič

We present LemmaTag, a featureless neural network architecture that jointly generates part-of-speech tags and lemmas for sentences by using bidirectional RNNs with character-level and word-level embeddings. We demonstrate that both tasks benefit from sharing the encoding part of the network, predicting tag subcategories, and using the tagger output as an input to the lemmatizer. We evaluate our model across several languages with complex morphology, which surpasses state-of-the-art accuracy in both part-of-speech tagging and lemmatization in Czech, German, and Arabic.

Recovering Missing Characters in Old Hawaiian Writing
Brendan Shillingford and Oiwi Parker Jones

In contrast to the older writing system of the 19th century, modern Hawaiian orthography employs characters for long vowels and glottal stops. These extra characters account for about one-third of the phonemes in Hawaiian, so including them makes a big difference to reading comprehension and pronunciation. However, transliterating between older and newer texts is a laborious task when performed manually. We introduce two related methods to help solve this transliteration problem automatically. One approach is implemented, end-to-end, using finite state transducers (FSTs). The other is a hybrid deep learning approach, which approximately composes an FST with a recurrent neural network language model.

When data permutations are pathological: the case of neural natural language inference
Natalie Schluter and Daniel Varab

Consider two competitive machine learning models, one of which was considered state-of-the art, and the other a competitive baseline. Suppose that by just permuting the examples of the training set, say by reversing the original order, by shuffling, or by mini-batching, you could report substantially better/worst performance for the system of your choice, by multiple percentage points. In this paper, we illustrate this scenario for a trending NLP task: Natural Language Inference (NLI). We show that for the two central NLI corpora today, the learning process of neural systems is far too sensitive to permutations of the data. In doing so we reopen the question of how to judge a good neural architecture for NLI, given the available dataset and perhaps, further, the soundness of the NLI task itself in its current state.

Bridging Knowledge Gaps in Neural Entailment via Symbolic Models
Dongyeop Kang, Tushar Khot, Ashish Sabharwal, and Peter Clark

Most textual entailment models focus on lexical gaps between the premise text and the hypothesis, but rarely on knowledge gaps. We focus on filling these knowledge gaps in the Science Entailment task, by leveraging an external structured knowledge base (KB) of science facts. Our new architecture combines standard neural entailment models with a knowledge lookup module. To facilitate this lookup, we propose a fact-level decomposition of the hypothesis, and verifying the resulting sub-facts against both the textual premise and the structured KB. Our model, NSNet, learns to aggregate predictions from these heterogeneous data formats. On the SciTail dataset, NSNet outperforms a simpler combination of the two predictions by 3% and the base entailment model by 5%.

The BQ Corpus: A Large-scale Domain-specific Chinese Corpus For Sentence Semantic Equivalence Identification
Jing Chen, Qingcai Chen, Xin Liu, Haijun Yang, Daohe Lu, and Buzhou Tang

This paper introduces the Bank Question (BQ) corpus, a Chinese corpus for sentence semantic equivalence identification (SSEI). The BQ corpus contains 120,000 question pairs from 1-year online bank custom service logs. To efficiently process and annotate questions from such a large scale of logs, this paper proposes a clustering based annotation method to achieve questions with the same intent. First, the deduplicated questions with the same answer are clustered into stacks by the Word Mover's Distance (WMD) based Affinity Propagation (AP) algorithm. Then, the annotators are asked to assign the clustered questions into different intent categories. Finally, the positive and negative question pairs for SSEI are selected in the same intent category and between different intent categories respectively. We also present six SSEI benchmark performance on our corpus, including state-of-the-art algorithms. As the largest manually annotated public Chinese SSEI corpus in the bank domain, the BQ corpus is not only useful for Chinese question semantic matching research, but also a significant resource for cross-lingual and cross-domain SSEI research. The corpus is available in public.

Toward Understanding and Explaining Complex Deep Models in NLP
Reza Ghaeini, Xiaoli Fern, and Prasad Tadepalli

Deep learning models have achieved remarkable success in natural language inference (NLI) tasks. While these models are widely explored, they are hard to interpret and it is often unclear how and why they actually work. In this paper, we take a step toward explaining such deep learning based models through a case study on a popular neural model for NLI. In particular, we propose to interpret the intermediate layers of NLI models by visualizing the saliency of attention and LSTM gating signals. We present several examples for which our methods are able to reveal interesting insights and identify the critical information contributing to the model decisions.

Towards Semi-Supervised Learning for Deep Semantic Role Labeling
Sanket Vaibhav Mehta, Jay Yoon Lee, and Jaime Carbonell

Neural models have shown several state-of-the-art performances on Semantic Role Labeling (SRL). However, the neural models require an immense amount of semantic-role corpora and are thus not well suited for low-resource languages or domains. The paper proposes a semi-supervised semantic role labeling method that outperforms the state-of-the-art in limited SRL training corpora. The method is based on explicitly enforcing syntactic constraints by augmenting the training objective with a syntactic-inconsistency loss component and uses SRL-unlabeled instances to train a joint-objective LSTM. On CoNLL-2012 English section, the proposed semi-supervised training with 1%, 10% SRL-labeled data and varying amounts of SRL-unlabeled data achieves +1.58, +0.78 F1, respectively, over the pre-trained models that were trained on SOTA architecture with ELMo on the same SRL-labeled data. Additionally, by using the syntactic-inconsistency loss on inference time, the proposed model achieves +3.67, +2.1 F1 over pre-trained model on 1%, 10% SRL-labeled data, respectively.

Identifying Domain Adjacent Instances for Semantic Parsers
James Ferguson, Janara Christensen, Edward Li, and Edgar Gonzàlez

When the semantics of a sentence are not representable in a semantic parser's output schema, parsing will inevitably fail. Detection of these instances is commonly treated as an out-of-domain classification problem. However, there is also a more subtle scenario in which the test data is drawn from the same domain. In addition to formalizing this problem of domain-adjacency, we present a comparison of various baselines that could be used to solve it. We also propose a new simple sentence representation that emphasizes words which are unexpected. This approach improves the performance of a downstream semantic parser run on in-domain and domain-adjacent instances.

Mapping natural language commands to web elements
Panupong Pasupat, Tian-Shun Jiang, Evan Liu, Kelvin Guu, and Percy Liang

The web provides a rich, open-domain environment with textual, structural, and spatial properties. We propose a new task for grounding language in this environment: given a natural language command (e.g., "click on the second article"), choose the correct element on the web page (e.g., a hyperlink or text box). We collected a dataset of over 50,000 commands that capture various phenomena such as functional references (e.g. "find who made this site"), relational reasoning (e.g. "article by john"), and visual reasoning (e.g. "top-most article"). We also implemented and analyzed three baseline models that capture different phenomena present in the dataset.

Wronging a Right: Generating Better Errors to Improve Grammatical Error Detection
Sudhanshu Kasewa, Pontus Stenetorp, and Sebastian Riedel

Grammatical error correction, like other machine learning tasks, greatly benefits from large quantities of high quality training data, which is typically expensive to produce. While writing a program to automatically generate realistic grammatical errors would be difficult, one could learn the distribution of naturally-occurring errors and attempt to introduce them into other datasets. Initial work on inducing errors in this way using statistical machine translation has shown promise; we investigate cheaply constructing synthetic samples, given a small corpus of human-annotated data, using an off-the-rack attentive sequence-to-sequence model and a straight-forward post-processing procedure. Our approach yields error-filled artificial data that helps a vanilla bi-directional LSTM to outperform the previous state of the art at grammatical error detection, and a previously introduced model to gain further improvements of over 5% F0.5 score. When attempting to determine if a given sentence is synthetic, a human annotator at best achieves 39.39 F1 score, indicating that our model generates mostly human-like instances.

Modeling Input Uncertainty in Neural Network Dependency Parsing
Rob van der Goot and Gertjan van Noord

Recently introduced neural network parsers allow for new approaches to circumvent data sparsity issues by modeling character level information and by exploiting raw data in a semi-supervised setting. Data sparsity is especially prevailing when transferring to non-standard domains. In this setting, lexical normalization has often been used in the past to circumvent data sparsity. In this paper, we investigate whether these new neural approaches provide similar functionality as lexical normalization, or whether they are complementary. We provide experimental results which show that a separate normalization component improves performance of a neural network parser even if it has access to character level information as well as external word embeddings. Further improvements are obtained by a straightforward but novel approach in which the top-N best candidates provided by the normalization component are available to the parser.

Parameter sharing between dependency parsers for related languages
Miryam de Lhoneux, Johannes Bjerva, Isabelle Augenstein, and Anders Søgaard

Previous work has suggested that parameter sharing between transition-based neural dependency parsers for related languages can lead to better performance, but there is no consensus on what parameters to share. We present an evaluation of 27 different parameter sharing strategies across 10 languages, representing five pairs of related languages, each pair from a different language family. We find that sharing transition classifier parameters always helps, whereas the usefulness of sharing word and/or character LSTM parameters varies. Based on this result, we propose an architecture where the transition classifier is shared, and the sharing of word and character parameters is controlled by a parameter that can be tuned on validation data. This model is linguistically motivated and obtains significant improvements over a monolingually trained baseline. We also find that sharing transition classifier parameters helps when training a parser on unrelated language pairs, but we find that, in the case of unrelated languages, sharing too many parameters does not help.

Grammar Induction with Neural Language Models: Flawed Experiments, Important Results
Phu Mon Htut, Kyunghyun Cho, and Samuel Bowman

A substantial thread of recent work on latent tree learning has attempted to develop neural network models with parse-valued latent variables and train them on non-parsing tasks, in the hope of having them discover interpretable tree structure. In a recent paper, Shen et al. (2018) introduce such a model and report near-state-of-the-art results on the target task of language modeling, and the first strong latent tree learning result on constituency parsing. In an attempt to reproduce these results, we discover issues that make the original results hard to trust, including tuning and even training on what is effectively the test set. Here, we attempt to reproduce these results in a fair experiment and to extend them to two new datasets. We find that the results of this work are robust: All variants of the model under study outperform all latent tree learning baselines, and perform competitively with symbolic grammar induction systems. We find that this model represents the first empirical success for latent tree learning, and that neural network language modeling warrants further study as a setting for grammar induction.

Data Augmentation via Dependency Tree Morphing for Low-Resource Languages
Gozde Gul Sahin and Mark Steedman
Neural NLP systems achieve high scores in the presence of sizable training dataset. Lack of such datasets leads to poor system performances in the case low-resource languages. We present two simple text augmentation techniques using dependency trees, inspired from image processing. We "crop" sentences by removing dependency links, and we "rotate" sentences by moving the tree fragments around the root. We apply these techniques to augment the training sets of low-resource languages in Universal Dependencies project. We implement a character-level sequence tagging model and evaluate the augmented datasets on part-of-speech tagging task. We show that crop and rotate provides improvements over the models trained with non-augmented data for majority of the languages, especially for languages with rich case marking systems.

How Much Reading Does Reading Comprehension Require? A Critical Investigation of Popular Benchmarks
Divyansh Kaushik and Zachary C. Lipton
Many recent papers address reading comprehension, where examples consist of (question, passage, answer) tuples. Presumably, a model must combine information from both questions and passages to predict corresponding answers. However, despite intense interest in the topic, with hundreds of published papers vying for leaderboard dominance, basic questions about the difficulty of many popular benchmarks remain unanswered. In this paper, we establish sensible baselines for the bAbI, SQuAD, CBT, CNN, and Who-did-What datasets, finding that question- and passage-only models often perform surprisingly well. On 14 out of 20 bAbI tasks, passage-only models achieve greater than 50% accuracy, sometimes matching the full model. Interestingly, while CBT provides 20-sentence passages, only the last is needed for accurate prediction. By comparison, SQuAD and CNN appear better-constructed.

MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling
Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic
Even though machine learning has become the major scene in dialogue research community, the real breakthrough has been blocked by the scale of data available. To address this fundamental obstacle, we introduce the Multi-Domain Wizard-of-Oz dataset (MultiWOZ), a fully-labeled collection of human-human written conversations spanning over multiple domains and topics. At a size of 10k dialogues, it is at least one order of magnitude larger than all previous annotated task-oriented corpora. The contribution of this work apart from the open-sourced dataset is two-fold: firstly, a detailed description of the data collection procedure along with a summary of data structure and analysis is provided. The proposed data-collection pipeline is entirely based on crowd-sourcing without the need of hiring professional annotators; secondly, a set of benchmark results of belief tracking, dialogue act and response generation is reported, which shows the usability of the data and sets a baseline for future studies.

Linguistically-Informed Self-Attention for Semantic Role Labeling
Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum
Current state-of-the-art semantic role labeling (SRL) uses a deep neural network with no explicit linguistic features. However, prior work has shown that gold syntax trees can dramatically improve SRL decoding, suggesting the possibility of increased accuracy from explicit modeling of syntax. In this work, we present linguistically-informed self-attention (LISA): a neural network model that combines multi-head self-attention with multi-task learning across dependency parsing, part-of-speech tagging, predicate detection and SRL. Unlike previous models which require significant pre-processing to prepare linguistic

features, LISA can incorporate syntax using merely raw tokens as input, encoding the sequence only once to simultaneously perform parsing, predicate detection and role labeling for all predicates. Syntax is incorporated by training one attention head to attend to syntactic parents for each token. Moreover, if a high-quality syntactic parse is already available, it can be beneficially injected at test time without re-training our SRL model. In experiments on CoNLL-2005 SRL, LISA achieves new state-of-the-art performance for a model using predicted predicates and standard word embeddings, attaining 2.5 F1 absolute higher than the previous state-of-the-art on newswire and more than 3.5 F1 on out-of-domain data, nearly 10% reduction in error. On ConLL-2012 English SRL we also show an improvement of more than 2.5 F1. LISA also out-performs the state-of-the-art with contextually-encoded (ELMo) word representations, by nearly 1.0 F1 on news and more than 2.0 F1 on out-of-domain text.

Phrase-Based & Neural Unsupervised Machine Translation
Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato

Machine translation systems achieve near human-level performance on some languages, yet their effectiveness strongly relies on the availability of large amounts of parallel sentences, which hinders their applicability to the majority of language pairs. This work investigates how to learn to translate when having access to only large monolingual corpora in each language. We propose two model variants, a neural and a phrase-based model. Both versions leverage a careful initialization of the parameters, the denoising effect of language models and automatic generation of parallel data by iterative back-translation. These models are significantly better than methods from the literature, while being simpler and having fewer hyper-parameters. On the widely used WMT'14 English-French and WMT'16 German-English benchmarks, our models respectively obtain 28.1 and 25.2 BLEU points without using a single parallel sentence, outperforming the state of the art by more than 11 BLEU points. On low-resource languages like English-Urdu and English-Romanian, our methods achieve even better results than semi-supervised and supervised approaches leveraging the paucity of available bitexts. Our code for NMT and PBSMT is publicly available.

Plenary Session. Best Paper

16:30–18:00

Hall 100

Chairs:

How Much Reading Does Reading Comprehension Require? A Critical Investigation of Popular Benchmarks

Divyansh Kaushik and Zachary C. Lipton

Many recent papers address reading comprehension, where examples consist of (question, passage, answer) tuples. Presumably, a model must combine information from both questions and passages to predict corresponding answers. However, despite intense interest in the topic, with hundreds of published papers vying for leaderboard dominance, basic questions about the difficulty of many popular benchmarks remain unanswered. In this paper, we establish sensible baselines for the bAbI, SQuAD, CBT, CNN, and Who-did-What datasets, finding that question- and passage-only models often perform surprisingly well. On 14 out of 20 bAbI tasks, passage-only models achieve greater than 50% accuracy, sometimes matching the full model. Interestingly, while CBT provides 20-sentence passages, only the last is needed for accurate prediction. By comparison, SQuAD and CNN appear better-constructed.

papers-71

papers-71

Even though machine learning has become the major scene in dialogue research community, the real breakthrough has been blocked by the scale of data available. To address this fundamental obstacle, we introduce the Multi-Domain Wizard-of-Oz dataset (MultiWOZ), a fully-labeled collection of human-human written conversations spanning over multiple domains and topics. At a size of 10k dialogues, it is at least one order of magnitude larger than all previous annotated task-oriented corpora. The contribution of this work apart from the open-sourced dataset is two-fold: firstly, a detailed description of the data collection procedure along with a summary of data structure and analysis is provided. The proposed data-collection pipeline is entirely based on crowd-sourcing without the need of hiring professional annotators; secondly, a set of benchmark results of belief tracking, dialogue act and response generation is reported, which shows the usability of the data and sets a baseline for future studies.

Linguistically-Informed Self-Attention for Semantic Role Labeling

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum

Current state-of-the-art semantic role labeling (SRL) uses a deep neural network with no explicit linguistic features. However, prior work has shown that gold syntax trees can dramatically improve SRL decoding, suggesting the possibility of increased accuracy from explicit modeling of syntax. In this work, we present linguistically-informed self-attention (LISA): a neural network model that combines multi-head self-attention with multi-task learning across dependency parsing, part-of-speech tagging, predicate detection and SRL. Unlike previous models which require significant pre-processing to prepare linguistic features, LISA can incorporate syntax using merely raw tokens as input, encoding the sequence only once to simultaneously perform parsing, predicate detection and role labeling for all predicates. Syntax is incorporated by training one attention head to attend to syntactic parents for each token. Moreover, if a high-quality syntactic parse is already available, it can be beneficially injected at test time without re-training our SRL model. In experiments on CoNLL-2005 SRL, LISA achieves new state-of-the-art performance for a model using predicted predicates and standard word embeddings, attaining 2.5 F1 absolute higher than the previous state-of-the-art on newswire and more than 3.5 F1 on out-of-domain data, nearly 10% reduction in error. On ConLL-2012 English SRL we also show an improvement of more than 2.5 F1. LISA also out-performs the state-of-the-art with contextually-encoded (ELMo) word representations, by nearly 1.0 F1 on news and more than 2.0 F1 on out-of-domain text.

Phrase-Based & Neural Unsupervised Machine Translation

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato

Machine translation systems achieve near human-level performance on some languages, yet their effectiveness strongly relies on the availability of large amounts of parallel sentences, which hinders their applicability to the majority of language pairs. This work investigates how to learn to translate when having access to only large monolingual corpora in each language. We propose two model variants, a neural and a phrase-based model. Both versions leverage a careful initialization of the parameters, the denoising effect of language models and automatic generation of parallel data by iterative back-translation. These models are significantly better than methods from the literature, while being simpler and having fewer hyper-parameters. On the widely used WMT'14 English-French and WMT'16 German-English benchmarks, our models respectively obtain 28.1 and 25.2 BLEU points without using a single parallel sentence, outperforming the state of the art by more than 11 BLEU points. On low-resource languages like English-Urdu and English-Romanian, our methods achieve even better results than semi-supervised and supervised approaches leveraging the paucity of available bitexts. Our code for NMT and PBSMT is publicly available.

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.

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, 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.

Approved by ACL Executive Committee in 2016.

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

Index

Local Guide to Brussels

Written by Nathan Schaltin
Picture by Cédric Mayence Photography

General

1. Brussels, capital of Europe

Brussels is a cosmopolitan city, with a liveliness and an appeal that are intimately related to its role as a crossroad for all of Europe. Architectural styles range from Gothic cathedrals and churches to the gracious classical facades of the Palais des Nations, the Royal Palace and to the many Art Nouveau and Art Deco houses in the comfortable neighborhoods where the Bruxellois live.

The heart of Brussels and the place to start getting to know the city is the Grand Place. This historic square, lined with exuberantly ornate guild houses and focused on the Gothic heights of the Hotel de Ville, is widely held to be one of Europe's finest.

2. Useful smartphone apps

Visit.brussels.be	The official Brussels tourism site and app, showing everything you need to know when visiting Brussels: museums, festivals, restaurants, bars, transport, etc.
Tales & tours	Bring along your personal tour guide with your smartphone filled with travel &am; audio guides, walks and routes.
STIB	The public transportation network (metro, tram and bus) is managed by the Brussels Intermunicipal Transport Company (STIB-MIVB). With the app, you can check the arrival times and routes in real-time.
Brussels Gardens	Find your nearest park or garden.
Spotted by locals	Experience Brussels like a local.
HappyCow	This app shows you nearby vegetarian and vegan restaurants.

Transport

1) Airport

Nearest airports for Brussels are:

- Brussels Airport (in Zaventem)
- Brussels South (in Charleroi)

2) Metro, tram and bus

The public transportation network (metro, tram and bus) is managed by the Brussels Intermunicipal Transport Company (STIB-MIVB). The network maps, schedules, frequencies, rates or route planners are available on its website:

- STIB-MIVB

The night network of STIB (night buses during the weekend) is called Noctis.

Some bus lines in Brussels are also managed by:

- De Lijn

3) Train

The trains are managed by:

- SNCB-NMBS

4) Metro and train map

The map of the underground (subway) and train stations in Brussels:

- Metro (and train) map of Brussels (PDF, 1.97 MB)

5) Waterbus

The Waterbus is a bus service on the water between Vilvoorde and Brussels, a 10-kilometer navigational route on the Senne Canal:

- Waterbus

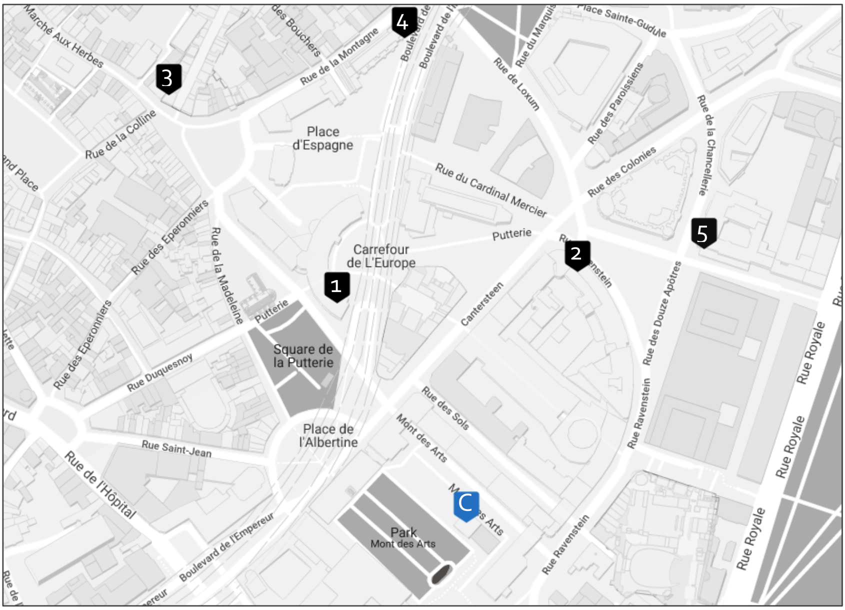
Public Holidays

All saints and All souls day:

All Saints' Day is a Christian festival celebrated in honor of all the saints, known and unknown, on the first of November. On the second of November, All Souls' Day commemorates All Souls, the Holy Souls, or the Faithful Departed; that is, the souls of Christians who have died. Christians typically remember deceased relatives on the day. **Therefore a lot of shops, restaurants and museums will be closed on the first and second of November.**

Banks and ATMs near the conference hall

There are several banks near the conference venue. You can find a selection on the map below. The blue marker with a "C" is the conference venue.



1: BNP Paribas Fortis	Carrefour De L'europe 1 1000 Brussels
2: ING	Rue Ravenstein 68 1000 Brussels
3: ING	Rue Du Marché Aux Herbes 90 1000 Brussels
4: BNP Paribas Fortis	Boulevard De L'impératrice 56 1000 Brussels
5: BNP Paribas Fortis	Rue De La Chancellerie 1 1000 Brussels

Exploring Brussels

1. Monuments and sites

Royal Palace Rue Brederode 16, 1000 Brussels

The Royal Palace in Brussels is the official seat of the King and Queen of the Belgians. It is however not used as their residence, as the King and his family reside in the Palace of Laeken, just outside of Brussels. The Royal Palace was built on the site of the former Palace of the Dukes of Brabant which was destroyed by fire in 1731. Started in 1820 under the reign of King William, it was modified in 1904 under Leopold II, who had it rebuilt in Louis XVI style. The side wings date from the 18th century and at the end of each wing there is a pavilion. Unfortunately the palace is only open for the public during summer.

Atomium Square de l'Atomium, 1020 Brussels

Unavoidable icon of Brussels and Belgium, important place for international tourism, unique creation in the history of architecture and emblematic vestige of the World fair in Brussels (Expo 58), the Atomium is today the most popular tourist attraction of Europe's Capital. The Atomium was constructed for the first post-war universal world exhibition (EXPO 58). The nine spheres represent an iron crystal magnified 165 billion times. They represent the faith one had in the power of science and moreover in nuclear power.
http://atomium.be/

Adults – 15€

Mon-Sun 10h-18h

Brussels' outdoor markets Various places

There are numerous general and specialized markets in Brussels, each with their own atmosphere. The most famous one is the flea market at the Vossenplein/Place Jeu du Balle. People from all over Belgium come here to buy and sell second-hand wares. But there's also the antiques market at Grand Sablon, the flower market at Place Sainte-Cathérine, even a used bicycle market at the Boulevard du Midi. These markets are often only open a single day per week (with the exception of the Vossenplein flea market, which is open all week from early in the morning till midday). For a list of markets and their opening hours check https://www.brussels.be/outdoor-markets.

The Europa building Rue de la Loi 175, 1048 Brussels

The Europa building is one of Brussels' most recent and remarkable examples of contemporary architecture. The gigantic cube is made up of 3,750 restored window frames and contains a glass lantern which, at night, is beautifully lit by 374 LED tubes. The Europa building and the Berlaymont, stand together as symbols of two of the three centres of EU decision making. This "House of European member states" with its multi-colored interior is the headquarters of the European Council and the Council of the European Union and will be where heads of state and minister of the European Union member states meet from now on. The visitors' centre opened august 2018.

Visits are free of charge.

Mon-Fri 9h30-11h & 13h30-16h

Closed on weekends and public holidays

Grand Place Grand Place 1, 1000 Brussels

The Grand-Place is a UNESCO World Heritage site. Construction on it began in the 15th century; first of all, some market halls, trade guild houses, and a town hall to establish the authority of this trading centre. Bombarded for 3 days by the French army in 1695, it was almost completely destroyed but, like a phoenix, it went on to rise from the ashes in less than 5 years. That's why four styles stand alongside each other or sometimes even overlap; it's a wonderful hotch-potch of Gothic, opulent baroque, neoclassical and neogothic.

Royal Galleries of Saint-Hubert Galerie du Roi 5, 1000 Brussels

Initiated by the architect Jean-Pierre Cluysenaer, the spectacular Galeries Royales project took shape in the 1830's. The works, begun in 1846, were almost completed in time for the official opening on 20 June 1847. The gallery included shops, auditoriums, cafés, restaurants and apartments. The place to be seen for the fashionable, right away, the Royal Galleries of Saint Hubert drew a wide audience, attracted by its luxury brands, elegant cafés and cultural spaces. The colony of French refugees, like Victor Hugo, Alexandre Dumas, Edgar Quinet and others, would also gather there. The Surrealist painters and artists from the Cobra group were regulars at the venue.

Église notre dame des victoires du Sablon Rue de la Régence, 38, 1000 Brussels

This church is a precious example of the development of ogival art from Brabant. It owes its beauty to the city's crossbowmen, who built it in the 15th century. Several members of the Von Thurn und Taxis family are buried there in sarcophagi placed under the chapel of Saint Ursula, which is their burial site. If you pass by this church, do not hesitate to enter and contemplate its architecture and numerous stained glass windows.

Manneken Pis Stoofstraat, 1000 Brussels

Literally just a 60 cm tall statue of a little peeing boy, occasionally dressed up in festive clothes. Please don't expect too much from it.

2. Museums

Please note that most museums will be closed November 1st!

Bozar - Museum of Fine Arts Rue Ravenstein 23, 1000 Brussels

It's impossible to talk about culture in Belgium without mentioning the Centre for Fine Arts located in Brussels. Whether it comes to exhibitions, concerts, theater performances, movies or projects concerning literature and architecture, BOZAR will always put originality and innovation first. They prove to keep a finger on the pulse with their varied and sophisticated program.
https://www.bozar.be/

Mon	Closed
Tue-Wed	10h-18h
Thu	10h-21h
Fri-Sun	10h-18h

Coudenberg Palace Place des Palais 7, 1000 Brussels

An underground tour discovering the remains of the palace of Charles V. From the middle ages, a castle overlooked Brussels from Coudenberg hill. From the 12th century, the successive monarchs and their representatives transformed a small fortified castle into a sumptuous residential palace, one of the most beautiful palace of Europe and one of Charles V's main residences. This prestigious building is severely damaged by fire in 1731. Some forty years later, the ruins of the palace are pulled down and the ground flattened out for the construction of the new royal district. The remains of this palace make up the Coudenberg archaeological site. Entrance is through the BELvue museum.
https://coudenberg.brussels/en

Adults – 7€

Mon	Closed
Tue-Fri	9h30-17h
Sat-Sun	10h-18h

BELvue Museum Place des Palais 7, 1000 Brussels

Democracy, prosperity, solidarity, pluralism, migration, languages and Europe: these are the themes used by the BELvue Museum to help you to discover Belgium. Each one is developed through present-day testimonials and then traced right through history. This is a modern, original way of gaining a fuller understanding of modern-day Belgium. Have a look at the gallery too: there are more than 200 objects to view, each one forming a tangible memory of Belgium. Here are just a few examples: a lithograph by Magritte, a motorcycle, crystal vases by Val Saint-Lambert, and a football signed by the Red Devils.
https://www.belvue.be/en

Adults – 7€

Mon	Closed
Tue-Fri	9h30-17h
Sat-Sun	10h-17h

Comic Strip Centre Rue des Sables 20, 1000 Brussels

Belgium has a rich arsenal of comic book artists, and comic books are seen as an integral part of Belgian Culture. This is easily noticed when you take a stroll through Brussels and find the immense wall paintings of Tintin, Thorgal, The Smurfs, Lucky Luke and many others scattered throughout the city. The Comics Art Museum was founded more than 25 years ago to honor these comic books and their creators. The regularly renewed permanent exhibitions and a diversified program of temporary exhibitions enable visitors to discover the countless aspects of comics art.
https://www.comicscenter.net/en/home

Adults – 10€

Mon-Sun	10h-18h

Victor Horta Museum Rue Américane, 1060 Brussels

Victor Horta (1861-1947) was one of the most talented and well-known architect in the Art Nouveau scene. The museum, located at his former house and studio, is the archetypical example of Art Nouveau culture: high arches, stained glass and beautiful mosaics.
http://www.hortamuseum.be/en/
Adults – 10€

Mon	Closed
Tue-Sun	14h-17h30

Museum of Erotics and Mythology Rue Sainte-Anne 32, 1000 Brussels

In a charming XVIIIth century house, at the Sablon, you will discover the fantastic history of Erotic art from Antiquity until today. This private collection, one of the most attractive in Europe, offers you the chance to see some rare pieces of paintings, sculptures, Greco-roman antiquities, Ivory, Japanese prints and other curiosities. The purpose of this museum is to show the existence of the erotism since the sexual pleasure has been discovered by humanity.
http://www.m-e-m.be/

Adults – 10€

Mon	14h-22h
Tue-Wed	Closed
Thu-Fri	14h-20h
Sat-Sun	11h-17h30

Musical Instruments Museum Rue Montagne de la Cour 2, 1000 Brussels

The entrance to the building is a jewel of the Art Nouveau style, while the rest is in perfect harmony with the neoclassical ensemble on the Place Royale. With a wealth of some 9.000 instruments, some 1.200 of which are on permanent exhibit, the museum's reputation is mainly built on its extraordinary collections.
http://www.mim.be/en

Adults – 10€

Mon	Closed
Tue-Fri	9h30-17h
Sat-Sun	10h-17h

Royal Belgian Institute of Natural Sciences Rue Vautier 29, 1000 Brussels

Nearly every person in Belgium has visited this museum at least once as a child, and will still be able to tell you about Stan, the Tyrannosaurus Rex. This museum is dedicated to the history of the earth, nature, and evolution. It hosts the largest dinosaur gallery of Europe. In other galleries you will find rare minerals, ancient skeletons and life-size replicas of animals that roamed the earth millions of years ago.

https://www.naturalsciences.be/

Adults – 7€

Mon	Closed
Tue-Fri	9h30-17h
Sat-Sun	10h-18h

Millenial Iconoclast Museum of Arts Quai du Hainaut 41, 1080 Molenbeek

The MIMA is situated on the remarkable site of the former brewery Belle-Vue beside the canal in the center of Brussels. It is a new museum with an international vocation that advances contemporary, accessible and viral creations such as Street Art. The MIMA presents temporary exhibitions, encounters, projections, concerts, performances as well as a permanent collection. It also houses an art shop, a restaurant and two magnificent panoramic viewing points over the canal and Brussels in a magnificent building covering four different floors.
http://www.mimamuseum.eu/

Adults – 9,5€

Mon-Tue	Closed
Wed-Sun	10h-18h

Weirtz Museum Rue vautier 62, 1050 Brussels

The museum, which has preserved its original atmosphere, is dedicated to painter, sculptor and writer Antoine Wiertz (1806-1865), a somewhat controversial artistic figure of the Belgian Romantic movement. Wiertz loved the spectacular, painting giant canvasses to indulge his thirst for the excessive, like the more than 8 metres long Greeks and Trojans fighting over the body of Patroclus.
https://www.fine-arts-museum.be/en/museums/musee-wiertz-museum

Free admission

Mon	closed
Tue-Fri	10h-12h & 12h45-17h

3. Parks

Kunstberg/Mont des arts

Besides housing the Square Brussels Conference Venue where EMNLP 2018 will take place, this park is home to many other influential institutions such as the Royal Library and the Museum of Fine Arts. In the 15th-19th century this hill was densely populated, but by the end of the 19th century, King Leopold II had the idea of converting the site to an arts quarter and bought the entire neighborhood. The construction of the quarter took many decades and a series of redesigns, and was finally opened in 1969 - exactly sixty years after King Leopold II's death. It offers a magnificent view of downtown Brussels and the Town Hall.

Cinquantenaire Park

The Cinquantenaire (50th anniversary) site is comprised of a vast set of gardens dotted with monuments and museums. It is dominated by a triumphal arch with three arches. The park hosts numerous activities throughout the year: events, celebrations, firework displays, sporting events, concerts, etc. This place of interest was built in 1880 for the 50th anniversary of the independence of Belgium. The broad pathways lead to the Pavilion of Human Passions designed by Victor Horta, the Royal Museum of the Armed Forces & Military History, the Royal Museums of Art and History and to Autoworld. At the top of the three triumphal arches there's a bronze quadriga and an unbeatable sweeping view over the whole of Brussels.

Park of Brussels

Known as the "Parc Royal" for many years, it was the first public park in Brussels. This large quadrilateral planted with trees and standing proud between the Federal Parliament and the Royal Palace spreads before you broad pathways, symmetrical alignments and surprising perspectives. As you wend your way along the paths, you'll come across secluded little leafy clearings, a bandstand, a theatre, two fountains and a few kiosks serving refreshments. And yet, this spot, which radiates tranquillity, was once the site of fierce battles for the people of Brussels who fought their revolution here in 1830.

Leopold Park

A small island of tranquillity on the edge of the big sites of the European district. A landscaped, undulating park, which houses the Museum of the Institute of Natural Sciences and other scientific institutes, such as the magnificent Solvay Library, the work of Henry van de Velde. Today the library hosts cultural events.

Petit Sablon Park

Inaugurated in 1890, this little park is surrounded by a superb wrought-iron balustrade, decorated with 48 bronze statuettes which represent the ancient crafts once practised in Brussels. There are ten statues decorating the rear of the park, intended to illustrate the 16th century in Belgium through the country's celebrated scholars and humanists, such as Mercator, Abraham Ortelius, Louis Van Bodeghem, Rombaud Dodonée, amongst others. In the centre, the statues of the Counts of Egmont and Hornes symbolise the fight against Spanish tyranny in the 16th century. The pedestal is decorated with the coats of arms of the two lords.

Eating & Drinking

1. Restaurants close to the conference venue

There are numerous restaurants to be found near the conference hall, both Belgian and international kitchens. The Sablon borough - just south of the conference hall - is a lively area of chic bars, fashionable restaurants and chocolate shops. To the northwest you will find the touristic heart of Brussels with the Grand Place and its neighboring small cobbled streets. Both these boroughs are excellent places to look for something to eat. If you plan to go with a large group, we advise you to book in advance.

Jef *(0.5 km)* Rue Haute 20, 1000 Brussels

Belgian-French cuisine in charming little bistro. Also renowned for its fine selection of wines.
http://jefresto.be/

Mon	Closed
Tue-Fri	12h-14h & 19h-22h
Sat	19-22h
Sun	12-14h

Al Barmarki *(0.4 km)* Rue des Éperonniers 67, 1000 Brussels

According to the owners "the first Lebanese restaurant in Belgium". Serves traditional Lebanese meals for a reasonable price.
https://www.albarmaki.be/

Mon-Sat	19h-00h
Sun	Closed

Love Ciabatta *(0.2 km)* Coudenbreg 70, 1000 Brussels

Casual but cozy sandwich bar located close to the Magritte Museum. The focaccias, vegetarian pastas and salads make for a great light lunch.
https://love-ciabatta.be/en

Mon	11h-16h
Tue-Sun	9h-18h45

PEÏ & MEÏ *(0.6 km)* Rue de Rollebeek 15, 1000 Brussels

Original restaurant that succeeds in finding the perfect balance between traditional and modern cooking. Loved by local food critics.
http://www.peietmei.be/

Mon-Fri	12-14h & 19h30-21h30
Sat	19h30-21h30
Sun	Closed

C'est bon c'est belge *(0.6 km)* Rue de Rollebeek 3-5 , 1000 Brussels

Traditional Belgian meals and delicatessen: "Stoemp" and sausages, "Waterzooi" stewed vegetables, or "Vol au Vent". Accommodating Belgian beers are also available.
https://www.cestboncestbelge.be/en/

Mon	12h-21h
Tue-Wed	Closed
Thu-Sat	12h-21h
Sun	10h30-21h

Atelier des pates *(0.4 km)* Rue de Namur 3, 1000 Brussels

Fresh home-made Italian pasta, made by three young entrepreneurs. Located right next to the impressive Royal Square.
http://www.atelierdespates.be/?lang=en

Mon-Tue	12h-14h30
Wed-Fri	12h-14h30 & 18h30-22h30
Sat	12h30-15h & 19h-22h30
Sun	Closed

Kobob *(0.8 km)* Rue des Grands Carmes 10, 1000 Brussels

Authentic Ethiopian food for those who like something different. Meals are always served on shared plates and to be eaten without cutlery. It also serves a large selection of vegetarian and vegan dishes.
https://www.kokob.be/en/

Mon-Thu	18h – 23h
Friday	18h - 23h30
Sat	12h - 15h & 18h - 23h30
Sun	12h - 15h & 18h - 23h

La Table de Mus *(0.5 km)* Oud Korenhuis 31, 1000 Brussels

For those who do not mind to spend a little more on their food, "La table de Mus" is the best place to go. Carefully chosen seasonal ingredients, served by a warm and welcoming owner.
http://www.latabledemus.be/

Mon-Tue	Open
Wed	Closed
Thu-Sat	Open
Sun	Closed

POP-up Sablon *(0.6 km)* Grand Sablon 15

Much like the borough of Sablon itself, this restaurant/bar is young and vibrant without losing its roots. The heated terrace and large modern interior are open all week.
https://www.popupsablon.be/

Mon-Sun 12h-23h

Viva M'Boma *(1.1 km)* Rue de Flandre 17

When tripe goes gourmet, (re)discover the old recipes that grandma used to make.
www.vivamboma.be

Mon- Sun 12h-14h30 & 19h–22h30

Brasserie Ploegmans *(0.8 km)* Rue Haute 148

Old-world brasserie setting in the heart of the Marolles, where Belgian specialties are on the menu.
https://www.ploegmans.be/

Mon	Closed
Tue-Thu	12h-14h30 & 19h-22h
Fri-Sat	12h-14h30 & 19h-22h30
Sun	12h-14h30

Les Brigittines *(0.6 km)* Place de la Chapelle 5

In the Marolles district, a tasty brasserie with a strong Brussels accent.
https://www.lesbrigittines.com/

Mon-Fri	12h-14h & 19h-22h
Sat	19h-22h
Sun	Closed

Belga Queen *(0.9 km)* Rue Fossé aux Loups 32

This splendid former bank has become a temple to good eating with classic Belgo-French recipes adapted to today's tastes.
www.belgaqueen.be

Mon-Thu	12h-14h30 & 18h30-23h
Fri	12h-14h30 & 18h30-24h
Sat	18h30-24h
Sun	Closed

Madame Chapeau *(0.7 km)* Rue du Marché au Charbon 94

Stoemp, the staple of Belgian cuisine, is offered here in all its forms.
www.madamechapeau.be

Mon-Tue	Closed
Wed-Sun	12h-24h

Au Vieux Saint-Martin *(0.45 km)* Place du Grand Sablon 38

A Brussels institution, whose founder Joseph Niels is famous for creating the recipe for the famous Americain Préparé, the original steak tartare.
www.auvieuxsaintmartin.be

Mon-Sat	12h-24h
Sun	11h30-23h

Restobières *(1.1 km)* Rue des Renards 9

Simple, tasty Belgian cuisine perfectly complemented by an impressive list of beers.
https://www.restobieres.eu/

Mon	Closed
Tue-Thu	12h-15h &18h-22h
Fri	12h-15h & 18-22h30
Sat	12h-15h30 & 18h-22h30
Sun	12h-15h30 & 18h-22h

2. Bars and cafés in the area

Below you will find some of the best places in the area of the conference venue to grab a beer, a glass of wine, a cocktail or any other drink. But feel free to differ from this list and explore the stunning 540 bars and cafés that Brussels is home to.

Het Goudblommeke *(0.6 km)* Rue des Alexiens 55, 1000 Brussels

An artistic and literary café with a rich past. In the early 20th century, this place was frequented by artists such as Magritte and Lecomte. During world war II it was forced to close its doors. When it was reopened in late 1944, this café blossomed again the meeting place of the post-war surrealist movement.

Mon	Closed
Tue-Sat	11h-24h
Sun	11h-19h

Vertigo *(0.6 km)* Rue de Rollebeek 7, 1000 Brussels

Vertigo is somewhat hidden away in a tiny alley, but all the more rewarding when you find it. This bar serves excellent cocktails and tapas on its well preserved patio.

Mon-Thu	12h-00h
Fri-Sun	12h-01h

Poechenellekelder *(0.6 km)* Rue du chêne 5, 1000 Brussels

A typical Bruxellois café where time seems to have stopped in the 1950's. Enjoy a selection of 150 beers.

Mon	Closed
Tue-Thu	11h-01h
Fri-Sat	11h-02h
Sun	11h-01h

La Machine *(1.0 km)* Place Saint-Géry 2, 1000 Brussels

This popular bar in the Saint-Géry area, with a steampunk-inspired decor, serves no fewer than 59 beers, 70 cocktails, and 60 selected spirits. It also lays on great rock, jazz, and soul concerts.

Mon-Tue	15h-01h
Wed-Thu	15h-02h
Fri-Sun	15h-04h

COCO Donuts *(0.6 km)* Rue Sainte-Anne 36, 1000 Brussels

COCO's home-made, artisanal donuts are a treat like no other in the city. Enjoy them with a cup of coffee at their shop in the little cobbled street of Sainte-Anne – the patron saint of grandmothers and single women.

Mon-Sun 10h30-18h

La Pharmacie Anglaise *(0.2 Km)* Koudenberg 66, 1000 Brussels

“Cocktails & Curiosity”. This old pharmacy was revived as a cocktail bar. It still has the original decoration to show off its past. The cocktails can be pricey, but the unique interior is definitely worth a look.

Mon	Closed
Tue-Thu	17h30-01h
Fri-Sat	17h30-02h
Sun	Closed

Delirium Café *(0.7 km)* Getrouwheidsgang 4, 1000 Brussels

This bar holds the world record for largest variety of beers available in a single bar. It serves a stunning amount of 3,162 different beers. There is no better place to explore the Belgian beer culture.

Mon-Sat	22h-04h
Sun	22h-02h

Mug Brussels *(0.3 km)* Rue Ravenstein 20-22, 1000 Brussels

A great place to unwind from the rush of the city. Have a coffee or a slice of cake and enjoy the quirky interior.

Mon-Fri	*9h-15h*
Sat-Sun	*Closed*

L’Archiduc *(1.0 km)* Rue Antoine Dansaert 6, 1000 Brussels

A timeless institution of Brussels nightlife, l’Archiduc is a place that you must visit to get a real sense of Brussels’ DNA. Celebrities and regular people have been rubbing shoulders there for decades.

Mon-Sun 16h-05h

Benelux *(0.6 km)* Petite rue des Bouchers 6, 1000 Brussels

Contemporary yet cozy bar serving a bit of everything: Belgian beers, wine, and cocktails. Located close the beautiful Royal Galleries and Grand Place.

Au Darigman *(1.2 km)* Rue de Flandre 37, 1000 Brussel

A little bit further away than the other bars on this list, but a place that’s on The Guardians list of “Best bars in the world” is worth the walk. Small and cozy café, with a fine selection of Trappist and local beers.

Celtica *(0.7 km)* Kiekenmarkt 55, 1000 Brussel

Authentical Irish pub with live music every Friday, Saturday and Sunday. Prices are one of the lowest in the area.

Nightlife

1. Theater, concerts & movies

Brussels offers a great variety of shows, concerts and theater performances. For a complete list of every event available in Brussels, please see https://www.uitinvlaanderen.be/agenda/concert/brussel?date_range=30/10/2018%20-%204/11/2018.

Bozar Rue Ravenstein 23, 1000 Brussels

The Bozar Fine Arts Centre hosts several classical music concerts during the time of the conference. At only a five minute walk from the conference venue, the neoclassical 'Henry Le Boeuf Hall' itself – inaugurated in 1929 and regarded as one of the five best concert halls in the world – is reason enough to attend to one the concerts. Ticket prices range from 38€ to 70€ per person. It is advised to book well ahead, as these shows are often rapidly sold out. The following concerts are already confirmed by the Bozar event organisers, and more will likely follow: https://www.bozar.be/en

31st october 2018:	Salvatore Adamo
1st november 2018:	Katie Melua (featuring Gori womens choir)
2nd November 2018:	Angelo Branduardi "In concerto 2018"

Ancienne Belgique Boulevard Anspach 110, 1000 Brussels

One of Belgium's most renowned venues for contemporary music, this concert hall has welcomed famous artists such as Amy Winehouse, Coldplay, Lou Reed and Queens of the Stone Age. Several upcoming artists have seen their breakthrough after performing here. The website offers an extensive list of the artists performing during the period of the conference, but it is advised to check if tickets are still available. See https://www.abconcerts.be/en/agenda/

Cinematek Rue Baron Horta 9, 1000 Brussels

Film buffs will be delighted to hear that Belgium's greatest cinema is located only a few metres from the conference venue. This isn't a regular cinema showing today's latest blockbusters: Cinematek is the Royal Belgian film archive, and shows several films of its collection every day. These range from silent films - with live piano accompaniment - to more modern movies from all around the world. Be warned when you would like to watch one of the many foreign films: these are only subtitled in Dutch and French. http://cinematek.be/ (unfortunately this site is not available in English)

The music village Rue des Pierres 50, 1000 Brussels

Enjoying a 17th century location at the very heart of Brussels, only a few metres away from the legendary Grand'Place, the Music Village has opened its doors on the 1st of September 2000. With acoustics, sound system, lighting, stage, warm welcome, comfort, and food & drink that live up to the *most prestigious clubs of New-York and London, the Music Village offers all music fans, listeners and musicians alike, a return to the Golden Age of international jazz clubs. On the live program (250 concerts a year), the best Belgian and international jazz musicians, who have adopted this cosmopolitan venue that does justice to its home in the Capital of Europe.*

Doors – 19h

Concert start	Fri - Sat	21h
	Other days	20h30

2. Nightclubs and Dancehalls

Nostalgia Club Greepstraat 49, 1000 Brussels

The Nostalgia Club touts itself as one of the few clubs in Brussels to play the 60s, 70s, 80s, and 90s music. The club's atmosphere allows you to relive the nostalgic times of a bygone age. Drinks are a little pricey but are as extensive as the varied repertoire of music. Open from 10pm on Fridays, Saturdays and nights before public holidays.

DYoukes Rue Duquesnoy 18, 1000 Brussels

This dancehall offers the best of hip hop and the promise of great memories. Only open on Saturday.

Bloody louis Avenue Louise 32, 1050 Brussels

Only five years old, Bloody louis is already one of the main spots in the Brussels' music scene. It has had the opportunity to host Drake, Major Lazer, Martin Solveig and many other live artists and DJs. Tickets can be bought online or at the venue. See https://www.bloodylouis.be/tickets/ *for upcoming events.*

Fuse Rue Blaes 208, 1000 Brussels

Fuse is one of the first nightclubs in Belgium that prominently featured electronic music. Opened in 1994, it has a solid reputation for memorable parties with top DJs. Open every Saturday from 23h and during special events. It's advised to buy tickets in advance as there can be long lines at the entrance of the club. See http://fuse.be/agenda/ *for their events.*

Police, emergency numbers and guard

1) Emergency Services

European emergency number
Valid in Brussels and throughout Europe, for all medical emergencies
Phone: 112
Online: www.sos112.be

Firemen services
Phone: 100
Online: firebru.brussels

Centre Antipoisons (antipoison centre)
Phone: +32 (0)70 245 245
Online: www.poisoncentre.be (FR/NL)

2) Guard duties

Next services assure a guard duty for health problems in Brussels: doctors, dentists and pharmacies.

a) Doctors on duty

- Tel. 02 201 22 22 (in French and Dutch, 7 days a week, 24 hours a day) - www.gbbw.be
- Tel. 02 479 18 18 (in French, 7 days a week, 24 hours a day) - www.medi-garde.be
- Tel. 02 513 02 02 (in French) - www.sosmedecins.be
- Tel. 02 242 43 44 (in Dutch, only during the weekend, www.bhak.be)

This number allows to obtain the address and phone number of a doctor who is close to your place of residence.

b) Dentists on duty

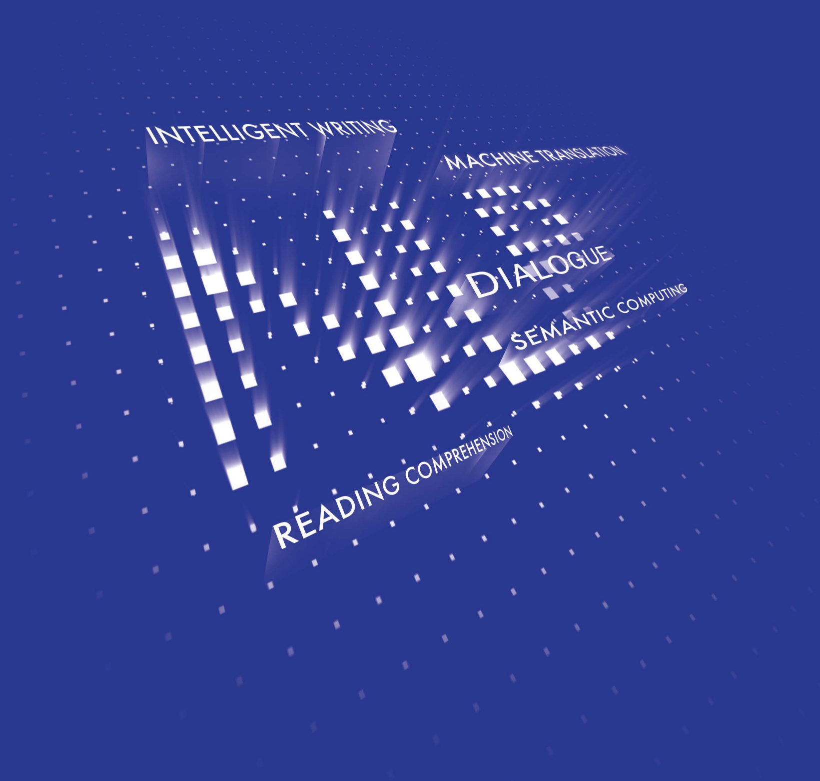
- Tel. 02 426 10 26
- Tel. 0903 399 69 (in Dutch, on Saturday, Sunday and holidays from 9 am to 6 pm)

This number supplies the list of dentists on duty in the evening or during the weekend.

c) Pharmacies on duty

- Tel. 0900 40 090
- Tel. 0903 99 000 (24 hours a day and 7 days a week)
- Tel. 070 660 160 (guard duties pharmacies of Brussels)


Ready to build the platform of tomorrow right now?
We're building the world's most trusted information network for financial professionals. Our 5,000+ engineers, developers, and data scientists are dedicated to advancing and building new solutions and systems for the Bloomberg Terminal in order to solve complex, real-world problems.
Bloomberg.com/careers
Architect on purpose.
Bloomberg
Engineering
© Bloomberg L.P. 2018 277445 0918


INTELLIGENT WRITING
MACHINE TRANSLATION
DIALOGUE
SEMANTIC COMPUTING
READING COMPREHENSION

From the everyday to the exceptional, Microsoft Research pushes boundaries to help you achieve more.
microsoft.com/research
Microsoft

MAXHUB — Efficient Conference Platform

In this diverse and open era,MAXHUB breaks the limitation of previous product design logic , develop all-in-one solution of HD display, touch writing, wireless presenting instead single feature, also compatible with different teleconference software& hardware solution and many office applications, aiming to improve meeting efficiency of financial institutions, technology industries, real estate companies, manufacturing industries, consulting services and administrative organizations.

- Whiteboard fluent writing
- Dual pen writing
- scan to take and share meeting minutes
- Email meeting records
- Annotate in any input channel

- Wireless Mirroring
- Voice Assistant
- Dual camera
- 4K / 1080P high definition display
- High performance modular design

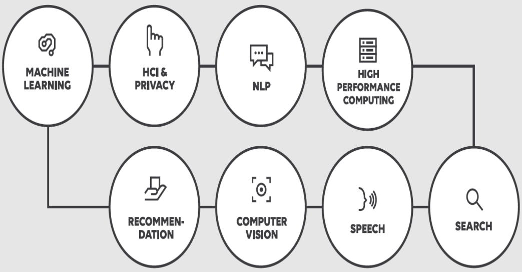
For more information, http://www.maxhub.vip

* Different types have different functions.

CVTE 视源股份 Sub-brand

Clova

AI PLATFORM JOINTLY DEVELOPED BY NAVER AND LINE



NAVER

No 1. search portal in Korea

LINE

Global mobile messenger and content platform

Visit **http://clova.ai** for more information

The EMNLP organizers gratefully acknowledge the support from the following sponsors.

Diamond Tier

Google

facebook

Bloomberg

salesforce

Platinum Tier

amazon

Bai du 百度

grammarly

京东

NAVER LABS EUROPE

visit.brussels

fwo Research Foundation Flanders Opening new horizons

KU LEUVEN Department of Computer Science

Gold Tier

CVTE 视源股份

ebay

Microsoft

NAVER | LINE

ORACLE

POLYAI

Silver Tier

HUAWEI

duolingo

figure eight

NUANCE

Bronze Tier & Supporter Tier

NEXT CANADA

AI2 ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

textkernel Machine Intelligence for People and Jobs

text IQ