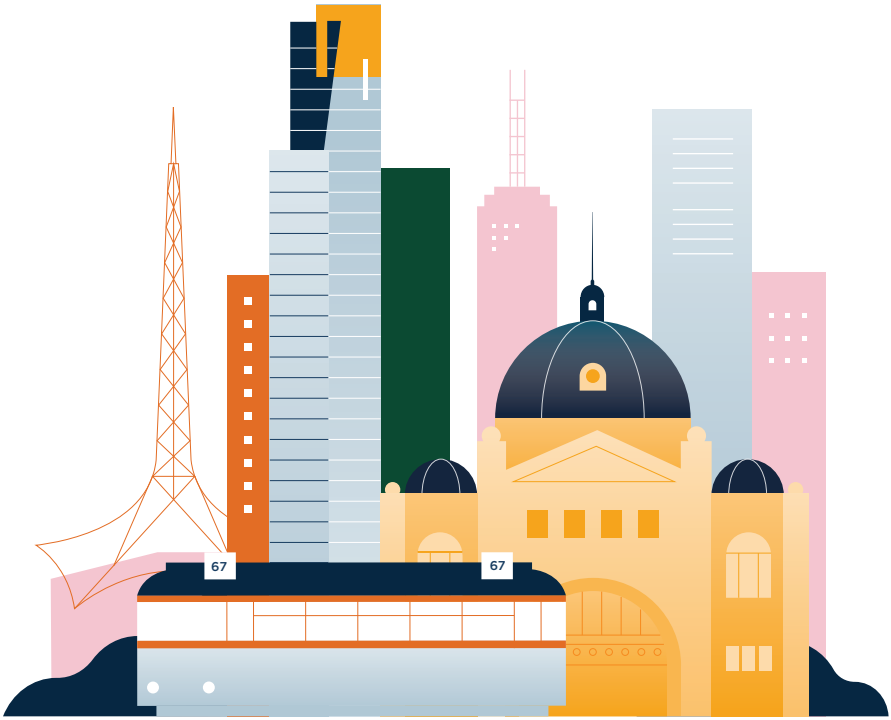
ACL2018

15 – 20 JULY / MELBOURNE AUSTRALIA



CONFERENCE HANDBOOK

56TH ANNUAL MEETING

ASSOCIATION FOR COMPUTATIONAL LINGUISTICS

†: *Start, end and lunch times vary between workshops, the above is a rough guide.*

Cover design by Carrie Chilton
Handbook assembled by Jey Han Lau and Trevor Cohn
Printing by Minuteman Press

Contents

1

Conference Information

Message from the CEO of the Melbourne Convention Bureau

Melbourne Convention Bureau
Collins Square
Tower Two
Level 28, 727 Collins Street
Melbourne VIC 3008

GPO Box 2219T
Melbourne VIC 3001 Australia

Tel +61 3 9002 2222
info@melbournecb.com.au
melbournecb.com.au
ABN 62 072 324 933

On behalf of the Melbourne Convention Bureau, it is my great pleasure to welcome you to Melbourne - officially the world's most liveable city and host of the 56th Annual Meeting of the Association for Computational Linguistics.

Melbourne is renowned as a city that is full of surprises; around every corner, laneway, building and outdoor space, Melbourne is alive with fashion, food, art and culture, and I am excited for you to make these discoveries.

Every Melburnian absolutely loves their city and in the next few days you will find out why. I wish you a very successful and enjoyable visit to Melbourne.

Yours sincerely

KAREN BOLINGER
CHIEF EXECUTIVE OFFICER
MELBOURNE CONVENTION BUREAU

a subsidiary of VISIT VICTORIA

Message from the General Chair

It is an honor to write the initial words of this proceedings as General Chair of the 56th Annual Meeting of the Association for Computational Linguistics! This is only the second time that an ACL conference has been held in Australia — the first time was for the joint COLING/ACL conference in June of 2006 in Sydney, and I was one of its Program Chairs. For ACL 2018 we have tried to maintain the welcoming and intimate spirit and the relaxed and genial character of the much smaller ACL conferences of the past in spite of the ever-growing number of researchers in the field and participants in our conferences.

It is my pleasure here to express gratitude to all those without whom this conference would not exist. My biggest thanks go to the Program Chairs Iryna Gurevych and Yusuke Miyao, as well as to Local Chairs Tim Baldwin, Trevor Cohn and Karin Verspoor. They have done a tremendous job to manage the submission and review process, and the local arrangement details, respectively.

I also want to thank all of the other chairs for their very hard work: Workshops Chairs Brendan O'Connor and Eva Maria Vecchi; Tutorials Chairs Yoav Artzi and Jacob Eisenstein; Demo Chairs Fei Liu and Thamar Solorio; Student Research Workshop Organizers Vered Shwartz, Jeniya Tabassum and Rob Voigt; Faculty Advisors to the Student Research Workshop Marie-Catherine de Marneffe, Wanxiang Che and Malvina Nissim; Publications Chairs Shay Cohen, Kevin Gimpel and Wei Lu; Exhibits Coordinator Karin Verspoor; Student Volunteer Coordinator Karin Vespoor; Conference Handbook Chairs Jey Han Lau and Trevor Cohn; Publicity Chair Sarvnaz Karimi; Local Sponsorship Chair Cecile Paris; Webmaster Andrew MacKinlay; and Priscilla Rasmussen, giver of advice and wisdom to all of us as ACL Business Manager.

I also warmly thank the ACL Executive Committee for its guidance and advice on many important issues and concerns as they arose.

I am also extremely grateful to all the sponsors for their great support to the conference.

Many thanks to the area chairs, the reviewers, the invited speakers, the authors of the various papers, posters and presentations.

And, finally, many many thanks to all the participants who will put the final touches on making ACL 2018 an exciting, stimulating and inspiring event!

ACL 2018 General Chair
Claire Cardie, Cornell University

Message from the Program Committee Co-Chairs

Welcome to the 56th Annual Meeting of the Association for Computational Linguistics 2018 – or ACL 2018 for short.

In September 2017, Program Committee Co-Chairs (PCs) posted the call for nominations of Area Chairs (AC), Reviewers and Invited Speakers. We received 752 responses in total. Overall, out of 388 valid nominations for area chairs, 299 unique persons were suggested; 110 persons were self-nominations. About 70% of the 56 selected area chairs (later expanded to 61 area chairs due to the high number of submissions) were nominated by the community. For the reviewers, we collected 936 valid nominations. At the PhD level, 139 persons were self-nominations and 129 were nominated by others. At the Postdoc/Ass.Prof. level, 160 were self-nominated, 112 nominated by others. At the Prof. level, 221 persons were self-nominated, 175 nominated by others.

We received 138 unique nominations for invited speakers, from which two invited speakers of the conference were selected:

- Carolyn Penstein Rosé, Language Technologies Institute at Carnegie Mellon University, USA
- Anton van den Hengel, Australian Centre for Visual Technologies at University of Adelaide, Australia

Our community is steadily growing: in total, 1621 submissions were received right after the submission deadline: 1045 long, 576 short papers. 13 erroneous submissions were deleted or withdrawn in the preliminary checks by PCs. 25 papers were rejected without review (16 long, 9 short); the reasons are the violation of the ACL 2018 style and dual submission guidelines. 32 papers were withdrawn before the review period started; the main reason was that the papers have been accepted as the short papers at NAACL HLT 2018. In total, 1551 papers went into the reviewing phase: 1021 long, 530 short papers. 1610 reviewers (1473 primary and 137 secondary reviewers) were involved in the reviewing process; each reviewer has reviewed about 3 papers on average. 3 long and 4 short papers were withdrawn during the reviewing period, and finally 1018 long and 526 short papers were considered during the acceptance decision phase.

The assignment of papers to areas and reviewers has been done in multiple rounds. First round: Initial assignments of papers to areas were determined automatically with the help of the authors' input, while PCs went through all submissions and moved papers to other areas, considering COI and the topical fit. PCs assigned one AC as a meta-reviewer to each paper using Toronto Paper Matching System (TPMS) scores. Second round: ACs looked into the papers in their area, and adjusted meta-reviewer assignments. ACs sent a report to PCs if they found any problems. Third round: PCs made the final decision, considering the workload balance, possible COIs and the topical fit. Fourth round: ACs decided which reviewers would review each paper, based on AC's knowledge about the reviewers, TPMS scores, reviewers' bids, and COI.

We have introduced several innovations to the reviewing process. One of them is an argument-based review form. The reviewers were asked to provide arguments for and against the paper. This has been tremendously helpful for ACs and PCs to analyze the reviews and come up with final recommendations. The authors were asked to respond to the con arguments during the rebuttal. In coordination with the NAACL HLT 2018 PCs, we plan to do some analytics on anonymized reviews and rebuttal statements, with the consent of the reviewers and authors. Our purpose is to improve the quality of the review process. The data will be compiled into a unique corpus for NLP, and will be made available to the research community after appropriate anonymization checks, at the earliest in 2 years after ACL 2018. We hope to provide data on *how to review* to younger researchers, and to improve the transparency of the reviewing process in general.

The ACL 2018 conference is super-competitive: We accepted 256 out of 1018 submitted long papers and 125 out of 526 short papers, with an overall acceptance rate of 24.7%. The details of the review process are available at the conference homepage. Criteria of acceptance were mainly:

- strengths/weaknesses raised by reviewers and their significance;
- the result of discussions and author responses;
- contribution to CL as the science of language: whether the paper advances (or contributes to) our understanding of language in any way;
- diversity: we do not want to fill ACL with similar papers like achieving 1% improvement on a well-known task.

We also considered the balance of paper types, topics and contributions and re-considered the acceptance when reviewers reported any problem in preliminary checks (*Appropriateness* to *Handling of Human Participants*).

Continuing the tradition, ACL 2018 will feature 20 papers which were accepted for publication in the Transactions of the Association for Computational Linguistics (TACL). The TACL papers were split into 10 oral presentations and 10 poster presentations.

There are many people to thank for who have worked diligently to make ACL 2018 possible. All names are listed in the Program Committee section of the Front Matter.

Since the conference size continues to grow and the organizational complexity increases, we have introduced the role of Program Committee Co-Chair Assistants. In total, 5 senior researchers have supported the PCs during most intensive work phases to handle the communication in a timely manner, draft various documents and effectively prepare decisions.

Thanks to our area chairs for their hard work on recruiting reviewers, managing reviews, leading discussions, and making recommendations.

This program certainly would not be possible without the help of the 1610 reviewers. In particular, 192 reviewers from this list were recognized by the area chairs as outstanding reviewers who have turned in exceptionally well-written and constructive reviews and who have actively engaged themselves in the post-rebuttal discussions.

We are also deeply indebted to the best paper selection committee which consists of 22 members. They had to additionally review 6-8 papers according to the best paper criteria on short notice. Their time and effort in recommending the best paper awards is much appreciated.

We also would like to thank many colleagues for generously sharing their experience in organizing prior ACL conferences and for their advice. We are grateful for the guidance and the support of the ACL presidents Joakim Nivre and Marti Hearst, and the ACL board. We also would like to thank the publication co-chairs Shay Cohen, Kevin Gimpel and Wei Lu (Advisory) and the handbook chair Jey Han Lau for putting together the proceedings and the conference handbook; and Rich Gerber from Softconf for always being responsive to our requests. We would like to thank the ACL Business Manager Priscilla Rasmussen for helping us to sort important things out. Finally, this conference could not have happened without the efforts of the general chair, Claire Cardie. We thank her for the leadership and advice, especially when matters got complicated.

We hope you will enjoy ACL 2018 and contribute to the future success of our community!

ACL 2018 Program Committee Co-Chairs
Iryna Gurevych, TU Darmstadt, Germany
Yusuke Miyao, National Institute of Informatics, Japan

Organizing Committee

General Chair
Claire Cardie, Cornell University

Program Chairs
Iryna Gurevych, TU Darmstadt
Yusuke Miyao, National Institute of Informatics

Workshop Chairs
Brendan O'Connor, University of Massachusetts Amherst
Eva Maria Vecchi, University of Cambridge

Tutorial Chairs
Yoav Artzi, Cornell University
Jacob Eisenstein, Georgia Institute of Technology

Demo Chairs
Fei Liu, University of Central Florida
Thamar Solorio, University of Houston

Publication Chairs
Shay Cohen, University of Edinburgh
Kevin Gimpel, Toyota Technological Institute at Chicago
Wei Lu, Singapore University of Technology and Design (Advisory)

Exhibits Coordinator
Karin Verspoor, University of Melbourne

Conference Handbook Chairs
Jey Han Lau, IBM Research
Trevor Cohn, University of Melbourne

Publicity Chair
Sarvnaz Karimi, CSIRO

Local Sponsorship Chair
Cecile Paris, CSIRO

Student Volunteer Coordinator
Karin Verspoor, University of Melbourne

Local Chairs
Tim Baldwin, University of Melbourne
Karin Verspoor, University of Melbourne
Trevor Cohn, University of Melbourne

Student Research Workshop Organisers
Vered Shwartz, Bar-Ilan University
Jeniya Tabassum, Ohio State University
Rob Voigt, Stanford University

Faculty Advisors to the Student Research Workshop
Marie-Catherine de Marneffe, Ohio State University
Wanxiang Che, Harbin Institute of Technology
Malvina Nissim, University of Groningen

Webmaster
Andrew MacKinlay, Culture Amp / University of Melbourne

Program Committee

Program Committee Co-chairs
Iryna Gurevych, TU Darmstadt, Germany
Yusuke Miyao, National Institute of Informatics, Japan

Program Committee Co-Chair Assistants
Yang Gao, TU Darmstadt, Germany
Ivan Habernal, TU Darmstadt, Germany
Sang Phan, National Institute of Informatics, Japan
Steffen Eger, TU Darmstadt, Germany
Christian Meyer, TU Darmstadt, Germany

Area Chairs
(Senior Area Chairs are indicated in boldface.)

Dialogue and Interactive Systems
Asli Celikyilmaz
Verena Rieser
Milica Gasic
Jason Williams

Discourse and Pragmatics
Manfred Stede
Ani Nenkova

Document Analysis
Hang Li
Yiqun Liu
Eugene Agichtein

Generation
Ioannis Konstas
Claire Gardent

Information Extraction and Text Mining
Feiyu Xu
Kevin Cohen
Zhiyuan Liu
Ralph Grishman
Yi Yang
Nazli Goharian

Linguistic Theories, Cognitive Modeling and Psycholinguistics
Shuly Wintner
Tim O'Donnell

Machine Learning
Andre Martins
Ariadna Quattoni
Jun Suzuki

Machine Translation
Yang Liu (Tsinghua University)
Matt Post
Lucia Specia
Dekai Wu

Multidisciplinary and Area Chair COI
Yoav Goldberg
Anders Søgaard
Mirella Lapata

Multilinguality

Bernardo Magnini
Tristan Miller

Phonology, Morphology, and Word Segmentation
Graham Neubig
Hai Zhao

Question Answering
Lluís Màrquez
Teruko Mitamura
Zornitsa Kozareva
Richard Socher

Resources and Evaluation
Gerard de Melo
Sara Tonelli
Karën Fort

Sentence-level Semantics
Luke Zettlemoyer
Ellie Pavlick
Jacob Uszkoreit

Sentiment Analysis and Argument Mining
Smaranda Muresan
Benno Stein
Yulan He

Social Media
David Jurgens
Jing Jiang

Summarization
Kathleen McKeown
Xiaodan Zhu

Tagging, Chunking, Syntax, and Parsing
Liang Huang
Weiwei Sun
Željko Agić
Yue Zhang

Textual Inference and Other Areas of Semantics
Michael Roth
Fabio Massimo Zanzotto

Vision, Robotics, Multimodal, Grounding and Speech
Yoav Artzi
Shinji Watanabe
Timothy Hospedales

Word-level Semantics
Ekaterina Shutova
Roberto Navigli

Best Paper Selection Committee
Timothy Baldwin
Pushpak Bhattacharyya
Phil Blunsom
Johan Bos
Jordan Boyd-Graber
Trevor Cohn
Vera Demberg
Kevin Duh

Katrin Erk
Mark Johnson
Yang Liu (Tsinghua University)
Yuji Matsumoto
Jong Park
Ellie Pavlick
Simone Paolo Ponzetto
Sebastian Riedel
Carolyn Penstein Rosé
Noah A. Smith
Anders Søgaard
Ivan Titov
Benjamin Van Durme
Ming Zhou

Meal Info

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

- Mid-morning breaks include coffee, tea, water, and snacks in the Melbourne Room and level 2 foyer
- A light lunch of sandwiches and a muffin is provided during each of the midday poster sesions on Monday–Wednesday, in the Melbourne Room and level 2 foyer
- Mid-afternoon breaks include coffee, tea, water, and snacks in the Melbourne Room and level 2 foyer

Light refreshments will be provided for students on Monday at the recruitment event, and there is also the Welcome Reception and the Social Event, but you are otherwise on your own.

2

Tutorials: Sunday, July 15

Overview

07:30–18:00	**Registration**	*Level 2 Foyer*
09:00–12:30	**Morning Tutorials**	
	100 Things You Always Wanted to Know about Semantics & Pragmatics But Were Afraid to Ask *Emily M. Bender*	*216*
	Neural Approaches to Conversational AI *Jianfeng Gao, Michel Galley, and Lihong Li*	*219*
	Variational Inference and Deep Generative Models *Wilker Aziz and Philip Schulz*	*217*
	Connecting Language and Vision to Actions *Peter Anderson, Abhishek Das, and Qi Wu*	*218*
10:30–11:00	**Coffee Break**	*Level 2 Foyer*
12:30–13:30	**Lunch Break**	
13:30–17:00	**Afternoon Tutorials**	
	Beyond Multiword Expressions: Processing Idioms and Metaphors *Valia Kordoni*	*216*
	Neural Semantic Parsing *Luke Zettlemoyer, Matt Gardner, Pradeep Dasigi, Srinivasan Iyer, and Alane Suhr*	*217*
	Deep Reinforcement Learning for NLP *William Yang Wang, Jiwei Li, and Xiaodong He*	*219*
	Multi-lingual Entity Discovery and Linking *Avirup Sil, Heng Ji, Dan Roth, and Silviu-Petru Cucerzan*	*218*
15:00–15:30	**Coffee Break**	*Level 2 Foyer*
18:00–20:00	**Welcome Reception**	*Melbourne Room 1*

Message from the Tutorial Co-Chairs

This volume contains the abstracts of the ACL 2018 tutorials. Tutorials were selected from 49 submissions to a joint call, which was coordinated with NAACL, COLING, and EMNLP. From these submissions, eight half-day tutorials were selected for ACL, on the criteria of quality, relevance, interest, and balance. We thank Mausam for coordinating this process across all four conferences, and we wish to acknowledge support from the publications chairs Kevin Gimpel, Shay Cohen, and Wei Lu (ACL publications chairs) and Jey Han Lau and Trevor Cohn (ACL handbook chairs), as well as Stephanie Lukin (NAACL publications co-chair). Most importantly, we thank the tutorial presenters for their contributions, which we hope that you will enjoy.

ACL 2018 Tutorial Co-Chairs
Yoav Artzi, Cornell University Jacob Eisenstein, Georgia Institute of Technology

Tutorial 1

100 Things You Always Wanted to Know about Semantics & Pragmatics But Were Afraid to Ask

Emily M. Bender

Sunday, July 15, 2018, 9:00–12:30

216

Meaning is a fundamental concept in Natural Language Processing (NLP), given its aim to build systems that mean what they say to you, and understand what you say to them. In order for NLP to scale beyond partial, task-specific solutions, it must be informed by what is known about how humans use language to express and understand communicative intents. The purpose of this tutorial is to present a selection of useful information about semantics and pragmatics, as understood in linguistics, in a way that's accessible to and useful for NLP practitioners with minimal (or even no) prior training in linguistics. The tutorial content is based on a manuscript in progress I am co-authoring with Prof. Alex Lascarides of the University of Edinburgh.

Emily M. Bender is a Professor in the Department of Linguistics and Adjunct Professor in the Paul G. Allen School of Computer Science & Engineering at the University of Washington. She is also the past chair (2016–2017) of NAACL. Her research interests lie in multilingual grammar engineering, computational semantics, and the incorporation of linguistic knowledge in natural language processing. She is the primary developer of the Grammar Matrix grammar customization system, which is developed in the context of the DELPH-IN Consortium (Deep Linguistic Processing with HPSG Initiative). More generally, she is interested in the intersection of linguistics and computational linguistics, from both directions: bringing computational methodologies to linguistic science and linguistic science to natural language processing.

Tutorial 2

Neural Approaches to Conversational AI

Jianfeng Gao, Michel Galley, and Lihong Li

Sunday, July 15, 2018, 9:00–12:30

219

This tutorial surveys neural approaches to conversational AI that were developed in the last few years. We group conversational systems into three categories: (1) question answering agents, (2) task-oriented dialogue agents, and (3) social bots. For each category, we present a review of state-of-the-art neural approaches, draw the connection between neural approaches and traditional symbolic approaches, and discuss the progress we have made and challenges we are facing, using specific systems and models as case studies.

Jianfeng Gao is Partner Research Manager at Microsoft AI and Research, Redmond. He leads the development of AI systems for machine reading comprehension, question answering, chitchat bots, task-oriented dialogue, and business applications. From 2014 to 2017, he was Partner Research Manager and Principal Researcher at Deep Learning Technology Center at Microsoft Research, Redmond, where he was leading the research on deep learning for text and image processing. From 2006 to 2014, he was Researcher, Senior Researcher, and Principal Researcher at Natural Language Processing Group at Microsoft Research, Redmond, where he worked on the Bing search engine, improving its core relevance engine and query spelling, understanding and reformulation engines, MS ads relevance and prediction, and statistical machine translation. From 2005 to 2006, he was a Research Lead in Natural Interactive Services Division at Microsoft, where he worked on Project X, an effort of developing natural user interface for Windows. From 2000 to 2005, he was Research Lead in Natural Language Computing Group at Microsoft Research Asia, where he and his colleagues developed the first Chinese speech recognition system released with Microsoft Office, the Chinese/Japanese Input Method Editors (IME) which were the leading products in the market, and the natural language platform for Microsoft Windows.

Michel Galley is a Senior Researcher at Microsoft Research. His research interests are in the areas of natural language processing and machine learning, with a particular focus on conversational AI, text generation, and machine translation. From 2007 to 2010, he was a Postdoctoral Scholar then Research Associate in the Computer Science department at Stanford University, working primarily on Machine Translation. In 2007, he obtained his Ph.D. from the Computer Science department at Columbia University, with research focusing on summarization, discourse, and dialogue. From 2003 to 2005, he did several internships at USC/ISI and Language Weaver on machine translation, which included work that won several NIST MT competitions. From 2000-2001, he did an 8-month internship and undergraduate thesis work in the Spoken Dialog Systems group at Bell Labs, working on generation for dialogue systems.

Lihong Li is a Research Scientist at Google Inc. He obtained a PhD degree in Computer Science from Rutgers University, specializing in reinforcement learning theory and algorithms. After that, he has held Researcher, Senior Researcher, and Principal Researcher positions in Yahoo! Research (2009-2012) and Microsoft Research (2012-2017), before joining Google. His main research interests are reinforcement learning (in both Markov decision processes and contextual bandits) and other related problems in AI (including active leaning, online learning and large-scale machine learning). His work has found applications in recommendation, advertising, Web search and conversation systems, and has won best paper awards at ICML, AISTATS and WSDM. In recent years, he served as area chairs or senior program

committee members at AAAI, AISTATS, ICML, IJCAI and NIPS. More information can be found on his homepage: `http://lihongli.github.io`.

Tutorial 3

Variational Inference and Deep Generative Models

Wilker Aziz and Philip Schulz

Sunday, July 15, 2018, 9:00–12:30

217

NLP has seen a surge in neural network models in recent years. These models provide state-of-the-art performance on many supervised tasks. Unsupervised and semi-supervised learning has only been addressed scarcely, however. Deep generative models (DGMs) make it possible to integrate neural networks with probabilistic graphical models. Using DGMs one can easily design latent variable models that account for missing observations and thereby enable unsupervised and semi-supervised learning with neural networks. The method of choice for training these models is variational inference.

This tutorial offers a general introduction to variational inference followed by a thorough and example-driven discussion of how to use variational methods for training DGMs. It provides both the mathematical background necessary for deriving the learning algorithms as well as practical implementation guidelines. Importantly, the tutorial will cover models with continuous and discrete variables.

We provide practical coding exercises implemented in IPython notebooks as well as short notes on the more intricate mathematical details that the audience can use as a reference after the tutorial. We expect that with these additional materials the tutorial will have a long-lasting impact on the community.

Wilker Aziz is a research associate at the University of Amsterdam (UvA) working on natural language processing problems such as machine translation, textual entailment, and paraphrasing. His research interests include statistical learning, probabilistic models, and methods for approximate inference. Before joining UvA, Wilker worked on exact sampling and optimisation for statistical machine translation at the University of Sheffield (UK) and at the University of Wolverhampton (UK) where he obtained his PhD. Wilker's background is in Computer Engineering which he studied at the Engineering School of the University of São Paulo (Brazil).

Philip Schulz is an applied scientist at Amazon Research. Before joining Amazon, Philip did his PhD at the University of Amsterdam. During the last months of his PhD trajectory, he visited the University of Melbourne. Philip's background is in Linguistics which he studied at the University of Tübingen and UCL in London. These days, his research interests revolve around statistical learning. He has worked on Bayesian graphical models for machine translation. More recently he has extended this line of work towards deep generative models. More broadly, Philip is interested in probabilistic modeling, approximate inference methods and statistical theory.

Tutorial 4

Connecting Language and Vision to Actions

Peter Anderson, Abhishek Das, and Qi Wu

Sunday, July 15, 2018, 9:00–12:30

218

A long-term goal of AI research is to build intelligent agents that can see the rich visual environment around us, communicate this understanding in natural language to humans and other agents, and act in a physical or embodied environment. To this end, recent advances at the intersection of language and vision have made incredible progress – from being able to generate natural language descriptions of images/videos, to answering questions about them, to even holding free-form conversations about visual content! However, while these agents can passively describe images or answer (a sequence of) questions about them, they cannot act in the world (what if I cannot answer a question from my current view, or I am asked to move or manipulate something?). Thus, the challenge now is to extend this progress in language and vision to embodied agents that take actions and actively interact with their visual environments. To reduce the entry barrier for new researchers, this tutorial will provide an overview of the growing number of multimodal tasks and datasets that combine textual and visual understanding. We will comprehensively review existing state-of-the-art approaches to selected tasks such as image captioning, visual question answering (VQA) and visual dialog, presenting the key architectural building blocks (such as co-attention) and novel algorithms (such as cooperative/adversarial games) used to train models for these tasks. We will then discuss some of the current and upcoming challenges of combining language, vision and actions, and introduce some recently-released interactive 3D simulation environments designed for this purpose.

Peter Anderson is a final year PhD candidate in Computer Science at the Australian National University, supervised by Dr Stephen Gould, and a researcher within the Australian Centre for Robotic Vision (ACRV). His PhD focuses on deep learning for visual understanding in natural language. He was an integral member of the team that won first place in the 2017 Visual Question Answering (VQA) challenge at CVPR, and he has made several contributions in image captioning, including achieving first place on the COCO leaderboard in July 2017. He has published at CVPR, ECCV, EMNLP and ICRA, and spent time at numerous universities and research labs including Adelaide University, Macquarie University, and Microsoft Research. His research is currently focused on vision-and-language understanding in complex 3D environments.

Abhishek Das is a Computer Science PhD student at Georgia Institute of Technology, advised by Dhruv Batra, and working closely with Devi Parikh. He is interested in deep learning and its applications in building agents that can see (computer vision), think (reasoning and interpretability), talk (language modeling) and act (reinforcement learning). He is a recipient of an Adobe Research Fellowship and a Snap Research Fellowship. He has published at CVPR, ICCV, EMNLP, HCOMP and CVIU, co-organized the NIPS 2017 workshop on Visually-Grounded Interaction and Language, and has held visiting positions at Virginia Tech, Queensland Brain Institute and Facebook AI Research. He graduated from Indian Institute of Technology Roorkee in 2015 with a Bachelor's in Electrical Engineering.

Qi Wu is a research fellow in the Australia Centre for Robotic Vision (ACRV) in the University of Adelaide. Before that, he was a postdoc researcher in the Australia Centre for Visual Technologies (ACVT) in the University of Adelaide. He obtained his PhD degree in 2015 and MSc degree in 2011, in Computer Science from University of Bath, United Kingdom. His research interests are mainly in Computer Vision

and Machine Learning. Currently, he is working on the vision to language problem and he is especially an expert in the area of Image Captioning and Visual Question Answering (VQA). His attributes-based image captioning model got first place on the COCO Image Captioning Challenge Leader Board in the October of 2015. He has published several papers in prestigious conferences and journals, such as TPAMI, CVPR, ICCV, ECCV, IJCAI and AAAI.

Tutorial 5

Beyond Multiword Expressions: Processing Idioms and Metaphors

Valia Kordoni

Sunday, July 15, 2018, 13:30–17:00

216

Idioms and metaphors are characteristic to all areas of human activity and to all types of discourse. Their processing is a rapidly growing area in NLP, since they have become a big challenge for NLP systems. Their omnipresence in language has been established in a number of corpus studies and the role they play in human reasoning has also been confirmed in psychological experiments. This makes idioms and metaphors an important research area for computational and cognitive linguistics, and their automatic identification and interpretation indispensable for any semantics-oriented NLP application.

This tutorial aims to provide attendees with a clear notion of the linguistic characteristics of idioms and metaphors, computational models of idioms and metaphors using state-of-the-art NLP techniques, their relevance for the intersection of deep learning and natural language processing, what methods and resources are available to support their use, and what more could be done in the future. Our target audience are researchers and practitioners in machine learning, parsing (syntactic and semantic) and language technology, not necessarily experts in idioms and metaphors, who are interested in tasks that involve or could benefit from considering idioms and metaphors as a pervasive phenomenon in human language and communication.

Valia Kordoni is a professor at Humboldt University Berlin (Deputy Chair for the subject area "English Linguistics"). She is a leader in EU-funded research in Machine Translation, Computational Semantics, and Machine Learning. She has organized conferences and workshops dedicated to research on MWEs, recently including the EACL 2014 10th Workshop on Multiword Expressions (MWE 2014) in Gothenburg, Sweden, the NAACL 2015 11th Workshop on Multiword Expressions in Denver, Colorado, and the ACL 2016 12th Workshop on Multiword Expressions in Berlin, Germany, among others. She has been the Local Chair of ACL 2016 - The 54th Annual Meeting of the Association for Computational Linguistics which took place at the Humboldt University Berlin in August 2016. Recent activities of hers include a tutorial on Robust Automated Natural Language Processing with Multiword Expressions and Collocations in ACL 2013, as well as a tutorial on Beyond Words: Deep Learning for Multiword Expressions and Collocations in ACL 2017. She is the author of Multiword Expressions - From Linguistic Analysis to Language Technology Applications (to appear, Springer).

Tutorial 6

Neural Semantic Parsing

Luke Zettlemoyer, Matt Gardner, Pradeep Dasigi, Srinivasan Iyer, and Alane Suhr

Sunday, July 15, 2018, 13:30–17:00

217

Semantic parsing, the study of translating natural language utterances into machine-executable programs, is a well-established research area and has applications in question answering, instruction following, voice assistants, and code generation. In the last two years, the models used for semantic parsing have changed dramatically with the introduction of neural encoder-decoder methods that allow us to rethink many of the previous assumptions underlying semantic parsing. We aim to inform those already interested in semantic parsing research of these new developments in the field, as well as introduce the topic as an exciting research area to those who are unfamiliar with it.

Current approaches for neural semantic parsing share several similarities with neural machine translation, but the key difference between the two fields is that semantic parsing translates natural language into a formal language, while machine translation translates it into a different natural language. The formal language used in semantic parsing allows for constrained decoding, where the model is constrained to only produce outputs that are valid formal statements. We will describe the various approaches researchers have taken to do this. We will also discuss the choice of formal languages used by semantic parsers, and describe why much recent work has chosen to use standard programming languages instead of more linguistically-motivated representations. We will then describe a particularly challenging setting for semantic parsing, where there is additional context or interaction that the parser must take into account when translating natural language to formal language, and give an overview of recent work in this direction. Finally, we will introduce some tools available in AllenNLP for doing semantic parsing research.

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 maintainer of the AllenNLP toolkit and a host of the NLP Highlights podcast.

Pradeep Dasigi is a PhD student at the Language Technologies Institute in Carnegie Mellon University. His research interest lies in building knowledge-aware language understanding systems, with a recent focus on neural semantic parsing.

Srinivasan Iyer is a graduate student in the Natural Language Processing group at the University of Washington, Seattle. His main research area is context dependent semantic parsing directly from natural language to general purpose programming source code. Other aspects of his research are learning semantic parsers from massive online resources and incorporating user feedback for model improvement.

Alane Suhr is a PhD student in Computer Science at Cornell University. Alane's research interests include developing machine learning methods for understanding natural language grounded in complex environments and interactions. She is a recipient of an NSF Graduate Research Fellowship, the Best Resource Paper award at ACL 2017, and an Outstanding Paper Award at NAACL 2018.

Luke Zettlemoyer is an Associate Professor in the Paul G. Allen School of Computer Science & Engineering at the University of Washington. He has a been doing research in semantic parsing for many years, and recently shifted to studying neural models for this problem. Luke's honors include multiple best paper awards, a PECASE award, and an Allen Distinguished Investigator award.

Tutorial 7

Deep Reinforcement Learning for NLP

William Yang Wang, Jiwei Li, and Xiaodong He

Sunday, July 15, 2018, 13:30–17:00

219

Many Natural Language Processing (NLP) tasks (including generation, language grounding, reasoning, information extraction, coreference resolution, and dialog) can be formulated as deep reinforcement learning (DRL) problems. However, since language is often discrete and the space for all sentences is infinite, there are many challenges for formulating reinforcement learning problems of NLP tasks. In this tutorial, we provide a gentle introduction to the foundation of deep reinforcement learning, as well as some practical DRL solutions in NLP. We describe recent advances in designing deep reinforcement learning for NLP, with a special focus on generation, dialogue, and information extraction. Finally, we discuss why they succeed, and when they may fail, aiming at providing some practical advice about deep reinforcement learning for solving real-world NLP problems.

William Yang Wang is an Assistant Professor at the Department of Computer Science, University of California, Santa Barbara. He received his PhD from School of Computer Science, Carnegie Mellon University. He focuses on information extraction and he is the faculty author of DeepPath—the first deep reinforcement learning system for multi-hop reasoning. He has published more than 50 papers at leading conferences and journals including ACL, EMNLP, NAACL, CVPR, COLING, IJCAI, CIKM, ICWSM, SIGDIAL, IJCNLP, INTERSPEECH, ICASSP, ASRU, SLT, Machine Learning, and Computer Speech & Language, and he has received paper awards and honors from CIKM, ASRU, and EMNLP. Website: `http://www.cs.ucsb.edu/~william/`.

Jiwei Li recently spent three years and received his PhD in Computer Science from Stanford University. His research interests are deep learning and dialogue. He is the most prolific NLP/ML first author during 2012-2016, and the lead author of the first study in deep reinforcement learning for dialogue generation. He is the recipient of a Facebook Fellowship in 2015. Website: `https://web.stanford.edu/~jiweil/`.

Xiaodong He is the Deputy Managing Director of JD AI Research and Head of the Deep learning, NLP and Speech Lab, and a Technical Vice President of JD.com. He is also an Affiliate Professor at the University of Washington (Seattle), serves in doctoral supervisory committees. Before joining JD.com, He was with Microsoft for about 15 years, served as Principal Researcher and Research Manager of the DLTC at Microsoft Research, Redmond. His research interests are mainly in artificial intelligence areas including deep learning, natural language, computer vision, speech, information retrieval, and knowledge representation. He has published more than 100 papers in ACL, EMNLP, NAACL, CVPR, SIGIR, WWW, CIKM, NIPS, ICLR, ICASSP, Proc. IEEE, IEEE TASLP, IEEE SPM, and other venues. He received several awards including the Outstanding Paper Award at ACL 2015. Website: `http://air.jd.com/people2.html`.

Tutorial 8

Multi-lingual Entity Discovery and Linking

Avirup Sil, Heng Ji, Dan Roth, and Silviu-Petru Cucerzan

Sunday, July 15, 2018, 13:30–17:00

218

The primary goals of this tutorial are to review the framework of cross-lingual EL and motivate it as a broad paradigm for the Information Extraction task. We will start by discussing the traditional EL techniques and metrics and address questions relevant to the adequacy of these to across domains and languages. We will then present more recent approaches such as Neural EL, discuss the basic building blocks of a state-of-the-art neural EL system and analyze some of the current results on English EL. We will then proceed to Cross-lingual EL and discuss methods that work across languages. In particular, we will discuss and compare multiple methods that make use of multi-lingual word embeddings. We will also present EL methods that work for both name tagging and linking in very low resource languages. Finally, we will discuss the uses of cross-lingual EL in a variety of applications like search engines and commercial product selling applications. Also, contrary to the 2014 EL tutorial, we will also focus on Entity Discovery which is an essential component of EL.

Avirup Sil is a Research Staff Member and the chair of the NLP community at IBM Research AI. His research interests are in multi-lingual information extraction from large text collection (cross-lingual entity extraction, disambiguation and slot filling), machine learning and knowledge representation. Avi has published several papers on Entity Linking and his systems at IBM have obtained top scores in TAC-KBP annual multi-lingual entity linking evaluations. Avi is an area chair for Information Extraction at NAACL 2018 and also for COLING 2018. He is also organizing the workshop on the "Relevance of Linguistic Structure in Neural NLP" at ACL 2018.

Heng Ji is the Edward G. Hamilton Development Chair Professor in Computer Science Department of Rensselaer Polytechnic Institute. Her research interests focus on Natural Language Processing, especially on Crosssource Information Extraction and Knowledge Base Population. She coordinated the NIST TAC Knowledge Base Population task since 2010 and has published many papers on entity discovery and linking. Heng has co-taught the "Wikification and Beyond: The Challenges of Entity and Concept Grounding" tutorial with Dan Roth at ACL 2014.

Dan Roth is the Eduardo D. Glandt Distinguished Professor at the Department of Computer and Information Science, University of Pennsylvania. He is a fellow of AAAS, AAAI, ACL, and the ACM and the winner of the IJCAI-2017 John McCarthy Award, for "major conceptual and theoretical advances in the modeling of natural language understanding, machine learning, and reasoning." Roth has published broadly in machine learning, natural language processing, knowledge representation and reasoning, and has developed several machine learning based natural language processing systems that are widely used in the computational linguistics community and in industry. Over the last few years he has worked on Entity Linking and Wikification. He has taught several tutorials at ACL/NAACL/ECL and other forums. Dan has co-taught the "Wikification and Beyond: The Challenges of Entity and Concept Grounding" tutorial with Heng Ji at ACL 2014.

Silviu-Petru Cucerzan is a Principal Researcher at Microsoft Research and the Bing Knowledge Graph group. His research has focused on topics at the intersection of NLP and IR with concrete applications to industry, including multilingual spelling correction, question answering, entity recognition and linking, query suggestion, vertical search, and ads selection. Many of the technologies developed by Silviu

have been shipped with Microsoft products. The NEMO entity linking system developed by Silviu has scored the top performance during the four consecutive years it participated in the TAC-KBP evaluations organized by NIST and LDC.

Welcome Reception

Sunday, July 15, 2018, 6:00pm – 8:00pm

Melbourne Convention and Exhibition Centre (conference venue)
Melbourne Room 1

Catch up with your colleagues at the **Welcome Reception**! It will be held immediately following the Tutorials on Sunday, July 15 at 6:00pm in the Melbourne Room (level 2) of the Melbourne Convention and Exhibition Centre (the conference venue). Light refreshments will be provided, along with an open bar.

3

Main Conference: Monday, July 16

Overview

07:30–18:00 **Registration** *Level 2 Foyer*

09:00–10:00 **Welcome Session & Presidential Address** (Sponsors: Facebook & Baidu) *Plenary*

10:00–10:30 **Coffee Break** *Level 2 Foyer and Melbourne Room*

Session 1

10:30–12:10						
	Word Semantics 1	Machine Translation 1	Information Extraction 1	Summarization	Resource, Annotation	Argument Mining
	Plenary	*203–204*	*210–211*	*212–213*	*219*	*220*

12:10–12:30 **Short Break**

12:30–14:00 **Poster Session 1** *Melbourne Room 1 & 2*

12:30–14:00 **Demo Poster Session 1** *Melbourne Room 1 & 2*

Session 2

14:00–15:40						
	Semantic Parsing 1	Multilinguality	Question Answering 1	Generation 1	Vision	Sentiment
	Plenary	*203–204*	*210–211*	*212–213*	*219*	*220*

15:40–16:10 **Coffee Break** *Level 2 Foyer and Melbourne Room*

Session 3

16:10–17:50						
	Inference, Reasoning	Machine Learning 1	Text Mining and Applications	Dialog System 1	Linguistics, Psycholinguistics and Cognitive Modeling	Parsing 1
	Plenary	*203–204*	*210–211*	*212–213*	*219*	*220*

18:15–19:45 **Student Recruitment Event** *Showtime Events Centre and Common Man Lawn*

Session 1 Overview – Monday, July 16, 2018

Track A *Word Semantics 1*	**Track B** *Machine Translation 1*	**Track C** *Information Extraction 1*	**Track D** *Summarization*	**Track E** *Resource, Annotation*	**Track F** *Argument Mining*	
Plenary	203–204	210–211	212–213	219	220	
Probabilistic FastText for Multi-Sense Word Embeddings *Athiwaratkun, Wilson, and Anandkumar*	Unsupervised Neural Machine Translation with Weight Sharing *Yang, Chen, Wang, and Xu*	Ultra-Fine Entity Typing *Choi, Levy, Choi, and Zettlemoyer*	A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss *Hsu, Lin, Lee, Min, Tang, and Sun*	Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words *Mohammad*	Neural Argument Generation Augmented with Externally Retrieved Evidence *Hua and Wang*	10:30
A La Carte Embedding: Cheap but Effective Induction of Semantic Feature Vectors *Khodak, Saunshi, Liang, Ma, Stewart, and Arora*	Triangular Architecture for Rare Language Translation *Ren, Chen, Liu, Li, Zhou, and Ma*	Hierarchical Losses and New Resources for Fine-grained Entity Typing and Linking *Murty, Verga, Vilnis, Radovanovic, and McCallum*	Extractive Summarization with SWAP-NET: Sentences and Words from Alternating Pointer Networks *Jadhav and Rajan*	Comprehensive Supersense Disambiguation of English Prepositions and Possessives *Schneider, Hwang, Srikumar, Prange, Blodgett, Moeller, Stern, Bitan, and Abend*	A Stylometric Inquiry into Hyperpartisan and Fake News *Potthast, Kiesel, Reinartz, Bevendorff, and Stein*	10:55
Unsupervised Learning of Distributional Relation Vectors *Jameel, Bouraoui, and Schockaert*	Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates *Kudo*	Improving Knowledge Graph Embedding Using Simple Constraints *Ding, Wang, Wang, and Guo*	Retrieve, Rerank and Rewrite: Soft Template Based Neural Summarization *Cao, Li, Li, and Wei*	A Corpus with Multi-Level Annotations of Patients, Interventions and Outcomes to Support Language Processing for Medical Literature *Nye, Li, Patel, Yang, Marshall, Nenkova, and Wallace*	Retrieval of the Best Counterargument without Prior Topic Knowledge *Wachsmuth, Syed, and Stein*	11:20
Explicit Retrofitting of Distributional Word Vectors *Glavaš and Vulić*	The Best of Both Worlds: Combining Recent Advances in Neural Machine Translation *Chen, Firat, Bapna, Johnson, Macherey, Foster, Jones, Schuster, Shazeer, Parmar, Vaswani, Uszkoreit, Kaiser, Chen, Wu, and Hughes*	Towards Understanding the Geometry of Knowledge Graph Embeddings *-, Sharma, and Talukdar*	Simple and Effective Text Simplification Using Semantic and Neural Methods *Sulem, Abend, and Rappoport*	Efficient Online Scalar Annotation with Bounded Support *Sakaguchi and Van Durme*	[TACL] Finding convincing arguments using scalable Bayesian preference learning *Simpson and Gurevych*	11:45

Parallel Session 1

Session 1A: Word Semantics 1

Plenary *Chair: Gerard de Melo*

Probabilistic FastText for Multi-Sense Word Embeddings
Ben Athiwaratkun, Andrew Wilson, and Anima Anandkumar 10:30–10:55

We introduce Probabilistic FastText, a new model for word embeddings that can capture multiple word senses, sub-word structure, and uncertainty information. In particular, we represent each word with a Gaussian mixture density, where the mean of a mixture component is given by the sum of n-grams. This representation allows the model to share the "strength" across sub-word structures (e.g. Latin roots), producing accurate representations of rare, misspelt, or even unseen words. Moreover, each component of the mixture can capture a different word sense. Probabilistic FastText outperforms both FastText, which has no probabilistic model, and dictionary-level probabilistic embeddings, which do not incorporate subword structures, on several word-similarity benchmarks, including English RareWord and foreign language datasets. We also achieve state-of-art performance on benchmarks that measure ability to discern different meanings. Thus, our model is the first to achieve best of both the worlds: multi-sense representations while having enriched semantics on rare words.

A La Carte Embedding: Cheap but Effective Induction of Semantic Feature Vectors
Mikhail Khodak, Nikunj Saunshi, Yingyu Liang, Tengyu Ma, Brandon Stewart, and Sanjeev Arora 10:55–11:20

Motivations like domain adaptation, transfer learning, and feature learning have fueled interest in inducing embeddings for rare or unseen words, n-grams, synsets, and other textual features. This paper introduces a la carte embedding, a simple and general alternative to the usual word2vec-based approaches for building such representations that is based upon recent theoretical results for GloVe-like embeddings. Our method relies mainly on a linear transformation that is efficiently learnable using pretrained word vectors and linear regression. This transform is applicable on the fly in the future when a new text feature or rare word is encountered, even if only a single usage example is available. We introduce a new dataset showing how the a la carte method requires fewer examples of words in context to learn high-quality embeddings and we obtain state-of-the-art results on a nonce task and some unsupervised document classification tasks.

Unsupervised Learning of Distributional Relation Vectors
Shoaib Jameel, Zied Bouraoui, and Steven Schockaert 11:20–11:45

Word embedding models such as GloVe rely on co-occurrence statistics to learn vector representations of word meaning. While we may similarly expect that co-occurrence statistics can be used to capture rich information about the relationships between different words, existing approaches for modeling such relationships are based on manipulating pre-trained word vectors. In this paper, we introduce a novel method which directly learns relation vectors from co-occurrence statistics. To this end, we first introduce a variant of GloVe, in which there is an explicit connection between word vectors and PMI weighted co-occurrence vectors. We then show how relation vectors can be naturally embedded into the resulting vector space.

Explicit Retrofitting of Distributional Word Vectors
Goran Glavaš and Ivan Vulić 11:45–12:10

Semantic specialization of distributional word vectors, referred to as retrofitting, is a process of fine-tuning word vectors using external lexical knowledge in order to better embed some semantic relation. Existing retrofitting models integrate linguistic constraints directly into learning objectives and, consequently, specialize only the vectors of words from the constraints. In this work, in contrast, we transform external lexico-semantic relations into training examples which we use to learn an explicit retrofitting model (ER). The ER model allows us to learn a global specialization function and specialize the vectors of words unobserved in the training data as well. We report large gains over original distributional vector spaces in (1) intrinsic word similarity evaluation and on (2) two downstream tasks - lexical simplification and dialog state tracking. Finally, we also successfully specialize vector spaces of new languages (i.e., unseen in the training data) by coupling ER with shared multilingual distributional vector spaces.

Session 1B: Machine Translation 1

203–204 *Chair: Graham Neubig*

Unsupervised Neural Machine Translation with Weight Sharing
Zhen Yang, Wei Chen, Feng Wang, and Bo Xu 10:30–10:55

Unsupervised neural machine translation (NMT) is a recently proposed approach for machine translation which aims to train the model without using any labeled data. The models proposed for unsupervised NMT often use only one shared encoder to map the pairs of sentences from different languages to a shared-latent space, which is weak in keeping the unique and internal characteristics of each language, such as the style, terminology, and sentence structure. To address this issue, we introduce an extension by utilizing two independent encoders but sharing some partial weights which are responsible for extracting high-level representations of the input sentences. Besides, two different generative adversarial networks (GANs), namely the local GAN and global GAN, are proposed to enhance the cross-language translation. With this new approach, we achieve significant improvements on English-German, English-French and Chinese-to-English translation tasks.

Triangular Architecture for Rare Language Translation
Shuo Ren, Wenhu Chen, Shujie Liu, Mu Li, Ming Zhou, and Shuai Ma 10:55–11:20

Neural Machine Translation (NMT) performs poor on the low-resource language pair (X,Z), especially when Z is a rare language. By introducing another rich language Y, we propose a novel triangular training architecture (TA-NMT) to leverage bilingual data (Y,Z) (may be small) and (X,Y) (can be rich) to improve the translation performance of low-resource pairs. In this triangular architecture, Z is taken as the intermediate latent variable, and translation models of Z are jointly optimized with an unified bidirectional EM algorithm under the goal of maximizing the translation likelihood of (X,Y). Empirical results demonstrate that our method significantly improves the translation quality of rare languages on MultiUN and IWSLT2012 datasets, and achieves even better performance combining back-translation methods.

Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates
Taku Kudo 11:20–11:45

Subword units are an effective way to alleviate the open vocabulary problems in neural machine translation (NMT). While sentences are usually converted into unique subword sequences, subword segmentation is potentially ambiguous and multiple segmentations are possible even with the same vocabulary. The question addressed in this paper is whether it is possible to harness the segmentation ambiguity as a noise to improve the robustness of NMT. We present a simple regularization method, subword regularization, which trains the model with multiple subword segmentations probabilistically sampled during training. In addition, for better subword sampling, we propose a new subword segmentation algorithm based on a unigram language model. We experiment with multiple corpora and report consistent improvements especially on low resource and out-of-domain settings.

The Best of Both Worlds: Combining Recent Advances in Neural Machine Translation
Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Mike Schuster, Noam Shazeer, Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Zhifeng Chen, Yonghui Wu, and Macduff Hughes 11:45–12:10

The past year has witnessed rapid advances in sequence-to-sequence (seq2seq) modeling for Machine Translation (MT). The classic RNN-based approaches to MT were first out-performed by the convolutional seq2seq model, which was then out-performed by the more recent Transformer model. Each of these new approaches consists of a fundamental architecture accompanied by a set of modeling and training techniques that are in principle applicable to other seq2seq architectures. In this paper, we tease apart the new architectures and their accompanying techniques in two ways. First, we identify several key modeling and training techniques, and apply them to the RNN architecture, yielding a new RNMT+ model that outperforms all of the three fundamental architectures on the benchmark WMT'14 English to French and English to German tasks. Second, we analyze the properties of each fundamental seq2seq architecture and devise new hybrid architectures intended to combine their strengths. Our hybrid models obtain further improvements, outperforming the RNMT+ model on both benchmark datasets.

Session 1C: Information Extraction 1

210–211 *Chair: William Wang*

Ultra-Fine Entity Typing

Eunsol Choi, Omer Levy, Yejin Choi, and Luke Zettlemoyer 10:30–10:55

We introduce a new entity typing task: given a sentence with an entity mention, the goal is to predict a set of free-form phrases (e.g. skyscraper, songwriter, or criminal) that describe appropriate types for the target entity. This formulation allows us to use a new type of distant supervision at large scale: head words, which indicate the type of the noun phrases they appear in. We show that these ultra-fine types can be crowd-sourced, and introduce new evaluation sets that are much more diverse and fine-grained than existing benchmarks. We present a model that can predict ultra-fine types, and is trained using a multitask objective that pools our new head-word supervision with prior supervision from entity linking. Experimental results demonstrate that our model is effective in predicting entity types at varying granularity; it achieves state of the art performance on an existing fine-grained entity typing benchmark, and sets baselines for our newly-introduced datasets.

Hierarchical Losses and New Resources for Fine-grained Entity Typing and Linking

Shikhar Murty, Patrick Verga, Luke Vilnis, Irena Radovanovic, and Andrew McCallum 10:55–11:20

Extraction from raw text to a knowledge base of entities and fine-grained types is often cast as prediction into a flat set of entity and type labels, neglecting the rich hierarchies over types and entities contained in curated ontologies. Previous attempts to incorporate hierarchical structure have yielded little benefit and are restricted to shallow ontologies. This paper presents new methods using real and complex bilinear mappings for integrating hierarchical information, yielding substantial improvement over flat predictions in entity linking and fine-grained entity typing, and achieving new state-of-the-art results for end-to-end models on the benchmark FIGER dataset. We also present two new human-annotated datasets containing wide and deep hierarchies which we will release to the community to encourage further research in this direction: *MedMentions*, a collection of PubMed abstracts in which 246k mentions have been mapped to the massive UMLS ontology; and *TypeNet*, which aligns Freebase types with the WordNet hierarchy to obtain nearly 2k entity types. In experiments on all three datasets we show substantial gains from hierarchy-aware training.

Improving Knowledge Graph Embedding Using Simple Constraints

Boyang Ding, Quan Wang, Bin Wang, and Li Guo 11:20–11:45

Embedding knowledge graphs (KGs) into continuous vector spaces is a focus of current research. Early works performed this task via simple models developed over KG triples. Recent attempts focused on either designing more complicated triple scoring models, or incorporating extra information beyond triples. This paper, by contrast, investigates the potential of using very simple constraints to improve KG embedding. We examine non-negativity constraints on entity representations and approximate entailment constraints on relation representations. The former help to learn compact and interpretable representations for entities. The latter further encode regularities of logical entailment between relations into their distributed representations. These constraints impose prior beliefs upon the structure of the embedding space, without negative impacts on efficiency or scalability. Evaluation on WordNet, Freebase, and DBpedia shows that our approach is simple yet surprisingly effective, significantly and consistently outperforming competitive baselines. The constraints imposed indeed improve model interpretability, leading to a substantially increased structuring of the embedding space. Code and data are available at `https://github.com/iieir-km/ComplEx-NNE_AER`.

Towards Understanding the Geometry of Knowledge Graph Embeddings

Chandrahas -, Aditya Sharma, and Partha Talukdar 11:45–12:10

Knowledge Graph (KG) embedding has emerged as a very active area of research over the last few years, resulting in the development of several embedding methods. These KG embedding methods represent KG entities and relations as vectors in a high-dimensional space. Despite this popularity and effectiveness of KG embeddings in various tasks (e.g., link prediction), geometric understanding of such embeddings (i.e., arrangement of entity and relation vectors in vector space) is unexplored — we fill this gap in the paper. We initiate a study to analyze the geometry of KG embeddings and correlate it with task performance and other hyperparameters. To the best of our knowledge, this is the first study of its kind. Through extensive experiments on real-world datasets, we discover several insights. For example, we find that there are sharp differences between the geometry of embeddings learnt by different classes of KG embeddings methods. We hope that this initial study will inspire other follow-up research on this important but unexplored problem.

Session 1D: Summarization

212–213 *Chair: Dragomir Radev*

A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss
Wan-Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, Jing Tang, and Min Sun 10:30–10:55

We propose a unified model combining the strength of extractive and abstractive summarization. On the one hand, a simple extractive model can obtain sentence-level attention with high ROUGE scores but less readable. On the other hand, a more complicated abstractive model can obtain word-level dynamic attention to generate a more readable paragraph. In our model, sentence-level attention is used to modulate the word-level attention such that words in less attended sentences are less likely to be generated. Moreover, a novel inconsistency loss function is introduced to penalize the inconsistency between two levels of attentions. By end-to-end training our model with the inconsistency loss and original losses of extractive and abstractive models, we achieve state-of-the-art ROUGE scores while being the most informative and readable summarization on the CNN/Daily Mail dataset in a solid human evaluation.

Extractive Summarization with SWAP-NET: Sentences and Words from Alternating Pointer Networks
Aishwarya Jadhav and Vaibhav Rajan 10:55–11:20

We present a new neural sequence-to-sequence model for extractive summarization called SWAP-NET (Sentences and Words from Alternating Pointer Networks). Extractive summaries comprising a salient subset of input sentences, often also contain important key words. Guided by this principle, we design SWAP-NET that models the interaction of key words and salient sentences using a new two-level pointer network based architecture. SWAP-NET identifies both salient sentences and key words in an input document, and then combines them to form the extractive summary. Experiments on large scale benchmark corpora demonstrate the efficacy of SWAP-NET that outperforms state-of-the-art extractive summarizers.

Retrieve, Rerank and Rewrite: Soft Template Based Neural Summarization
Ziqiang Cao, Wenjie Li, Sujian Li, and Furu Wei 11:20–11:45

Most previous seq2seq summarization systems purely depend on the source text to generate summaries, which tends to work unstably. Inspired by the traditional template-based summarization approaches, this paper proposes to use existing summaries as soft templates to guide the seq2seq model. To this end, we use a popular IR platform to Retrieve proper summaries as candidate templates. Then, we extend the seq2seq framework to jointly conduct template Reranking and template-aware summary generation (Rewriting). Experiments show that, in terms of informativeness, our model significantly outperforms the state-of-the-art methods, and even soft templates themselves demonstrate high competitiveness. In addition, the import of high-quality external summaries improves the stability and readability of generated summaries.

Simple and Effective Text Simplification Using Semantic and Neural Methods
Elior Sulem, Omri Abend, and Ari Rappoport 11:45–12:10

Sentence splitting is a major simplification operator. Here we present a simple and efficient splitting algorithm based on an automatic semantic parser. After splitting, the text is amenable for further fine-tuned simplification operations. In particular, we show that neural Machine Translation can be effectively used in this situation. Previous application of Machine Translation for simplification suffers from a considerable disadvantage in that they are over-conservative, often failing to modify the source in any way. Splitting based on semantic parsing, as proposed here, alleviates this issue. Extensive automatic and human evaluation shows that the proposed method compares favorably to the state-of-the-art in combined lexical and structural simplification.

Session 1E: Resource, Annotation

219 *Chair: Vered Shwartz*

Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words
Saif Mohammad 10:30–10:55

Words play a central role in language and thought. Factor analysis studies have shown that the primary dimensions of meaning are valence, arousal, and dominance (VAD). We present the NRC VAD Lexicon, which has human ratings of valence, arousal, and dominance for more than 20,000 English words. We use Best–Worst Scaling to obtain fine-grained scores and address issues of annotation consistency that plague traditional rating scale methods of annotation. We show that the ratings obtained are vastly more reliable than those in existing lexicons. We also show that there exist statistically significant differences in the shared understanding of valence, arousal, and dominance across demographic variables such as age, gender, and personality.

Comprehensive Supersense Disambiguation of English Prepositions and Possessives
Nathan Schneider, Jena D. Hwang, Vivek Srikumar, Jakob Prange, Austin Blodgett, Sarah R. Moeller, Aviram Stern, Adi Bitan, and Omri Abend 10:55–11:20

Semantic relations are often signaled with prepositional or possessive marking—but extreme polysemy bedevils their analysis and automatic interpretation. We introduce a new annotation scheme, corpus, and task for the disambiguation of prepositions and possessives in English. Unlike previous approaches, our annotations are comprehensive with respect to types and tokens of these markers; use broadly applicable supersense classes rather than fine-grained dictionary definitions; unite prepositions and possessives under the same class inventory; and distinguish between a marker's lexical contribution and the role it marks in the context of a predicate or scene. Strong interannotator agreement rates, as well as encouraging disambiguation results with established supervised methods, speak to the viability of the scheme and task.

A Corpus with Multi-Level Annotations of Patients, Interventions and Outcomes to Support Language Processing for Medical Literature
Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron Wallace 11:20–11:45

We present a corpus of 5,000 richly annotated abstracts of medical articles describing clinical randomized controlled trials. Annotations include demarcations of text spans that describe the Patient population enrolled, the Interventions studied and to what they were Compared, and the Outcomes measured (the 'PICO' elements). These spans are further annotated at a more granular level, e.g., individual interventions within them are marked and mapped onto a structured medical vocabulary. We acquired annotations from a diverse set of workers with varying levels of expertise and cost. We describe our data collection process and the corpus itself in detail. We then outline a set of challenging NLP tasks that would aid searching of the medical literature and the practice of evidence-based medicine.

Efficient Online Scalar Annotation with Bounded Support
Keisuke Sakaguchi and Benjamin Van Durme 11:45–12:10

We describe a novel method for efficiently eliciting scalar annotations for dataset construction and system quality estimation by human judgments. We contrast direct assessment (annotators assign scores to items directly), online pairwise ranking aggregation (scores derive from annotator comparison of items), and a hybrid approach (EASL: Efficient Annotation of Scalar Labels) proposed here. Our proposal leads to increased correlation with ground truth, at far greater annotator efficiency, suggesting this strategy as an improved mechanism for dataset creation and manual system evaluation.

Session 1F: Argument Mining

220 *Chair: Vincent Ng*

Neural Argument Generation Augmented with Externally Retrieved Evidence
Xinyu Hua and Lu Wang 10:30–10:55

High quality arguments are essential elements for human reasoning and decision-making processes. However, effective argument construction is a challenging task for both human and machines. In this work, we study a novel task on automatically generating arguments of a different stance for a given statement. We propose an encoder-decoder style neural network-based argument generation model enriched with externally retrieved evidence from Wikipedia. Our model first generates a set of talking point phrases as intermediate representation, followed by a separate decoder producing the final argument based on both input and the keyphrases. Experiments on a large-scale dataset collected from Reddit show that our model constructs arguments with more topic-relevant content than popular sequence-to-sequence generation models according to automatic evaluation and human assessments.

A Stylometric Inquiry into Hyperpartisan and Fake News
Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein 10:55–11:20

We report on a comparative style analysis of hyperpartisan (extremely one-sided) news and fake news. A corpus of 1,627 articles from 9 political publishers, three each from the mainstream, the hyperpartisan left, and the hyperpartisan right, have been fact-checked by professional journalists at BuzzFeed: 97% of the 299 fake news articles identified are also hyperpartisan. We show how a style analysis can distinguish hyperpartisan news from the mainstream (F1 = 0.78), and satire from both (F1 = 0.81). But stylometry is no silver bullet as style-based fake news detection does not work (F1 = 0.46). We further reveal that left-wing and right-wing news share significantly more stylistic similarities than either does with the mainstream. This result is robust: it has been confirmed by three different modeling approaches, one of which employs Unmasking in a novel way. Applications of our results include partisanship detection and pre-screening for semi-automatic fake news detection.

Retrieval of the Best Counterargument without Prior Topic Knowledge
Henning Wachsmuth, Shahbaz Syed, and Benno Stein 11:20–11:45

Given any argument on any controversial topic, how to counter it? This question implies the challenging retrieval task of finding the best counterargument. Since prior knowledge of a topic cannot be expected in general, we hypothesize the best counterargument to invoke the same aspects as the argument while having the opposite stance. To operationalize our hypothesis, we simultaneously model the similarity and dissimilarity of pairs of arguments, based on the words and embeddings of the arguments' premises and conclusions. A salient property of our model is its independence from the topic at hand, i.e., it applies to arbitrary arguments. We evaluate different model variations on millions of argument pairs derived from the web portal idebate.org. Systematic ranking experiments suggest that our hypothesis is true for many arguments: For 7.6 candidates with opposing stance on average, we rank the best counterargument highest with 60% accuracy. Even among all 2801 test set pairs as candidates, we still find the best one about every third time.

[TACL] Finding convincing arguments using scalable Bayesian preference learning
Edwin Simpson and Iryna Gurevych 11:45–12:10

We introduce a scalable Bayesian preference learning method for identifying convincing ar- guments in the absence of gold-standard rat- ings or rankings. In contrast to previous work, we avoid the need for separate methods to perform quality control on training data, pre- dict rankings and perform pairwise classifica- tion. Bayesian approaches are an effective so- lution when faced with sparse or noisy train- ing data, but have not previously been used to identify convincing arguments. One issue is scalability, which we address by develop- ing a stochastic variational inference method for Gaussian process (GP) preference learn- ing. We show how our method can be ap- plied to predict argument convincingness from crowdsourced data, outperforming the previ- ous state-of-the-art, particularly when trained with small amounts of unreliable data. We demonstrate how the Bayesian approach en- ables more effective active learning, thereby reducing the amount of data required to iden- tify convincing arguments for new users and domains. While word embeddings are princi- pally used with neural networks, our results show that word embeddings in combination with linguistic features also benefit GPs when predicting argument convincingness.

Poster Session 1

Time: *12:30–14:00* Location: Melbourne Room 1 & 2

Poster Session 1A: Machine Learning

LinkNBed: Multi-Graph Representation Learning with Entity Linkage
Rakshit Trivedi, Bunyamin Sisman, Xin Luna Dong, Christos Faloutsos, Jun Ma, and Hongyuan Zha

Knowledge graphs have emerged as an important model for studying complex multi-relational data. This has given rise to the construction of numerous large scale but incomplete knowledge graphs encoding information extracted from various resources. An effective and scalable approach to jointly learn over multiple graphs and eventually construct a unified graph is a crucial next step for the success of knowledge-based inference for many downstream applications. To this end, we propose LinkNBed, a deep relational learning framework that learns entity and relationship representations across multiple graphs. We identify entity linkage across graphs as a vital component to achieve our goal. We design a novel objective that leverage entity linkage and build an efficient multi-task training procedure. Experiments on link prediction and entity linkage demonstrate substantial improvements over the state-of-the-art relational learning approaches.

Probabilistic Embedding of Knowledge Graphs with Box Lattice Measures
Luke Vilnis, Xiang Li, Shikhar Murty, and Andrew McCallum

Embedding methods which enforce a partial order or lattice structure over the concept space, such as Order Embeddings (OE), are a natural way to model transitive relational data (e.g. entailment graphs). However, OE learns a deterministic knowledge base, limiting expressiveness of queries and the ability to use uncertainty for both prediction and learning (e.g. learning from expectations). Probabilistic extensions of OE have provided the ability to somewhat calibrate these denotational probabilities while retaining the consistency and inductive bias of ordered models, but lack the ability to model the negative correlations found in real-world knowledge. In this work we show that a broad class of models that assign probability measures to OE can never capture negative correlation, which motivates our construction of a novel box lattice and accompanying probability measure to capture anti-correlation and even disjoint concepts, while still providing the benefits of probabilistic modeling, such as the ability to perform rich joint and conditional queries over arbitrary sets of concepts, and both learning from and predicting calibrated uncertainty. We show improvements over previous approaches in modeling the Flickr and WordNet entailment graphs, and investigate the power of the model.

Graph-to-Sequence Learning using Gated Graph Neural Networks
Daniel Beck, Gholamreza Haffari, and Trevor Cohn

Many NLP applications can be framed as a graph-to-sequence learning problem. Previous work proposing neural architectures on graph-to-sequence obtained promising results compared to grammar-based approaches but still rely on linearisation heuristics and/or standard recurrent networks to achieve the best performance. In this work propose a new model that encodes the full structural information contained in the graph. Our architecture couples the recently proposed Gated Graph Neural Networks with an input transformation that allows nodes and edges to have their own hidden representations, while tackling the parameter explosion problem present in previous work. Experimental results shows that our model outperforms strong baselines in generation from AMR graphs and syntax-based neural machine translation.

Sharp Nearby, Fuzzy Far Away: How Neural Language Models Use Context
Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky

We know very little about how neural language models (LM) use prior linguistic context. In this paper, we investigate the role of context in an LSTM LM, through ablation studies. Specifically, we analyze the increase in perplexity when prior context words are shuffled, replaced, or dropped. On two standard datasets, Penn Treebank and WikiText-2, we find that the model is capable of using about 200 tokens of context on average, but sharply distinguishes nearby context (recent 50 tokens) from the distant history. The model is highly sensitive to the order of words within the most recent sentence, but ignores word order in the long-range context (beyond 50 tokens), suggesting the distant past is modeled only as a rough semantic field or topic. We further find that the neural caching model (Grave et al., 2017b) especially helps the LSTM to copy words from within this distant context. Overall, our analysis not only provides a better understanding of how neural LMs use their context, but also sheds light on recent success from cache-based models.

Bridging CNNs, RNNs, and Weighted Finite-State Machines
Roy Schwartz, Sam Thomson, and Noah A. Smith

Recurrent and convolutional neural networks comprise two distinct families of models that have proven to be useful for encoding natural language utterances. In this paper we present SoPa, a new model that aims to bridge these two approaches. SoPa combines neural representation learning with weighted finite-state automata (WFSAs) to learn a soft version of traditional surface patterns. We show that SoPa is an extension of a one-layer CNN, and that such CNNs are equivalent to a restricted version of SoPa, and accordingly, to a restricted form of WFSA. Empirically, on three text classification tasks, SoPa is comparable or better than both a BiLSTM (RNN) baseline and a CNN baseline, and is particularly useful in small data settings.

Zero-shot Learning of Classifiers from Natural Language Quantification
Shashank Srivastava, Igor Labutov, and Tom Mitchell

Humans can efficiently learn new concepts using language. We present a framework through which a set of explanations of a concept can be used to learn a classifier without access to any labeled examples. We use semantic parsing to map explanations to probabilistic assertions grounded in latent class labels and observed attributes of unlabeled data, and leverage the differential semantics of linguistic quantifiers (e.g., 'usually' vs 'always') to drive model training. Experiments on three domains show that the learned classifiers outperform previous approaches for learning with limited data, and are comparable with fully supervised classifiers trained from a small number of labeled examples.

Sentence-State LSTM for Text Representation
Yue Zhang, Qi Liu, and Linfeng Song

Bi-directional LSTMs are a powerful tool for text representation. On the other hand, they have been shown to suffer various limitations due to their sequential nature. We investigate an alternative LSTM structure for encoding text, which consists of a parallel state for each word. Recurrent steps are used to perform local and global information exchange between words simultaneously, rather than incremental reading of a sequence of words. Results on various classification and sequence labelling benchmarks show that the proposed model has strong representation power, giving highly competitive performances compared to stacked BiLSTM models with similar parameter numbers.

Universal Language Model Fine-tuning for Text Classification
Jeremy Howard and Sebastian Ruder

Inductive transfer learning has greatly impacted computer vision, but existing approaches in NLP still require task-specific modifications and training from scratch. We propose Universal Language Model Fine-tuning (ULMFiT), an effective transfer learning method that can be applied to any task in NLP, and introduce techniques that are key for fine-tuning a language model. Our method significantly outperforms the state-of-the-art on six text classification tasks, reducing the error by 18-24% on the majority of datasets. Furthermore, with only 100 labeled examples, it matches the performance of training from scratch on 100 times more data. We open-source our pretrained models and code.

Evaluating neural network explanation methods using hybrid documents and morphosyntactic agreement
Nina Poerner, Hinrich Schütze, and Benjamin Roth

The behavior of deep neural networks (DNNs) is hard to understand. This makes it necessary to explore post hoc explanation methods. We conduct the first comprehensive evaluation of explanation methods for NLP. To this end, we design two novel evaluation paradigms that cover two important classes of NLP problems: small context and large context problems. Both paradigms require no manual annotation and are therefore broadly applicable. We also introduce LIMSSE, an explanation method inspired by LIME that is designed for NLP. We show empirically that LIMSSE, LRP and DeepLIFT are the most effective explanation methods and recommend them for explaining DNNs in NLP.

Continuous Learning in a Hierarchical Multiscale Neural Network
Thomas Wolf, Julien Chaumond, and Clement Delangue

We reformulate the problem of encoding a multi-scale representation of a sequence in a language model by casting it in a continuous learning framework. We propose a hierarchical multi-scale language model in which short time-scale dependencies are encoded in the hidden state of a lower-level recurrent neural network while longer time-scale dependencies are encoded in the dynamic of the lower-level network by having a meta-learner update the weights of the lower-level neural network in an online meta-learning fashion. We use elastic weights consolidation as a higher-level to prevent catastrophic forgetting in our continuous learning framework.

Restricted Recurrent Neural Tensor Networks: Exploiting Word Frequency and Compositionality
Alexandre Salle and Aline Villavicencio

Increasing the capacity of recurrent neural networks (RNN) usually involves augmenting the size of the hidden layer, with significant increase of computational cost. Recurrent neural tensor networks (RNTN) increase capacity using distinct hidden layer weights for each word, but with greater costs in memory usage. In this paper, we introduce restricted recurrent neural tensor networks (r-RNTN) which reserve distinct hidden layer weights for frequent vocabulary words while sharing a single set of weights for infrequent words. Perplexity evaluations show that for fixed hidden layer sizes, r-RNTNs improve language model performance over RNNs using only a small fraction of the parameters of unrestricted RNTNs. These results hold for r-RNTNs using Gated Recurrent Units and Long Short-Term Memory.

Deep RNNs Encode Soft Hierarchical Syntax
Terra Blevins, Omer Levy, and Luke Zettlemoyer

We present a set of experiments to demonstrate that deep recurrent neural networks (RNNs) learn internal representations that capture soft hierarchical notions of syntax from highly varied supervision. We consider four syntax tasks at different depths of the parse tree; for each word, we predict its part of speech as well as the first (parent), second (grandparent) and third level (great-grandparent) constituent labels that appear above it. These predictions are made from representations produced at different depths in networks that are pretrained with one of four objectives: dependency parsing, semantic role labeling, machine translation, or language modeling. In every case, we find a correspondence between network depth and syntactic depth, suggesting that a soft syntactic hierarchy emerges. This effect is robust across all conditions, indicating that the models encode significant amounts of syntax even in the absence of an explicit syntactic training supervision.

Word Error Rate Estimation for Speech Recognition: e-WER
Ahmed Ali and Steve Renals

Measuring the performance of automatic speech recognition (ASR) systems requires manually transcribed data in order to compute the word error rate (WER), which is often time-consuming and expensive. In this paper, we propose a novel approach to estimate WER, or e-WER, which does not require a gold-standard transcription of the test set. Our e-WER framework uses a comprehensive set of features: ASR recognised text, character recognition results to complement recognition output, and internal decoder features. We report results for the two features; black-box and glass-box using unseen 24 Arabic broadcast programs. Our system achieves 16.9% WER root mean squared error (RMSE) across 1,400 sentences. The estimated overall WER e-WER was 25.3% for the three hours test set, while the actual WER was 28.5%.

Towards Robust and Privacy-preserving Text Representations
Yitong Li, Timothy Baldwin, and Trevor Cohn

Written text often provides sufficient clues to identify the author, their gender, age, and other important attributes. Consequently, the authorship of training and evaluation corpora can have unforeseen impacts, including differing model performance for different user groups, as well as privacy implications. In this paper, we propose an approach to explicitly obscure important author characteristics at training time, such that representations learned are invariant to these attributes. Evaluating on two tasks, we show that this leads to increased privacy in the learned representations, as well as more robust models to varying evaluation conditions, including out-of-domain corpora.

HotFlip: White-Box Adversarial Examples for Text Classification
Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou

We propose an efficient method to generate white-box adversarial examples to trick a character-level neural classifier. We find that only a few manipulations are needed to greatly decrease the accuracy. Our method relies on an atomic flip operation, which swaps one token for another, based on the gradients of the one-hot input vectors. Due to efficiency of our method, we can perform adversarial training which makes the model more robust to attacks at test time. With the use of a few semantics-preserving constraints, we demonstrate that HotFlip can be adapted to attack a word-level classifier as well.

Domain Adapted Word Embeddings for Improved Sentiment Classification
Prathusha K Sarma, Yingyu Liang, and Bill Sethares

Generic word embeddings are trained on large-scale generic corpora; Domain Specific (DS) word embeddings are trained only on data from a domain of interest. This paper proposes a method to combine the breadth of generic embeddings with the specificity of domain specific embeddings. The resulting embeddings, called Domain Adapted (DA) word embeddings, are formed by aligning corresponding word vectors using Canonical Correlation Analysis (CCA) or the related nonlinear Kernel CCA. Evaluation results on sentiment classification tasks show that the DA embeddings substantially outperform both generic, DS embeddings when used as input features to standard or state-of-the-art sentence encoding algorithms for classification.

Poster Session 1B: Semantics

Improving Text-to-SQL Evaluation Methodology
Catherine Finegan-Dollak, Jonathan K. Kummerfeld, Li Zhang, Karthik Ramanathan, Sesh Sadasivam, Rui Zhang, and Dragomir Radev

To be informative, an evaluation must measure how well systems generalize to realistic unseen data. We identify limitations of and propose improvements to current evaluations of text-to-SQL systems. First, we compare human-generated and automatically generated questions, characterizing properties of queries necessary for real-world applications. To facilitate evaluation on multiple datasets, we release standardized and improved versions of seven existing datasets and one new text-to-SQL dataset. Second, we show that the current division of data into training and test sets measures robustness to variations in the way questions are asked, but only partially tests how well systems generalize to new queries; therefore, we propose a complementary dataset split for evaluation of future work. Finally, we demonstrate how the common practice of anonymizing variables during evaluation removes an important challenge of the task. Our observations highlight key difficulties, and our methodology enables effective measurement of future development.

Semantic Parsing with Syntax- and Table-Aware SQL Generation
Yibo Sun, Duyu Tang, Nan Duan, Jianshu Ji, Guihong Cao, Xiaocheng Feng, Bing Qin, Ting Liu, and Ming Zhou

We present a generative model to map natural language questions into SQL queries. Existing neural network based approaches typically generate a SQL query word-by-word, however, a large portion of the generated results is incorrect or not executable due to the mismatch between question words and table contents. Our approach addresses this problem by considering the structure of table and the syntax of SQL language. The quality of the generated SQL query is significantly improved through (1) learning to replicate content from column names, cells or SQL keywords; and (2) improving the generation of WHERE clause by leveraging the column-cell relation. Experiments are conducted on WikiSQL, a recently released dataset with the largest question- SQL pairs. Our approach significantly improves the state-of-the-art execution accuracy from 69.0% to 74.4%.

Multitask Parsing Across Semantic Representations
Daniel Hershcovich, Omri Abend, and Ari Rappoport

The ability to consolidate information of different types is at the core of intelligence, and has tremendous practical value in allowing learning for one task to benefit from generalizations learned for others. In this paper we tackle the challenging task of improving semantic parsing performance, taking UCCA parsing as a test case, and AMR, SDP and Universal Dependencies (UD) parsing as auxiliary tasks. We experiment on three languages, using a uniform transition-based system and learning architecture for all parsing tasks. Despite notable conceptual, formal and domain differences, we show that multitask learning significantly improves UCCA parsing in both in-domain and out-of-domain settings.

Character-Level Models versus Morphology in Semantic Role Labeling
Gozde Gul Sahin and Mark Steedman

Character-level models have become a popular approach specially for their accessibility and ability to handle unseen data. However, little is known on their ability to reveal the underlying morphological structure of a word, which is a crucial skill for high-level semantic analysis tasks, such as semantic role labeling (SRL). In this work, we train various types of SRL models that use word, character and morphology level information and analyze how performance of characters compare to words and morphology for several languages. We conduct an in-depth error analysis for each morphological typology and analyze the strengths and limitations of character-level models that relate to out-of-domain data, training data size, long range dependencies and model complexity. Our exhaustive analyses shed light on important characteristics of character-level models and their semantic capability.

AMR Parsing as Graph Prediction with Latent Alignment
Chunchuan Lyu and Ivan Titov

Abstract meaning representations (AMRs) are broad-coverage sentence-level semantic representations. AMRs represent sentences as rooted labeled directed acyclic graphs. AMR parsing is challenging partly due to the lack of annotated alignments between nodes in the graphs and words in the corresponding sentences. We introduce a neural parser which treats alignments as latent variables within a joint probabilistic model of concepts, relations and alignments. As exact inference requires marginalizing over alignments and is infeasible, we use the variational autoencoding framework and a continuous relaxation of the discrete alignments. We show that joint modeling is preferable to using a pipeline of align and parse. The parser achieves the best reported results on the standard benchmark (74.4%

on LDC2016E25).

Accurate SHRG-Based Semantic Parsing
Yufei Chen, Weiwei Sun, and Xiaojun Wan

We demonstrate that an SHRG-based parser can produce semantic graphs much more accurately than previously shown, by relating synchronous production rules to the syntacto-semantic composition process. Our parser achieves an accuracy of 90.35 for EDS (89.51 for DMRS) in terms of elementary dependency match, which is a 4.87 (5.45) point improvement over the best existing data-driven model, indicating, in our view, the importance of linguistically-informed derivation for data-driven semantic parsing. This accuracy is equivalent to that of English Resource Grammar guided models, suggesting that (recurrent) neural network models are able to effectively learn deep linguistic knowledge from annotations.

Using Intermediate Representations to Solve Math Word Problems
Danqing Huang, Jin-Ge Yao, Chin-Yew Lin, Qingyu Zhou, and Jian Yin

To solve math word problems, previous statistical approaches attempt at learning a direct mapping from a problem description to its corresponding equation system. However, such mappings do not include the information of a few higher-order operations that cannot be explicitly represented in equations but are required to solve the problem. The gap between natural language and equations makes it difficult for a learned model to generalize from limited data. In this work we present an intermediate meaning representation scheme that tries to reduce this gap. We use a sequence-to-sequence model with a novel attention regularization term to generate the intermediate forms, then execute them to obtain the final answers. Since the intermediate forms are latent, we propose an iterative labeling framework for learning by leveraging supervision signals from both equations and answers. Our experiments show using intermediate forms outperforms directly predicting equations.

Discourse Representation Structure Parsing
Jiangming Liu, Shay B. Cohen, and Mirella Lapata

We introduce an open-domain neural semantic parser which generates formal meaning representations in the style of Discourse Representation Theory (DRT; Kamp and Reyle 1993). We propose a method which transforms Discourse Representation Structures (DRSs) to trees and develop a structure-aware model which decomposes the decoding process into three stages: basic DRS structure prediction, condition prediction (i.e., predicates and relations), and referent prediction (i.e., variables). Experimental results on the Groningen Meaning Bank (GMB) show that our model outperforms competitive baselines by a wide margin.

Baseline Needs More Love: On Simple Word-Embedding-Based Models and Associated Pooling Mechanisms
Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, and Lawrence Carin

Many deep learning architectures have been proposed to model the *compositionality* in text sequences, requiring substantial number of parameters and expensive computations. However, there has not been a rigorous evaluation regarding the added value of sophisticated compositional functions. In this paper, we conduct a point-by-point comparative study between Simple Word-Embedding-based Models (SWEMs), consisting of parameter-free pooling operations, relative to word-embedding-based RNN/CNN models. Surprisingly, SWEMs exhibit comparable or even superior performance in the majority of cases considered. Based upon this understanding, we propose two additional pooling strategies over learned word embeddings: (i) a max-pooling operation for improved interpretability; and (ii) a hierarchical pooling operation, which preserves spatial (n-gram) information within text sequences. We present experiments on 17 datasets encompassing three tasks: (i) (long) document classification; (ii) text sequence matching; and (iii) short text tasks, including classification and tagging.

ParaNMT-50M: Pushing the Limits of Paraphrastic Sentence Embeddings with Millions of Machine Translations
John Wieting and Kevin Gimpel

We describe ParaNMT-50M, a dataset of more than 50 million English-English sentential paraphrase pairs. We generated the pairs automatically by using neural machine translation to translate the non-English side of a large parallel corpus, following Wieting et al. (2017). Our hope is that ParaNMT-50M can be a valuable resource for paraphrase generation and can provide a rich source of semantic knowledge to improve downstream natural language understanding tasks. To show its utility, we use ParaNMT-50M to train paraphrastic sentence embeddings that outperform all supervised systems on every SemEval semantic textual similarity competition, in addition to showing how it can be used for paraphrase generation.

Event2Mind: Commonsense Inference on Events, Intents, and Reactions
Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi

We investigate a new commonsense inference task: given an event described in a short free-form text ("X drinks coffee in the morning"), a system reasons about the likely intents ("X wants to stay awake") and reactions ("X feels alert") of the event's participants. To support this study, we construct a new crowdsourced corpus of 25,000 event phrases covering a diverse range of everyday events and situations. We report baseline performance on this task, demonstrating that neural encoder-decoder models can successfully compose embedding representations of previously unseen events and reason about the likely intents and reactions of the event participants. In addition, we demonstrate how commonsense inference on people's intents and reactions can help unveil the implicit gender inequality prevalent in modern movie scripts.

Neural Adversarial Training for Semi-supervised Japanese Predicate-argument Structure Analysis
Shuhei Kurita, Daisuke Kawahara, and Sadao Kurohashi

Japanese predicate-argument structure (PAS) analysis involves zero anaphora resolution, which is notoriously difficult. To improve the performance of Japanese PAS analysis, it is straightforward to increase the size of corpora annotated with PAS. However, since it is prohibitively expensive, it is promising to take advantage of a large amount of raw corpora. In this paper, we propose a novel Japanese PAS analysis model based on semi-supervised adversarial training with a raw corpus. In our experiments, our model outperforms existing state-of-the-art models for Japanese PAS analysis.

Active learning for deep semantic parsing
Long Duong, Hadi Afshar, Dominique Estival, Glen Pink, Philip Cohen, and Mark Johnson

Semantic parsing requires training data that is expensive and slow to collect. We apply active learning to both traditional and "overnight" data collection approaches. We show that it is possible to obtain good training hyperparameters from seed data which is only a small fraction of the full dataset. We show that uncertainty sampling based on least confidence score is competitive in traditional data collection but not applicable for overnight collection. We propose several active learning strategies for overnight data collection and show that different example selection strategies per domain perform best.

Learning Thematic Similarity Metric from Article Sections Using Triplet Networks
Liat Ein Dor, Yosi Mass, Alon Halfon, Elad Venezian, Ilya Shnayderman, Ranit Aharonov, and Noam Slonim

In this paper we suggest to leverage the partition of articles into sections, in order to learn thematic similarity metric between sentences. We assume that a sentence is thematically closer to sentences within its section than to sentences from other sections. Based on this assumption, we use Wikipedia articles to automatically create a large dataset of weakly labeled sentence triplets, composed of a pivot sentence, one sentence from the same section and one from another section. We train a triplet network to embed sentences from the same section closer. To test the performance of the learned embeddings, we create and release a sentence clustering benchmark. We show that the triplet network learns useful thematic metrics, that significantly outperform state-of-the-art semantic similarity methods and multipurpose embeddings on the task of thematic clustering of sentences. We also show that the learned embeddings perform well on the task of sentence semantic similarity prediction.

Unsupervised Semantic Frame Induction using Triclustering
Dmitry Ustalov, Alexander Panchenko, Andrey Kutuzov, Chris Biemann, and Simone Paolo Ponzetto

We use dependency triples automatically extracted from a Web-scale corpus to perform unsupervised semantic frame induction. We cast the frame induction problem as a triclustering problem that is a generalization of clustering for triadic data. Our replicable benchmarks demonstrate that the proposed graph-based approach, Triframes, shows state-of-the art results on this task on a FrameNet-derived dataset and performing on par with competitive methods on a verb class clustering task.

Poster Session 1C: Information Extraction, Text Mining

Improving Event Coreference Resolution by Modeling Correlations between Event Coreference Chains and Document Topic Structures
Prafulla Kumar Choubey and Ruihong Huang

This paper proposes a novel approach for event coreference resolution that models correlations between event coreference chains and document topical structures through an Integer Linear Programming formulation. We explicitly model correlations between the main event chains of a document with topic transition sentences, inter-coreference chain correlations, event mention distributional characteristics and sub-event structure, and use them with scores obtained from a local coreference relation classifier for jointly resolving multiple event chains in a document. Our experiments across KBP 2016 and 2017 datasets suggest that each of the structures contribute to improving event coreference resolution performance.

DSGAN: Generative Adversarial Training for Distant Supervision Relation Extraction
Pengda Qin, Weiran XU, and William Yang Wang

Distant supervision can effectively label data for relation extraction, but suffers from the noise labeling problem. Recent works mainly perform soft bag-level noise reduction strategies to find the relatively better samples in a sentence bag, which is suboptimal compared with making a hard decision of false positive samples in sentence level. In this paper, we introduce an adversarial learning framework, which we named DSGAN, to learn a sentence-level true-positive generator. Inspired by Generative Adversarial Networks, we regard the positive samples generated by the generator as the negative samples to train the discriminator. The optimal generator is obtained until the discrimination ability of the discriminator has the greatest decline. We adopt the generator to filter distant supervision training dataset and redistribute the false positive instances into the negative set, in which way to provide a cleaned dataset for relation classification. The experimental results show that the proposed strategy significantly improves the performance of distant supervision relation extraction comparing to state-of-the-art systems.

Extracting Relational Facts by an End-to-End Neural Model with Copy Mechanism
Xiangrong Zeng, Daojian Zeng, Shizhu He, Kang Liu, and Jun Zhao

The relational facts in sentences are often complicated. Different relational triplets may have overlaps in a sentence. We divided the sentences into three types according to triplet overlap degree, including Normal, EntityPairOverlap and SingleEntiyOverlap. Existing methods mainly focus on Normal class and fail to extract relational triplets precisely. In this paper, we propose an end-to-end model based on sequence-to-sequence learning with copy mechanism, which can jointly extract relational facts from sentences of any of these classes. We adopt two different strategies in decoding process: employing only one united decoder or applying multiple separated decoders. We test our models in two public datasets and our model outperform the baseline method significantly.

Self-regulation: Employing a Generative Adversarial Network to Improve Event Detection
Yu Hong, Wenxuan Zhou, Jingli Zhang, Guodong Zhou, and Qiaoming Zhu

Due to the ability of encoding and mapping semantic information into a high-dimensional latent feature space, neural networks have been successfully used for detecting events to a certain extent. However, such a feature space can be easily contaminated by spurious features inherent in event detection. In this paper, we propose a self-regulated learning approach by utilizing a generative adversarial network to generate spurious features. On the basis, we employ a recurrent network to eliminate the fakes. Detailed experiments on the ACE 2005 and TAC-KBP 2015 corpora show that our proposed method is highly effective and adaptable.

Context-Aware Neural Model for Temporal Information Extraction
Yuanliang Meng and Anna Rumshisky

We propose a context-aware neural network model for temporal information extraction. This model has a uniform architecture for event-event, event-timex and timex-timex pairs. A Global Context Layer (GCL), inspired by Neural Turing Machine (NTM), stores processed temporal relations in narrative order, and retrieves them for use when relevant entities come in. Relations are then classified in context. The GCL model has long-term memory and attention mechanisms to resolve irregular long-distance dependencies that regular RNNs such as LSTM cannot recognize. It does not require any new input features, while outperforming the existing models in literature. To our knowledge it is also the first model to use NTM-like architecture to process the information from global context in discourse-scale natural text processing. We are going to release the source code in the future.

Temporal Event Knowledge Acquisition via Identifying Narratives
Wenlin Yao and Ruihong Huang

Inspired by the double temporality characteristic of narrative texts, we propose a novel approach for acquiring rich temporal "before/after" event knowledge across sentences in narrative stories. The double temporality states that a narrative story often describes a sequence of events following the chronological order and therefore, the temporal order of events matches with their textual order. We explored narratology principles and built a weakly supervised approach that identifies 287k narrative paragraphs from three large corpora. We then extracted rich temporal event knowledge from these narrative paragraphs. Such event knowledge is shown useful to improve temporal relation classification and outperforms several recent neural network models on the narrative cloze task.

Textual Deconvolution Saliency (TDS) : a deep tool box for linguistic analysis
Laurent Vanni, Mélanie Ducoffe, Carlos Aguilar, Frederic Precioso, and Damon Mayaffre

In this paper, we propose a new strategy, called Text Deconvolution Saliency (TDS), to visualize linguistic information detected by a CNN for text classification. We extend Deconvolution Networks to text in order to present a new perspective on text analysis to the linguistic community. We empirically demonstrated the efficiency of our Text Deconvolution Saliency on corpora from three different languages: English, French, and Latin. For every tested dataset, our Text Deconvolution Saliency automatically encodes complex linguistic patterns based on co-occurrences and possibly on grammatical and syntax analysis.

[TACL] Learning Distributed Representations of Texts and Entities from Knowledge Base
Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji

We describe a neural network model that jointly learns distributed representations of texts and knowledge base (KB) entities. Given a text in the KB, we train our proposed model to predict entities that are relevant to the text. Our model is designed to be generic with the ability to address various NLP tasks with ease. We train the model using a large corpus of texts and their entity annotations extracted from Wikipedia. We evaluated the model on three important NLP tasks (i.e., sentence textual similarity, entity linking, and factoid question answering) involving both unsupervised and supervised settings. As a result, we achieved state-of-the-art results on all three of these tasks. Our code and trained models are publicly available for further academic research.

[TACL] From Characters to Time Intervals: New Paradigms for Evaluation and Neural Parsing of Time Normalizations
Egoitz Laparra, Dongfang Xu, and Steven Bethard

This paper presents the first model for time normalization trained on the SCATE corpus. In the SCATE schema, time expressions are annotated as a semantic composition of time entities. This novel schema favors machine learning approaches, as it can be viewed as a semantic parsing task. In this work, we propose a character level multi-output neural network that outperforms previous state-of-the-art built on the TimeML schema. To compare predictions of systems that follow both SCATE and TimeML, we present a new scoring metric for time intervals. We also apply this new metric to carry out a comparative analysis of the annotations of both schemes in the same corpus.

Identification of Alias Links among Participants in Narratives
Sangameshwar Patil, Sachin Pawar, Swapnil Hingmire, Girish Palshikar, Vasudeva Varma, and Pushpak Bhattacharyya

Identification of distinct and independent participants (entities of interest) in a narrative is an important task for many NLP applications. This task becomes challenging because these participants are often referred to using multiple aliases. In this paper, we propose an approach based on linguistic knowledge for identification of aliases mentioned using proper nouns, pronouns or noun phrases with common noun headword. We use Markov Logic Network (MLN) to encode the linguistic knowledge for identification of aliases. We evaluate on four diverse history narratives of varying complexity. Our approach performs better than the state-of-the-art approach as well as a combination of standard named entity recognition and coreference resolution techniques.

Named Entity Recognition With Parallel Recurrent Neural Networks
Andrej Zukov Gregoric, Yoram Bachrach, and Sam Coope

We present a new architecture for named entity recognition. Our model employs multiple independent bidirectional LSTM units across the same input and promotes diversity among them by employing an inter-model regularization term. By distributing computation across multiple smaller LSTMs we find a significant reduction in the total number of parameters. We find our architecture achieves state-of-the-art performance on the CoNLL 2003 NER dataset.

Type-Sensitive Knowledge Base Inference Without Explicit Type Supervision
Prachi Jain, Pankaj Kumar, Mausam -, and Soumen Chakrabarti

State-of-the-art knowledge base completion (KBC) models predict a score for every known or unknown fact via a latent factorization over entity and relation embeddings. We observe that when they fail, they often make entity pre-

dictions that are incompatible with the type required by the relation. In response, we enhance each base factorization with two type-compatibility terms between entity-relation pairs, and combine the signals in a novel manner. Without explicit supervision from a type catalog, our proposed modification obtains up to 7% MRR gains over base models, and new state-of-the-art results on several datasets. Further analysis reveals that our models better represent the latent types of entities and their embeddings also predict supervised types better than the embeddings fitted by baseline models.

A Walk-based Model on Entity Graphs for Relation Extraction
Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou

We present a novel graph-based neural network model for relation extraction. Our model treats multiple pairs in a sentence simultaneously and considers interactions among them. All the entities in a sentence are placed as nodes in a fully-connected graph structure. The edges are represented with position-aware contexts around the entity pairs. In order to consider different relation paths between two entities, we construct up to l-length walks between each pair. The resulting walks are merged and iteratively used to update the edge representations into longer walks representations. We show that the model achieves performance comparable to the state-of-the-art systems on the ACE 2005 dataset without using any external tools.

Ranking-Based Automatic Seed Selection and Noise Reduction for Weakly Supervised Relation Extraction
Van-Thuy Phi, Joan Santoso, Masashi Shimbo, and Yuji Matsumoto

This paper addresses the tasks of automatic seed selection for bootstrapping relation extraction, and noise reduction for distantly supervised relation extraction. We first point out that these tasks are related. Then, inspired by ranking relation instances and patterns computed by the HITS algorithm, and selecting cluster centroids using the K-means, LSA, or NMF method, we propose methods for selecting the initial seeds from an existing resource, or reducing the level of noise in the distantly labeled data. Experiments show that our proposed methods achieve a better performance than the baseline systems in both tasks.

Automatic Extraction of Commonsense LocatedNear Knowledge
Frank F. Xu, Bill Yuchen Lin, and Kenny Zhu

LocatedNear relation is a kind of commonsense knowledge describing two physical objects that are typically found near each other in real life. In this paper, we study how to automatically extract such relationship through a sentence-level relation classifier and aggregating the scores of entity pairs from a large corpus. Also, we release two benchmark datasets for evaluation and future research.

Poster Session 1D: Discourse, Linguistics, Cognitive Modeling

Coherence Modeling of Asynchronous Conversations: A Neural Entity Grid Approach
Shafiq Joty, Muhammad Tasnim Mohiuddin, and Dat Tien Nguyen

We propose a novel coherence model for written asynchronous conversations (e.g., forums, emails), and show its applications in coherence assessment and thread reconstruction tasks. We conduct our research in two steps. First, we propose improvements to the recently proposed neural entity grid model by lexicalizing its entity transitions. Then, we extend the model to asynchronous conversations by incorporating the underlying conversational structure in the entity grid representation and feature computation. Our model achieves state of the art results on standard coherence assessment tasks in monologue and conversations outperforming existing models. We also demonstrate its effectiveness in reconstructing thread structures.

Deep Reinforcement Learning for Chinese Zero Pronoun Resolution
Qingyu Yin, Yu Zhang, Wei-Nan Zhang, Ting Liu, and William Yang Wang

Recent neural network models for Chinese zero pronoun resolution gain great performance by capturing semantic information for zero pronouns and candidate antecedents, but tend to be short-sighted, operating solely by making local decisions. They typically predict coreference links between the zero pronoun and one single candidate antecedent at a time while ignoring their influence on future decisions. Ideally, modeling useful information of preceding potential antecedents is crucial for classifying later zero pronoun-candidate antecedent pairs, a need which leads traditional models of zero pronoun resolution to draw on reinforcement learning. In this paper, we show how to integrate these goals, applying deep reinforcement learning to deal with the task. With the help of the reinforcement learning agent, our system learns the policy of selecting antecedents in a sequential manner, where useful information provided by earlier predicted antecedents could be utilized for making later coreference decisions. Experimental results on OntoNotes 5.0 show that our approach substantially outperforms the state-of-the-art methods under three experimental settings.

Entity-Centric Joint Modeling of Japanese Coreference Resolution and Predicate Argument Structure Analysis
Tomohide Shibata and Sadao Kurohashi

Predicate argument structure analysis is a task of identifying structured events. To improve this field, we need to identify a salient entity, which cannot be identified without performing coreference resolution and predicate argument structure analysis simultaneously. This paper presents an entity-centric joint model for Japanese coreference resolution and predicate argument structure analysis. Each entity is assigned an embedding, and when the result of both analyses refers to an entity, the entity embedding is updated. The analyses take the entity embedding into consideration to access the global information of entities. Our experimental results demonstrate the proposed method can improve the performance of the inter-sentential zero anaphora resolution drastically, which is a notoriously difficult task in predicate argument structure analysis.

Neural Coreference Resolution with Deep Biaffine Attention by Joint Mention Detection and Mention Clustering
Rui Zhang, Cicero Nogueira dos Santos, Michihiro Yasunaga, Bing Xiang, and Dragomir Radev

Coreference resolution aims to identify in a text all mentions that refer to the same real world entity. The state-of-the-art end-to-end neural coreference model considers all text spans in a document as potential mentions and learns to link an antecedent for each possible mention. In this paper, we propose to improve the end-to-end coreference resolution system by (1) using a biaffine attention model to get antecedent scores for each possible mention, and (2) jointly optimizing the mention detection accuracy and mention clustering accuracy given the mention cluster labels. Our model achieves the state-of-the-art performance on the CoNLL-2012 shared task English test set.

Fully Statistical Neural Belief Tracking
Nikola Mrkšić and Ivan Vulić

This paper proposes an improvement to the existing data-driven Neural Belief Tracking (NBT) framework for Dialogue State Tracking (DST). The existing NBT model uses a hand-crafted belief state update mechanism which involves an expensive manual retuning step whenever the model is deployed to a new dialogue domain. We show that this update mechanism can be learned jointly with the semantic decoding and context modelling parts of the NBT model, eliminating the last rule-based module from this DST framework. We propose two different statistical update mechanisms and show that dialogue dynamics can be modelled with a very small number of additional model parameters. In our DST evaluation over three languages, we show that this model achieves competitive performance and provides a robust framework for building resource-light DST models.

Constraining MGbank: Agreement, L-Selection and Supertagging in Minimalist Grammars
John Torr

This paper reports on two strategies that have been implemented for improving the efficiency and precision of wide-coverage Minimalist Grammar (MG) parsing. The first extends the formalism presented in Torr and Stabler (Torr & Stabler, 2016) with a mechanism for enforcing fine-grained selectional restrictions and agreements. The second is a method for factoring computationally costly null heads out from bottom-up MG parsing; this has the additional benefit of rendering the formalism fully compatible for the first time with highly efficient Markovian supertaggers. These techniques aided in the task of generating MGbank, the first wide-coverage corpus of Minimalist Grammar derivation trees.

Not that much power: Linguistic alignment is influenced more by low-level linguistic features rather than social power
Yang Xu, Jeremy Cole, and David Reitter

Linguistic alignment between dialogue partners has been claimed to be affected by their relative social power. A common finding has been that interlocutors of higher power tend to receive more alignment than those of lower power. However, these studies overlook some low-level linguistic features that can also affect alignment, which casts doubts on these findings. This work characterizes the effect of power on alignment with logistic regression models in two datasets, finding that the effect vanishes or is reversed after controlling for low-level features such as utterance length. Thus, linguistic alignment is explained better by low-level features than by social power. We argue that a wider range of factors, especially cognitive factors, need to be taken into account for future studies on observational data when social factors of language use are in question.

Some of Them Can be Guessed! Exploring the Effect of Linguistic Context in Predicting Quantifiers
Sandro Pezzelle, Shane Steinert-Threlkeld, Raffaella Bernardi, and Jakub Szymanik

We study the role of linguistic context in predicting quantifiers ('few', 'all'). We collect crowdsourced data from human participants and test various models in a local (single-sentence) and a global context (multi-sentence) condition. Models significantly out-perform humans in the former setting and are only slightly better in the latter. While human performance improves with more linguistic context (especially on proportional quantifiers), model performance suffers. Models are very effective in exploiting lexical and morpho-syntactic patterns; humans are better at genuinely understanding the meaning of the (global) context.

Poster Session 1E: Resources and Evaluation

TutorialBank: A Manually-Collected Corpus for Prerequisite Chains, Survey Extraction and Resource Recommendation
Alexander Fabbri, Irene Li, Prawat Trairatvorakul, Yijiao He, Weitai Ting, Robert Tung, Caitlin Westerfield, and Dragomir Radev

The field of Natural Language Processing (NLP) is growing rapidly, with new research published daily along with an abundance of tutorials, codebases and other online resources. In order to learn this dynamic field or stay up-to-date on the latest research, students as well as educators and researchers must constantly sift through multiple sources to find valuable, relevant information. To address this situation, we introduce TutorialBank, a new, publicly available dataset which aims to facilitate NLP education and research. We have manually collected and categorized over 5,600 resources on NLP as well as the related fields of Artificial Intelligence (AI), Machine Learning (ML) and Information Retrieval (IR). Our dataset is notably the largest manually-picked corpus of resources intended for NLP education which does not include only academic papers. Additionally, we have created both a search engine and a command-line tool for the resources and have annotated the corpus to include lists of research topics, relevant resources for each topic, prerequisite relations among topics, relevant sub-parts of individual resources, among other annotations. We are releasing the dataset and present several avenues for further research.

Give Me More Feedback: Annotating Argument Persuasiveness and Related Attributes in Student Essays
Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng

While argument persuasiveness is one of the most important dimensions of argumentative essay quality, it is relatively little studied in automated essay scoring research. Progress on scoring argument persuasiveness is hindered in part by the scarcity of annotated corpora. We present the first corpus of essays that are simultaneously annotated with argument components, argument persuasiveness scores, and attributes of argument components that impact an argument's persuasiveness. This corpus could trigger the development of novel computational models concerning argument persuasiveness that provide useful feedback to students on why their arguments are (un)persuasive in addition to how persuasive they are.

Inherent Biases in Reference-based Evaluation for Grammatical Error Correction
Leshem Choshen and Omri Abend

The prevalent use of too few references for evaluating text-to-text generation is known to bias estimates of their quality (henceforth, low coverage bias or LCB). This paper shows that overcoming LCB in Grammatical Error Correction (GEC) evaluation cannot be attained by re-scaling or by increasing the number of references in any feasible range, contrary to previous suggestions. This is due to the long-tailed distribution of valid corrections for a sentence. Concretely, we show that LCB incentivizes GEC systems to avoid correcting even when they can generate a valid correction. Consequently, existing systems obtain comparable or superior performance compared to humans, by making few but targeted changes to the input. Similar effects on Text Simplification further support our claims.

The price of debiasing automatic metrics in natural language evalaution
Arun Chaganty, Stephen Mussmann, and Percy Liang

For evaluating generation systems, automatic metrics such as BLEU cost nothing to run but have been shown to correlate poorly with human judgment, leading to systematic bias against certain model improvements. On the other hand, averaging human judgments, the unbiased gold standard, is often too expensive. In this paper, we use control variates to combine automatic metrics with human evaluation to obtain an unbiased estimator with lower cost than human evaluation alone. In practice, however, we obtain only a 7-13% cost reduction on evaluating summarization and open-response question answering systems. We then prove that our estimator is optimal: there is no unbiased estimator with lower cost. Our theory further highlights the two fundamental bottlenecks—the automatic metric and the prompt shown to human evaluators—both of which need to be improved to obtain greater cost savings.

A Named Entity Recognition Shootout for German
Martin Riedl and Sebastian Padó

We ask how to practically build a model for German named entity recognition (NER) that performs at the state of the art for both contemporary and historical texts, i.e., a big-data and a small-data scenario. The two best-performing model families are pitted against each other (linear-chain CRFs and BiLSTM) to observe the trade-off between expressiveness and data requirements. BiLSTM outperforms the CRF when large datasets are available and performs inferior for the smallest dataset. BiLSTMs profit substantially from transfer learning, which enables them to be trained on multiple corpora, resulting in a new state-of-the-art model for German NER on two contemporary German corpora (CoNLL 2003 and GermEval 2014) and two historic corpora.

A dataset for identifying actionable feedback in collaborative software development
Benjamin S Meyers, Nuthan Munaiah, Emily T. Prud'hommeaux, Andrew Meneely, Josephine Wolff, Cecilia Ovesdotter Alm, and Pradeep Murukannaiah

Software developers and testers have long struggled with how to elicit proactive responses from their coworkers when reviewing code for security vulnerabilities and errors. For a code review to be successful, it must not only identify potential problems but also elicit an active response from the colleague responsible for modifying the code. To understand the factors that contribute to this outcome, we analyze a novel dataset of more than one million code reviews for the Google Chromium project, from which we extract linguistic features of feedback that elicited responsive actions from coworkers. Using a manually-labeled subset of reviewer comments, we trained a highly accurate classifier to identify acted-upon comments (AUC = 0.85). Our results demonstrate the utility of our dataset, the feasibility of using NLP for this new task, and the potential of NLP to improve our understanding of how communications between colleagues can be authored to elicit positive, proactive responses.

SNAG: Spoken Narratives and Gaze Dataset
Preethi Vaidyanathan, Emily T. Prud'hommeaux, Jeff B. Pelz, and Cecilia O. Alm

Humans rely on multiple sensory modalities when examining and reasoning over images. In this paper, we describe a new multimodal dataset that consists of gaze measurements and spoken descriptions collected in parallel during an image inspection task. The task was performed by multiple participants on 100 general-domain images showing everyday objects and activities. We demonstrate the usefulness of the dataset by applying an existing visual-linguistic data fusion framework in order to label important image regions with appropriate linguistic labels.

Analogical Reasoning on Chinese Morphological and Semantic Relations
Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du

Analogical reasoning is effective in capturing linguistic regularities. This paper proposes an analogical reasoning task on Chinese. After delving into Chinese lexical knowledge, we sketch 68 implicit morphological relations and 28 explicit semantic relations. A big and balanced dataset CA8 is then built for this task, including 17813 questions. Furthermore, we systematically explore the influences of vector representations, context features, and corpora on analogical reasoning. With the experiments, CA8 is proved to be a reliable benchmark for evaluating Chinese word embeddings.

Construction of a Chinese Corpus for the Analysis of the Emotionality of Metaphorical Expressions
Dongyu Zhang, Hongfei Lin, Liang Yang, Shaowu Zhang, and BO XU

Metaphors are frequently used to convey emotions. However, there is little research on the construction of metaphor corpora annotated with emotion for the analysis of emotionality of metaphorical expressions. Furthermore, most studies focus on English, and few in other languages, particularly Sino-Tibetan languages such as Chinese, for emotion analysis from metaphorical texts, although there are likely to be many differences in emotional expressions of metaphorical usages across different languages. We therefore construct a significant new corpus on metaphor, with 5,605 manually annotated sentences in Chinese. We present an annotation scheme that contains annotations of linguistic metaphors, emotional categories (joy, anger, sadness, fear, love, disgust and surprise), and intensity. The annotation agreement analyses for multiple annotators are described. We also use the corpus to explore and analyze the emotionality of metaphors. To the best of our knowledge, this is the first relatively large metaphor corpus with an annotation of emotions in Chinese.

Automatic Article Commenting: the Task and Dataset
Lianhui Qin, Lemao Liu, Wei Bi, Yan Wang, Xiaojiang Liu, Zhiting Hu, Hai Zhao, and Shuming Shi

Comments of online articles provide extended views and improve user engagement. Automatically making comments thus become a valuable functionality for online forums, intelligent chatbots, etc. This paper proposes the new task of automatic article commenting, and introduces a large-scale Chinese dataset with millions of real comments and a human-annotated subset characterizing the comments' varying quality. Incorporating the human bias of comment quality, we further develop automatic metrics that generalize a broad set of popular reference-based metrics and exhibit greatly improved correlations with human evaluations.

Improved Evaluation Framework for Complex Plagiarism Detection
Anton Belyy, Marina Dubova, and Dmitry Nekrasov

Plagiarism is a major issue in science and education. Complex plagiarism, such as plagiarism of ideas, is hard to detect, and therefore it is especially important to track improvement of methods correctly. In this paper, we study the performance of plagdet, the main measure for plagiarim detection, on manually paraphrased datasets (such as PAN Summary). We reveal its fallibility under certain conditions and propose an evaluation framework with normalization of inner terms, which is resilient to the dataset imbalance. We conclude with the experimental justification of the

proposed measure. The implementation of the new framework is made publicly available as a Github repository.

Poster Session 1F: Summarization, Social Media

Neural Document Summarization by Jointly Learning to Score and Select Sentences
Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao

Sentence scoring and sentence selection are two main steps in extractive document summarization systems. However, previous works treat them as two separated subtasks. In this paper, we present a novel end-to-end neural network framework for extractive document summarization by jointly learning to score and select sentences. It first reads the document sentences with a hierarchical encoder to obtain the representation of sentences. Then it builds the output summary by extracting sentences one by one. Different from previous methods, our approach integrates the selection strategy into the scoring model, which directly predicts the relative importance given previously selected sentences. Experiments on the CNN/Daily Mail dataset show that the proposed framework significantly outperforms the state-of-the-art extractive summarization models.

Unsupervised Abstractive Meeting Summarization with Multi-Sentence Compression and Budgeted Submodular Maximization
Guokan Shang, Wensi Ding, Zekun Zhang, Antoine Tixier, Polykarpos Meladianos, Michalis Vazirgiannis, and Jean-Pierre Lorré

We introduce a novel graph-based framework for abstractive meeting speech summarization that is fully unsupervised and does not rely on any annotations. Our work combines the strengths of multiple recent approaches while addressing their weaknesses. Moreover, we leverage recent advances in word embeddings and graph degeneracy applied to NLP to take exterior semantic knowledge into account, and to design custom diversity and informativeness measures. Experiments on the AMI and ICSI corpus show that our system improves on the state-of-the-art. Code and data are publicly available, and our system can be interactively tested.

Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting
Yen-Chun Chen and Mohit Bansal

Inspired by how humans summarize long documents, we propose an accurate and fast summarization model that first selects salient sentences and then rewrites them abstractively (i.e., compresses and paraphrases) to generate a concise overall summary. We use a novel sentence-level policy gradient method to bridge the non-differentiable computation between these two neural networks in a hierarchical way, while maintaining language fluency. Empirically, we achieve the new state-of-the-art on all metrics (including human evaluation) on the CNN/Daily Mail dataset, as well as significantly higher abstractiveness scores. Moreover, by first operating at the sentence-level and then the word-level, we enable parallel decoding of our neural generative model that results in substantially faster (10-20x) inference speed as well as 4x faster training convergence than previous long-paragraph encoder-decoder models. We also demonstrate the generalization of our model on the test-only DUC-2002 dataset, where we achieve higher scores than a state-of-the-art model.

Soft Layer-Specific Multi-Task Summarization with Entailment and Question Generation
Han Guo, Ramakanth Pasunuru, and Mohit Bansal

An accurate abstractive summary of a document should contain all its salient information and should be logically entailed by the input document. We improve these important aspects of abstractive summarization via multi-task learning with the auxiliary tasks of question generation and entailment generation, where the former teaches the summarization model how to look for salient questioning-worthy details, and the latter teaches the model how to rewrite a summary which is a directed-logical subset of the input document. We also propose novel multi-task architectures with high-level (semantic) layer-specific sharing across multiple encoder and decoder layers of the three tasks, as well as soft-sharing mechanisms (and show performance ablations and analysis examples of each contribution). Overall, we achieve statistically significant improvements over the state-of-the-art on both the CNN/DailyMail and Gigaword datasets, as well as on the DUC-2002 transfer setup. We also present several quantitative and qualitative analysis studies of our model's learned saliency and entailment skills.

Global Encoding for Abstractive Summarization
Junyang Lin, Xu Sun, Shuming Ma, and Qi Su

In neural abstractive summarization, the conventional sequence-to-sequence (seq2seq) model often suffers from repetition and semantic irrelevance. To tackle the problem, we propose a global encoding framework, which controls the information flow from the encoder to the decoder based on the global information of the source context. It consists of a convolutional gated unit to perform global encoding to improve the representations of the source-side information. Evaluations on the LCSTS and the English Gigaword both demonstrate that our model outperforms the baseline models, and the analysis shows that our model is capable of generating summary of higher quality and reducing repetition.

A Language Model based Evaluator for Sentence Compression
Yang Zhao, Zhiyuan Luo, and Akiko Aizawa

We herein present a language-model-based evaluator for deletion-based sentence compression and view this task as a series of deletion-and-evaluation operations using the evaluator. More specifically, the evaluator is a syntactic neural language model that is first built by learning the syntactic and structural collocation among words. Subsequently, a series of trial-and-error deletion operations are conducted on the source sentences via a reinforcement learning framework to obtain the best target compression. An empirical study shows that the proposed model can effectively generate more readable compression, comparable or superior to several strong baselines. Furthermore, we introduce a 200-sentence test set for a large-scale dataset, setting a new baseline for the future research.

Modeling and Prediction of Online Product Review Helpfulness: A Survey
Gerardo Ocampo Diaz and Vincent Ng

As the amount of free-form user-generated reviews in e-commerce websites continues to increase, there is an increasing need for automatic mechanisms that sift through the vast amounts of user reviews and identify quality content. Review helpfulness modeling is a task which studies the mechanisms that affect review helpfulness and attempts to accurately predict it. This paper provides an overview of the most relevant work in helpfulness prediction and understanding in the past decade, discusses the insights gained from said work, and provides guidelines for future research.

Mining Cross-Cultural Differences and Similarities in Social Media
Bill Yuchen Lin, Frank F. Xu, Kenny Zhu, and Seung-won Hwang

Cross-cultural differences and similarities are common in cross-lingual natural language understanding, especially for research in social media. For instance, people of distinct cultures often hold different opinions on a single named entity. Also, understanding slang terms across languages requires knowledge of cross-cultural similarities. In this paper, we study the problem of computing such cross-cultural differences and similarities. We present a lightweight yet effective approach, and evaluate it on two novel tasks: 1) mining cross-cultural differences of named entities and 2) finding similar terms for slang across languages. Experimental results show that our framework substantially outperforms a number of baseline methods on both tasks. The framework could be useful for machine translation applications and research in computational social science.

Classification of Moral Foundations in Microblog Political Discourse
Kristen Johnson and Dan Goldwasser

Previous works in computer science, as well as political and social science, have shown correlation in text between political ideologies and the moral foundations expressed within that text. Additional work has shown that policy frames, which are used by politicians to bias the public towards their stance on an issue, are also correlated with political ideology. Based on these associations, this work takes a first step towards modeling both the language and how politicians frame issues on Twitter, in order to predict the moral foundations that are used by politicians to express their stances on issues. The contributions of this work includes a dataset annotated for the moral foundations, annotation guidelines, and probabilistic graphical models which show the usefulness of jointly modeling abstract political slogans, as opposed to the unigrams of previous works, with policy frames for the prediction of the morality underlying political tweets.

Identifying and Understanding User Reactions to Deceptive and Trusted Social News Sources
Maria Glenski, Tim Weninger, and Svitlana Volkova

In the age of social news, it is important to understand the types of reactions that are evoked from news sources with various levels of credibility. In the present work we seek to better understand how users react to trusted and deceptive news sources across two popular, and very different, social media platforms. To that end, (1) we develop a model to classify user reactions into one of nine types, such as answer, elaboration, and question, etc, and (2) we measure the speed and the type of reaction for trusted and deceptive news sources for 10.8M Twitter posts and 6.2M Reddit comments. We show that there are significant differences in the speed and the type of reactions between trusted and deceptive news sources on Twitter, but far smaller differences on Reddit.

Content-based Popularity Prediction of Online Petitions Using a Deep Regression Model
Shivashankar Subramanian, Timothy Baldwin, and Trevor Cohn

Online petitions are a cost-effective way for citizens to collectively engage with policy-makers in a democracy. Predicting the popularity of a petition — commonly measured by its signature count — based on its textual content has utility for policymakers as well as those posting the petition. In this work, we model this task using CNN regression with an auxiliary ordinal regression objective. We demonstrate the effectiveness of our proposed approach using UK and US government petition datasets.

Fighting Offensive Language on Social Media with Unsupervised Text Style Transfer
Cicero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi

We introduce a new approach to tackle the problem of offensive language in online social media. Our approach uses unsupervised text style transfer to translate offensive sentences into non-offensive ones. We propose a new method for training encoder-decoders using non-parallel data that combines a collaborative classifier, attention and the cycle consistency loss. Experimental results on data from Twitter and Reddit show that our method outperforms a state-of-the-art text style transfer system in two out of three quantitative metrics and produces reliable non-offensive transferred sentences.

Diachronic degradation of language models: Insights from social media
Kokil Jaidka, Niyati Chhaya, and Lyle Ungar

Natural languages change over time because they evolve to the needs of their users and the socio-technological environment. This study investigates the diachronic accuracy of pre-trained language models for downstream tasks in machine learning and user profiling. It asks the question: given that the social media platform and its users remain the same, how is language changing over time? How can these differences be used to track the changes in the affect around a particular topic? To our knowledge, this is the first study to show that it is possible to measure diachronic semantic drifts within social media and within the span of a few years.

Demo Poster Session 1

Time: *12:30–14:00* Location: Melbourne Room 1 & 2

Platforms for Non-speakers Annotating Names in Any Language
Ying Lin, Cash Costello, Boliang Zhang, Di Lu, Heng Ji, James Mayfield, and Paul McNamee

We demonstrate two annotation platforms that allow an English speaker to annotate names for any language without knowing the language. These platforms provided high-quality ''silver standard" annotations for low-resource language name taggers (Zhang et al., 2017) that achieved state-of-the-art performance on two surprise languages (Oromo and Tigrinya) at LoreHLT20171 and ten languages at TAC-KBP EDL2017 (Ji et al., 2017). We discuss strengths and limitations and compare other methods of creating silver- and gold- standard annotations using native speakers. We will make our tools publicly available for research use.

NovelPerspective: Identifying Point of View Characters
Lyndon White, Roberto Togneri, Wei Liu, and Mohammed Bennamoun

We present NovelPerspective: a tool to allow consumers to subset their digital literature, based on point of view (POV) character. Many novels have multiple main characters each with their own storyline running in parallel. A well-known example is George R. R. Martin's novel: "A Game of Thrones", and others from that series. Our tool detects the main character that each section is from the POV of, and allows the user to generate a new ebook with only those sections. This gives consumers new options in how they consume their media; allowing them to pursue the storylines sequentially, or skip chapters about characters they find boring. We present two heuristic-based baselines, and two machine learning based methods for the detection of the main character.

Out-of-the-box Universal Romanization Tool uroman
Ulf Hermjakob, Jonathan May, and Kevin Knight

We present uroman, a tool for converting text in myriads of languages and scripts such as Chinese, Arabic and Cyrillic into a common Latin-script representation. The tool relies on Unicode data and other tables, and handles nearly all character sets, including some that are quite obscure such as Tibetan and Tifinagh. uroman converts digital numbers in various scripts to Western Arabic numerals. Romanization enables the application of string-similarity metrics to texts from different scripts without the need and complexity of an intermediate phonetic representation. The tool is freely and publicly available as a Perl script suitable for inclusion in data processing pipelines and as an interactive demo web page.

HarriGT: A Tool for Linking News to Science
James Ravenscroft, Amanda Clare, and Maria Liakata

Being able to reliably link scientific works to the newspaper articles that discuss them could provide a breakthrough in the way we rationalise and measure the impact of science on our society. Linking these articles is challenging because the language used in the two domains is very different, and the gathering of online resources to align the two is a substantial information retrieval endeavour. We present HarriGT, a semi-automated tool for building corpora of news articles linked to the scientific papers that they discuss. Our aim is to facilitate future development of information-retrieval tools for newspaper/scientific work citation linking. HarriGT retrieves newspaper articles from an archive containing 17 years of UK web content. It also integrates with 3 large external citation networks, leveraging named entity extraction, and document classification to surface relevant examples of scientific literature to the user. We also provide a tuned candidate ranking algorithm to highlight potential links between scientific papers and newspaper articles to the user, in order of likelihood. HarriGT is provided as an open source tool (http://harrigt.xyz).

Jack the Reader – A Machine Reading Framework
Dirk Weissenborn, Pasquale Minervini, Isabelle Augenstein, Johannes Welbl, Tim Rocktäschel, Matko Bosnjak, Jeff Mitchell, Thomas Demeester, Tim Dettmers, Pontus Stenetorp, and Sebastian Riedel

Many Machine Reading and Natural Language Understanding tasks require reading supporting text in order to answer questions. For example, in Question Answering, the supporting text can be newswire or Wikipedia articles; in Natural Language Inference, premises can be seen as the supporting text and hypotheses as questions. Providing a set of useful primitives operating in a single framework of related tasks would allow for expressive modelling, and easier model comparison and replication. To that end, we present Jack the Reader (JACK), a framework for Machine Reading that allows for quick model prototyping by component reuse, evaluation of new models on existing datasets as well as integrating new datasets and applying them on a growing set of implemented baseline models. JACK is currently supporting (but not limited to) three tasks: Question Answering, Natural Language Inference, and Link Prediction. It is developed with the aim of increasing research efficiency and code reuse.

YEDDA: A Lightweight Collaborative Text Span Annotation Tool
Jie Yang, Yue Zhang, Linwei Li, and Xingxuan Li

In this paper, we introduce YEDDA, a lightweight but efficient and comprehensive open-source tool for text span annotation. YEDDA provides a systematic solution for text span annotation, ranging from collaborative user annotation to administrator evaluation and analysis. It overcomes the low efficiency of traditional text annotation tools by annotating entities through both command line and shortcut keys, which are configurable with custom labels. YEDDA also gives intelligent recommendations by learning the up-to-date annotated text. An administrator client is developed to evaluate annotation quality of multiple annotators and generate detailed comparison report for each annotator pair. Experiments show that the proposed system can reduce the annotation time by half compared with existing annotation tools. And the annotation time can be further compressed by 16.47% through intelligent recommendation.

NextGen AML: Distributed Deep Learning based Language Technologies to Augment Anti Money Laundering Investigation
Jingguang Han, Utsab Barman, Jeremiah Hayes, Jinhua Du, Edward Burgin, and Dadong Wan

Most of the current anti money laundering (AML) systems, using handcrafted rules, are heavily reliant on existing structured databases, which are not capable of effectively and efficiently identifying hidden and complex ML activities, especially those with dynamic and time-varying characteristics, resulting in a high percentage of false positives. Therefore, analysts are engaged for further investigation which significantly increases human capital cost and processing time. To alleviate these issues, this paper presents a novel framework for the next generation AML by applying and visualizing deep learning-driven natural language processing (NLP) technologies in a distributed and scalable manner to augment AML monitoring and investigation. The proposed distributed framework performs news and tweet sentiment analysis, entity recognition, relation extraction, entity linking and link analysis on different data sources (e.g. news articles and tweets) to provide additional evidence to human investigators for final decision-making. Each NLP module is evaluated on a task-specific data set, and the overall experiments are performed on synthetic and real-world datasets. Feedback from AML practitioners suggests that our system can reduce approximately 30% time and cost compared to their previous manual approaches of AML investigation.

NLP Web Services for Resource-Scarce Languages
Martin Puttkammer, Roald Eiselen, Justin Hocking, and Frederik Koen

In this paper, we present a project where existing text-based core technologies were ported to Java-based web services from various architectures. These technologies were developed over a period of eight years through various government funded projects for 10 resource-scarce languages spoken in South Africa. We describe the API and a simple web front-end capable of completing various predefined tasks.

Session 2 Overview – Monday, July 16, 2018

Track A *Semantic Parsing 1* Plenary	**Track B** *Multilinguality* 203–204	**Track C** *Question Answering 1* 210–211	**Track D** *Generation 1* 212–213	**Track E** *Vision* 219	**Track F** *Sentiment* 220	
Coarse-to-Fine Decoding for Neural Semantic Parsing *Dong and Lapata*	On the Limitations of Unsupervised Bilingual Dictionary Induction *Søgaard, Ruder, and Vulić*	Knowledgeable Reader: Enhancing Cloze-Style Reading Comprehension with External Commonsense Knowledge *Mihaylov and Frank*	Style Transfer Through Back-Translation *Prabhumoye, Tsvetkov, Salakhutdinov, and Black*	No Metrics Are Perfect: Adversarial Reward Learning for Visual Storytelling *Wang, Chen, Wang, and Wang*	Transformation Networks for Target-Oriented Sentiment Classification *Li, Bing, Lam, and Shi*	14:00
Confidence Modeling for Neural Semantic Parsing *Dong, Quirk, and Lapata*	A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings *Artetxe, Labaka, and Agirre*	Multi-Relational Question Answering from Narratives: Machine Reading and Reasoning in Simulated Worlds *Labutov, Yang, Prakash, and Azaria*	Generating Fine-Grained Open Vocabulary Entity Type Descriptions *Bhowmik and Melo*	Bridging Languages through Images with Deep Partial Canonical Correlation Analysis *Rotman, Vulić, and Reichart*	Target-Sensitive Memory Networks for Aspect Sentiment Classification *Wang, Mazumder, Liu, Zhou, and Chang*	14:25
StructVAE: Tree-structured Latent Variable Models for Semi-supervised Semantic Parsing *Yin, Zhou, He, and Neubig*	A Multi-lingual Multi-task Architecture for Low-resource Sequence Labeling *Lin, Yang, Stoyanov, and Ji*	Simple and Effective Multi-Paragraph Reading Comprehension *Clark and Gardner*	Hierarchical Neural Story Generation *Fan, Lewis, and Dauphin*	Illustrative Language Understanding: Large-Scale Visual Grounding with Image Search *Kiros, Chan, and Hinton*	Identifying Transferable Information Across Domains for Cross-domain Sentiment Classification *Sharma, Bhattacharyya, Dandapat, and Bhatt*	14:50
Sequence-to-Action: End-to-End Semantic Graph Generation for Semantic Parsing *Chen, Sun, and Han*	Two Methods for Domain Adaptation of Bilingual Tasks: Delightfully Simple and Broadly Applicable *Hangya, Braune, Fraser, and Schütze*	Semantically Equivalent Adversarial Rules for Debugging NLP models *Ribeiro, Singh, and Guestrin*	[TACL] Generating Sentences by Editing Prototypes *Guu, Hashimoto, Oren, and Liang*	What Action Causes This? Towards Naive Physical Action-Effect Prediction *Gao, Yang, Chai, and Vanderwende*	Unpaired Sentiment-to-Sentiment Translation: A Cycled Reinforcement Learning Approach *Xu, Sun, Zeng, Zhang, Ren, Wang, and Li*	15:15

Parallel Session 2

Session 2A: Semantic Parsing 1

Plenary *Chair: Jonathan Berant*

Coarse-to-Fine Decoding for Neural Semantic Parsing
Li Dong and Mirella Lapata 14:00–14:25

Semantic parsing aims at mapping natural language utterances into structured meaning representations. In this work, we propose a structure-aware neural architecture which decomposes the semantic parsing process into two stages. Given an input utterance, we first generate a rough sketch of its meaning, where low-level information (such as variable names and arguments) is glossed over. Then, we fill in missing details by taking into account the natural language input and the sketch itself. Experimental results on four datasets characteristic of different domains and meaning representations show that our approach consistently improves performance, achieving competitive results despite the use of relatively simple decoders.

Confidence Modeling for Neural Semantic Parsing
Li Dong, Chris Quirk, and Mirella Lapata 14:25–14:50

In this work we focus on confidence modeling for neural semantic parsers which are built upon sequence-to-sequence models. We outline three major causes of uncertainty, and design various metrics to quantify these factors. These metrics are then used to estimate confidence scores that indicate whether model predictions are likely to be correct. Beyond confidence estimation, we identify which parts of the input contribute to uncertain predictions allowing users to interpret their model, and verify or refine its input. Experimental results show that our confidence model significantly outperforms a widely used method that relies on posterior probability, and improves the quality of interpretation compared to simply relying on attention scores.

StructVAE: Tree-structured Latent Variable Models for Semi-supervised Semantic Parsing
Pengcheng Yin, Chunting Zhou, Junxian He, and Graham Neubig 14:50–15:15

Semantic parsing is the task of transducing natural language (NL) utterances into formal meaning representations (MRs), commonly represented as tree structures. Annotating NL utterances with their corresponding MRs is expensive and time-consuming, and thus the limited availability of labeled data often becomes the bottleneck of data-driven, supervised models. We introduce StructVAE, a variational auto-encoding model for semi-supervised semantic parsing, which learns both from limited amounts of parallel data, and readily-available unlabeled NL utterances. StructVAE models latent MRs not observed in the unlabeled data as tree-structured latent variables. Experiments on semantic parsing on the ATIS domain and Python code generation show that with extra unlabeled data, StructVAE outperforms strong supervised models.

Sequence-to-Action: End-to-End Semantic Graph Generation for Semantic Parsing
Bo Chen, Le Sun, and Xianpei Han 15:15–15:40

This paper proposes a neural semantic parsing approach – Sequence-to-Action, which models semantic parsing as an end-to-end semantic graph generation process. Our method simultaneously leverages the advantages from two recent promising directions of semantic parsing. Firstly, our model uses a semantic graph to represent the meaning of a sentence, which has a tight-coupling with knowledge bases. Secondly, by leveraging the powerful representation learning and prediction ability of neural network models, we propose a RNN model which can effectively map sentences to action sequences for semantic graph generation. Experiments show that our method achieves state-of-the-art performance on Overnight dataset and gets competitive performance on Geo and Atis datasets.

Session 2B: Multilinguality

203–204 *Chair: Shuly Wintner*

On the Limitations of Unsupervised Bilingual Dictionary Induction
Anders Søgaard, Sebastian Ruder, and Ivan Vulić 14:00–14:25

Unsupervised machine translation - i.e., not assuming any cross-lingual supervision signal, whether a dictionary, translations, or comparable corpora - seems impossible, but nevertheless, Lample et al. (2017) recently proposed a fully unsupervised machine translation (MT) model. The model relies heavily on an adversarial, unsupervised cross-lingual word embedding technique for bilingual dictionary induction (Conneau et al., 2017), which we examine here. Our results identify the limitations of current unsupervised MT: unsupervised bilingual dictionary induction performs much worse on morphologically rich languages that are not dependent marking, when monolingual corpora from different domains or different embedding algorithms are used. We show that a simple trick, exploiting a weak supervision signal from identical words, enables more robust induction and establish a near-perfect correlation between unsupervised bilingual dictionary induction performance and a previously unexplored graph similarity metric.

A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings
Mikel Artetxe, Gorka Labaka, and Eneko Agirre 14:25–14:50

Recent work has managed to learn cross-lingual word embeddings without parallel data by mapping monolingual embeddings to a shared space through adversarial training. However, their evaluation has focused on favorable conditions, using comparable corpora or closely-related languages, and we show that they often fail in more realistic scenarios. This work proposes an alternative approach based on a fully unsupervised initialization that explicitly exploits the structural similarity of the embeddings, and a robust self-learning algorithm that iteratively improves this solution. Our method succeeds in all tested scenarios and obtains the best published results in standard datasets, even surpassing previous supervised systems. Our implementation is released as an open source project at https://github.com/artetxem/vecmap.

A Multi-lingual Multi-task Architecture for Low-resource Sequence Labeling
Ying Lin, Shengqi Yang, Veselin Stoyanov, and Heng Ji 14:50–15:15

We propose a multi-lingual multi-task architecture to develop supervised models with a minimal amount of labeled data for sequence labeling. In this new architecture, we combine various transfer models using two layers of parameter sharing. On the first layer, we construct the basis of the architecture to provide universal word representation and feature extraction capability for all models. On the second level, we adopt different parameter sharing strategies for different transfer schemes. This architecture proves to be particularly effective for low-resource settings, when there are less than 200 training sentences for the target task. Using Name Tagging as a target task, our approach achieved 4.3%-50.5% absolute F-score gains compared to the mono-lingual single-task baseline model.

Two Methods for Domain Adaptation of Bilingual Tasks: Delightfully Simple and Broadly Applicable
Viktor Hangya, Fabienne Braune, Alexander Fraser, and Hinrich Schütze 15:15–15:40

Bilingual tasks, such as bilingual lexicon induction and cross-lingual classification, are crucial for overcoming data sparsity in the target language. Resources required for such tasks are often out-of-domain, thus domain adaptation is an important problem here. We make two contributions. First, we test a delightfully simple method for domain adaptation of bilingual word embeddings. We evaluate these embeddings on two bilingual tasks involving different domains: cross-lingual twitter sentiment classification and medical bilingual lexicon induction. Second, we tailor a broadly applicable semi-supervised classification method from computer vision to these tasks. We show that this method also helps in low-resource setups. Using both methods together we achieve large improvements over our baselines, by using only additional unlabeled data.

Session 2C: Question Answering 1

210–211 *Chair: Diane J. Litman*

Knowledgeable Reader: Enhancing Cloze-Style Reading Comprehension with External Commonsense Knowledge
Todor Mihaylov and Anette Frank 14:00–14:25

We introduce a neural reading comprehension model that integrates external commonsense knowledge, encoded as a key-value memory, in a cloze-style setting. Instead of relying only on document-to-question interaction or discrete features as in prior work, our model attends to relevant external knowledge and combines this knowledge with the context representation before inferring the answer. This allows the model to attract and imply knowledge from an external knowledge source that is not explicitly stated in the text, but that is relevant for inferring the answer. Our model improves results over a very strong baseline on a hard Common Nouns dataset, making it a strong competitor of much more complex models. By including knowledge explicitly, our model can also provide evidence about the background knowledge used in the RC process.

Multi-Relational Question Answering from Narratives: Machine Reading and Reasoning in Simulated Worlds
Igor Labutov, Bishan Yang, Anusha Prakash, and Amos Azaria 14:25–14:50

Question Answering (QA), as a research field, has primarily focused on either knowledge bases (KBs) or free text as a source of knowledge. These two sources have historically shaped the kinds of questions that are asked over these sources, and the methods developed to answer them. In this work, we look towards a practical use-case of QA over user-instructed knowledge that uniquely combines elements of both structured QA over knowledge bases, and unstructured QA over narrative, introducing the task of multi-relational QA over personal narrative. As a first step towards this goal, we make three key contributions: (i) we generate and release TextWorldsQA, a set of five diverse datasets, where each dataset contains dynamic narrative that describes entities and relations in a simulated world, paired with variably compositional questions over that knowledge, (ii) we perform a thorough evaluation and analysis of several state-of-the-art QA models and their variants at this task, and (iii) we release a lightweight Python-based framework we call TextWorlds for easily generating arbitrary additional worlds and narrative, with the goal of allowing the community to create and share a growing collection of diverse worlds as a test-bed for this task.

Simple and Effective Multi-Paragraph Reading Comprehension
Christopher Clark and Matt Gardner 14:50–15:15

We introduce a method of adapting neural paragraph-level question answering models to the case where entire documents are given as input. Most current question answering models cannot scale to document or multi-document input, and naively applying these models to each paragraph independently often results in them being distracted by irrelevant text. We show that it is possible to significantly improve performance by using a modified training scheme that teaches the model to ignore non-answer containing paragraphs. Our method involves sampling multiple paragraphs from each document, and using an objective function that requires the model to produce globally correct output. We additionally identify and improve upon a number of other design decisions that arise when working with document-level data. Experiments on TriviaQA and SQuAD shows our method advances the state of the art, including a 10 point gain on TriviaQA.

Semantically Equivalent Adversarial Rules for Debugging NLP models
Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin 15:15–15:40

Complex machine learning models for NLP are often brittle, making different predictions for input instances that are extremely similar semantically. To automatically detect this behavior for individual instances, we present semantically equivalent adversaries (SEAs) – semantic-preserving perturbations that induce changes in the model's predictions. We generalize these adversaries into semantically equivalent adversarial rules (SEARs) – simple, universal replacement rules that induce adversaries on many instances. We demonstrate the usefulness and flexibility of SEAs and SEARs by detecting bugs in black-box state-of-the-art models for three domains: machine comprehension, visual question-answering, and sentiment analysis. Via user studies, we demonstrate that we generate high-quality local adversaries for more instances than humans, and that SEARs induce four times as many mistakes as the bugs discovered by human experts. SEARs are also actionable: retraining models using data augmentation significantly reduces bugs, while maintaining accuracy.

Session 2D: Generation 1

212–213 *Chair: Lu Wang*

Style Transfer Through Back-Translation
Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W. Black 14:00–14:25

Style transfer is the task of rephrasing the text to contain specific stylistic properties without changing the intent or affect within the context. This paper introduces a new method for automatic style transfer. We first learn a latent representation of the input sentence which is grounded in a language translation model in order to better preserve the meaning of the sentence while reducing stylistic properties. Then adversarial generation techniques are used to make the output match the desired style. We evaluate this technique on three different style transformations: sentiment, gender and political slant. Compared to two state-of-the-art style transfer modeling techniques we show improvements both in automatic evaluation of style transfer and in manual evaluation of meaning preservation and fluency.

Generating Fine-Grained Open Vocabulary Entity Type Descriptions
Rajarshi Bhowmik and Gerard de Melo 14:25–14:50

While large-scale knowledge graphs provide vast amounts of structured facts about entities, a short textual description can often be useful to succinctly characterize an entity and its type. Unfortunately, many knowledge graphs entities lack such textual descriptions. In this paper, we introduce a dynamic memory-based network that generates a short open vocabulary description of an entity by jointly leveraging induced fact embeddings as well as the dynamic context of the generated sequence of words. We demonstrate the ability of our architecture to discern relevant information for more accurate generation of type description by pitting the system against several strong baselines.

Hierarchical Neural Story Generation
Angela Fan, Mike Lewis, and Yann Dauphin 14:50–15:15

We explore story generation: creative systems that can build coherent and fluent passages of text about a topic. We collect a large dataset of 300K human-written stories paired with writing prompts from an online forum. Our dataset enables hierarchical story generation, where the model first generates a premise, and then transforms it into a passage of text. We gain further improvements with a novel form of model fusion that improves the relevance of the story to the prompt, and adding a new gated multi-scale self-attention mechanism to model long-range context. Experiments show large improvements over strong baselines on both automated and human evaluations. Human judges prefer stories generated by our approach to those from a strong non-hierarchical model by a factor of two to one.

[TACL] Generating Sentences by Editing Prototypes
Kelvin Guu, Tatsunori Hashimoto, Yonatan Oren, and Percy Liang 15:15–15:40

We propose a new generative model of sentences that first samples a prototype sentence from the training corpus and then edits it into a new sentence. Compared to traditional models that generate from scratch either left-to-right or by first sampling a latent sentence vector, our prototype-then-edit model improves perplexity on language modeling and generates higher quality outputs according to human evaluation. Furthermore, the model gives rise to a latent edit vector that captures interpretable semantics such as sentence similarity and sentence-level analogies.

Session 2E: Vision

219 *Chair: Richard Socher*

No Metrics Are Perfect: Adversarial Reward Learning for Visual Storytelling
Xin Wang, Wenhu Chen, Yuan-Fang Wang, and William Yang Wang 14:00–14:25

Though impressive results have been achieved in visual captioning, the task of generating abstract stories from photo streams is still a little-tapped problem. Different from captions, stories have more expressive language styles and contain many imaginary concepts that do not appear in the images. Thus it poses challenges to behavioral cloning algorithms. Furthermore, due to the limitations of automatic metrics on evaluating story quality, reinforcement learning methods with hand-crafted rewards also face difficulties in gaining an overall performance boost. Therefore, we propose an Adversarial REward Learning (AREL) framework to learn an implicit reward function from human demonstrations, and then optimize policy search with the learned reward function. Though automatic evaluation indicates slight performance boost over state-of-the-art (SOTA) methods in cloning expert behaviors, human evaluation shows that our approach achieves significant improvement in generating more human-like stories than SOTA systems.

Bridging Languages through Images with Deep Partial Canonical Correlation Analysis
Guy Rotman, Ivan Vulić, and Roi Reichart 14:25–14:50

We present a deep neural network that leverages images to improve bilingual text embeddings. Relying on bilingual image tags and descriptions, our approach conditions text embedding induction on the shared visual information for both languages, producing highly correlated bilingual embeddings. In particular, we propose a novel model based on Partial Canonical Correlation Analysis (PCCA). While the original PCCA finds linear projections of two views in order to maximize their canonical correlation conditioned on a shared third variable, we introduce a non-linear Deep PCCA (DPCCA) model, and develop a new stochastic iterative algorithm for its optimization. We evaluate PCCA and DPCCA on multilingual word similarity and cross-lingual image description retrieval. Our models outperform a large variety of previous methods, despite not having access to any visual signal during test time inference.

Illustrative Language Understanding: Large-Scale Visual Grounding with Image Search
Jamie Kiros, William Chan, and Geoffrey Hinton 14:50–15:15

We introduce Picturebook, a large-scale lookup operation to ground language via 'snapshots' of our physical world accessed through image search. For each word in a vocabulary, we extract the top-k images from Google image search and feed the images through a convolutional network to extract a word embedding. We introduce a multimodal gating function to fuse our Picturebook embeddings with other word representations. We also introduce Inverse Picturebook, a mechanism to map a Picturebook embedding back into words. We experiment and report results across a wide range of tasks: word similarity, natural language inference, semantic relatedness, sentiment/topic classification, image-sentence ranking and machine translation. We also show that gate activations corresponding to Picturebook embeddings are highly correlated to human judgments of concreteness ratings.

What Action Causes This? Towards Naive Physical Action-Effect Prediction
Qiaozi Gao, Shaohua Yang, Joyce Chai, and Lucy Vanderwende 15:15–15:40

Despite recent advances in knowledge representation, automated reasoning, and machine learning, artificial agents still lack the ability to understand basic action-effect relations regarding the physical world, for example, the action of cutting a cucumber most likely leads to the state where the cucumber is broken apart into smaller pieces. If artificial agents (e.g., robots) ever become our partners in joint tasks, it is critical to empower them with such action-effect understanding so that they can reason about the state of the world and plan for actions. Towards this goal, this paper introduces a new task on naive physical action-effect prediction, which addresses the relations between concrete actions (expressed in the form of verb-noun pairs) and their effects on the state of the physical world as depicted by images. We collected a dataset for this task and developed an approach that harnesses web image data through distant supervision to facilitate learning for action-effect prediction. Our empirical results have shown that web data can be used to complement a small number of seed examples (e.g., three examples for each action) for model learning. This opens up possibilities for agents to learn physical action-effect relations for tasks at hand through communication with humans with a few examples.

Session 2F: Sentiment

220 *Chair: Yue Zhang*

Transformation Networks for Target-Oriented Sentiment Classification
Xin Li, Lidong Bing, Wai Lam, and Bei Shi 14:00–14:25

Target-oriented sentiment classification aims at classifying sentiment polarities over individual opinion targets in a sentence. RNN with attention seems a good fit for the characteristics of this task, and indeed it achieves the state-of-the-art performance. After re-examining the drawbacks of attention mechanism and the obstacles that block CNN to perform well in this classification task, we propose a new model that achieves new state-of-the-art results on a few benchmarks. Instead of attention, our model employs a CNN layer to extract salient features from the transformed word representations originated from a bi-directional RNN layer. Between the two layers, we propose a component which first generates target-specific representations of words in the sentence, and then incorporates a mechanism for preserving the original contextual information from the RNN layer.

Target-Sensitive Memory Networks for Aspect Sentiment Classification
Shuai Wang, Sahisnu Mazumder, Bing Liu, Mianwei Zhou, and Yi Chang 14:25–14:50

Aspect sentiment classification (ASC) is a fundamental task in sentiment analysis. Given an aspect/target and a sentence, the task classifies the sentiment polarity expressed on the target in the sentence. Memory networks (MNs) have been used for this task recently and have achieved state-of-the-art results. In MNs, attention mechanism plays a crucial role in detecting the sentiment context for the given target. However, we found an important problem with the current MNs in performing the ASC task. Simply improving the attention mechanism will not solve it. The problem is referred to as target-sensitive sentiment, which means that the sentiment polarity of the (detected) context is dependent on the given target and it cannot be inferred from the context alone. To tackle this problem, we propose the target-sensitive memory networks (TMNs). Several alternative techniques are designed for the implementation of TMNs and their effectiveness is experimentally evaluated.

Identifying Transferable Information Across Domains for Cross-domain Sentiment Classification
Raksha Sharma, Pushpak Bhattacharyya, Sandipan Dandapat, and Himanshu Sharad Bhatt 14:50–15:15

Getting manually labeled data in each domain is always an expensive and a time consuming task. Cross-domain sentiment analysis has emerged as a demanding concept where a labeled source domain facilitates a sentiment classifier for an unlabeled target domain. However, polarity orientation (positive or negative) and the significance of a word to express an opinion often differ from one domain to another domain. Owing to these differences, cross-domain sentiment classification is still a challenging task. In this paper, we propose that words that do not change their polarity and significance represent the transferable (usable) information across domains for cross-domain sentiment classification. We present a novel approach based on χ^2 test and cosine-similarity between context vector of words to identify polarity preserving significant words across domains. Furthermore, we show that a weighted ensemble of the classifiers enhances the cross-domain classification performance.

Unpaired Sentiment-to-Sentiment Translation: A Cycled Reinforcement Learning Approach
Jingjing Xu, Xu Sun, Qi Zeng, Xiaodong Zhang, Xuancheng Ren, Houfeng Wang, and Wenjie Li 15:15–15:40

The goal of sentiment-to-sentiment "translation" is to change the underlying sentiment of a sentence while keeping its content. The main challenge is the lack of parallel data. To solve this problem, we propose a cycled reinforcement learning method that enables training on unpaired data by collaboration between a neutralization module and an emotionalization module. We evaluate our approach on two review datasets, Yelp and Amazon. Experimental results show that our approach significantly outperforms the state-of-the-art systems. Especially, the proposed method substantially improves the content preservation performance. The BLEU score is improved from 1.64 to 22.46 and from 0.56 to 14.06 on the two datasets, respectively.

Session 3 Overview – Monday, July 16, 2018

	Track A *Inference, Reasoning*	**Track B** *Machine Learning 1*	**Track C** *Text Mining and Applications*	**Track D** *Dialog System 1*	**Track E** *Linguistics, Psycholinguistics and Cognitive Modeling*	**Track F** *Parsing 1*
	Plenary	203–204	210–211	212–213	219	220
16:10	Discourse Marker Augmented Network with Reinforcement Learning for Natural Language Inference *Pan, Yang, Zhao, Zhuang, Cai, and He*	Adversarial Contrastive Estimation *Bose, Ling, and Cao*	A Neural Architecture for Automated ICD Coding *Xie and Xing*	Unsupervised Discrete Sentence Representation Learning for Interpretable Neural Dialog Generation *Zhao, Lee, and Eskenazi*	Taylor's law for Human Linguistic Sequences *Kobayashi and Tanaka-Ishii*	Prefix Lexicalization of Synchronous CFGs using Synchronous TAG *Born and Sarkar*
16:35	Working Memory Networks: Augmenting Memory Networks with a Relational Reasoning Module *Pavez, Alllende, and Allende-Cid*	Adaptive Scaling for Sparse Detection in Information Extraction *Lin, Lu, Han, and Sun*	Domain Adaptation with Adversarial Training and Graph Embeddings *Alam, Joty, and Imran*	Learning to Control the Specificity in Neural Response Generation *Zhang, Guo, Fan, Lan, Xu, and Cheng*	A Framework for Representing Language Acquisition in a Population Setting *Kodner and Cerezo Falco*	Straight to the Tree: Constituency Parsing with Neural Syntactic Distance *Shen, Lin, Jacob, Sordoni, Courville, and Bengio*
17:00	Reasoning with Sarcasm by Reading In-Between *Tay, Luu, Hui, and Su*	Strong Baselines for Neural Semi-Supervised Learning under Domain Shift *Ruder and Plank*	TDNN: A Two-stage Deep Neural Network for Prompt-independent Automated Essay Scoring *Jin, He, Hui, and Sun*	Multi-Turn Response Selection for Chatbots with Deep Attention Matching Network *Zhou, Li, Dong, Liu, Chen, Zhao, Yu, and Wu*	[TACL] On the Complexity and Typology of Inflectional Morphological Systems *Cotterell, Kirov, Hulden, and Eisner*	Gaussian Mixture Latent Vector Grammars *Zhao, Zhang, and Tu*
17:25	[TACL] Representation Learning for Grounded Spatial Reasoning *Janner, Narasimhan, and Barzilay*	Fluency Boost Learning and Inference for Neural Grammatical Error Correction *Ge, Wei, and Zhou*	[TACL] Measuring the Evolution of a Scientific Field through Citation Frames *Jurgens, Kumar, Hoover, McFarland, and Jurafsky*	MojiTalk: Generating Emotional Responses at Scale *Zhou and Wang*	[TACL] Native Language Cognate Effects on Second Language Lexical Choice *Rabinovich, Tsvetkov, and Wintner*	Extending a Parser to Distant Domains Using a Few Dozen Partially Annotated Examples *Joshi, Peters, and Hopkins*

Parallel Session 3

Session 3A: Inference, Reasoning

Plenary *Chair: Xiaodan Zhu*

Discourse Marker Augmented Network with Reinforcement Learning for Natural Language Inference
Boyuan Pan, Yazheng Yang, Zhou Zhao, Yueting Zhuang, Deng Cai, and Xiaofei He 16:10–16:35

Natural Language Inference (NLI), also known as Recognizing Textual Entailment (RTE), is one of the most important problems in natural language processing. It requires to infer the logical relationship between two given sentences. While current approaches mostly focus on the interaction architectures of the sentences, in this paper, we propose to transfer knowledge from some important discourse markers to augment the quality of the NLI model. We observe that people usually use some discourse markers such as "so" or "but" to represent the logical relationship between two sentences. These words potentially have deep connections with the meanings of the sentences, thus can be utilized to help improve the representations of them. Moreover, we use reinforcement learning to optimize a new objective function with a reward defined by the property of the NLI datasets to make full use of the labels information. Experiments show that our method achieves the state-of-the-art performance on several large-scale datasets.

Working Memory Networks: Augmenting Memory Networks with a Relational Reasoning Module
Juan Pavez, Hector Alllende, and Hector Allende-Cid 16:35–17:00

During the last years, there has been a lot of interest in achieving some kind of complex reasoning using deep neural networks. To do that, models like Memory Networks (MemNNs) have combined external memory storages and attention mechanisms. These architectures, however, lack of more complex reasoning mechanisms that could allow, for instance, relational reasoning. Relation Networks (RNs), on the other hand, have shown outstanding results in relational reasoning tasks. Unfortunately, their computational cost grows quadratically with the number of memories, something prohibitive for larger problems. To solve these issues, we introduce the Working Memory Network, a MemNN architecture with a novel working memory storage and reasoning module. Our model retains the relational reasoning abilities of the RN while reducing its computational complexity from quadratic to linear. We tested our model on the text QA dataset bAbI and the visual QA dataset NLVR. In the jointly trained bAbI-10k, we set a new state-of-the-art, achieving a mean error of less than 0.5%. Moreover, a simple ensemble of two of our models solves all 20 tasks in the joint version of the benchmark.

Reasoning with Sarcasm by Reading In-Between
Yi Tay, Anh Tuan Luu, Siu Cheung Hui, and Jian Su 17:00–17:25

Sarcasm is a sophisticated speech act which commonly manifests on social communities such as Twitter and Reddit. The prevalence of sarcasm on the social web is highly disruptive to opinion mining systems due to not only its tendency of polarity flipping but also usage of figurative language. Sarcasm commonly manifests with a contrastive theme either between positive-negative sentiments or between literal-figurative scenarios. In this paper, we revisit the notion of modeling contrast in order to reason with sarcasm. More specifically, we propose an attention-based neural model that looks in-between instead of across, enabling it to explicitly model contrast and incongruity. We conduct extensive experiments on six benchmark datasets from Twitter, Reddit and the Internet Argument Corpus. Our proposed model not only achieves state-of-the-art performance on all datasets but also enjoys improved interpretability.

[TACL] Representation Learning for Grounded Spatial Reasoning
Michael Janner, Karthik Narasimhan, and Regina Barzilay 17:25–17:50

The interpretation of spatial references is highly contextual, requiring joint inference over both language and the environment. We consider the task of spatial reasoning in a simulated environment, where an agent can act and receive rewards. The proposed model learns a representation of the world steered by instruction text. This design allows for precise alignment of local neighborhoods with corresponding verbalizations, while also handling global references in the instructions. We train our model with reinforcement learning using a variant of generalized value iteration. The model outperforms state-of-the-art approaches on several metrics, yielding a 45% reduction in goal localization error.

Session 3B: Machine Learning 1

203–204 *Chair: Sujith Ravi*

Adversarial Contrastive Estimation
Avishek Joey Bose, Huan Ling, and Yanshuai Cao 16:10–16:35

Learning by contrasting positive and negative samples is a general strategy adopted by many methods. Noise contrastive estimation (NCE) for word embeddings and translating embeddings for knowledge graphs are examples in NLP employing this approach. In this work, we view contrastive learning as an abstraction of all such methods and augment the negative sampler into a mixture distribution containing an adversarially learned sampler. The resulting adaptive sampler finds harder negative examples, which forces the main model to learn a better representation of the data. We evaluate our proposal on learning word embeddings, order embeddings and knowledge graph embeddings and observe both faster convergence and improved results on multiple metrics.

Adaptive Scaling for Sparse Detection in Information Extraction
Hongyu Lin, Yaojie Lu, Xianpei Han, and Le Sun 16:35–17:00

This paper focuses on detection tasks in information extraction, where positive instances are sparsely distributed and models are usually evaluated using F-measure on positive classes. These characteristics often result in deficient performance of neural network based detection models. In this paper, we propose *adaptive scaling*, an algorithm which can handle the positive sparsity problem and directly optimize over F-measure via dynamic cost-sensitive learning. To this end, we borrow the idea of marginal utility from economics and propose a theoretical framework for instance importance measuring without introducing any additional hyper-parameters. Experiments show that our algorithm leads to a more effective and stable training of neural network based detection models.

Strong Baselines for Neural Semi-Supervised Learning under Domain Shift
Sebastian Ruder and Barbara Plank 17:00–17:25

Novel neural models have been proposed in recent years for learning under domain shift. Most models, however, only evaluate on a single task, on proprietary datasets, or compare to weak baselines, which makes comparison of models difficult. In this paper, we re-evaluate classic general-purpose bootstrapping approaches in the context of neural networks under domain shifts vs. recent neural approaches and propose a novel multi-task tri-training method that reduces the time and space complexity of classic tri-training. Extensive experiments on two benchmarks for part-of-speech tagging and sentiment analysis are negative: while our novel method establishes a new state-of-the-art for sentiment analysis, it does not fare consistently the best. More importantly, we arrive at the somewhat surprising conclusion that classic tri-training, with some additions, outperforms the state-of-the-art for NLP. Hence classic approaches constitute an important and strong baseline.

Fluency Boost Learning and Inference for Neural Grammatical Error Correction
Tao Ge, Furu Wei, and Ming Zhou 17:25–17:50

Most of the neural sequence-to-sequence (seq2seq) models for grammatical error correction (GEC) have two limitations: (1) a seq2seq model may not be well generalized with only limited error-corrected data; (2) a seq2seq model may fail to completely correct a sentence with multiple errors through normal seq2seq inference. We attempt to address these limitations by proposing a fluency boost learning and inference mechanism. Fluency boosting learning generates fluency-boost sentence pairs during training, enabling the error correction model to learn how to improve a sentence's fluency from more instances, while fluency boosting inference allows the model to correct a sentence incrementally with multiple inference steps until the sentence's fluency stops increasing. Experiments show our approaches improve the performance of seq2seq models for GEC, achieving state-of-the-art results on both CoNLL-2014 and JFLEG benchmark datasets.

Session 3C: Text Mining and Applications

210–211 *Chair: Sadao Kurohashi*

A Neural Architecture for Automated ICD Coding
Pengtao Xie and Eric Xing 16:10–16:35

The International Classification of Diseases (ICD) provides a hierarchy of diagnostic codes for classifying diseases. Medical coding – which assigns a subset of ICD codes to a patient visit – is a mandatory process that is crucial for patient care and billing. Manual coding is time-consuming, expensive, and error prone. In this paper, we build a neural architecture for automated coding. It takes the diagnosis descriptions (DDs) of a patient as inputs and selects the most relevant ICD codes. This architecture contains four major ingredients: (1) tree-of-sequences LSTM encoding of code descriptions (CDs), (2) adversarial learning for reconciling the different writing styles of DDs and CDs, (3) isotonic constraints for incorporating the importance order among the assigned codes, and (4) attentional matching for performing many-to-one and one-to-many mappings from DDs to CDs. We demonstrate the effectiveness of the proposed methods on a clinical datasets with 59K patient visits.

Domain Adaptation with Adversarial Training and Graph Embeddings
Firoj Alam, Shafiq Joty, and Muhammad Imran 16:35–17:00

The success of deep neural networks (DNNs) is heavily dependent on the availability of labeled data. However, obtaining labeled data is a big challenge in many real-world problems. In such scenarios, a DNN model can leverage labeled and unlabeled data from a related domain, but it has to deal with the shift in data distributions between the source and the target domains. In this paper, we study the problem of classifying social media posts during a crisis event (e.g., Earthquake). For that, we use labeled and unlabeled data from past similar events (e.g., Flood) and unlabeled data for the current event. We propose a novel model that performs adversarial learning based domain adaptation to deal with distribution drifts and graph based semi-supervised learning to leverage unlabeled data within a single unified deep learning framework. Our experiments with two real-world crisis datasets collected from Twitter demonstrate significant improvements over several baselines.

TDNN: A Two-stage Deep Neural Network for Prompt-independent Automated Essay Scoring
Cancan Jin, Ben He, Kai Hui, and Le Sun 17:00–17:25

Existing automated essay scoring (AES) models rely on rated essays for the target prompt as training data. Despite their successes in prompt-dependent AES, how to effectively predict essay ratings under a prompt-independent setting remains a challenge, where the rated essays for the target prompt are not available. To close this gap, a two-stage deep neural network (TDNN) is proposed. In particular, in the first stage, using the rated essays for non-target prompts as the training data, a shallow model is learned to select essays with an extreme quality for the target prompt, serving as pseudo training data; in the second stage, an end-to-end hybrid deep model is proposed to learn a prompt-dependent rating model consuming the pseudo training data from the first step. Evaluation of the proposed TDNN on the standard ASAP dataset demonstrates a promising improvement for the prompt-independent AES task.

[TACL] Measuring the Evolution of a Scientific Field through Citation Frames
David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky 17:25–17:50

Citations have long been used to characterize the state of a scientific field and to identify influential works. However, writers use citations for different purposes, and this varied purpose influences uptake by future scholars. Unfortunately, our understanding of how scholars use and frame citations has been limited to small-scale manual citation analysis of individual papers. We perform the largest behavioral study of citations to date, analyzing how scientific works frame their contributions through different types of citations and how this framing affects the field as a whole. We introduce a new dataset of nearly 2,000 citations annotated for their function, and use it to develop a state-of-the-art classifier and label the papers of an entire field: Natural Language Processing. We then show how differences in framing affect scientific uptake and reveal the evolution of the publication venues and the field as a whole. We demonstrate that authors are sensitive to discourse structure and publication venue when citing, and that how a paper frames its work through citations is predictive of the citation count it will receive. Finally, we use changes in citation framing to show that the field of NLP is undergoing a significant increase in consensus.

Session 3D: Dialog System 1

212–213 *Chair: Zornitsa Kozareva*

Unsupervised Discrete Sentence Representation Learning for Interpretable Neural Dialog Generation
Tiancheng Zhao, Kyusong Lee, and Maxine Eskenazi 16:10–16:35

The encoder-decoder dialog model is one of the most prominent methods used to build dialog systems in complex domains. Yet it is limited because it cannot output interpretable actions as in traditional systems, which hinders humans from understanding its generation process. We present an unsupervised discrete sentence representation learning method that can integrate with any existing encoder-decoder dialog models for interpretable response generation. Building upon variational autoencoders (VAEs), we present two novel models, DI-VAE and DI-VST that improve VAEs and can discover interpretable semantics via either auto encoding or context predicting. Our methods have been validated on real-world dialog datasets to discover semantic representations and enhance encoder-decoder models with interpretable generation.

Learning to Control the Specificity in Neural Response Generation
Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, Jun Xu, and Xueqi Cheng 16:35–17:00

In conversation, a general response (e.g., "I don't know") could correspond to a large variety of input utterances. Previous generative conversational models usually employ a single model to learn the relationship between different utterance-response pairs, thus tend to favor general and trivial responses which appear frequently. To address this problem, we propose a novel controlled response generation mechanism to handle different utterance-response relationships in terms of specificity. Specifically, we introduce an explicit specificity control variable into a sequence-to-sequence model, which interacts with the usage representation of words through a Gaussian Kernel layer, to guide the model to generate responses at different specificity levels. We describe two ways to acquire distant labels for the specificity control variable in learning. Empirical studies show that our model can significantly outperform the state-of-the-art response generation models under both automatic and human evaluations.

Multi-Turn Response Selection for Chatbots with Deep Attention Matching Network
Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu
17:00–17:25

Human generates responses relying on semantic and functional dependencies, including coreference relation, among dialogue elements and their context. In this paper, we investigate matching a response with its multi-turn context using dependency information based entirely on attention. Our solution is inspired by the recently proposed Transformer in machine translation (Vaswani et al., 2017) and we extend the attention mechanism in two ways. First, we construct representations of text segments at different granularities solely with stacked self-attention. Second, we try to extract the truly matched segment pairs with attention across the context and response. We jointly introduce those two kinds of attention in one uniform neural network. Experiments on two large-scale multi-turn response selection tasks show that our proposed model significantly outperforms the state-of-the-art models.

MojiTalk: Generating Emotional Responses at Scale
Xianda Zhou and William Yang Wang 17:25–17:50

Generating emotional language is a key step towards building empathetic natural language processing agents. However, a major challenge for this line of research is the lack of large-scale labeled training data, and previous studies are limited to only small sets of human annotated sentiment labels. Additionally, explicitly controlling the emotion and sentiment of generated text is also difficult. In this paper, we take a more radical approach: we exploit the idea of leveraging Twitter data that are naturally labeled with emojis. We collect a large corpus of Twitter conversations that include emojis in the response and assume the emojis convey the underlying emotions of the sentence. We investigate several conditional variational autoencoders training on these conversations, which allow us to use emojis to control the emotion of the generated text. Experimentally, we show in our quantitative and qualitative analyses that the proposed models can successfully generate high-quality abstractive conversation responses in accordance with designated emotions.

Session 3E: Linguistics, Psycholinguistics and Cognitive Modeling

219 *Chair: Kevin Cohen*

Taylor's law for Human Linguistic Sequences
Tatsuru Kobayashi and Kumiko Tanaka-Ishii 16:10–16:35

Taylor's law describes the fluctuation characteristics underlying a system in which the variance of an event within a time span grows by a power law with respect to the mean. Although Taylor's law has been applied in many natural and social systems, its application for language has been scarce. This article describes a new way to quantify Taylor's law in natural language and conducts Taylor analysis of over 1100 texts across 14 languages. We found that the Taylor exponents of natural language written texts exhibit almost the same value. The exponent was also compared for other language-related data, such as the child-directed speech, music, and programming languages. The results show how the Taylor exponent serves to quantify the fundamental structural complexity underlying linguistic time series. The article also shows the applicability of these findings in evaluating language models.

A Framework for Representing Language Acquisition in a Population Setting
Jordan Kodner and Christopher Cerezo Falco 16:35–17:00

Language variation and change are driven both by individuals' internal cognitive processes and by the social structures through which language propagates. A wide range of computational frameworks have been proposed to connect these drivers. We compare the strengths and weaknesses of existing approaches and propose a new analytic framework which combines previous network models' ability to capture realistic social structure with practically and more elegant computational properties. The framework privileges the process of language acquisition and embeds learners in a social network but is modular so that population structure can be combined with different acquisition models. We demonstrate two applications for the framework: a test of practical concerns that arise when modeling acquisition in a population setting and an application of the framework to recent work on phonological mergers in progress.

[TACL] On the Complexity and Typology of Inflectional Morphological Systems
Ryan Cotterell, Christo Kirov, Mans Hulden, and Jason Eisner 17:00–17:25

We quantify the linguistic complexity of different languages' morphological systems. We verify that there is an empirical trade-off between paradigm size and irregularity: a language's inflectional paradigms may be either large in size or highly irregular, but never both. Our methodology quantifies paradigm irregularity as the entropy of the surface realization of a paradigm—how hard it is to jointly predict all the surface forms of a paradigm—which we estimate by a variational approximation. Our measurements are taken on large morphological paradigms from 31 typologically diverse languages.

[TACL] Native Language Cognate Effects on Second Language Lexical Choice
Ella Rabinovich, Yulia Tsvetkov, and Shuly Wintner 17:25–17:50

We present a computational analysis of cognate effects on the spontaneous linguistic productions of advanced non-native speakers. Introducing a large corpus of highly competent non-native English speakers, and using a set of carefully selected lexical items, we show that the lexical choices of non-natives are affected by cognates in their native language. This effect is so powerful that we are able to reconstruct the phylogenetic language tree of the Indo-European language family solely from the frequencies of specific lexical items in the English of authors with various native languages. We quantitatively analyze non-native lexical choice, highlighting cognate facilitation as one of the important phenomena shaping the language of non-native speakers.

Session 3F: Parsing 1

220 *Chair: Liang Huang*

Prefix Lexicalization of Synchronous CFGs using Synchronous TAG
Logan Born and Anoop Sarkar 16:10–16:35

We show that an epsilon-free, chain-free synchronous context-free grammar (SCFG) can be converted into a weakly equivalent synchronous tree-adjoining grammar (STAG) which is prefix lexicalized. This transformation at most doubles the grammar's rank and cubes its size, but we show that in practice the size increase is only quadratic. Our results extend Greibach normal form from CFGs to SCFGs and prove new formal properties about SCFG, a formalism with many applications in natural language processing.

Straight to the Tree: Constituency Parsing with Neural Syntactic Distance
Yikang Shen, Zhouhan Lin, Athul Paul Jacob, Alessandro Sordoni, Aaron Courville, and Yoshua Bengio
16:35–17:00

In this work, we propose a novel constituency parsing scheme. The model first predicts a real-valued scalar, named syntactic distance, for each split position in the sentence. The topology of grammar tree is then determined by the values of syntactic distances. Compared to traditional shift-reduce parsing schemes, our approach is free from the potentially disastrous compounding error. It is also easier to parallelize and much faster. Our model achieves the state-of-the-art single model F1 score of 92.1 on PTB and 86.4 on CTB dataset, which surpasses the previous single model results by a large margin.

Gaussian Mixture Latent Vector Grammars
Yanpeng Zhao, Liwen Zhang, and Kewei Tu 17:00–17:25

We introduce Latent Vector Grammars (LVeGs), a new framework that extends latent variable grammars such that each nonterminal symbol is associated with a continuous vector space representing the set of (infinitely many) subtypes of the nonterminal. We show that previous models such as latent variable grammars and compositional vector grammars can be interpreted as special cases of LVeGs. We then present Gaussian Mixture LVeGs (GM-LVeGs), a new special case of LVeGs that uses Gaussian mixtures to formulate the weights of production rules over subtypes of nonterminals. A major advantage of using Gaussian mixtures is that the partition function and the expectations of subtype rules can be computed using an extension of the inside-outside algorithm, which enables efficient inference and learning. We apply GM-LVeGs to part-of-speech tagging and constituency parsing and show that GM-LVeGs can achieve competitive accuracies.

Extending a Parser to Distant Domains Using a Few Dozen Partially Annotated Examples
Vidur Joshi, Matthew Peters, and Mark Hopkins 17:25–17:50

We revisit domain adaptation for parsers in the neural era. First we show that recent advances in word representations greatly diminish the need for domain adaptation when the target domain is syntactically similar to the source domain. As evidence, we train a parser on the Wall Street Journal alone that achieves over 90% F1 on the Brown corpus. For more syntactically distant domains, we provide a simple way to adapt a parser using only dozens of partial annotations. For instance, we increase the percentage of error-free geometry-domain parses in a held-out set from 45% to 73% using approximately five dozen training examples. In the process, we demonstrate a new state-of-the-art single model result on the Wall Street Journal test set of 94.3%. This is an absolute increase of 1.7% over the previous state-of-the-art of 92.6%.

Student Recruitment Event

Monday, July 16, 2018, 6:15pm–7:45pm

Showtime Events Centre and Common Man Lawn

Join your fellow students for a free student-only networking event, on Monday, July 16 from 6:15pm, straight after the end of the final session of Day 1 of the main conference. The event will be split across two adjoining venues (Showtime and Common Man), just outside the Melbourne Convention and Exhibition Centre on the Yarra River. See map on page 216. Soft drinks and canapés will be provided, and there will be a cash bar for those who want to imbibe. Tickets to the event have been allocated to student registrants on a first-come-first-served basis, and can be found inside your conference badge. This is a chance to get to know other students in a casual atmosphere, and also speak to recruiters about possible future graduate destinations.

4

Main Conference: Tuesday, July 17

Overview

08:00–17:00 **Registration** *Level 2 Foyer*

09:00–10:00 **Invited Talk 1: Carolyn Penstein Rosé** (sponsors: Samsung & Megagon Labs) *Plenary*

10:00–10:30 **Coffee Break** *Level 2 Foyer and Melbourne Room*

Session 4

10:30–12:10	Word Semantics 2	Machine Translation 2	Information Extraction 2	Dialog System 2	Evaluation	Parsing 2
	Plenary	*203–204*	*210–211*	*212–213*	*219*	*220*

12:10–12:30 **Short Break**

12:30–14:00 **Poster Session 2** *Melbourne Room 1 & 2*

12:30–14:00 **Demo Poster Session 2** *Melbourne Room 1 & 2*

Session 5

14:00–15:00	Semantics 1 (Short)	Machine Translation, Multilinguality 1 (Short)	Information Extraction 1 (Short)	Dialog System, Discourse (Short)	Vision, Linguistics, Resource and Evaluation (Short)	Parsing, Morphology (Short)
	Plenary	*203–204*	*210–211*	*212–213*	*219*	*220*

15:00–15:30 **Coffee Break** *Level 2 Foyer and Melbourne Room*

Session 6

15:30–17:10	Semantic Parsing 2	Machine Learning 2	Question Answering 2	Generation 2	Social Media	Information Retrieval
	Plenary	*203–204*	*210–211*	*212–213*	*219*	*220*

17:10–17:20 **Short Break**

17:20–18:50 **ACL Business Meeting** *Plenary*

19:30–22:30 **Social Event** *Melbourne Sealife Aquarium*

Keynote Address: Carolyn Penstein Rosé (Sponsored by Samsung & Megagon Labs)

Who is the Bridge Between the What and the How

Tuesday, July 17, 2018, 9:00am–10:00am

Plenary

Abstract: This talk reports on over a decade of research where theoretical foundations motivate computational models that produce real world impact in online spaces. Both the earliest philosophers of language and the most recent researchers in computational approaches to social media analysis have acknowledged the distinction between the what of language, namely its propositional content, and the how of language, or its form, style, or framing. What bridges between these realms are social processes that motivate the linguistic choices that result in specific realizations of propositional content situated within social interactions, designed to achieve social goals. These insights allow researchers to make sense of the connection between discussion processes and outcomes from those discussions. These findings motivate on the one hand design of computational approaches to real time monitoring of discussion processes and on the other hand the design of interventions that support interactions in online spaces with the goal of increasing desired outcomes, including learning, health, and wellbeing.

As an example, in this talk we probe into a specific quality of discussion referred to as Transactivity. Transactivity is the extent to which a contribution articulates the reasoning of the speaker, that of an interlocutor, and the relation between them. In different contexts, and within very distinct theoretical frameworks, this construct has been associated with solidarity, influence, expertise transfer, and learning. Within the construct of Transactivity, the cognitive and social underpinnings are inextricably linked such that modeling the who enables prediction of the connection between the what and the how.

Biography: Dr. Carolyn Rosé is a Professor of Language Technologies and Human-Computer Interaction in the School of Computer Science at Carnegie Mellon University. Her research program is focused on better understanding the social and pragmatic nature of conversation, and using this understanding to build computational systems that can improve the efficacy of conversation between people, and between people and computers. In order to pursue these goals, she invokes approaches from computational discourse analysis and text mining, conversational agents, and computer supported collaborative learning. Her research group's highly interdisciplinary work, published in 200 peer reviewed publications, is represented in the top venues in 5 fields: namely, Language Technologies, Learning Sciences, Cognitive Science, Educational Technology, and Human-Computer Interaction, with awards in 3 of these fields. She serves as Past President and Inaugural Fellow of the International Society of the Learning Sciences, Chair of the International Alliance to Advance Learning in the Digital Era, and Executive Editor of the International Journal of Computer-Supported Collaborative Learning.

Session 4 Overview – Tuesday, July 17, 2018

	Track A *Word Semantics 2*	**Track B** *Machine Translation 2*	**Track C** *Information Extraction 2*	**Track D** *Dialog System 2*	**Track E** *Evaluation*	**Track F** *Parsing 2*
	Plenary	203–204	210–211	212–213	219	220
10:30	Paraphrase to Explicate: Revealing Implicit Noun-Compound Relations *Shwartz and Dagan*	A Stochastic Decoder for Neural Machine Translation *Schulz, Aziz, and Cohn*	Which Melbourne? Augmenting Geocoding with Maps *Gritta, Pilehvar, and Collier*	Exemplar Encoder-Decoder for Neural Conversation Generation *Pandey, Contractor, Kumar, and Joshi*	Are BLEU and Meaning Representation in Opposition? *Cifka and Bojar*	Distilling Knowledge for Search-based Structured Prediction *Liu, Che, Zhao, Qin, and Liu*
10:55	Searching for the X-Factor: Exploring Corpus Subjectivity for Word Embeddings *Tkachenko, Chia, and Lauw*	Forest-Based Neural Machine Translation *Ma, Tamura, Utiyama, Zhao, and Sumita*	Learning Prototypical Goal Activities for Locations *Jiang and Riloff*	DialSQL: Dialogue Based Structured Query Generation *Gur, Yavuz, Su, and Yan*	Automatic Metric Validation for Grammatical Error Correction *Choshen and Abend*	Stack-Pointer Networks for Dependency Parsing *Ma, Hu, Liu, Peng, Neubig, and Hovy*
11:20	Word Embedding and WordNet Based Metaphor Identification and Interpretation *Mao, Lin, and Guerin*	Context-Aware Neural Machine Translation Learns Anaphora Resolution *Voita, Serdyukov, Sennrich, and Titov*	Guess Me if You Can: Acronym Disambiguation for Enterprises *Li, Zhao, Fuxman, and Tao*	Conversations Gone Awry: Detecting Early Signs of Conversational Failure *Zhang, Chang, Danescu-Niculescu-Mizil, Dixon, Hua, Taraborelli, and Thain*	The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing *Dror, Baumer, Shlomov, and Reichart*	Twitter Universal Dependency Parsing for African-American and Mainstream American English *Blodgett, Wei, and O'Connor*
11:45	Incorporating Latent Meanings of Morphological Compositions to Enhance Word Embeddings *Xu, Liu, Yang, and Huang*	Document Context Neural Machine Translation with Memory Networks *Maruf and Haffari*	A Multi-Axis Annotation Scheme for Event Temporal Relations *Ning, Wu, and Roth*	[TACL] Detecting Institutional Dialog Acts in Police Traffic Stops *Prabhakaran, Griffiths, Su, Verma, Morgan, Eberhardt, and Jurafsky*	[TACL] Replicability Analysis for Natural Language Processing: Testing Significance with Multiple Datasets *Dror, Baumer, Bogomolov, and Reichart*	LSTMs Can Learn Syntax-Sensitive Dependencies Well, But Modeling Structure Makes Them Better *Kuncoro, Dyer, Hale, Yogatama, Clark, and Blunsom*

Parallel Session 4

Session 4A: Word Semantics 2

Plenary *Chair: Valia Kordoni*

Paraphrase to Explicate: Revealing Implicit Noun-Compound Relations
Vered Shwartz and Ido Dagan 10:30–10:55

Revealing the implicit semantic relation between the constituents of a noun-compound is important for many NLP applications. It has been addressed in the literature either as a classification task to a set of pre-defined relations or by producing free text paraphrases explicating the relations. Most existing paraphrasing methods lack the ability to generalize, and have a hard time interpreting infrequent or new noun-compounds. We propose a neural model that generalizes better by representing paraphrases in a continuous space, generalizing for both unseen noun-compounds and rare paraphrases. Our model helps improving performance on both the noun-compound paraphrasing and classification tasks.

Searching for the X-Factor: Exploring Corpus Subjectivity for Word Embeddings
Maksim Tkachenko, Chong Cher Chia, and Hady Lauw 10:55–11:20

We explore the notion of subjectivity, and hypothesize that word embeddings learnt from input corpora of varying levels of subjectivity behave differently on natural language processing tasks such as classifying a sentence by sentiment, subjectivity, or topic. Through systematic comparative analyses, we establish this to be the case indeed. Moreover, based on the discovery of the outsized role that sentiment words play on subjectivity-sensitive tasks such as sentiment classification, we develop a novel word embedding SentiVec which is infused with sentiment information from a lexical resource, and is shown to outperform baselines on such tasks.

Word Embedding and WordNet Based Metaphor Identification and Interpretation
Rui Mao, Chenghua Lin, and Frank Guerin 11:20–11:45

Metaphoric expressions are widespread in natural language, posing a significant challenge for various natural language processing tasks such as Machine Translation. Current word embedding based metaphor identification models cannot identify the exact metaphorical words within a sentence. In this paper, we propose an unsupervised learning method that identifies and interprets metaphors at word-level without any preprocessing, outperforming strong baselines in the metaphor identification task. Our model extends to interpret the identified metaphors, paraphrasing them into their literal counterparts, so that they can be better translated by machines. We evaluated this with two popular translation systems for English to Chinese, showing that our model improved the systems significantly.

Incorporating Latent Meanings of Morphological Compositions to Enhance Word Embeddings
Yang Xu, Jiawei Liu, Wei Yang, and Liusheng Huang 11:45–12:10

Traditional word embedding approaches learn semantic information at word level while ignoring the meaningful internal structures of words like morphemes. Furthermore, existing morphology-based models directly incorporate morphemes to train word embeddings, but still neglect the latent meanings of morphemes. In this paper, we explore to employ the latent meanings of morphological compositions of words to train and enhance word embeddings. Based on this purpose, we propose three Latent Meaning Models (LMMs), named LMM-A, LMM-S and LMM-M respectively, which adopt different strategies to incorporate the latent meanings of morphemes during the training process. Experiments on word similarity, syntactic analogy and text classification are conducted to validate the feasibility of our models. The results demonstrate that our models outperform the baselines on five word similarity datasets. On Wordsim-353 and RG-65 datasets, our models nearly achieve 5% and 7% gains over the classic CBOW model, respectively. For the syntactic analogy and text classification tasks, our models also surpass all the baselines including a morphology-based model.

Session 4B: Machine Translation 2

203–204 *Chair: Stefan Riezler*

A Stochastic Decoder for Neural Machine Translation
Philip Schulz, Wilker Aziz, and Trevor Cohn 10:30–10:55

The process of translation is ambiguous, in that there are typically many valid translations for a given sentence. This gives rise to significant variation in parallel corpora, however, most current models of machine translation do not account for this variation, instead treating the problem as a deterministic process. To this end, we present a deep generative model of machine translation which incorporates a chain of latent variables, in order to account for local lexical and syntactic variation in parallel corpora. We provide an in-depth analysis of the pitfalls encountered in variational inference for training deep generative models. Experiments on several different language pairs demonstrate that the model consistently improves over strong baselines.

Forest-Based Neural Machine Translation
Chunpeng Ma, Akihiro Tamura, Masao Utiyama, Tiejun Zhao, and Eiichiro Sumita 10:55–11:20

Tree-based neural machine translation (NMT) approaches, although achieved impressive performance, suffer from a major drawback: they only use the 1-best parse tree to direct the translation, which potentially introduces translation mistakes due to parsing errors. For statistical machine translation (SMT), forest-based methods have been proven to be effective for solving this problem, while for NMT this kind of approach has not been attempted. This paper proposes a forest-based NMT method that translates a linearized packed forest under a simple sequence-to-sequence framework (i.e., a forest-to-sequence NMT model). The BLEU score of the proposed method is higher than that of the sequence-to-sequence NMT, tree-based NMT, and forest-based SMT systems.

Context-Aware Neural Machine Translation Learns Anaphora Resolution
Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov 11:20–11:45

Standard machine translation systems process sentences in isolation and hence ignore extra-sentential information, even though extended context can both prevent mistakes in ambiguous cases and improve translation coherence. We introduce a context-aware neural machine translation model designed in such way that the flow of information from the extended context to the translation model can be controlled and analyzed. We experiment with an English-Russian subtitles dataset, and observe that much of what is captured by our model deals with improving pronoun translation. We measure correspondences between induced attention distributions and coreference relations and observe that the model implicitly captures anaphora. It is consistent with gains for sentences where pronouns need to be gendered in translation. Beside improvements in anaphoric cases, the model also improves in overall BLEU, both over its context-agnostic version (+0.7) and over simple concatenation of the context and source sentences (+0.6).

Document Context Neural Machine Translation with Memory Networks
Sameen Maruf and Gholamreza Haffari 11:45–12:10

We present a document-level neural machine translation model which takes both source and target document context into account using memory networks. We model the problem as a structured prediction problem with interdependencies among the observed and hidden variables, i.e., the source sentences and their unobserved target translations in the document. The resulting structured prediction problem is tackled with a neural translation model equipped with two memory components, one each for the source and target side, to capture the documental interdependencies. We train the model end-to-end, and propose an iterative decoding algorithm based on block coordinate descent. Experimental results of English translations from French, German, and Estonian documents show that our model is effective in exploiting both source and target document context, and statistically significantly outperforms the previous work in terms of BLEU and METEOR.

Session 4C: Information Extraction 2

210–211 *Chair: Kai-Wei Chang*

Which Melbourne? Augmenting Geocoding with Maps
Milan Gritta, Mohammad Taher Pilehvar, and Nigel Collier 10:30–10:55

The purpose of text geolocation is to associate geographic information contained in a document with a set (or sets) of coordinates, either implicitly by using linguistic features and/or explicitly by using geographic metadata combined with heuristics. We introduce a geocoder (location mention disambiguator) that achieves state-of-the-art (SOTA) results on three diverse datasets by exploiting the implicit lexical clues. Moreover, we propose a new method for systematic encoding of geographic metadata to generate two distinct views of the same text. To that end, we introduce the Map Vector (MapVec), a sparse representation obtained by plotting prior geographic probabilities, derived from population figures, on a World Map. We then integrate the implicit (language) and explicit (map) features to significantly improve a range of metrics. We also introduce an open-source dataset for geoparsing of news events covering global disease outbreaks and epidemics to help future evaluation in geoparsing.

Learning Prototypical Goal Activities for Locations
Tianyu Jiang and Ellen Riloff 10:55–11:20

People go to different places to engage in activities that reflect their goals. For example, people go to restaurants to eat, libraries to study, and churches to pray. We refer to an activity that represents a common reason why people typically go to a location as a prototypical goal activity (goal-act). Our research aims to learn goal-acts for specific locations using a text corpus and semi-supervised learning. First, we extract activities and locations that co-occur in goal-oriented syntactic patterns. Next, we create an activity profile matrix and apply a semi-supervised label propagation algorithm to iteratively revise the activity strengths for different locations using a small set of labeled data. We show that this approach outperforms several baseline methods when judged against goal-acts identified by human annotators.

Guess Me if You Can: Acronym Disambiguation for Enterprises
Yang Li, Bo Zhao, Ariel Fuxman, and Fangbo Tao 11:20–11:45

Acronyms are abbreviations formed from the initial components of words or phrases. In enterprises, people often use acronyms to make communications more efficient. However, acronyms could be difficult to understand for people who are not familiar with the subject matter (new employees, etc.), thereby affecting productivity. To alleviate such troubles, we study how to automatically resolve the true meanings of acronyms in a given context. Acronym disambiguation for enterprises is challenging for several reasons. First, acronyms may be highly ambiguous since an acronym used in the enterprise could have multiple internal and external meanings. Second, there are usually no comprehensive knowledge bases such as Wikipedia available in enterprises. Finally, the system should be generic to work for any enterprise. In this work we propose an end-to-end framework to tackle all these challenges. The framework takes the enterprise corpus as input and produces a high-quality acronym disambiguation system as output. Our disambiguation models are trained via distant supervised learning, without requiring any manually labeled training examples. Therefore, our proposed framework can be deployed to any enterprise to support high-quality acronym disambiguation. Experimental results on real world data justified the effectiveness of our system.

A Multi-Axis Annotation Scheme for Event Temporal Relations
Qiang Ning, Hao Wu, and Dan Roth 11:45–12:10

Existing temporal relation (TempRel) annotation schemes often have low inter-annotator agreements (IAA) even between experts, suggesting that the current annotation task needs a better definition. This paper proposes a new multi-axis modeling to better capture the temporal structure of events. In addition, we identify that event end-points are a major source of confusion in annotation, so we also propose to annotate TempRels based on start-points only. A pilot expert annotation effort using the proposed scheme shows significant improvement in IAA from the conventional 60's to 80's (Cohen's Kappa). This better-defined annotation scheme further enables the use of crowdsourcing to alleviate the labor intensity for each annotator. We hope that this work can foster more interesting studies towards event understanding.

Session 4D: Dialog System 2

212–213 *Chair: Zornitsa Kozareva*

Exemplar Encoder-Decoder for Neural Conversation Generation
Gaurav Pandey, Danish Contractor, Vineet Kumar, and Sachindra Joshi 10:30–10:55

In this paper we present the Exemplar Encoder-Decoder network (EED), a novel conversation model that learns to utilize *similar* examples from training data to generate responses. Similar conversation examples (context-response pairs) from training data are retrieved using a traditional TF-IDF based retrieval model and the corresponding responses are used by our decoder to generate the ground truth response. The contribution of each retrieved response is weighed by the similarity of corresponding context with the input context. As a result, our model learns to assign higher similarity scores to those retrieved contexts whose responses are crucial for generating the final response. We present detailed experiments on two large data sets and we find that our method out-performs state of the art sequence to sequence generative models on several recently proposed evaluation metrics.

DialSQL: Dialogue Based Structured Query Generation
Izzeddin Gur, Semih Yavuz, Yu Su, and Xifeng Yan 10:55–11:20

The recent advance in deep learning and semantic parsing has significantly improved the translation accuracy of natural language questions to structured queries. However, further improvement of the existing approaches turns out to be quite challenging. Rather than solely relying on algorithmic innovations, in this work, we introduce DialSQL, a dialogue-based structured query generation framework that leverages human intelligence to boost the performance of existing algorithms via user interaction. DialSQL is capable of identifying potential errors in a generated SQL query and asking users for validation via simple multi-choice questions. User feedback is then leveraged to revise the query. We design a generic simulator to bootstrap synthetic training dialogues and evaluate the performance of DialSQL on the WikiSQL dataset. Using SQLNet as a black box query generation tool, DialSQL improves its performance from 61.3% to 69.0% using only 2.4 validation questions per dialogue.

Conversations Gone Awry: Detecting Early Signs of Conversational Failure
Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain 11:20–11:45

One of the main challenges online social systems face is the prevalence of antisocial behavior, such as harassment and personal attacks. In this work, we introduce the task of predicting from the very start of a conversation whether it will get out of hand. As opposed to detecting undesirable behavior after the fact, this task aims to enable early, actionable prediction at a time when the conversation might still be salvaged. To this end, we develop a framework for capturing pragmatic devices—such as politeness strategies and rhetorical prompts—used to start a conversation, and analyze their relation to its future trajectory. Applying this framework in a controlled setting, we demonstrate the feasibility of detecting early warning signs of antisocial behavior in online discussions.

[TACL] Detecting Institutional Dialog Acts in Police Traffic Stops
Vinodkumar Prabhakaran, Camilla Griffiths, Hang Su, Prateek Verma, Nelson Morgan, Jennifer Eberhardt, and Dan Jurafsky 11:45–12:10

We apply computational dialog methods on police body-worn camera footage to model conversations between police officers and community members in traffic stops. Relying on the theory of institutional talk, we develop a labeling scheme for police talk in traffic stops, and a tagger to detect institutional dialog acts (Reasons, Searches, Offering Help) from transcribed text at the turn (78% F-score) and stop (89% F-score) level. We then develop speech recognition and segmentation algorithms to detect these acts at the stop level from raw camera audio (81% F-score, with even higher accuracy for crucial acts like the stop Reason). We demonstrate that the dialog structures produced by our tagger could reveal whether officers follow law enforcement norms like introducing themselves, explaining the reason for the stop, and asking permission for searches, making our work an important step in improving police-community relations.

Session 4E: Evaluation

219 *Chair: Bonnie Webber*

Are BLEU and Meaning Representation in Opposition?
Ondřej Cífka and Ondřej Bojar 10:30–10:55

One of possible ways of obtaining continuous-space sentence representations is by training neural machine translation (NMT) systems. The recent attention mechanism however removes the single point in the neural network from which the source sentence representation can be extracted. We propose several variations of the attentive NMT architecture bringing this meeting point back. Empirical evaluation suggests that the better the translation quality, the worse the learned sentence representations serve in a wide range of classification and similarity tasks.

Automatic Metric Validation for Grammatical Error Correction
Leshem Choshen and Omri Abend 10:55–11:20

Metric validation in Grammatical Error Correction (GEC) is currently done by observing the correlation between human and metric-induced rankings. However, such correlation studies are costly, methodologically troublesome, and suffer from low inter-rater agreement. We propose maege, an automatic methodology for GEC metric validation, that overcomes many of the difficulties in the existing methodology. Experiments with maege shed a new light on metric quality, showing for example that the standard M^2 metric fares poorly on corpus-level ranking. Moreover, we use maege to perform a detailed analysis of metric behavior, showing that some types of valid edits are consistently penalized by existing metrics.

The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing
Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart 11:20–11:45

Statistical significance testing is a standard statistical tool designed to ensure that experimental results are not coincidental. In this opinion/ theoretical paper we discuss the role of statistical significance testing in Natural Language Processing (NLP) research. We establish the fundamental concepts of significance testing and discuss the specific aspects of NLP tasks, experimental setups and evaluation measures that affect the choice of significance tests in NLP research. Based on this discussion we propose a simple practical protocol for statistical significance test selection in NLP setups and accompany this protocol with a brief survey of the most relevant tests. We then survey recent empirical papers published in ACL and TACL during 2017 and show that while our community assigns great value to experimental results, statistical significance testing is often ignored or misused. We conclude with a brief discussion of open issues that should be properly addressed so that this important tool can be applied. in NLP research in a statistically sound manner.

[TACL] Replicability Analysis for Natural Language Processing: Testing Significance with Multiple Datasets
Rotem Dror, Gili Baumer, Marina Bogomolov, and Roi Reichart 11:45–12:10

With the ever growing amounts of textual data from a large variety of languages, domains and genres, it has become standard to evaluate NLP algorithms on multiple datasets in order to ensure consistent performance across heterogeneous setups. However, such multiple comparisons pose significant challenges to traditional statistical analysis methods in NLP and can lead to erroneous conclusions. In this paper we propose a Replicability Analysis framework for a statistically sound analysis of multiple comparisons between algorithms for NLP tasks. We discuss the theoretical advantages of this framework over the current, statistically unjustified, practice in the NLP literature, and demonstrate its empirical value across four applications: multi-domain dependency parsing, multilingual POS tagging, cross-domain sentiment classification and word similarity prediction.

Session 4F: Parsing 2

220 *Chair: Joakim Nivre*

Distilling Knowledge for Search-based Structured Prediction
Yijia Liu, Wanxiang Che, Huaipeng Zhao, Bing Qin, and Ting Liu 10:30–10:55

Many natural language processing tasks can be modeled into structured prediction and solved as a search problem. In this paper, we distill an ensemble of multiple models trained with different initialization into a single model. In addition to learning to match the ensemble's probability output on the reference states, we also use the ensemble to explore the search space and learn from the encountered states in the exploration. Experimental results on two typical search-based structured prediction tasks – transition-based dependency parsing and neural machine translation show that distillation can effectively improve the single model's performance and the final model achieves improvements of 1.32 in LAS and 2.65 in BLEU score on these two tasks respectively over strong baselines and it outperforms the greedy structured prediction models in previous literatures.

Stack-Pointer Networks for Dependency Parsing
Xuezhe Ma, Zecong Hu, Jingzhou Liu, Nanyun Peng, Graham Neubig, and Eduard Hovy 10:55–11:20

We introduce a novel architecture for dependency parsing: stack-pointer networks (StackPtr). Combining pointer networks (Vinyals et al., 2015) with an internal stack, the proposed model first reads and encodes the whole sentence, then builds the dependency tree top-down (from root-to-leaf) in a depth-first fashion. The stack tracks the status of the depth-first search and the pointer networks select one child for the word at the top of the stack at each step. The StackPtr parser benefits from the information of whole sentence and all previously derived subtree structures, and removes the left-to-right restriction in classical transition-based parsers. Yet the number of steps for building any (non-projective) parse tree is linear in the length of the sentence just as other transition-based parsers, yielding an efficient decoding algorithm with $O(n^2)$ time complexity. We evaluate our model on 29 treebanks spanning 20 languages and different dependency annotation schemas, and achieve state-of-the-art performances on 21 of them

Twitter Universal Dependency Parsing for African-American and Mainstream American English
Su Lin Blodgett, Johnny Wei, and Brendan O'Connor 11:20–11:45

Due to the presence of both Twitter-specific conventions and non-standard and dialectal language, Twitter presents a significant parsing challenge to current dependency parsing tools. We broaden English dependency parsing to handle social media English, particularly social media African-American English (AAE), by developing and annotating a new dataset of 500 tweets, 250 of which are in AAE, within the Universal Dependencies 2.0 framework. We describe our standards for handling Twitter- and AAE-specific features and evaluate a variety of cross-domain strategies for improving parsing with no, or very little, in-domain labeled data, including a new data synthesis approach. We analyze these methods' impact on performance disparities between AAE and Mainstream American English tweets, and assess parsing accuracy for specific AAE lexical and syntactic features. Our annotated data and a parsing model are available at: http://slanglab.cs.umass.edu/TwitterAAE/.

LSTMs Can Learn Syntax-Sensitive Dependencies Well, But Modeling Structure Makes Them Better
Adhiguna Kuncoro, Chris Dyer, John Hale, Dani Yogatama, Stephen Clark, and Phil Blunsom 11:45–12:10

Language exhibits hierarchical structure, but recent work using a subject-verb agreement diagnostic argued that state-of-the-art language models, LSTMs, fail to learn long-range syntax sensitive dependencies. Using the same diagnostic, we show that, in fact, LSTMs do succeed in learning such dependencies—provided they have enough capacity. We then explore whether models that have access to explicit syntactic information learn agreement more effectively, and how the way in which this structural information is incorporated into the model impacts performance. We find that the mere presence of syntactic information does not improve accuracy, but when model architecture is determined by syntax, number agreement is improved. Further, we find that the choice of how syntactic structure is built affects how well number agreement is learned: top-down construction outperforms left-corner and bottom-up variants in capturing non-local structural dependencies.

Poster Session 2

Time: *12:30–14:00* Location: Melbourne Room 1 & 2

Poster Session 2A: Student Research Workshop

Towards Opinion Summarization of Customer Reviews
Samuel Pecar
In recent years, the number of texts has grown rapidly. For example, most review-based portals, like Yelp or Amazon, contain thousands of user-generated reviews. It is impossible for any human reader to process even the most relevant of these documents. The most promising tool to solve this task is a text summarization. Most existing approaches, however, work on small, homogeneous, English datasets, and do not account to multi-linguality, opinion shift, and domain effects. In this paper, we introduce our research plan to use neural networks on user-generated travel reviews to generate summaries that take into account shifting opinions over time. We outline future directions in summarization to address all of these issues. By resolving the existing problems, we will make it easier for users of review-sites to make more informed decisions.

Sampling Informative Training Data for RNN Language Models
Jared Fernandez and Doug Downey
We propose an unsupervised importance sampling approach to selecting training data for recurrent neural network (RNNs) language models. To increase the information content of the training set, our approach preferentially samples high perplexity sentences, as determined by an easily queryable n-gram language model. We experimentally evaluate the heldout perplexity of models trained with our various importance sampling distributions. We show that language models trained on data sampled using our proposed approach outperform models trained over randomly sampled subsets of both the Billion Word (Chelba et al., 2014 Wikitext-103 benchmark corpora (Merity et al., 2016).

Learning-based Composite Metrics for Improved Caption Evaluation
Naeha Sharif, Lyndon White, Mohammed Bennamoun, and Syed Afaq Ali Shah
The evaluation of image caption quality is a challenging task, which requires the assessment of two main aspects in a caption: adequacy and fluency. These quality aspects can be judged using a combination of several linguistic features. However, most of the current image captioning metrics focus only on specific linguistic facets, such as the lexical or semantic, and fail to meet a satisfactory level of correlation with human judgements at the sentence-level. We propose a learning-based framework to incorporate the scores of a set of lexical and semantic metrics as features, to capture the adequacy and fluency of captions at different linguistic levels. Our experimental results demonstrate that composite metrics draw upon the strengths of stand-alone measures to yield improved correlation and accuracy.

Recursive Neural Network Based Preordering for English-to-Japanese Machine Translation
Yuki Kawara, Chenhui Chu, and Yuki Arase
The word order between source and target languages significantly influences the translation quality. Preordering can effectively address this problem. Previous preordering methods require a manual feature design, making language dependent design difficult. In this paper, we propose a preordering method with recursive neural networks that learn features from raw inputs. Experiments show the proposed method is comparable to the state-of-the-art method but without a manual feature design.

Pushing the Limits of Radiology with Joint Modeling of Visual and Textual Information
Sonit Singh
Recently, there has been increasing interest in the intersection of computer vision and natural language processing. Researchers have studied several interesting tasks, including generating text descriptions from images and videos and language embedding of images. More recent work has further extended the scope of this area to combine videos and language, learning to solve non-visual tasks using visual cues, visual question answering, and visual dialog. Despite a large body of research on the intersection of vision-language technology, its adaption to the medical domain is not fully explored. To address this research gap, we aim to develop machine learning models that can reason jointly on medical images and clinical text for advanced search, retrieval, annotation and description of medical images.

Recognizing Complex Entity Mentions: A Review and Future Directions
Xiang Dai

Standard named entity recognizers can effectively recognize entity mentions that consist of contiguous tokens and do not overlap with each other. However, in practice, there are many domains, such as the biomedical domain, in which there are nested, overlapping, and discontinuous entity mentions. These complex mentions cannot be directly recognized by conventional sequence tagging models because they may break the assumptions based on which sequence tagging techniques are built. We review the existing methods which are revised to tackle complex entity mentions and categorize them as tokenlevel and sentence-level approaches. We then identify the research gap, and discuss some directions that we are exploring.

Automatic Detection of Cross-Disciplinary Knowledge Associations
Menasha Thilakaratne, Katrina Falkner, and Thushari Atapattu

Detecting interesting, cross-disciplinary knowledge associations hidden in scientific publications can greatly assist scientists to formulate and validate scientifically sensible novel research hypotheses. This will also introduce new areas of research that can be successfully linked with their research discipline. Currently, this process is mostly performed manually by exploring the scientific publications, requiring a substantial amount of time and effort. Due to the exponential growth of scientific literature, it has become almost impossible for a scientist to keep track of all research advances. As a result, scientists tend to deal with fragments of the literature according to their specialisation. Consequently, important and hidden associations among these fragmented knowledge that can be linked to produce significant scientific discoveries remain unnoticed. This doctoral work aims to develop a novel knowledge discovery approach that suggests most promising research pathways by analysing the existing scientific literature.

Language Identification and Named Entity Recognition in Hinglish Code Mixed Tweets
Kushagra Singh, Indira Sen, and Ponnurangam Kumaraguru

While growing code-mixed content on Online Social Networks(OSN) provides a fertile ground for studying various aspects of code-mixing, the lack of automated text analysis tools render such studies challenging. To meet this challenge, a family of tools for analyzing code-mixed data such as language identifiers, parts-of-speech (POS) taggers, chunkers have been developed. Named Entity Recognition (NER) is an important text analysis task which is not only informative by itself, but is also needed for downstream NLP tasks such as semantic role labeling. In this work, we present an exploration of automatic NER of code-mixed data. We compare our method with existing off-the-shelf NER tools for social media content,and find that our systems outperforms the best baseline by 33.18 % (F1 score).

German and French Neural Supertagging Experiments for LTAG Parsing
Tatiana Bladier, Andreas van Cranenburgh, Younes Samih, and Laura Kallmeyer

We present ongoing work on data-driven parsing of German and French with Lexicalized Tree Adjoining Grammars. We use a supertagging approach combined with deep learning. We show the challenges of extracting LTAG supertags from the French Treebank, introduce the use of left- and right-sister-adjunction, present a neural architecture for the supertagger, and report experiments of n-best supertagging for French and German.

SuperNMT: Neural Machine Translation with Semantic Supersenses and Syntactic Supertags
Eva Vanmassenhove and Andy Way

In this paper we incorporate semantic supersensetags and syntactic supertag features into EN—FR and EN—DE factored NMT systems. In experiments on various test sets, we observe that such features (and particularly when combined) help the NMT model training to converge faster and improve the model quality according to the BLEU scores.

Unsupervised Semantic Abstractive Summarization
Shibhansh Dohare, Vivek Gupta, and Harish Karnick

Automatic abstractive summary generation remains a significant open problem for natural language processing. In this work, we develop a novel pipeline for Semantic Abstractive Summarization (SAS). SAS, as introduced by Liu et. al. (2015) first generates an AMR graph of an input story, through which it extracts a summary graph and finally, creates summary sentences from this summary graph. Compared to earlier approaches, we develop a more comprehensive method to generate the story AMR graph using state-of-the-art co-reference resolution and Meta Nodes. Which we then use in a novel unsupervised algorithm based on how humans summarize a piece of text to extract the summary sub-graph. Our algorithm outperforms the state of the art SAS method by 1.7% F1 score in node prediction.

Poster Session 2B: Dialog and Interactive Systems, Multilinguality

Sequicity: Simplifying Task-oriented Dialogue Systems with Single Sequence-to-Sequence Architectures
Wenqiang Lei, Xisen Jin, Min-Yen Kan, Zhaochun Ren, Xiangnan He, and Dawei Yin

Existing solutions to task-oriented dialogue systems follow pipeline designs which introduces architectural complexity and fragility. We propose a novel, holistic, extendable framework based on a single sequence-to-sequence (seq2seq) model which can be optimized with supervised or reinforcement learning. A key contribution is that we design text spans named belief spans to track dialogue believes, allowing task-oriented dialogue systems to be modeled in a seq2seq way. Based on this, we propose a simplistic Two Stage CopyNet instantiation which emonstrates good scalability: significantly reducing model complexity in terms of number of parameters and training time by a magnitude. It significantly outperforms state-of-the-art pipeline-based methods on large datasets and retains a satisfactory entity match rate on out-of-vocabulary (OOV) cases where pipeline-designed competitors totally fail.

An End-to-end Approach for Handling Unknown Slot Values in Dialogue State Tracking
Puyang Xu and Qi Hu

We highlight a practical yet rarely discussed problem in dialogue state tracking (DST), namely handling unknown slot values. Previous approaches generally assume predefined candidate lists and thus are not designed to output unknown values, especially when the spoken language understanding (SLU) module is absent as in many end-to-end (E2E) systems. We describe in this paper an E2E architecture based on the pointer network (PtrNet) that can effectively extract unknown slot values while still obtains state-of-the-art accuracy on the standard DSTC2 benchmark. We also provide extensive empirical evidence to show that tracking unknown values can be challenging and our approach can bring significant improvement with the help of an effective feature dropout technique.

Global-Locally Self-Attentive Encoder for Dialogue State Tracking
Victor Zhong, Caiming Xiong, and Richard Socher

Dialogue state tracking, which estimates user goals and requests given the dialogue context, is an essential part of task-oriented dialogue systems. In this paper, we propose the Global-Locally Self-Attentive Dialogue State Tracker (GLAD), which learns representations of the user utterance and previous system actions with global-local modules. Our model uses global modules to shares parameters between estimators for different types (called slots) of dialogue states, and uses local modules to learn slot-specific features. We show that this significantly improves tracking of rare states. GLAD obtains 88.3% joint goal accuracy and 96.4% request accuracy on the WoZ state tracking task, outperforming prior work by 3.9% and 4.8%. On the DSTC2 task, our model obtains 74.7% joint goal accuracy and 97.3% request accuracy, outperforming prior work by 1.3% and 0.8%

Mem2Seq: Effectively Incorporating Knowledge Bases into End-to-End Task-Oriented Dialog Systems
Andrea Madotto, Chien-Sheng Wu, and Pascale Fung

End-to-end task-oriented dialog systems usually suffer from the challenge of incorporating knowledge bases. In this paper, we propose a novel yet simple end-to-end differentiable model called memory-to-sequence (Mem2Seq) to address this issue. Mem2Seq is the first neural generative model that combines the multi-hop attention over memories with the idea of pointer network. We empirically show how Mem2Seq controls each generation step, and how its multi-hop attention mechanism helps in learning correlations between memories. In addition, our model is quite general without complicated task-specific designs. As a result, we show that Mem2Seq can be trained faster and attain the state-of-the-art performance on three different task-oriented dialog datasets.

Tailored Sequence to Sequence Models to Different Conversation Scenarios
Hainan Zhang, Yanyan Lan, Jiafeng Guo, Jun Xu, and Xueqi Cheng

Sequence to sequence (Seq2Seq) models have been widely used for response generation in the area of conversation. However, the requirements for different conversation scenarios are distinct. For example, customer service requires the generated responses to be specific and accurate, while chatbot prefers diverse responses so as to attract different users. The current Seq2Seq model fails to meet these diverse requirements, by using a general average likelihood as the optimization criteria. As a result, it usually generates safe and commonplace responses, such as 'I don't know'. In this paper, we propose two tailored optimization criteria for Seq2Seq to different conversation scenarios, i.e., the maximum generated likelihood for specific-requirement scenario, and the conditional value-at-risk for diverse-requirement scenario. Experimental results on the Ubuntu dialogue corpus (Ubuntu service scenario) and Chinese Weibo dataset (social chatbot scenario) show that our proposed models not only satisfies diverse requirements for different scenarios, but also yields better performances against traditional Seq2Seq models in terms of both

metric-based and human evaluations.

Knowledge Diffusion for Neural Dialogue Generation
Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin

End-to-end neural dialogue generation has shown promising results recently, but it does not employ knowledge to guide the generation and hence tends to generate short, general, and meaningless responses. In this paper, we propose a neural knowledge diffusion (NKD) model to introduce knowledge into dialogue generation. This method can not only match the relevant facts for the input utterance but diffuse them to similar entities. With the help of facts matching and entity diffusion, the neural dialogue generation is augmented with the ability of convergent and divergent thinking over the knowledge base. Our empirical study on a real-world dataset prove that our model is capable of generating meaningful, diverse and natural responses for both factoid-questions and knowledge grounded chi-chats. The experiment results also show that our model outperforms competitive baseline models significantly.

Generating Informative Responses with Controlled Sentence Function
Pei Ke, Jian Guan, Minlie Huang, and Xiaoyan Zhu

Sentence function is a significant factor to achieve the purpose of the speaker, which, however, has not been touched in large-scale conversation generation so far. In this paper, we present a model to generate informative responses with controlled sentence function. Our model utilizes a continuous latent variable to capture various word patterns that realize the expected sentence function, and introduces a type controller to deal with the compatibility of controlling sentence function and generating informative content. Conditioned on the latent variable, the type controller determines the type (i.e., function-related, topic, and ordinary word) of a word to be generated at each decoding position. Experiments show that our model outperforms state-of-the-art baselines, and it has the ability to generate responses with both controlled sentence function and informative content.

Sentiment Adaptive End-to-End Dialog Systems
Weiyan Shi and Zhou Yu

End-to-end learning framework is useful for building dialog systems for its simplicity in training and efficiency in model updating. However, current end-to-end approaches only consider user semantic inputs in learning and underutilize other user information. Therefore, we propose to include user sentiment obtained through multimodal information (acoustic, dialogic and textual), in the end-to-end learning framework to make systems more user-adaptive and effective. We incorporated user sentiment information in both supervised and reinforcement learning settings. In both settings, adding sentiment information reduced the dialog length and improved the task success rate on a bus information search task. This work is the first attempt to incorporate multimodal user information in the adaptive end-to-end dialog system training framework and attained state-of-the-art performance.

Task-oriented Dialogue System for Automatic Diagnosis
Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuanjing Huang, Kam-Fai Wong, and Xiangying Dai

In this paper, we make a move to build a dialogue system for automatic diagnosis. We first build a dataset collected from an online medical forum by extracting symptoms from both patients' self-reports and conversational data between patients and doctors. Then we propose a task-oriented dialogue system framework to make diagnosis for patients automatically, which can converse with patients to collect additional symptoms beyond their self-reports. Experimental results on our dataset show that additional symptoms extracted from conversation can greatly improve the accuracy for disease identification and our dialogue system is able to collect these symptoms automatically and make a better diagnosis.

Transfer Learning for Context-Aware Question Matching in Information-seeking Conversations in E-commerce
Minghui Qiu, Liu Yang, Feng Ji, Wei Zhou, Jun Huang, Haiqing Chen, Bruce Croft, and Wei Lin

Building multi-turn information-seeking conversation systems is an important and challenging research topic. Although several advanced neural text matching models have been proposed for this task, they are generally not efficient for industrial applications. Furthermore, they rely on a large amount of labeled data, which may not be available in real-world applications. To alleviate these problems, we study transfer learning for multi-turn information seeking conversations in this paper. We first propose an efficient and effective multi-turn conversation model based on convolutional neural networks. After that, we extend our model to adapt the knowledge learned from a resource-rich domain to enhance the performance. Finally, we deployed our model in an industrial chatbot called AliMe Assist and observed a significant improvement over the existing online model.

Embedding Learning Through Multilingual Concept Induction
Philipp Dufter, Mengjie Zhao, Martin Schmitt, Alexander Fraser, and Hinrich Schütze

We present a new method for estimating vector space representations of words: embedding learning by concept induction. We test this method on a highly parallel corpus and learn semantic representations of words in 1259 different languages in a single common space. An extensive experimental evaluation on crosslingual word similarity and sentiment analysis indicates that concept-based multilingual embedding learning performs better than previous approaches.

Isomorphic Transfer of Syntactic Structures in Cross-Lingual NLP
Edoardo Maria Ponti, Roi Reichart, Anna Korhonen, and Ivan Vulić

The transfer or share of knowledge between languages is a potential solution to resource scarcity in NLP. However, the effectiveness of cross-lingual transfer can be challenged by variation in syntactic structures. Frameworks such as Universal Dependencies (UD) are designed to be cross-lingually consistent, but even in carefully designed resources trees representing equivalent sentences may not always overlap. In this paper, we measure cross-lingual syntactic variation, or anisomorphism, in the UD treebank collection, considering both morphological and structural properties. We show that reducing the level of anisomorphism yields consistent gains in cross-lingual transfer tasks. We introduce a source language selection procedure that facilitates effective cross-lingual parser transfer, and propose a typologically driven method for syntactic tree processing which reduces anisomorphism. Our results show the effectiveness of this method for both machine translation and cross-lingual sentence similarity, demonstrating the importance of syntactic structure compatibility for boosting cross-lingual transfer in NLP.

Language Modeling for Code-Mixing: The Role of Linguistic Theory based Synthetic Data
Adithya Pratapa, Gayatri Bhat, Monojit Choudhury, Sunayana Sitaram, Sandipan Dandapat, and Kalika Bali

Training language models for Code-mixed (CM) language is known to be a difficult problem because of lack of data compounded by the increased confusability due to the presence of more than one language. We present a computational technique for creation of grammatically valid artificial CM data based on the Equivalence Constraint Theory. We show that when training examples are sampled appropriately from this synthetic data and presented in certain order (aka training curriculum) along with monolingual and real CM data, it can significantly reduce the perplexity of an RNN-based language model. We also show that randomly generated CM data does not help in decreasing the perplexity of the LMs.

A Multi-task Approach to Learning Multilingual Representations
Karan Singla, Dogan Can, and Shrikanth Narayanan

We present a novel multi-task modeling approach to learning multilingual distributed representations of text. Our system learns word and sentence embeddings jointly by training a multilingual skip-gram model together with a cross-lingual sentence similarity model. Our architecture can transparently use both monolingual and sentence aligned bilingual corpora to learn multilingual embeddings, thus covering a vocabulary significantly larger than the vocabulary of the bilingual corpora alone. Our model shows competitive performance in a standard cross-lingual document classification task. We also show the effectiveness of our method in a limited resource scenario.

Characterizing Departures from Linearity in Word Translation
Ndapa Nakashole and Raphael Flauger

We investigate the behavior of maps learned by machine translation methods. The maps translate words by projecting between word embedding spaces of different languages. We locally approximate these maps using linear maps, and find that they vary across the word embedding space. This demonstrates that the underlying maps are non-linear. Importantly, we show that the locally linear maps vary by an amount that is tightly correlated with the distance between the neighborhoods on which they are trained. Our results can be used to test non-linear methods, and to drive the design of more accurate maps for word translation.

Filtering and Mining Parallel Data in a Joint Multilingual Space
Holger Schwenk

We learn a joint multilingual sentence embedding and use the distance between sentences in different languages to filter noisy parallel data and to mine for parallel data in large news collections. We are able to improve a competitive baseline on the WMT'14 English to German task by 0.3 BLEU by filtering out 25% of the training data. The same approach is used to mine additional bitexts for the WMT'14 system and to obtain competitive results on the BUCC shared task to identify parallel sentences in comparable corpora. The approach is generic, it can be applied to many language pairs and it is independent of the architecture of the machine translation system.

Poster Session 2C: Information Extraction, Text Mining

Chinese NER Using Lattice LSTM
Yue Zhang and Jie Yang

We investigate a lattice-structured LSTM model for Chinese NER, which encodes a sequence of input characters as well as all potential words that match a lexicon. Compared with character-based methods, our model explicitly leverages word and word sequence information. Compared with word-based methods, lattice LSTM does not suffer from segmentation errors. Gated recurrent cells allow our model to choose the most relevant characters and words from a sentence for better NER results. Experiments on various datasets show that lattice LSTM outperforms both word-based and character-based LSTM baselines, achieving the best results.

Nugget Proposal Networks for Chinese Event Detection
Hongyu Lin, Yaojie Lu, Xianpei Han, and Le Sun

Neural network based models commonly regard event detection as a word-wise classification task, which suffer from the mismatch problem between words and event triggers, especially in languages without natural word delimiters such as Chinese. In this paper, we propose Nugget Proposal Networks (NPNs), which can solve the word-trigger mismatch problem by directly proposing entire trigger nuggets centered at each character regardless of word boundaries. Specifically, NPNs perform event detection in a character-wise paradigm, where a hybrid representation for each character is first learned to capture both structural and semantic information from both characters and words. Then based on learned representations, trigger nuggets are proposed and categorized by exploiting character compositional structures of Chinese event triggers. Experiments on both ACE2005 and TAC KBP 2017 datasets show that NPNs significantly outperform the state-of-the-art methods.

Higher-order Relation Schema Induction using Tensor Factorization with Back-off and Aggregation
Madhav Nimishakavi, Manish Gupta, and Partha Talukdar

Relation Schema Induction (RSI) is the problem of identifying type signatures of arguments of relations from unlabeled text. Most of the previous work in this area have focused only on binary RSI, i.e., inducing only the subject and object type signatures per relation. However, in practice, many relations are high-order, i.e., they have more than two arguments and inducing type signatures of all arguments is necessary. For example, in the sports domain, inducing a schema win(WinningPlayer, OpponentPlayer, Tournament, Location) is more informative than inducing just win(WinningPlayer, OpponentPlayer). We refer to this problem as Higher-order Relation Schema Induction (HRSI). In this paper, we propose Tensor Factorization with Back-off and Aggregation (TFBA), a novel framework for the HRSI problem. To the best of our knowledge, this is the first attempt at inducing higher-order relation schemata from unlabeled text. Using the experimental analysis on three real world datasets we show how TFBA helps in dealing with sparsity and induce higher-order schemata.

Discovering Implicit Knowledge with Unary Relations
Michael Glass and Alfio Gliozzo

State-of-the-art relation extraction approaches are only able to recognize relationships between mentions of entity arguments stated explicitly in the text and typically localized to the same sentence. However, the vast majority of relations are either implicit or not sententially localized. This is a major problem for Knowledge Base Population, severely limiting recall. In this paper we propose a new methodology to identify relations between two entities, consisting of detecting a very large number of unary relations, and using them to infer missing entities. We describe a deep learning architecture able to learn thousands of such relations very efficiently by using a common deep learning based representation. Our approach largely outperforms state of the art relation extraction technology on a newly introduced web scale knowledge base population benchmark, that we release to the research community.

Improving Entity Linking by Modeling Latent Relations between Mentions
Phong Le and Ivan Titov

Entity linking involves aligning textual mentions of named entities to their corresponding entries in a knowledge base. Entity linking systems often exploit relations between textual mentions in a document (e.g., coreference) to decide if the linking decisions are compatible. Unlike previous approaches, which relied on supervised systems or heuristics to predict these relations, we treat relations as latent variables in our neural entity-linking model. We induce the relations without any supervision while optimizing the entity-linking system in an end-to-end fashion. Our multi-relational model achieves the best reported scores on the standard benchmark (AIDA-CoNLL) and substantially outperforms its relation-agnostic version. Its training also converges much faster, suggesting that the injected structural bias helps to explain regularities in the training data.

Dating Documents using Graph Convolution Networks
Shikhar Vashishth, Shib Sankar Dasgupta, Swayambhu Nath Ray, and Partha Talukdar

Document date is essential for many important tasks, such as document retrieval, summarization, event detection, etc. While existing approaches for these tasks assume accurate knowledge of the document date, this is not always available, especially for arbitrary documents from the Web. Document Dating is a challenging problem which requires inference over the temporal structure of the document. Prior document dating systems have largely relied on hand-crafted features while ignoring such document-internal structures. In this paper, we propose NeuralDater, a Graph Convolutional Network (GCN) based document dating approach which jointly exploits syntactic and temporal graph structures of document in a principled way. To the best of our knowledge, this is the first application of deep learning for the problem of document dating. Through extensive experiments on real-world datasets, we find that NeuralDater significantly outperforms state-of-the-art baseline by 19% absolute (45% relative) accuracy points.

Hybrid semi-Markov CRF for Neural Sequence Labeling
Zhixiu Ye and Zhen-Hua Ling

This paper proposes hybrid semi-Markov conditional random fields (SCRFs) for neural sequence labeling in natural language processing. Based on conventional conditional random fields (CRFs), SCRFs have been designed for the tasks of assigning labels to segments by extracting features from and describing transitions between segments instead of words. In this paper, we improve the existing SCRF methods by employing word-level and segment-level information simultaneously. First, word-level labels are utilized to derive the segment scores in SCRFs. Second, a CRF output layer and an SCRF output layer are integrated into a unified neural network and trained jointly. Experimental results on CoNLL 2003 named entity recognition (NER) shared task show that our model achieves state-of-the-art performance when no external knowledge is used.

A Study of the Importance of External Knowledge in the Named Entity Recognition Task
Dominic Seyler, Tatiana Dembelova, Luciano Del Corro, Johannes Hoffart, and Gerhard Weikum

In this work, we discuss the importance of external knowledge for performing Named Entity Recognition (NER). We present a novel modular framework that divides the knowledge into four categories according to the depth of knowledge they convey. Each category consists of a set of features automatically generated from different information sources, such as a knowledge-base, a list of names, or document-specific semantic annotations. Further, we show the effects on performance when incrementally adding deeper knowledge and discuss effectiveness/efficiency trade-offs.

Improving Topic Quality by Promoting Named Entities in Topic Modeling
Katsiaryna Krasnashchok and Salim Jouili

News related content has been extensively studied in both topic modeling research and named entity recognition. However, expressive power of named entities and their potential for improving the quality of discovered topics has not received much attention. In this paper we use named entities as domain-specific terms for news-centric content and present a new weighting model for Latent Dirichlet Allocation. Our experimental results indicate that involving more named entities in topic descriptors positively influences the overall quality of topics, improving their interpretability, specificity and diversity.

Obligation and Prohibition Extraction Using Hierarchical RNNs
Ilias Chalkidis, Ion Androutsopoulos, and Achilleas Michos

We consider the task of detecting contractual obligations and prohibitions. We show that a self-attention mechanism improves the performance of a BILSTM classifier, the previous state of the art for this task, by allowing it to focus on indicative tokens. We also introduce a hierarchical BILSTM, which converts each sentence to an embedding, and processes the sentence embeddings to classify each sentence. Apart from being faster to train, the hierarchical BILSTM outperforms the flat one, even when the latter considers surrounding sentences, because the hierarchical model has a broader discourse view.

Poster Session 2D: Generation

A Graph-to-Sequence Model for AMR-to-Text Generation
Linfeng Song, Yue Zhang, Zhiguo Wang, and Daniel Gildea

The problem of AMR-to-text generation is to recover a text representing the same meaning as an input AMR graph. The current state-of-the-art method uses a sequence-to-sequence model, leveraging LSTM for encoding a linearized AMR structure. Although being able to model non-local semantic information, a sequence LSTM can lose information from the AMR graph structure, and thus facing challenges with large-graphs, which result in long sequences. We introduce a neural graph-to-sequence model, using a novel LSTM structure for directly encoding graph-level semantics. On a standard benchmark, our model shows superior results to existing methods in the literature.

GTR-LSTM: A Triple Encoder for Sentence Generation from RDF Data
Bayu Distiawan Trisedya, Jianzhong Qi, Rui Zhang, and Wei Wang

A knowledge base is a large repository of facts that are mainly represented as RDF triples, each of which consists of a subject, a predicate (relationship), and an object. The RDF triple representation offers a simple interface for applications to access the facts. However, this representation is not in a natural language form, which is difficult for humans to understand. We address this problem by proposing a system to translate a set of RDF triples into natural sentences based on an encoder-decoder framework. To preserve as much information from RDF triples as possible, we propose a novel graph-based triple encoder. The proposed encoder encodes not only the elements of the triples but also the relationships both within a triple and between the triples. Experimental results show that the proposed encoder achieves a consistent improvement over the baseline models by up to 17.6%, 6.0%, and 16.4% in three common metrics BLEU, METEOR, and TER, respectively.

Learning to Write with Cooperative Discriminators
Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi

Despite their local fluency, long-form text generated from RNNs is often generic, repetitive, and even self-contradictory. We propose a unified learning framework that collectively addresses all the above issues by composing a committee of discriminators that can guide a base RNN generator towards more globally coherent generations. More concretely, discriminators each specialize in a different principle of communication, such as Grice's maxims, and are collectively combined with the base RNN generator through a composite decoding objective. Human evaluation demonstrates that text generated by our model is preferred over that of baselines by a large margin, significantly enhancing the overall coherence, style, and information of the generations.

A Neural Approach to Pun Generation
Zhiwei Yu, Jiwei Tan, and Xiaojun Wan

Automatic pun generation is an interesting and challenging text generation task. Previous efforts rely on templates or laboriously manually annotated pun datasets, which heavily constrains the quality and diversity of generated puns. Since sequence-to-sequence models provide an effective technique for text generation, it is promising to investigate these models on the pun generation task. In this paper, we propose neural network models for homographic pun generation, and they can generate puns without requiring any pun data for training. We first train a conditional neural language model from a general text corpus, and then generate puns from the language model with an elaborately designed decoding algorithm. Automatic and human evaluations show that our models are able to generate homographic puns of good readability and quality.

Learning to Generate Move-by-Move Commentary for Chess Games from Large-Scale Social Forum Data
Harsh Jhamtani, Varun Gangal, Eduard Hovy, Graham Neubig, and Taylor Berg-Kirkpatrick

This paper examines the problem of generating natural language descriptions of chess games. We introduce a new large-scale chess commentary dataset and propose methods to generate commentary for individual moves in a chess game. The introduced dataset consists of more than 298K chess move-commentary pairs across 11K chess games. We highlight how this task poses unique research challenges in natural language generation: the data contain a large variety of styles of commentary and frequently depend on pragmatic context. We benchmark various baselines and propose an end-to-end trainable neural model which takes into account multiple pragmatic aspects of the game state that may be commented upon to describe a given chess move. Through a human study on predictions for a subset of the data which deals with direct move descriptions, we observe that outputs from our models are rated similar to ground truth commentary texts in terms of correctness and fluency.

From Credit Assignment to Entropy Regularization: Two New Algorithms for Neural Sequence Prediction
Zihang Dai, Qizhe Xie, and Eduard Hovy

In this work, we study the credit assignment problem in reward augmented maximum likelihood (RAML) learning, and establish a theoretical equivalence between the token-level counterpart of RAML and the entropy regularized reinforcement learning. Inspired by the connection, we propose two sequence prediction algorithms, one extending RAML with fine-grained credit assignment and the other improving Actor-Critic with a systematic entropy regularization. On two benchmark datasets, we show the proposed algorithms outperform RAML and Actor-Critic respectively, providing new alternatives to sequence prediction.

Paper Abstract Writing through Editing Mechanism
Qingyun Wang, Zhihao Zhou, Lifu Huang, Spencer Whitehead, Boliang Zhang, Heng Ji, and Kevin Knight

We present a paper abstract writing system based on an attentive neural sequence-to-sequence model that can take a title as input and automatically generate an abstract. We design a novel Writing-editing Network that can attend to both the title and the previously generated abstract drafts and then iteratively revise and polish the abstract. With two series of Turing tests, where the human judges are asked to distinguish the system-generated abstracts from human-written ones, our system passes Turing tests by junior domain experts at a rate up to 30% and by non-expert at a rate up to 80%.

Conditional Generators of Words Definitions
Artyom Gadetsky, Ilya Yakubovskiy, and Dmitry Vetrov

We explore recently introduced definition modeling technique that provided the tool for evaluation of different distributed vector representations of words through modeling dictionary definitions of words. In this work, we study the problem of word ambiguities in definition modeling and propose a possible solution by employing latent variable modeling and soft attention mechanisms. Our quantitative and qualitative evaluation and analysis of the model shows that taking into account words' ambiguity and polysemy leads to performance improvement.

Poster Session 2E: Question Answering

DuoRC: Towards Complex Language Understanding with Paraphrased Reading Comprehension
Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan

We propose DuoRC, a novel dataset for Reading Comprehension (RC) that motivates several new challenges for neural approaches in language understanding beyond those offered by existing RC datasets. DuoRC contains 186,089 unique question-answer pairs created from a collection of 7680 pairs of movie plots where each pair in the collection reflects two versions of the same movie - one from Wikipedia and the other from IMDb - written by two different authors. We asked crowdsourced workers to create questions from one version of the plot and a different set of workers to extract or synthesize answers from the other version. This unique characteristic of DuoRC where questions and answers are created from different versions of a document narrating the same underlying story, ensures by design, that there is very little lexical overlap between the questions created from one version and the segments containing the answer in the other version. Further, since the two versions have different levels of plot detail, narration style, vocabulary, etc., answering questions from the second version requires deeper language understanding and incorporating external background knowledge. Additionally, the narrative style of passages arising from movie plots (as opposed to typical descriptive passages in existing datasets) exhibits the need to perform complex reasoning over events across multiple sentences. Indeed, we observe that state-of-the-art neural RC models which have achieved near human performance on the SQuAD dataset, even when coupled with traditional NLP techniques to address the challenges presented in DuoRC exhibit very poor performance (F1 score of 37.42% on DuoRC v/s 86% on SQuAD dataset). This opens up several interesting research avenues wherein DuoRC could complement other RC datasets to explore novel neural approaches for studying language understanding.

Stochastic Answer Networks for Machine Reading Comprehension
Xiaodong Liu, Yelong Shen, Kevin Duh, and Jianfeng Gao

We propose a simple yet robust stochastic answer network (SAN) that simulates multi-step reasoning in machine reading comprehension. Compared to previous work such as ReasoNet which used reinforcement learning to determine the number of steps, the unique feature is the use of a kind of stochastic prediction dropout on the answer module (final layer) of the neural network during the training. We show that this simple trick improves robustness and achieves results competitive to the state-of-the-art on the Stanford Question Answering Dataset (SQuAD), the Adversarial SQuAD, and the Microsoft MAchine Reading COmprehension Dataset (MS MARCO).

Multi-Granularity Hierarchical Attention Fusion Networks for Reading Comprehension and Question Answering
Wei Wang, Ming Yan, and Chen Wu

This paper describes a novel hierarchical attention network for reading comprehension style question answering, which aims to answer questions for a given narrative paragraph. In the proposed method, attention and fusion are conducted horizontally and vertically across layers at different levels of granularity between question and paragraph. Specifically, it first encode the question and paragraph with fine-grained language embeddings, to better capture the respective representations at semantic level. Then it proposes a multi-granularity fusion approach to fully fuse information from both global and attended representations. Finally, it introduces a hierarchical attention network to focuses on the answer span progressively with multi-level soft-alignment. Extensive experiments on the large-scale SQuAD, TriviaQA dataset validate the effectiveness of the proposed method. At the time of writing the paper, our model achieves state-of-the-art on the both SQuAD and TriviaQA Wiki leaderboard as well as two adversarial SQuAD datasets.

Joint Training of Candidate Extraction and Answer Selection for Reading Comprehension
Zhen Wang, Jiachen Liu, Xinyan Xiao, Yajuan Lyu, and Tian Wu

While sophisticated neural-based techniques have been developed in reading comprehension, most approaches model the answer in an independent manner, ignoring its relations with other answer candidates. This problem can be even worse in open-domain scenarios, where candidates from multiple passages should be combined to answer a single question. In this paper, we formulate reading comprehension as an extract-then-select two-stage procedure. We first extract answer candidates from passages, then select the final answer by combining information from all the candidates. Furthermore, we regard candidate extraction as a latent variable and train the two-stage process jointly with reinforcement learning. As a result, our approach has improved the state-of-the-art performance significantly on two challenging open-domain reading comprehension datasets. Further analysis demonstrates the effectiveness of our model components, especially the information fusion of all the candidates and the joint training of the extract-then-select procedure.

Efficient and Robust Question Answering from Minimal Context over Documents
Sewon Min, Victor Zhong, Richard Socher, and Caiming Xiong

Neural models for question answering (QA) over documents have achieved significant performance improvements. Although effective, these models do not scale to large corpora due to their complex modeling of interactions between the document and the question. Moreover, recent work has shown that such models are sensitive to adversarial inputs. In this paper, we study the minimal context required to answer the question, and find that most questions in existing datasets can be answered with a small set of sentences. Inspired by this observation, we propose a simple sentence selector to select the minimal set of sentences to feed into the QA model. Our overall system achieves significant reductions in training (up to 15 times) and inference times (up to 13 times), with accuracy comparable to or better than the state-of-the-art on SQuAD, NewsQA, TriviaQA and SQuAD-Open. Furthermore, our experimental results and analyses show that our approach is more robust to adversarial inputs.

Denoising Distantly Supervised Open-Domain Question Answering
Yankai Lin, Haozhe Ji, Zhiyuan Liu, and Maosong Sun

Distantly supervised open-domain question answering (DS-QA) aims to find answers in collections of unlabeled text. Existing DS-QA models usually retrieve related paragraphs from a large-scale corpus and apply reading comprehension technique to extract answers from the most relevant paragraph. They ignore the rich information contained in other paragraphs. Moreover, distant supervision data inevitably accompanies with the wrong labeling problem, and these noisy data will substantially degrade the performance of DS-QA. To address these issues, we propose a novel DS-QA model which employs a paragraph selector to filter out those noisy paragraphs and a paragraph reader to extract the correct answer from those denoised paragraphs. Experimental results on real-world datasets show that our model can capture useful information from noisy data and achieve significant improvements on DS-QA as compared to all baselines.

Question Condensing Networks for Answer Selection in Community Question Answering
Wei Wu, Xu Sun, and Houfeng Wang

Answer selection is an important subtask of community question answering (CQA). In a real-world CQA forum, a question is often represented as two parts: a subject that summarizes the main points of the question, and a body that elaborates on the subject in detail. Previous researches on answer selection usually ignored the difference between these two parts and concatenated them as the question representation. In this paper, we propose the Question Condensing Networks (QCN) to make use of the subject-body relationship of community questions. In our model, the question subject is the primary part of the question representation, and the question body information is aggregated based on similarity and disparity with the question subject. Experimental results show that QCN outperforms all existing models on two CQA datasets.

[TACL] Constructing Datasets for Multi-hop Reading Comprehension Across Documents
Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel

Most Reading Comprehension methods limit themselves to queries which can be answered using a single sentence, paragraph, or document. Enabling models to combine disjoint pieces of textual evidence would extend the scope of machine comprehension methods, but currently no resources exist to train and test this capability. We propose a novel task to encourage the development of models for text understanding across multiple documents and to investigate the limits of existing methods. In our task, a model learns to seek and combine evidence – effectively performing multi-hop, alias multi-step, inference. We devise a methodology to produce datasets for this task, given a collection of query-answer pairs and thematically linked documents. Two datasets from different domains are induced, and we identify potential pitfalls and devise circumvention strategies. We evaluate two previously proposed competitive models and find that one can integrate information across documents. However, both models struggle to select relevant information; and providing documents guaranteed to be relevant greatly improves their performance. While the models outperform several strong baselines, their best accuracy reaches 54.5% on an annotated test set, compared to human performance at 85.0%, leaving ample room for improvement.

CNN for Text-Based Multiple Choice Question Answering
Akshay Chaturvedi, Onkar Pandit, and Utpal Garain

The task of Question Answering is at the very core of machine comprehension. In this paper, we propose a Convolutional Neural Network (CNN) model for text-based multiple choice question answering where questions are based on a particular article. Given an article and a multiple choice question, our model assigns a score to each question-option tuple and chooses the final option accordingly. We test our model on Textbook Question Answering (TQA) and SciQ dataset. Our model outperforms several LSTM-based baseline models on the two datasets.

Narrative Modeling with Memory Chains and Semantic Supervision
Fei Liu, Trevor Cohn, and Timothy Baldwin

Story comprehension requires a deep semantic understanding of the narrative, making it a challenging task. Inspired by previous studies on ROC Story Cloze Test, we propose a novel method, tracking various semantic aspects with external neural memory chains while encouraging each to focus on a particular semantic aspect. Evaluated on the task of story ending prediction, our model demonstrates superior performance to a collection of competitive baselines, setting a new state of the art.

Injecting Relational Structural Representation in Neural Networks for Question Similarity
Antonio Uva, Daniele Bonadiman, and Alessandro Moschitti

Effectively using full syntactic parsing information in Neural Networks (NNs) for solving relational tasks, e.g., question similarity, is still an open problem. In this paper, we propose to inject structural representations in NNs by (i) learning a model with Tree Kernels (TKs) on relatively few pairs of questions (few thousands) as gold standard (GS) training data is typically scarce, (ii) predicting labels on a very large corpus of question pairs, and (iii) pre-training NNs on such large corpus. The results on Quora and SemEval question similarity datasets show that NNs using our approach can learn more accurate models, especially after fine tuning on GS.

Poster Session 2F: Machine Translation

Towards Robust Neural Machine Translation
Yong Cheng, Zhaopeng Tu, Fandong Meng, Junjie Zhai, and Yang Liu

Small perturbations in the input can severely distort intermediate representations and thus impact translation quality of neural machine translation (NMT) models. In this paper, we propose to improve the robustness of NMT models with adversarial stability training. The basic idea is to make both the encoder and decoder in NMT models robust against input perturbations by enabling them to behave similarly for the original input and its perturbed counterpart. Experimental results on Chinese-English, English-German and English-French translation tasks show that our approaches can not only achieve significant improvements over strong NMT systems but also improve the robustness of NMT models.

Attention Focusing for Neural Machine Translation by Bridging Source and Target Embeddings
Shaohui Kuang, Junhui Li, António Branco, Weihua Luo, and Deyi Xiong

In neural machine translation, a source sequence of words is encoded into a vector from which a target sequence is generated in the decoding phase. Differently from statistical machine translation, the associations between source words and their possible target counterparts are not explicitly stored. Source and target words are at the two ends of a long information processing procedure, mediated by hidden states at both the source encoding and the target decoding phases. This makes it possible that a source word is incorrectly translated into a target word that is not any of its admissible equivalent counterparts in the target language. In this paper, we seek to somewhat shorten the distance between source and target words in that procedure, and thus strengthen their association, by means of a method we term bridging source and target word embeddings. We experiment with three strategies: (1) a source-side bridging model, where source word embeddings are moved one step closer to the output target sequence; (2) a target-side bridging model, which explores the more relevant source word embeddings for the prediction of the target sequence; and (3) a direct bridging model, which directly connects source and target word embeddings seeking to minimize errors in the translation of ones by the others. Experiments and analysis presented in this paper demonstrate that the proposed bridging models are able to significantly improve quality of both sentence translation, in general, and alignment and translation of individual source words with target words, in particular.

Reliability and Learnability of Human Bandit Feedback for Sequence-to-Sequence Reinforcement Learning
Julia Kreutzer, Joshua Uyheng, and Stefan Riezler

We present a study on reinforcement learning (RL) from human bandit feedback for sequence-to-sequence learning, exemplified by the task of bandit neural machine translation (NMT). We investigate the reliability of human bandit feedback, and analyze the influence of reliability on the learnability of a reward estimator, and the effect of the quality of reward estimates on the overall RL task. Our analysis of cardinal (5-point ratings) and ordinal (pairwise preferences) feedback shows that their intra- and inter-annotator α-agreement is comparable. Best reliability is obtained for standardized cardinal feedback, and cardinal feedback is also easiest to learn and generalize from. Finally, improvements of over 1 BLEU can be obtained by integrating a regression-based reward estimator trained on cardinal feedback for 800 translations into RL for NMT. This shows that RL is possible even from small amounts of fairly reliable human feedback, pointing to a great potential for applications at larger scale.

Accelerating Neural Transformer via an Average Attention Network
Biao Zhang, Deyi Xiong, and Jinsong Su

With parallelizable attention networks, the neural Transformer is very fast to train. However, due to the auto-regressive architecture and self-attention in the decoder, the decoding procedure becomes slow. To alleviate this issue, we propose an average attention network as an alternative to the self-attention network in the decoder of the neural Transformer. The average attention network consists of two layers, with an average layer that models dependencies on previous positions and a gating layer that is stacked over the average layer to enhance the expressiveness of the proposed attention network. We apply this network on the decoder part of the neural Transformer to replace the original target-side self-attention model. With masking tricks and dynamic programming, our model enables the neural Transformer to decode sentences over four times faster than its original version with almost no loss in training time and translation performance. We conduct a series of experiments on WMT17 translation tasks, where on 6 different language pairs, we obtain robust and consistent speed-ups in decoding.

How Much Attention Do You Need? A Granular Analysis of Neural Machine Translation Architectures
Tobias Domhan

With recent advances in network architectures for Neural Machine Translation (NMT) recurrent models have effec-

tively been replaced by either convolutional or self-attentional approaches, such as in the Transformer. While the main innovation of the Transformer architecture is its use of self-attentional layers, there are several other aspects, such as attention with multiple heads and the use of many attention layers, that distinguish the model from previous baselines. In this work we take a fine-grained look at the different architectures for NMT. We introduce an Architecture Definition Language (ADL) allowing for a flexible combination of common building blocks. Making use of this language we show in experiments that one can bring recurrent and convolutional models very close to the Transformer performance by borrowing concepts from the Transformer architecture, but not using self-attention. Additionally, we find that self-attention is much more important on the encoder side than on the decoder side, where it can be replaced by a RNN or CNN without a loss in performance in most settings. Surprisingly, even a model without any target side self-attention performs well.

[TACL] Scheduled Multi-Task Learning: From Syntax to Translation
Eliyahu Kiperwasser and Miguel Ballesteros

Neural encoder-decoder models of machine translation have achieved impressive results, while learning linguistic knowledge of both the source and target languages in an implicit end-to-end manner. We propose a framework in which our model begins learning syntax and translation interleaved and gradually puts more focus on translation. Using this approach, we achieve considerable improvements in terms of BLEU score on relatively large parallel corpus (WMT14 English to German) and a low-resource (WIT German to English) setup.

[TACL] Phrase Table Induction Using In-Domain Monolingual Data for Domain Adaptation in Statistical Machine Translation
Benjamin Marie and Atsushi Fujita

We present a new framework to induce an in-domain phrase table from in-domain monolingual data that can be used to adapt a general-domain statistical machine translation system to the targeted domain.Our method first compiles sets of phrases in source and target languages separately and generates candidate phrase pairs by taking the Cartesian product of the two phrase sets. It then computes inexpensive features for each candidate phrase pair and filters them using a supervised classifier in order to induce an in-domain phrase table. We experimented on the language pair English-French, both translation directions, in two domains and obtained consistently better results than a strong baseline system that uses an in-domain bilingual lexicon. We also conducted an error analysis that showed the induced phrase tables proposed useful translations, especially for words and phrases unseen in the parallel data used to train the general-domain baseline system.

[TACL] Learning to Remember Translation History with a Continuous Cache
Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang

Existing neural machine translation (NMT) models generally translate sentences in isolation, missing the opportunity to take advantage of document-level information. In this work, we propose to augment NMT models with a very light-weight cache-like memory network, which stores recent hidden representations as translation history. The probability distribution over generated words is updated online depending on the translation history retrieved from the memory, endowing NMT models with the capability to dynamically adapt over time. Experiments on multiple domains with different topics and styles show the effectiveness of the proposed approach with negligible impact on the computational cost.

[TACL] Modeling Past and Future for Neural Machine Translation
Zaixiang Zheng, Hao Zhou, Shujian Huang, Lili Mou, Xinyu Dai, Jiajun Chen, and Zhaopeng Tu

Existing neural machine translation systems do not explicitly model what has been translated and what has not during the decoding phase. To address this problem, we propose a novel mechanism that separates the source information into two parts: translated PAST contents and untranslated FUTURE contents, which are modeled by two additional recurrent layers. The PAST and FUTURE contents are fed to both the attention model and the decoder states, which provides Neural Machine Translation (NMT) systems with the knowledge of translated and untranslated contents. Experimental results show that the proposed approach significantly improves the performance in Chinese-English, German-English, and English-German translation tasks. Specifically, the proposed model outperforms the conventional coverage model in terms of both the translation quality and the alignment error rate.

[TACL] Leveraging Orthographic Similarity for Multilingual Neural Transliteration
Anoop Kunchukuttan, Mitesh M. Khapra, Gurneet Singh, and Pushpak Bhattacharyya

We address the task of joint training of transliteration models for multiple language pairs (multilingual transliteration). This is an instance of multitask learning, where individual tasks (language pairs) benefit from sharing knowledge with related tasks. We focus on transliteration involving related tasks i.e., languages sharing writing systems and phonetic properties (orthographically similar languages). We propose a modified neural encoder-decoder model that maxi-

mizes parameter sharing across language pairs in order to effectively leverage orthographic similarity. We show that multilingual transliteration significantly outperforms bilingual transliteration in different scenarios (average increase of 58% across a variety of languages we experimented with). We also show that multilingual transliteration models can generalize well to languages/language pairs not encountered during training and hence perform well on the zeroshot transliteration task. We show that further improvements can be achieved by using phonetic feature input.

A Simple and Effective Approach to Coverage-Aware Neural Machine Translation
Yanyang Li, Tong Xiao, Yinqiao Li, Qiang Wang, Changming Xu, and Jingbo Zhu

We offer a simple and effective method to seek a better balance between model confidence and length preference for Neural Machine Translation (NMT). Unlike the popular length normalization and coverage models, our model does not require training nor reranking the limited n-best outputs. Moreover, it is robust to large beam sizes, which is not well studied in previous work. On the Chinese-English and English-German translation tasks, our approach yields +0.4~1.5 BLEU improvements over the state-of-the-art baselines.

Dynamic Sentence Sampling for Efficient Training of Neural Machine Translation
Rui Wang, Masao Utiyama, and Eiichiro Sumita

Traditional Neural machine translation (NMT) involves a fixed training procedure where each sentence is sampled once during each epoch. In reality, some sentences are well-learned during the initial few epochs; however, using this approach, the well-learned sentences would continue to be trained along with those sentences that were not well learned for 10-30 epochs, which results in a wastage of time. Here, we propose an efficient method to dynamically sample the sentences in order to accelerate the NMT training. In this approach, a weight is assigned to each sentence based on the measured difference between the training costs of two iterations. Further, in each epoch, a certain percentage of sentences are dynamically sampled according to their weights. Empirical results based on the NIST Chinese-to-English and the WMT English-to-German tasks show that the proposed method can significantly accelerate the NMT training and improve the NMT performance.

Compositional Representation of Morphologically-Rich Input for Neural Machine Translation
Duygu Ataman and Marcello Federico

Neural machine translation (NMT) models are typically trained with fixed-size input and output vocabularies, which creates an important bottleneck on their accuracy and generalization capability. As a solution, various studies proposed segmenting words into sub-word units and performing translation at the sub-lexical level. However, statistical word segmentation methods have recently shown to be prone to morphological errors, which can lead to inaccurate translations. In this paper, we propose to overcome this problem by replacing the source-language embedding layer of NMT with a bi-directional recurrent neural network that generates compositional representations of the input at any desired level of granularity. We test our approach in a low-resource setting with five languages from different morphological typologies, and under different composition assumptions. By training NMT to compose word representations from character n-grams, our approach consistently outperforms (from 1.71 to 2.48 BLEU points) NMT learning embeddings of statistically generated sub-word units.

Extreme Adaptation for Personalized Neural Machine Translation
Paul Michel and Graham Neubig

Every person speaks or writes their own flavor of their native language, influenced by a number of factors: the content they tend to talk about, their gender, their social status, or their geographical origin. When attempting to perform Machine Translation (MT), these variations have a significant effect on how the system should perform translation, but this is not captured well by standard one-size-fits-all models. In this paper, we propose a simple and parameter-efficient adaptation technique that only requires adapting the bias of the output softmax to each particular user of the MT system, either directly or through a factored approximation. Experiments on TED talks in three languages demonstrate improvements in translation accuracy, and better reflection of speaker traits in the target text.

Multi-representation ensembles and delayed SGD updates improve syntax-based NMT
Danielle Saunders, Felix Stahlberg, Adrià de Gispert, and Bill Byrne

We explore strategies for incorporating target syntax into Neural Machine Translation. We specifically focus on syntax in ensembles containing multiple sentence representations. We formulate beam search over such ensembles using WFSTs, and describe a delayed SGD update training procedure that is especially effective for long representations like linearized syntax. Our approach gives state-of-the-art performance on a difficult Japanese-English task.

Learning from Chunk-based Feedback in Neural Machine Translation
Pavel Petrushkov, Shahram Khadivi, and Evgeny Matusov

We empirically investigate learning from partial feedback in neural machine translation (NMT), when partial feedback is collected by asking users to highlight a correct chunk of a translation. We propose a simple and effective way of utilizing such feedback in NMT training. We demonstrate how the common machine translation problem of domain mismatch between training and deployment can be reduced solely based on chunk-level user feedback. We conduct a series of simulation experiments to test the effectiveness of the proposed method. Our results show that chunk-level feedback outperforms sentence based feedback by up to 2.61% BLEU absolute.

Bag-of-Words as Target for Neural Machine Translation
Shuming Ma, Xu Sun, Yizhong Wang, and Junyang Lin

A sentence can be translated into more than one correct sentences. However, most of the existing neural machine translation models only use one of the correct translations as the targets, and the other correct sentences are punished as the incorrect sentences in the training stage. Since most of the correct translations for one sentence share the similar bag-of-words, it is possible to distinguish the correct translations from the incorrect ones by the bag-of-words. In this paper, we propose an approach that uses both the sentences and the bag-of-words as targets in the training stage, in order to encourage the model to generate the potentially correct sentences that are not appeared in the training set. We evaluate our model on a Chinese-English translation dataset, and experiments show our model outperforms the strong baselines by the BLEU score of 4.55.

Improving Beam Search by Removing Monotonic Constraint for Neural Machine Translation
Raphael Shu and Hideki Nakayama

To achieve high translation performance, neural machine translation models usually rely on the beam search algorithm for decoding sentences. The beam search finds good candidate translations by considering multiple hypotheses of translations simultaneously. However, as the algorithm produces hypotheses in a monotonic left-to-right order, a hypothesis can not be revisited once it is discarded. We found such monotonicity forces the algorithm to sacrifice some good decoding paths. To mitigate this problem, we relax the monotonic constraint of the beam search by maintaining all found hypotheses in a single priority queue and using a universal score function for hypothesis selection. The proposed algorithm allows discarded hypotheses to be recovered in a later step. Despite its simplicity, we show that the proposed decoding algorithm enhances the quality of selected hypotheses and improve the translations even for high-performance models in English-Japanese translation task.

Demo Poster Session 2

Time: *12:30–14:00* Location: Melbourne Room 1 & 2

DCFEE: A Document-level Chinese Financial Event Extraction System based on Automatically Labeled Training Data
Hang Yang, Yubo Chen, Kang Liu, Yang Xiao, and Jun Zhao

We present an event extraction framework to detect event mentions and extract events from the document-level financial news. Up to now, methods based on supervised learning paradigm gain the highest performance in public datasets (such as ACE2005, KBP2015). These methods heavily depend on the manually labeled training data. However, in particular areas, such as financial, medical and judicial domains, there is no enough labeled data due to the high cost of data labeling process. Moreover, most of the current methods focus on extracting events from one sentence, but an event is usually expressed by multiple sentences in one document. To solve these problems, we propose a Document-level Chinese Financial Event Extraction (DCFEE) system which can automatically generate a large scaled labeled data and extract events from the whole document. Experimental results demonstrate the effectiveness of it

Sentence Suggestion of Japanese Functional Expressions for Chinese-speaking Learners
Jun Liu, Hiroyuki Shindo, and Yuji Matsumoto

We present a computer-assisted learning system, Jastudy, which is particularly designed for Chinese-speaking learners of Japanese as a second language (JSL) to learn Japanese functional expressions with suggestion of appropriate example sentences. The system automatically recognizes Japanese functional expressions using a free Japanese morphological analyzer MeCab, which is retrained on a new Conditional Random Fields (CRF) model. In order to select appropriate example sentences, we apply a pairwise-based machine learning tool, Support Vector Machine for Ranking (SVMrank) to estimate the complexity of the example sentences using Japanese—Chinese homographs as an important feature. In addition, we cluster the example sentences that contain Japanese functional expressions with two or more meanings and usages, based on part-of-speech, conjugation forms of verbs and semantic attributes, using the K-means clustering algorithm in Scikit-Learn. Experimental results demonstrate the effectiveness of our approach.

Translating a Language You Don't Know In the Chinese Room
Ulf Hermjakob, Jonathan May, Michael Pust, and Kevin Knight

In a corruption of John Searle's famous AI thought experiment, the Chinese Room (Searle, 1980), we twist its original intent by enabling humans to translate text, e.g. from Uyghur to English, even if they don't have any prior knowledge of the source language. Our enabling tool, which we call the Chinese Room, is equipped with the same resources made available to a machine translation engine. We find that our superior language model and world knowledge allows us to create perfectly fluent and nearly adequate translations, with human expertise required only for the target language. The Chinese Room tool can be used to rapidly create small corpora of parallel data when bilingual translators are not readily available, in particular for low-resource languages.

SANTO: A Web-based Annotation Tool for Ontology-driven Slot Filling
Matthias Hartung, Hendrik ter Horst, Frank Grimm, Tim Diekmann, Roman Klinger, and Philipp Cimiano

Supervised machine learning algorithms require training data whose generation for complex relation extraction tasks tends to be difficult. Being optimized for relation extraction at sentence level, many annotation tools lack in facilitating the annotation of relational structures that are widely spread across the text. This leads to non-intuitive and cumbersome visualizations, making the annotation process unnecessarily time-consuming. We propose SANTO, an easy-to-use, domain-adaptive annotation tool specialized for complex slot filling tasks which may involve problems of cardinality and referential grounding. The web-based architecture enables fast and clearly structured annotation for multiple users in parallel. Relational structures are formulated as templates following the conceptualization of an underlying ontology. Further, import and export procedures of standard formats enable interoperability with external sources and tools.

NCRF++: An Open-source Neural Sequence Labeling Toolkit
Jie Yang and Yue Zhang

This paper describes NCRF++, a toolkit for neural sequence labeling. NCRF++ is designed for quick implementation of different neural sequence labeling models with a CRF inference layer. It provides users with an inference for building the custom model structure through configuration file with flexible neural feature design and utilization. Built on PyTorch, the core operations are calculated in batch, making the toolkit efficient with the acceleration of GPU. It also includes the implementations of most state-of-the-art neural sequence labeling models such as LSTM-CRF,

facilitating reproducing and refinement on those methods.

TALEN: Tool for Annotation of Low-resource ENtities
Stephen Mayhew and Dan Roth

We present a new web-based interface, TALEN, designed for named entity annotation in low-resource settings where the annotators do not speak the language. To address this non-traditional scenario, TALEN includes such features as in-place lexicon integration, TF-IDF token statistics, Internet search, and entity propagation, all implemented so as to make this difficult task efficient and frictionless. We conduct a small user study to compare against a popular annotation tool, showing that TALEN achieves higher precision and recall against ground-truth annotations, and that users strongly prefer it over the alternative. TALEN is available at: github.com/CogComp/talen.

A Web-scale system for scientific knowledge exploration
Zhihong Shen, Hao Ma, and Kuansan Wang

To enable efficient exploration of Web-scale scientific knowledge, it is necessary to organize scientific publications into a hierarchical concept structure. In this work, we present a large-scale system to (1) identify hundreds of thousands of scientific concepts, (2) tag these identified concepts to hundreds of millions of scientific publications by leveraging both text and graph structure, and (3) build a six-level concept hierarchy with a subsumption-based model. The system builds the most comprehensive cross-domain scientific concept ontology published to date, with more than 200 thousand concepts and over one million relationships.

ScoutBot: A Dialogue System for Collaborative Navigation
Stephanie M. Lukin, Felix Gervits, Cory Hayes, Pooja Moolchandani, Anton Leuski, John Rogers, Carlos Sanchez Amaro, Matthew Marge, Clare Voss, and David Traum

ScoutBot is a dialogue interface to physical and simulated robots that supports collaborative exploration of environments. The demonstration will allow users to issue unconstrained spoken language commands to ScoutBot. ScoutBot will prompt for clarification if the user's instruction needs additional input. It is trained on human-robot dialogue collected from Wizard-of-Oz experiments, where robot responses were initiated by a human wizard in previous interactions. The demonstration will show a simulated ground robot (Clearpath Jackal) in a simulated environment supported by ROS (Robot Operating System).

Session 5 Overview – Tuesday, July 17, 2018

	Track A *Semantics 1 (Short)*	**Track B** *Machine Translation, Multilinguality 1 (Short)*	**Track C** *Information Extraction 1 (Short)*	**Track D** *Dialog System, Discourse (Short)*	**Track E** *Vision, Linguistics, Resource and Evaluation (Short)*	**Track F** *Parsing, Morphology (Short)*	
	Plenary	203–204	210–211	212–213	219	220	
14:00	Leveraging distributed representations and lexico-syntactic fixedness for token-level prediction of the idiomaticity of English verb-noun combinations *King and Cook*	Sparse and Constrained Attention for Neural Machine Translation *Malaviya, Ferreira, and Martins*	Neural Cross-Lingual Coreference Resolution And Its Application To Entity Linking *Kundu, Sil, Florian, and Hamza*	Learning Matching Models with Weak Supervision for Response Selection in Retrieval-based Chatbots *Wu, Wu, Li, and Zhou*	SciDTB: Discourse Dependency TreeBank for Scientific Abstracts *Yang and Li*	Policy Gradient as a Proxy for Dynamic Oracles in Constituency Parsing *Fried and Klein*	
14:15	Using pseudo-senses for improving the extraction of synonyms from word embeddings *Ferret*	Neural Hidden Markov Model for Machine Translation *Wang, Zhu, Alkhouli, Gan, and Ney*	Judicious Selection of Training Data in Assisting Language for Multilingual Neural NER *Murthy, Kunchukuttan, and Bhattacharyya*	Improving Slot Filling in Spoken Language Understanding with Joint Pointer and Attention *Zhao and Feng*	Predicting accuracy on large datasets from smaller pilot data *Johnson, Anderson, Dras, and Steedman*	Linear-time Constituency Parsing with RNNs and Dynamic Programming *Hong and Huang*	
14:30	Hearst Patterns Revisited: Automatic Hypernym Detection from Large Text Corpora *Roller, Kiela, and Nickel*	Bleaching Text: Abstract Features for Cross-lingual Gender Prediction *Goot, Ljubešić, Matroos, Nissim, and Plank*	Neural Open Information Extraction *Cui, Wei, and Zhou*	Large-Scale Multi-Domain Belief Tracking with Knowledge Sharing *Ramadan, Budzianowski, and Gasic*	The Influence of Context on Sentence Acceptability Judgements *Bernardy, Lappin, and Lau*	Simpler but More Accurate Semantic Dependency Parsing *Dozat and Manning*	
14:45	Jointly Predicting Predicates and Arguments in Neural Semantic Role Labeling *He, Lee, Levy, and Zettlemoyer*	Orthographic Features for Bilingual Lexicon Induction *Riley and Gildea*	Document Embedding Enhanced Event Detection with Hierarchical and Supervised Attention *Zhao, Jin, Wang, and Cheng*	Modeling discourse cohesion for discourse parsing via memory network *Jia, Ye, Feng, Lai, Yan, and Zhao*	Do Neural Network Cross-Modal Mappings Really Bridge Modalities? *Collell and Moens*	Simplified Abugidas *Ding, Utiyama, and Sumita*	

Parallel Session 5

Session 5A: Semantics 1 (Short)

Plenary *Chair: Kevin Duh*

Leveraging distributed representations and lexico-syntactic fixedness for token-level prediction of the idiomaticity of English verb-noun combinations
Milton King and Paul Cook 14:00–14:15

Verb-noun combinations (VNCs) - e.g., blow the whistle, hit the roof, and see stars - are a common type of English idiom that are ambiguous with literal usages. In this paper we propose and evaluate models for classifying VNC usages as idiomatic or literal, based on a variety of approaches to forming distributed representations. Our results show that a model based on averaging word embeddings performs on par with, or better than, a previously-proposed approach based on skip-thoughts. Idiomatic usages of VNCs are known to exhibit lexico-syntactic fixedness. We further incorporate this information into our models, demonstrating that this rich linguistic knowledge is complementary to the information carried by distributed representations.

Using pseudo-senses for improving the extraction of synonyms from word embeddings
Olivier Ferret 14:15–14:30

The methods proposed recently for specializing word embeddings according to a particular perspective generally rely on external knowledge. In this article, we propose Pseudofit, a new method for specializing word embeddings according to semantic similarity without any external knowledge. Pseudofit exploits the notion of pseudo-sense for building several representations for each word and uses these representations for making the initial embeddings more generic. We illustrate the interest of Pseudofit for acquiring synonyms and study several variants of Pseudofit according to this perspective.

Hearst Patterns Revisited: Automatic Hypernym Detection from Large Text Corpora
Stephen Roller, Douwe Kiela, and Maximilian Nickel 14:30–14:45

Methods for unsupervised hypernym detection may broadly be categorized according to two paradigms: pattern-based and distributional methods. In this paper, we study the performance of both approaches on several hypernymy tasks and find that simple pattern-based methods consistently outperform distributional methods on common benchmark datasets. Our results show that pattern-based models provide important contextual constraints which are not yet captured in distributional methods.

Jointly Predicting Predicates and Arguments in Neural Semantic Role Labeling
Luheng He, Kenton Lee, Omer Levy, and Luke Zettlemoyer 14:45–15:00

Recent BIO-tagging-based neural semantic role labeling models are very high performing, but assume gold predicates as part of the input and cannot incorporate span-level features. We propose an end-to-end approach for jointly predicting all predicates, arguments spans, and the relations between them. The model makes independent decisions about what relationship, if any, holds between every possible word-span pair, and learns contextualized span representations that provide rich, shared input features for each decision. Experiments demonstrate that this approach sets a new state of the art on PropBank SRL without gold predicates.

Session 5B: Machine Translation, Multilinguality 1 (Short)

203–204 *Chair: Valia Kordoni*

Sparse and Constrained Attention for Neural Machine Translation
Chaitanya Malaviya, Pedro Ferreira, and André F. T. Martins 14:00–14:15

In neural machine translation, words are sometimes dropped from the source or generated repeatedly in the translation. We explore novel strategies to address the coverage problem that change only the attention transformation. Our approach allocates fertilities to source words, used to bound the attention each word can receive. We experiment with various sparse and constrained attention transformations and propose a new one, constrained sparsemax, shown to be differentiable and sparse. Empirical evaluation is provided in three languages pairs.

Neural Hidden Markov Model for Machine Translation
Weiyue Wang, Derui Zhu, Tamer Alkhouli, Zixuan Gan, and Hermann Ney 14:15–14:30

Attention-based neural machine translation (NMT) models selectively focus on specific source positions to produce a translation, which brings significant improvements over pure encoder-decoder sequence-to-sequence models. This work investigates NMT while replacing the attention component. We study a neural hidden Markov model (HMM) consisting of neural network-based alignment and lexicon models, which are trained jointly using the forward-backward algorithm. We show that the attention component can be effectively replaced by the neural network alignment model and the neural HMM approach is able to provide comparable performance with the state-of-the-art attention-based models on the WMT 2017 German<->English and Chinese->English translation tasks.

Bleaching Text: Abstract Features for Cross-lingual Gender Prediction
Rob van der Goot, Nikola Ljubešić, Ian Matroos, Malvina Nissim, and Barbara Plank 14:30–14:45

Gender prediction has typically focused on lexical and social network features, yielding good performance, but making systems highly language-, topic-, and platform dependent. Cross-lingual embeddings circumvent some of these limitations, but capture gender-specific style less. We propose an alternative: bleaching text, i.e., transforming lexical strings into more abstract features. This study provides evidence that such features allow for better transfer across languages. Moreover, we present a first study on the ability of humans to perform cross-lingual gender prediction. We find that human predictive power proves similar to that of our bleached models, and both perform better than lexical models.

Orthographic Features for Bilingual Lexicon Induction
Parker Riley and Daniel Gildea 14:45–15:00

Recent embedding-based methods in bilingual lexicon induction show good results, but do not take advantage of orthographic features, such as edit distance, which can be helpful for pairs of related languages. This work extends embedding-based methods to incorporate these features, resulting in significant accuracy gains for related languages.

Session 5C: Information Extraction 1 (Short)

210–211 *Chair: Feiyu Xu*

Neural Cross-Lingual Coreference Resolution And Its Application To Entity Linking
Gourab Kundu, Avi Sil, Radu Florian, and Wael Hamza 14:00–14:15

We propose an entity-centric neural crosslingual coreference model that builds on multi-lingual embeddings and language independent features. We perform both intrinsic and extrinsic evaluations of our model. In the intrinsic evaluation, we show that our model, when trained on English and tested on Chinese and Spanish, achieves competitive results to the models trained directly on Chinese and Spanish respectively. In the extrinsic evaluation, we show that our English model helps achieve superior entity linking accuracy on Chinese and Spanish test sets than the top 2015 TAC system without using any annotated data from Chinese or Spanish.

Judicious Selection of Training Data in Assisting Language for Multilingual Neural NER
Rudra Murthy, Anoop Kunchukuttan, and Pushpak Bhattacharyya 14:15–14:30

Multilingual learning for Neural Named Entity Recognition (NNER) involves jointly training a neural network for multiple languages. Typically, the goal is improving the NER performance of one of the languages (the primary language) using the other assisting languages. We show that the divergence in the tag distributions of the common named entities between the primary and assisting languages can reduce the effectiveness of multilingual learning. To alleviate this problem, we propose a metric based on symmetric KL divergence to filter out the highly divergent training instances in the assisting language. We empirically show that our data selection strategy improves NER performance in many languages, including those with very limited training data.

Neural Open Information Extraction
Lei Cui, Furu Wei, and Ming Zhou 14:30–14:45

Conventional Open Information Extraction (Open IE) systems are usually built on hand-crafted patterns from other NLP tools such as syntactic parsing, yet they face problems of error propagation. In this paper, we propose a neural Open IE approach with an encoder-decoder framework. Distinct from existing methods, the neural Open IE approach learns highly confident arguments and relation tuples bootstrapped from a state-of-the-art Open IE system. An empirical study on a large benchmark dataset shows that the neural Open IE system significantly outperforms several baselines, while maintaining comparable computational efficiency.

Document Embedding Enhanced Event Detection with Hierarchical and Supervised Attention
Yue Zhao, Xiaolong Jin, Yuanzhuo Wang, and Xueqi Cheng 14:45–15:00

Document-level information is very important for event detection even at sentence level. In this paper, we propose a novel Document Embedding Enhanced Bi-RNN model, called DEEB-RNN, to detect events in sentences. This model first learns event detection oriented embeddings of documents through a hierarchical and supervised attention based RNN, which pays word-level attention to event triggers and sentence-level attention to those sentences containing events. It then uses the learned document embedding to enhance another bidirectional RNN model to identify event triggers and their types in sentences. Through experiments on the ACE-2005 dataset, we demonstrate the effectiveness and merits of the proposed DEEB-RNN model via comparison with state-of-the-art methods.

Session 5D: Dialog System, Discourse (Short)

212–213 *Chair: Daisuke Kawahara*

Learning Matching Models with Weak Supervision for Response Selection in Retrieval-based Chatbots

Yu Wu, Wei Wu, Zhoujun Li, and Ming Zhou 14:00–14:15

We propose a method that can leverage unlabeled data to learn a matching model for response selection in retrieval-based chatbots. The method employs a sequence-to-sequence architecture (Seq2Seq) model as a weak annotator to judge the matching degree of unlabeled pairs, and then performs learning with both the weak signals and the unlabeled data. Experimental results on two public data sets indicate that matching models get significant improvements when they are learned with the proposed method.

Improving Slot Filling in Spoken Language Understanding with Joint Pointer and Attention

Lin Zhao and Zhe Feng 14:15–14:30

We present a generative neural network model for slot filling based on a sequence- to-sequence (Seq2Seq) model together with a pointer network, in the situation where only sentence-level slot annotations are available in the spoken dialogue data. This model predicts slot values by jointly learning to copy a word which may be out-of-vocabulary (OOV) from an input utterance through a pointer network, or generate a word within the vocabulary through an attentional Seq2Seq model. Experimental results show the effective- ness of our slot filling model, especially at addressing the OOV problem. Addi- tionally, we integrate the proposed model into a spoken language understanding sys- tem and achieve the state-of-the-art perfor- mance on the benchmark data.

Large-Scale Multi-Domain Belief Tracking with Knowledge Sharing

Osman Ramadan, Paweł Budzianowski, and Milica Gasic 14:30–14:45

Robust dialogue belief tracking is a key component in maintaining good quality dialogue systems. The tasks that dialogue systems are trying to solve are becoming increasingly complex, requiring scalability to multi-domain, semantically rich dialogues. However, most current approaches have difficulty scaling up with domains because of the dependency of the model parameters on the dialogue ontology. In this paper, a novel approach is introduced that fully utilizes semantic similarity between dialogue utterances and the ontology terms, allowing the information to be shared across domains. The evaluation is performed on a recently collected multi-domain dialogues dataset, one order of magnitude larger than currently available corpora. Our model demonstrates great capability in handling multi-domain dialogues, simultaneously outperforming existing state-of-the-art models in single-domain dialogue tracking tasks.

Modeling discourse cohesion for discourse parsing via memory network

Yanyan Jia, Yuan Ye, Yansong Feng, Yuxuan Lai, Rui Yan, and Dongyan Zhao 14:45–15:00

Identifying long-span dependencies between discourse units is crucial to improve discourse parsing performance. Most existing approaches design sophisticated features or exploit various off-the-shelf tools, but achieve little success. In this paper, we propose a new transition-based discourse parser that makes use of memory networks to take discourse cohesion into account. The automatically captured discourse cohesion benefits discourse parsing, especially for long span scenarios. Experiments on the RST discourse treebank show that our method outperforms traditional featured based methods, and the memory based discourse cohesion can improve the overall parsing performance significantly.

Session 5E: Vision, Linguistics, Resource and Evaluation (Short)

219 *Chair: Gerard de Melo*

SciDTB: Discourse Dependency TreeBank for Scientific Abstracts
An Yang and Sujian Li 14:00–14:15

Annotation corpus for discourse relations benefits NLP tasks such as machine translation and question answering. In this paper, we present SciDTB, a domain-specific discourse treebank annotated on scientific articles. Different from widely-used RST-DT and PDTB, SciDTB uses dependency trees to represent discourse structure, which is flexible and simplified to some extent but do not sacrifice structural integrity. We discuss the labeling framework, annotation workflow and some statistics about SciDTB. Furthermore, our treebank is made as a benchmark for evaluating discourse dependency parsers, on which we provide several baselines as fundamental work.

Predicting accuracy on large datasets from smaller pilot data
Mark Johnson, Peter Anderson, Mark Dras, and Mark Steedman 14:15–14:30

Because obtaining training data is often the most difficult part of an NLP or ML project, we develop methods for predicting how much data is required to achieve a desired test accuracy by extrapolating results from models trained on a small pilot training dataset. We model how accuracy varies as a function of training size on subsets of the pilot data, and use that model to predict how much training data would be required to achieve the desired accuracy. We introduce a new performance extrapolation task to evaluate how well different extrapolations predict accuracy on larger training sets. We show that details of hyperparameter optimisation and the extrapolation models can have dramatic effects in a document classification task. We believe this is an important first step in developing methods for estimating the resources required to meet specific engineering performance targets.

The Influence of Context on Sentence Acceptability Judgements
Jean-Philippe Bernardy, Shalom Lappin, and Jey Han Lau 14:30–14:45

We investigate the influence that document context exerts on human acceptability judgements for English sentences, via two sets of experiments. The first compares ratings for sentences presented on their own with ratings for the same set of sentences given in their document contexts. The second assesses the accuracy with which two types of neural models — one that incorporates context during training and one that does not — predict these judgements. Our results indicate that: (1) context improves acceptability ratings for ill-formed sentences, but also reduces them for well-formed sentences; and (2) context helps unsupervised systems to model acceptability.

Do Neural Network Cross-Modal Mappings Really Bridge Modalities?
Guillem Collell and Marie-Francine Moens 14:45–15:00

Feed-forward networks are widely used in cross-modal applications to bridge modalities by mapping distributed vectors of one modality to the other, or to a shared space. The predicted vectors are then used to perform e.g., retrieval or labeling. Thus, the success of the whole system relies on the ability of the mapping to make the neighborhood structure (i.e., the pairwise similarities) of the predicted vectors akin to that of the target vectors. However, whether this is achieved has not been investigated yet. Here, we propose a new similarity measure and two ad hoc experiments to shed light on this issue. In three cross-modal benchmarks we learn a large number of language-to-vision and vision-to-language neural network mappings (up to five layers) using a rich diversity of image and text features and loss functions. Our results reveal that, surprisingly, the neighborhood structure of the predicted vectors consistently resembles more that of the input vectors than that of the target vectors. In a second experiment, we further show that untrained nets do not significantly disrupt the neighborhood (i.e., semantic) structure of the input vectors.

Session 5F: Parsing, Morphology (Short)

220 *Chair: Emily Pitler*

Policy Gradient as a Proxy for Dynamic Oracles in Constituency Parsing
Daniel Fried and Dan Klein 14:00–14:15

Dynamic oracles provide strong supervision for training constituency parsers with exploration, but must be custom defined for a given parser's transition system. We explore using a policy gradient method as a parser-agnostic alternative. In addition to directly optimizing for a tree-level metric such as F1, policy gradient has the potential to reduce exposure bias by allowing exploration during training; moreover, it does not require a dynamic oracle for supervision. On four constituency parsers in three languages, the method substantially outperforms static oracle likelihood training in almost all settings. For parsers where a dynamic oracle is available (including a novel oracle which we define for the transition system of Dyer et al., 2016), policy gradient typically recaptures a substantial fraction of the performance gain afforded by the dynamic oracle.

Linear-time Constituency Parsing with RNNs and Dynamic Programming
Juneki Hong and Liang Huang 14:15–14:30

Recently, span-based constituency parsing has achieved competitive accuracies with extremely simple models by using bidirectional RNNs to model "spans". However, the minimal span parser of Stern et al. (2017a) which holds the current state of the art accuracy is a chart parser running in cubic time, $O(n^3)$, which is too slow for longer sentences and for applications beyond sentence boundaries such as end-to-end discourse parsing and joint sentence boundary detection and parsing. We propose a linear-time constituency parser with RNNs and dynamic programming using graph-structured stack and beam search, which runs in time $O(nb^2)$ where b is the beam size. We further speed this up to $O(nblogb)$ by integrating cube pruning. Compared with chart parsing baselines, this linear-time parser is substantially faster for long sentences on the Penn Treebank and orders of magnitude faster for discourse parsing, and achieves the highest F1 accuracy on the Penn Treebank among single model end-to-end systems.

Simpler but More Accurate Semantic Dependency Parsing
Timothy Dozat and Christopher D. Manning 14:30–14:45

While syntactic dependency annotations concentrate on the surface or functional structure of a sentence, semantic depen- dency annotations aim to capture between- word relationships that are more closely related to the meaning of a sentence, using graph-structured representations. We ex- tend the LSTM-based syntactic parser of Dozat and Manning (2017) to train on and generate these graph structures. The re- sulting system on its own achieves state-of-the-art performance, beating the pre- vious, substantially more complex state- of-the-art system by 0.6% labeled F1. Adding linguistically richer input repre- sentations pushes the margin even higher, allowing us to beat it by 1.9% labeled F1.

Simplified Abugidas
Chenchen Ding, Masao Utiyama, and Eiichiro Sumita 14:45–15:00

An abugida is a writing system where the consonant letters represent syllables with a default vowel and other vowels are denoted by diacritics. We investigate the feasibility of recovering the original text written in an abugida after omitting subordinate diacritics and merging consonant letters with similar phonetic values. This is crucial for developing more efficient input methods by reducing the complexity in abugidas. Four abugidas in the southern Brahmic family, i.e., Thai, Burmese, Khmer, and Lao, were studied using a newswire 20,000-sentence dataset. We compared the recovery performance of a support vector machine and an LSTM-based recurrent neural network, finding that the abugida graphemes could be recovered with 94% - 97% accuracy at the top-1 level and 98% - 99% at the top-4 level, even after omitting most diacritics (10 - 30 types) and merging the remaining 30 - 50 characters into 21 graphemes.

Session 6 Overview – Tuesday, July 17, 2018

	Track A *Semantic Parsing 2* Plenary	**Track B** *Machine Learning 2* 203–204	**Track C** *Question Answering 2* 210–211	**Track D** *Generation 2* 212–213	**Track E** *Social Media* 219	**Track F** *Information Retrieval* 220
15:30	Weakly Supervised Semantic Parsing with Abstract Examples *Goldman, Latcinnik, Nave, Globerson, and Berant*	Batch IS NOT Heavy: Learning Word Representations From All Samples *Xin, Yuan, He, and Jose*	Did the Model Understand the Question? *Mudrakarta, Taly, Sundararajan, and Dhamdhere*	Language Generation via DAG Transduction *Ye, Sun, and Wan*	Stock Movement Prediction from Tweets and Historical Prices *Xu and Cohen*	Semi-supervised User Geolocation via Graph Convolutional Networks *Rahimi, Cohn, and Baldwin*
15:55	Improving a Neural Semantic Parser by Counterfactual Learning from Human Bandit Feedback *Lawrence and Riezler*	Backpropagating through Structured Argmax using a SPIGOT *Peng, Thomson, and Smith*	Harvesting Paragraph-level Question-Answer Pairs from Wikipedia *Du and Cardie*	A Distributional and Orthographic Aggregation Model for English Derivational Morphology *Deutsch, Hewitt, and Roth*	Rumor Detection on Twitter with Tree-structured Recursive Neural Networks *Ma, Gao, and Wong*	Document Modeling with External Attention for Sentence Extraction *Narayan, Cardenas, Papasarantopoulos, Cohen, Lapata, Yu, and Chang*
16:20	AMR dependency parsing with a typed semantic algebra *Groschwitz, Lindemann, Fowlie, Johnson, and Koller*	Learning How to Actively Learn: A Deep Imitation Learning Approach *Liu, Buntine, and Haffari*	Multi-Passage Machine Reading Comprehension with Cross-Passage Answer Verification *Wang, Liu, Liu, He, Lyu, Wu, Li, and Wang*	Deep-speare: A joint neural model of poetic language, meter and rhyme *Lau, Cohn, Baldwin, Brooke, and Hammond*	Visual Attention Model for Name Tagging in Multimodal Social Media *Lu, Neves, Carvalho, Zhang, and Ji*	Neural Models for Documents with Metadata *Card, Tan, and Smith*
16:45	Sequence-to-sequence Models for Cache Transition Systems *Peng, Song, Gildea, and Satta*	Training Classifiers with Natural Language Explanations *Hancock, Varma, Wang, Bringmann, Liang, and Ré*	[TACL] The NarrativeQA Reading Comprehension Challenge *Kočiský, Schwarz, Blunsom, Dyer, Hermann, Melis, and Grefenstette*	NeuralREG: An end-to-end approach to referring expression generation *Castro Ferreira, Moussallem, Kádár, Wubben, and Krahmer*	Multimodal Named Entity Disambiguation for Noisy Social Media Posts *Moon, Neves, and Carvalho*	NASH: Toward End-to-End Neural Architecture for Generative Semantic Hashing *Shen, Su, Chapfuwa, Wang, Wang, Henao, and Carin*

Parallel Session 6

Session 6A: Semantic Parsing 2

Plenary *Chair: Luke Zettlemoyer*

Weakly Supervised Semantic Parsing with Abstract Examples
Omer Goldman, Veronica Latcinnik, Ehud Nave, Amir Globerson, and Jonathan Berant 15:30–15:55

Training semantic parsers from weak supervision (denotations) rather than strong supervision (programs) complicates training in two ways. First, a large search space of potential programs needs to be explored at training time to find a correct program. Second, spurious programs that accidentally lead to a correct denotation add noise to training. In this work we propose that in closed worlds with clear semantic types, one can substantially alleviate these problems by utilizing an abstract representation, where tokens in both the language utterance and program are lifted to an abstract form. We show that these abstractions can be defined with a handful of lexical rules and that they result in sharing between different examples that alleviates the difficulties in training. To test our approach, we develop the first semantic parser for CNLVR, a challenging visual reasoning dataset, where the search space is large and overcoming spuriousness is critical, because denotations are either TRUE or FALSE, and thus random programs are likely to lead to a correct denotation. Our method substantially improves performance, and reaches 82.5% accuracy, a 14.7% absolute accuracy improvement compared to the best reported accuracy so far.

Improving a Neural Semantic Parser by Counterfactual Learning from Human Bandit Feedback
Carolin Lawrence and Stefan Riezler 15:55–16:20

Counterfactual learning from human bandit feedback describes a scenario where user feedback on the quality of outputs of a historic system is logged and used to improve a target system. We show how to apply this learning framework to neural semantic parsing. From a machine learning perspective, the key challenge lies in a proper reweighting of the estimator so as to avoid known degeneracies in counterfactual learning, while still being applicable to stochastic gradient optimization. To conduct experiments with human users, we devise an easy-to-use interface to collect human feedback on semantic parses. Our work is the first to show that semantic parsers can be improved significantly by counterfactual learning from logged human feedback data.

AMR dependency parsing with a typed semantic algebra
Jonas Groschwitz, Matthias Lindemann, Meaghan Fowlie, Mark Johnson, and Alexander Koller 16:20–16:45

We present a semantic parser for Abstract Meaning Representations which learns to parse strings into tree representations of the compositional structure of an AMR graph. This allows us to use standard neural techniques for supertagging and dependency tree parsing, constrained by a linguistically principled type system. We present two approximative decoding algorithms, which achieve state-of-the-art accuracy and outperform strong baselines.

Sequence-to-sequence Models for Cache Transition Systems
Xiaochang Peng, Linfeng Song, Daniel Gildea, and Giorgio Satta 16:45–17:10

In this paper, we present a sequence-to-sequence based approach for mapping natural language sentences to AMR semantic graphs. We transform the sequence to graph mapping problem to a word sequence to transition action sequence problem using a special transition system called a cache transition system. To address the sparsity issue of neural AMR parsing, we feed feature embeddings from the transition state to provide relevant local information for each decoder state. We present a monotonic hard attention model for the transition framework to handle the strictly left-to-right alignment between each transition state and the current buffer input focus. We evaluate our neural transition model on the AMR parsing task, and our parser outperforms other sequence-to-sequence approaches and achieves competitive results in comparison with the best-performing models.

Session 6B: Machine Learning 2

203–204 *Chair: Yi Yang*

Batch IS NOT Heavy: Learning Word Representations From All Samples
Xin Xin, Fajie Yuan, Xiangnan He, and Joemon M Jose 15:30–15:55

Stochastic Gradient Descent (SGD) with negative sampling is the most prevalent approach to learn word representations. However, it is known that sampling methods are biased especially when the sampling distribution deviates from the true data distribution. Besides, SGD suffers from dramatic fluctuation due to the one-sample learning scheme. In this work, we propose AllVec that uses batch gradient learning to generate word representations from all training samples. Remarkably, the time complexity of AllVec remains at the same level as SGD, being determined by the number of positive samples rather than all samples. We evaluate AllVec on several benchmark tasks. Experiments show that AllVec outperforms sampling-based SGD methods with comparable efficiency, especially for small training corpora.

Backpropagating through Structured Argmax using a SPIGOT
Hao Peng, Sam Thomson, and Noah A. Smith 15:55–16:20

We introduce structured projection of intermediate gradients (SPIGOT), a new method for backpropagating through neural networks that include hard-decision structured predictions (e.g., parsing) in intermediate layers. SPIGOT requires no marginal inference, unlike structured attention networks and reinforcement learning-inspired solutions. Like so-called straight-through estimators, SPIGOT defines gradient-like quantities associated with intermediate nondifferentiable operations, allowing backpropagation before and after them; SPIGOT's proxy aims to ensure that, after a parameter update, the intermediate structure will remain well-formed. We experiment on two structured NLP pipelines: syntactic-then-semantic dependency parsing, and semantic parsing followed by sentiment classification. We show that training with SPIGOT leads to a larger improvement on the downstream task than a modularly-trained pipeline, the straight-through estimator, and structured attention, reaching a new state of the art on semantic dependency parsing.

Learning How to Actively Learn: A Deep Imitation Learning Approach
Ming Liu, Wray Buntine, and Gholamreza Haffari 16:20–16:45

Heuristic-based active learning (AL) methods are limited when the data distribution of the underlying learning problems vary. We introduce a method that learns an AL "policy" using "imitation learning" (IL). Our IL-based approach makes use of an efficient and effective "algorithmic expert", which provides the policy learner with good actions in the encountered AL situations. The AL strategy is then learned with a feedforward network, mapping situations to most informative query datapoints. We evaluate our method on two different tasks: text classification and named entity recognition. Experimental results show that our IL-based AL strategy is more effective than strong previous methods using heuristics and reinforcement learning.

Training Classifiers with Natural Language Explanations
Braden Hancock, Paroma Varma, Stephanie Wang, Martin Bringmann, Percy Liang, and Christopher Ré
16:45–17:10

Training accurate classifiers requires many labels, but each label provides only limited information (one bit for binary classification). In this work, we propose BabbleLabble, a framework for training classifiers in which an annotator provides a natural language explanation for each labeling decision. A semantic parser converts these explanations into programmatic labeling functions that generate noisy labels for an arbitrary amount of unlabeled data, which is used to train a classifier. On three relation extraction tasks, we find that users are able to train classifiers with comparable F1 scores from 5-100 faster by providing explanations instead of just labels. Furthermore, given the inherent imperfection of labeling functions, we find that a simple rule-based semantic parser suffices.

Session 6C: Question Answering 2

210–211 *Chair: Richard Socher*

Did the Model Understand the Question?
Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere 15:30–15:55

We analyze state-of-the-art deep learning models for three tasks: question answering on (1) images, (2) tables, and (3) passages of text. Using the notion of "attribution" (word importance), we find that these deep networks often ignore important question terms. Leveraging such behavior, we perturb questions to craft a variety of adversarial examples. Our strongest attacks drop the accuracy of a visual question answering model from 61.1% to 19%, and that of a tabular question answering model from 33.5% to 3.3%. Additionally, we show how attributions can strengthen attacks proposed by Jia and Liang (2017) on paragraph comprehension models. Our results demonstrate that attributions can augment standard measures of accuracy and empower investigation of model performance. When a model is accurate but for the wrong reasons, attributions can surface erroneous logic in the model that indicates inadequacies in the test data.

Harvesting Paragraph-level Question-Answer Pairs from Wikipedia
Xinya Du and Claire Cardie 15:55–16:20

We study the task of generating from Wikipedia articles question-answer pairs that cover content beyond a single sentence. We propose a neural network approach that incorporates coreference knowledge via a novel gating mechanism. As compared to models that only take into account sentence-level information (Heilman and Smith, 2010; Du et al., 2017; Zhou et al., 2017), we find that the linguistic knowledge introduced by the coreference representation aids question generation significantly, producing models that outperform the current state-of-the-art. We apply our system (composed of an answer span extraction system and the passage-level QG system) to the 10,000 top ranking Wikipedia articles and create a corpus of over one million question-answer pairs. We provide qualitative analysis for the this large-scale generated corpus from Wikipedia.

Multi-Passage Machine Reading Comprehension with Cross-Passage Answer Verification
Yizhong Wang, Kai Liu, Jing Liu, Wei He, Yajuan Lyu, Hua Wu, Sujian Li, and Haifeng Wang 16:20–16:45

Machine reading comprehension (MRC) on real web data usually requires the machine to answer a question by analyzing multiple passages retrieved by search engine. Compared with MRC on a single passage, multi-passage MRC is more challenging, since we are likely to get multiple confusing answer candidates from different passages. To address this problem, we propose an end-to-end neural model that enables those answer candidates from different passages to verify each other based on their content representations. Specifically, we jointly train three modules that can predict the final answer based on three factors: the answer boundary, the answer content and the cross-passage answer verification. The experimental results show that our method outperforms the baseline by a large margin and achieves the state-of-the-art performance on the English MS-MARCO dataset and the Chinese DuReader dataset, both of which are designed for MRC in real-world settings.

[TACL] The NarrativeQA Reading Comprehension Challenge
Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette 16:45–17:10

Reading comprehension (RC)—in contrast to information retrieval—requires integrating information and reasoning about events, entities, and their relations across a full document. Question answering is conventionally used to assess RC ability, in both artificial agents and children learning to read. However, existing RC datasets and tasks are dominated by questions that can be solved by selecting answers using superficial information (e.g., local context similarity or global term frequency); they thus fail to test for the essential integrative aspect of RC. To encourage progress on deeper comprehension of language, we present a new dataset and set of tasks in which the reader must answer questions about stories by reading entire books or movie scripts. These tasks are designed so that successfully answering their questions requires understanding the underlying narrative rather than relying on shallow pattern matching or salience. We show that although humans solve the tasks easily, standard RC models struggle on the tasks presented here. We provide an analysis of the dataset and the challenges it presents.

Session 6D: Generation 2

212–213 *Chair: Jackie Chi Kit Cheung*

Language Generation via DAG Transduction
Yajie Ye, Weiwei Sun, and Xiaojun Wan 15:30–15:55

A DAG automaton is a formal device for manipulating graphs. By augmenting a DAG automaton with transduction rules, a DAG transducer has potential applications in fundamental NLP tasks. In this paper, we propose a novel DAG transducer to perform graph-to-program transformation. The target structure of our transducer is a program licensed by a declarative programming language rather than linguistic structures. By executing such a program, we can easily get a surface string. Our transducer is designed especially for natural language generation (NLG) from type-logical semantic graphs. Taking Elementary Dependency Structures, a format of English Resource Semantics, as input, our NLG system achieves a BLEU-4 score of 68.07. This remarkable result demonstrates the feasibility of applying a DAG transducer to resolve NLG, as well as the effectiveness of our design.

A Distributional and Orthographic Aggregation Model for English Derivational Morphology
Daniel Deutsch, John Hewitt, and Dan Roth 15:55–16:20

Modeling derivational morphology to generate words with particular semantics is useful in many text generation tasks, such as machine translation or abstractive question answering. In this work, we tackle the task of derived word generation. That is, we attempt to generate the word "runner" for "someone who runs." We identify two key problems in generating derived words from root words and transformations. We contribute a novel aggregation model of derived word generation that learns derivational transformations both as orthographic functions using sequence-to-sequence models and as functions in distributional word embedding space. The model then learns to choose between the hypothesis of each system. We also present two ways of incorporating corpus information into derived word generation.

Deep-speare: A joint neural model of poetic language, meter and rhyme
Jey Han Lau, Trevor Cohn, Timothy Baldwin, Julian Brooke, and Adam Hammond 16:20–16:45

In this paper, we propose a joint architecture that captures language, rhyme and meter for sonnet modelling. We assess the quality of generated poems using crowd and expert judgements. The stress and rhyme models perform very well, as generated poems are largely indistinguishable from human-written poems. Expert evaluation, however, reveals that a vanilla language model captures meter implicitly, and that machine-generated poems still underperform in terms of readability and emotion. Our research shows the importance expert evaluation for poetry generation, and that future research should look beyond rhyme/meter and focus on poetic language.

NeuralREG: An end-to-end approach to referring expression generation
Thiago Castro Ferreira, Diego Moussallem, Ákos Kádár, Sander Wubben, and Emiel Krahmer 16:45–17:10

Traditionally, Referring Expression Generation (REG) models first decide on the form and then on the content of references to discourse entities in text, typically relying on features such as salience and grammatical function. In this paper, we present a new approach (NeuralREG), relying on deep neural networks, which makes decisions about form and content in one go without explicit feature extraction. Using a delexicalized version of the WebNLG corpus, we show that the neural model substantially improves over two strong baselines.

Session 6E: Social Media

219 *Chair: David Jurgens*

Stock Movement Prediction from Tweets and Historical Prices
Yumo Xu and Shay B. Cohen 15:30–15:55

Stock movement prediction is a challenging problem: the market is highly stochastic, and we make temporally-dependent predictions from chaotic data. We treat these three complexities and present a novel deep generative model jointly exploiting text and price signals for this task. Unlike the case with discriminative or topic modeling, our model introduces recurrent, continuous latent variables for a better treatment of stochasticity, and uses neural variational inference to address the intractable posterior inference. We also provide a hybrid objective with temporal auxiliary to flexibly capture predictive dependencies. We demonstrate the state-of-the-art performance of our proposed model on a new stock movement prediction dataset which we collected.

Rumor Detection on Twitter with Tree-structured Recursive Neural Networks
Jing Ma, Wei Gao, and Kam-Fai Wong 15:55–16:20

Automatic rumor detection is technically very challenging. In this work, we try to learn discriminative features from tweets content by following their non-sequential propagation structure and generate more powerful representations for identifying different type of rumors. We propose two recursive neural models based on a bottom-up and a top-down tree-structured neural networks for rumor representation learning and classification, which naturally conform to the propagation layout of tweets. Results on two public Twitter datasets demonstrate that our recursive neural models 1) achieve much better performance than state-of-the-art approaches; 2) demonstrate superior capacity on detecting rumors at very early stage.

Visual Attention Model for Name Tagging in Multimodal Social Media
Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji 16:20–16:45

Everyday billions of multimodal posts containing both images and text are shared in social media sites such as Snapchat, Twitter or Instagram. This combination of image and text in a single message allows for more creative and expressive forms of communication, and has become increasingly common in such sites. This new paradigm brings new challenges for natural language understanding, as the textual component tends to be shorter, more informal, and often is only understood if combined with the visual context. In this paper, we explore the task of name tagging in multimodal social media posts. We start by creating two new multimodal datasets: the first based on Twitter posts and the second based on Snapchat captions (exclusively submitted to public and crowd-sourced stories). We then propose a novel model architecture based on Visual Attention that not only provides deeper visual understanding on the decisions of the model, but also significantly outperforms other state-of-the-art baseline methods for this task.

Multimodal Named Entity Disambiguation for Noisy Social Media Posts
Seungwhan Moon, Leonardo Neves, and Vitor Carvalho 16:45–17:10

We introduce the new Multimodal Named Entity Disambiguation (MNED) task for multimodal social media posts such as Snapchat or Instagram captions, which are composed of short captions with accompanying images. Social media posts bring significant challenges for disambiguation tasks because 1) ambiguity not only comes from polysemous entities, but also from inconsistent or incomplete notations, 2) very limited context is provided with surrounding words, and 3) there are many emerging entities often unseen during training. To this end, we build a new dataset called SnapCaptionsKB, a collection of Snapchat image captions submitted to public and crowd-sourced stories, with named entity mentions fully annotated and linked to entities in an external knowledge base. We then build a deep zeroshot multimodal network for MNED that 1) extracts contexts from both text and image, and 2) predicts correct entity in the knowledge graph embeddings space, allowing for zeroshot disambiguation of entities unseen in training set as well. The proposed model significantly outperforms the state-of-the-art text-only NED models, showing efficacy and potentials of the MNED task.

Session 6F: Information Retrieval

220 *Chair: Hang Li*

Semi-supervised User Geolocation via Graph Convolutional Networks
Afshin Rahimi, Trevor Cohn, and Timothy Baldwin 15:30–15:55

Social media user geolocation is vital to many applications such as event detection. In this paper, we propose GCN, a multiview geolocation model based on Graph Convolutional Networks, that uses both text and network context. We compare GCN to the state-of-the-art, and to two baselines we propose, and show that our model achieves or is competitive with the state-of-the-art over three benchmark geolocation datasets when sufficient supervision is available. We also evaluate GCN under a minimal supervision scenario, and show it outperforms baselines. We find that high- way network gates are essential for control- ling the amount of useful neighbourhood expansion in GCN.

Document Modeling with External Attention for Sentence Extraction
Shashi Narayan, Ronald Cardenas, Nikos Papasarantopoulos, Shay B. Cohen, Mirella Lapata, Jiangsheng Yu, and Yi Chang 15:55–16:20

Document modeling is essential to a variety of natural language understanding tasks. We propose to use external information to improve document modeling for problems that can be framed as sentence extraction. We develop a framework composed of a hierarchical document encoder and an attention-based extractor with attention over external information. We evaluate our model on extractive document summarization (where the external information is image captions and the title of the document) and answer selection (where the external information is a question). We show that our model consistently outperforms strong baselines, in terms of both informativeness and fluency (for CNN document summarization) and achieves state-of-the-art results for answer selection on WikiQA and NewsQA.

Neural Models for Documents with Metadata
Dallas Card, Chenhao Tan, and Noah A. Smith 16:20–16:45

Most real-world document collections involve various types of metadata, such as author, source, and date, and yet the most commonly-used approaches to modeling text corpora ignore this information. While specialized models have been developed for particular applications, few are widely used in practice, as customization typically requires derivation of a custom inference algorithm. In this paper, we build on recent advances in variational inference methods and propose a general neural framework, based on topic models, to enable flexible incorporation of metadata and allow for rapid exploration of alternative models. Our approach achieves strong performance, with a manageable tradeoff between perplexity, coherence, and sparsity. Finally, we demonstrate the potential of our framework through an exploration of a corpus of articles about US immigration.

NASH: Toward End-to-End Neural Architecture for Generative Semantic Hashing
Dinghan Shen, Qinliang Su, Paidamoyo Chapfuwa, Wenlin Wang, Guoyin Wang, Ricardo Henao, and Lawrence Carin 16:45–17:10

Semantic hashing has become a powerful paradigm for fast similarity search in many information retrieval systems. While fairly successful, previous techniques generally require two-stage training, and the binary constraints are handled *ad-hoc*. In this paper, we present an *end-to-end* Neural Architecture for Semantic Hashing (NASH), where the binary hashing codes are treated as *Bernoulli* latent variables. A neural variational inference framework is proposed for training, where gradients are directly backpropagated through the discrete latent variable to optimize the hash function. We also draw the connections between proposed method and *rate-distortion theory*, which provides a theoretical foundation for the effectiveness of our framework. Experimental results on three public datasets demonstrate that our method significantly outperforms several state-of-the-art models on both *unsupervised* and *supervised* scenarios.

Social Event

Tuesday, July 17, 2018, 7:30pm–10:30pm

Melbourne Sealife Aquarium
King St, Melbourne
VIC 3000
`https://www.melbourneaquarium.com.au/`

In keeping with recent ACLs, the ACL 2018 Social Event will be held in an aquarium. The Melbourne Sealife Aquarium is a short 10 minute walk upriver along the Yarra from the convention centre (see map on page 216). The event starts at 19:30 and you will be served deserts and drinks, including an open bar. Be sure to see the massive 'Mermaid Garden', a walk-through glass passageway amongst hundreds of sharks and thousands more marine creatures, just a taste of the extraordinary marine diversity of Australia. Bring your partners and children to enjoy this spectacular place.

Note that dinner is not provided, and there is a limited time for dinner between the end of the conference programme and the start of the social event. Please plan accordingly. Note there are many dinner venues along the river foreshore as part of the South Wharf, Casino (which is directly opposite the Acquarium), and Southbank precincts.

5

Main Conference: Wednesday, July 18

Overview

08:30–17:00 **Registration** *Level 2 Foyer*

09:00–10:00 **Invited Talk 2: Anton van den Hengel** (sponsors: ByteDance & Tencent) *Plenary*

10:00–10:30 **Coffee Break** *Level 2 Foyer and Melbourne Room*

Session 7

10:30–12:10	Semantic Parsing 3	Language / Document Model	Information Extraction 3	Dialog System 3	Multimodal	Discourse
	Plenary	*203–204*	*210–211*	*212–213*	*219*	*220*

12:10–12:30 **Short Break**

12:30–14:00 **Poster Session 3** *Melbourne Room 1 & 2*

12:30–14:00 **Demo Poster Session 3** *Melbourne Room 1 & 2*

Session 8

14:00–15:00	Semantics 2 (Short)	Machine Translation, Multilinguality 2 (Short)	Information Extraction 2 (Short)	Generation, Summarization (Short)	Machine Learning, Question Answering (Short)	Sentiment (Short)
	Plenary	*203–204*	*210–211*	*212–213*	*219*	*220*

15:00–15:30 **Coffee Break** *Level 2 Foyer and Melbourne Room*

15:30–17:15 **Session 9: Best Paper Session** (sponsors: Apple & Amazon) *Plenary*

17:15–17:30 **Short Break**

17:30–18:30 **Lifetime Achievement Award** (sponsors: Google & JD.com) *Plenary*

18:30–18:45 **Closing Session** *Plenary*

Keynote Address: Anton van den Hengel (sponsored by ByteDance & Tencent)

Deep Neural Networks, and What They're Not Very Good at

Wednesday, July 18, 2018, 9:00am–10:00am

Plenary

Abstract: Deep Neural Networks have had an incredible impact in a variety of areas within machine learning, including computer vision and natural language processing. Deep Neural Networks use implicit representations that are very high-dimensional, however, and are thus particularly well suited to problems that can be solved by associative recall of previous solutions. They are ill-suited to problems that require human-interpretable representations, explicit manipulation of symbols, or reasoning. The dependency of Deep Neural Networks on large volumes of training data, also means that they are typically only applicable when the problem itself, and the nature of the test data, are predictable long in advance.

The application of Deep Neural Networks to Visual Question Answering has achieved results that would have been thought impossible only a few years ago. It has also thrown a spotlight on the shortcomings of current Deep Nets in solving problems that require explicit reasoning, the use of a knowledge base, or the ability to learn on the fly. In this talk I will illustrate some of the steps being taken to address these problems, and a new learning-to-learn approach that we hope will combine the power of Deep Learning with the significant benefits of explicit-reasoning-based methods.

Biography: Anton van den Hengel is a Professor in the School of Computer Science at the University of Adelaide, the Director of the Australian Institute for Machine Learning, and a Chief Investigator of the Australian Centre for Robotic Vision. Prof. van den Hengel has been a CI on over $60m in external research funding from sources including Google, Canon, BHP Billiton and the ARC, and has won a number of awards, including the Pearcey Foundation Entrepreneur Award, the SA Science Excellence Award for Research Collaboration, and the CVPR Best Paper prize in 2010. He has authored over 300 publications, had 8 patents commercialised, formed 2 start-ups, and has recently had a medical technology achieve first-in-class FDA approval. Current research interests include Deep Learning, vison and language problems, interactive image-based modelling, large-scale video surveillance, and learning from large image databases.

Session 7 Overview – Wednesday, July 18, 2018

	Track A *Semantic Parsing 3* Plenary	**Track B** *Language / Document Model* 203–204	**Track C** *Information Extraction 3* 210–211	**Track D** *Dialog System 3* 212–213	**Track E** *Multimodal* 219	**Track F** *Discourse* 220
10:30	Large-Scale QA-SRL Parsing *FitzGerald, Michael, He, and Zettlemoyer*	Token-level and sequence-level loss smoothing for RNN language models *ELBAYAD, Besacier, and Verbeek*	Robust Distant Supervision Relation Extraction via Deep Reinforcement Learning *Qin, XU, and Wang*	Deep Dyna-Q: Integrating Planning for Task-Completion Dialogue Policy Learning *Peng, Li, Gao, Liu, and Wong*	Multimodal Affective Analysis Using Hierarchical Attention Strategy with Word-Level Alignment *Gu, Yang, Fu, Chen, Li, and Marsic*	Discourse Coherence: Concurrent Explicit and Implicit Relations *Rohde, Johnson, Schneider, and Webber*
10:55	Syntax for Semantic Role Labeling, To Be, Or Not To Be *He, Li, Zhao, and Bai*	Numeracy for Language Models: Evaluating and Improving their Ability to Predict Numbers *Spithourakis and Riedel*	Interpretable and Compositional Relation Learning by Joint Training with an Autoencoder *Takahashi, Tian, and Inui*	Learning to Ask Questions in Open-domain Conversational Systems with Typed Decoders *Wang, Liu, Huang, and Nie*	Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph *Bagher Zadeh, Liang, Poria, Cambria, and Morency*	A Spatial Model for Extracting and Visualizing Latent Discourse Structure in Text *Srivastava and Jojic*
11:20	Situated Mapping of Sequential Instructions to Actions with Single-step Reward Observation *Suhr and Artzi*	To Attend or not to Attend: A Case Study on Syntactic Structures for Semantic Relatedness *Gupta and Zhang*	Zero-Shot Transfer Learning for Event Extraction *Huang, Ji, Cho, Dagan, Riedel, and Voss*	Personalizing Dialogue Agents: I have a dog, do you have pets too? *Zhang, Dinan, Urbanek, Szlam, Kiela, and Weston*	Efficient Low-rank Multimodal Fusion With Modality-Specific Factors *Liu, Shen, Lakshminarasimhan, Liang, Bagher Zadeh, and Morency*	Joint Reasoning for Temporal and Causal Relations *Ning, Feng, Wu, and Roth*
11:45	Marrying Up Regular Expressions with Neural Networks: A Case Study for Spoken Language Understanding *Luo, Feng, Wang, Huang, Yan, and Zhao*	What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties *Conneau, Kruszewski, Lample, Barrault, and Baroni*	Recursive Neural Structural Correspondence Network for Cross-domain Aspect and Opinion Co-Extraction *Wang and Pan*	Efficient Large-Scale Neural Domain Classification with Personalized Attention *Kim, Kim, Kumar, and Sarikaya*	[TACL] Whodunnit? Crime Drama as a Case for Natural Language Understanding *Frermann, Cohen, and Lapata*	Modeling Naive Psychology of Characters in Simple Commonsense Stories *Rashkin, Bosselut, Sap, Knight, and Choi*

Parallel Session 7

Session 7A: Semantic Parsing 3

Plenary *Chair: Lluís Màrquez*

Large-Scale QA-SRL Parsing
Nicholas FitzGerald, Julian Michael, Luheng He, and Luke Zettlemoyer 10:30–10:55

We present a new large-scale corpus of Question-Answer driven Semantic Role Labeling (QA-SRL) annotations, and the first high-quality QA-SRL parser. Our corpus, QA-SRL Bank 2.0, consists of over 250,000 question-answer pairs for over 64,000 sentences across 3 domains and was gathered with a new crowd-sourcing scheme that we show has high precision and good recall at modest cost. We also present neural models for two QA-SRL subtasks: detecting argument spans for a predicate and generating questions to label the semantic relationship. The best models achieve question accuracy of 82.6% and span-level accuracy of 77.6% (under human evaluation) on the full pipelined QA-SRL prediction task. They can also, as we show, be used to gather additional annotations at low cost.

Syntax for Semantic Role Labeling, To Be, Or Not To Be
Shexia He, Zuchao Li, Hai Zhao, and Hongxiao Bai 10:55–11:20

Semantic role labeling (SRL) is dedicated to recognizing the predicate-argument structure of a sentence. Previous studies have shown syntactic information has a remarkable contribution to SRL performance. However, such perception was challenged by a few recent neural SRL models which give impressive performance without a syntactic backbone. This paper intends to quantify the importance of syntactic information to dependency SRL in deep learning framework. We propose an enhanced argument labeling model companying with an extended korder argument pruning algorithm for effectively exploiting syntactic information. Our model achieves state-of-the-art results on the CoNLL-2008, 2009 benchmarks for both English and Chinese, showing the quantitative significance of syntax to neural SRL together with a thorough empirical survey over existing models.

Situated Mapping of Sequential Instructions to Actions with Single-step Reward Observation
Alane Suhr and Yoav Artzi 11:20–11:45

We propose a learning approach for mapping context-dependent sequential instructions to actions. We address the problem of discourse and state dependencies with an attention-based model that considers both the history of the interaction and the state of the world. To train from start and goal states without access to demonstrations, we propose SESTRA, a learning algorithm that takes advantage of single-step reward observations and immediate expected reward maximization. We evaluate on the SCONE domains, and show absolute accuracy improvements of 9.8%-25.3% across the domains over approaches that use high-level logical representations.

Marrying Up Regular Expressions with Neural Networks: A Case Study for Spoken Language Understanding
Bingfeng Luo, Yansong Feng, Zheng Wang, Songfang Huang, Rui Yan, and Dongyan Zhao 11:45–12:10

The success of many natural language processing (NLP) tasks is bound by the number and quality of annotated data, but there is often a shortage of such training data. In this paper, we ask the question: "Can we combine a neural network (NN) with regular expressions (RE) to improve supervised learning for NLP?". In answer, we develop novel methods to exploit the rich expressiveness of REs at different levels within a NN, showing that the combination significantly enhances the learning effectiveness when a small number of training examples are available. We evaluate our approach by applying it to spoken language understanding for intent detection and slot filling. Experimental results show that our approach is highly effective in exploiting the available training data, giving a clear boost to the RE-unaware NN.

Session 7B: Language/Document Model

203–204 *Chair: Chris Dyer*

Token-level and sequence-level loss smoothing for RNN language models
Maha ELBAYAD, Laurent Besacier, and Jakob Verbeek 10:30–10:55

Despite the effectiveness of recurrent neural network language models, their maximum likelihood estimation suffers from two limitations. It treats all sentences that do not match the ground truth as equally poor, ignoring the structure of the output space. Second, it suffers from 'exposure bias': during training tokens are predicted given ground-truth sequences, while at test time prediction is conditioned on generated output sequences. To overcome these limitations we build upon the recent reward augmented maximum likelihood approach that encourages the model to predict sentences that are close to the ground truth according to a given performance metric. We extend this approach to token-level loss smoothing, and propose improvements to the sequence-level smoothing approach. Our experiments on two different tasks, image captioning and machine translation, show that token-level and sequence-level loss smoothing are complementary, and significantly improve results.

Numeracy for Language Models: Evaluating and Improving their Ability to Predict Numbers
Georgios Spithourakis and Sebastian Riedel 10:55–11:20

Numeracy is the ability to understand and work with numbers. It is a necessary skill for composing and understanding documents in clinical, scientific, and other technical domains. In this paper, we explore different strategies for modelling numerals with language models, such as memorisation and digit-by-digit composition, and propose a novel neural architecture that uses a continuous probability density function to model numerals from an open vocabulary. Our evaluation on clinical and scientific datasets shows that using hierarchical models to distinguish numerals from words improves a perplexity metric on the subset of numerals by 2 and 4 orders of magnitude, respectively, over non-hierarchical models. A combination of strategies can further improve perplexity. Our continuous probability density function model reduces mean absolute percentage errors by 18% and 54% in comparison to the second best strategy for each dataset, respectively.

To Attend or not to Attend: A Case Study on Syntactic Structures for Semantic Relatedness
Amulya Gupta and Zhu Zhang 11:20–11:45

With the recent success of Recurrent Neural Networks (RNNs) in Machine Translation (MT), attention mechanisms have become increasingly popular. The purpose of this paper is two-fold; firstly, we propose a novel attention model on Tree Long Short-Term Memory Networks (Tree-LSTMs), a tree-structured generalization of standard LSTM. Secondly, we study the interaction between attention and syntactic structures, by experimenting with three LSTM variants: bidirectional-LSTMs, Constituency Tree-LSTMs, and Dependency Tree-LSTMs. Our models are evaluated on two semantic relatedness tasks: semantic relatedness scoring for sentence pairs (SemEval 2012, Task 6 and SemEval 2014, Task 1) and paraphrase detection for question pairs (Quora, 2017).

What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties
Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni 11:45–12:10

Although much effort has recently been devoted to training high-quality sentence embeddings, we still have a poor understanding of what they are capturing. "Downstream" tasks, often based on sentence classification, are commonly used to evaluate the quality of sentence representations. The complexity of the tasks makes it however difficult to infer what kind of information is present in the representations. We introduce here 10 probing tasks designed to capture simple linguistic features of sentences, and we use them to study embeddings generated by three different encoders trained in eight distinct ways, uncovering intriguing properties of both encoders and training methods.

Session 7C: Information Extraction 3

210–211 *Chair: Sebastian Riedel*

Robust Distant Supervision Relation Extraction via Deep Reinforcement Learning
Pengda Qin, Weiran XU, and William Yang Wang 10:30–10:55

Distant supervision has become the standard method for relation extraction. However, even though it is an efficient method, it does not come at no cost—The resulted distantly-supervised training samples are often very noisy. To combat the noise, most of the recent state-of-the-art approaches focus on selecting one-best sentence or calculating soft attention weights over the set of the sentences of one specific entity pair. However, these methods are suboptimal, and the false positive problem is still a key stumbling bottleneck for the performance. We argue that those incorrectly-labeled candidate sentences must be treated with a hard decision, rather than being dealt with soft attention weights. To do this, our paper describes a radical solution—We explore a deep reinforcement learning strategy to generate the false-positive indicator, where we automatically recognize false positives for each relation type without any supervised information. Unlike the removal operation in the previous studies, we redistribute them into the negative examples. The experimental results show that the proposed strategy significantly improves the performance of distant supervision comparing to state-of-the-art systems.

Interpretable and Compositional Relation Learning by Joint Training with an Autoencoder
Ryo Takahashi, Ran Tian, and Kentaro Inui 10:55–11:20

Embedding models for entities and relations are extremely useful for recovering missing facts in a knowledge base. Intuitively, a relation can be modeled by a matrix mapping entity vectors. However, relations reside on low dimension sub-manifolds in the parameter space of arbitrary matrices — for one reason, composition of two relations M1, M2 may match a third M3 (e.g. composition of relations currency_of_country and country_of_film usually matches currency_of_film_budget), which imposes compositional constraints to be satisfied by the parameters (i.e. M1*M2=M3). In this paper we investigate a dimension reduction technique by training relations jointly with an autoencoder, which is expected to better capture compositional constraints. We achieve state-of-the-art on Knowledge Base Completion tasks with strongly improved Mean Rank, and show that joint training with an autoencoder leads to interpretable sparse codings of relations, helps discovering compositional constraints and benefits from compositional training. Our source code is released at github.com/tianran/glimvec.

Zero-Shot Transfer Learning for Event Extraction
Lifu Huang, Heng Ji, Kyunghyun Cho, Ido Dagan, Sebastian Riedel, and Clare Voss 11:20–11:45

Most previous supervised event extraction methods have relied on features derived from manual annotations, and thus cannot be applied to new event types without extra annotation effort. We take a fresh look at event extraction and model it as a generic grounding problem: mapping each event mention to a specific type in a target event ontology. We design a transferable architecture of structural and compositional neural networks to jointly represent and map event mentions and types into a shared semantic space. Based on this new framework, we can select, for each event mention, the event type which is semantically closest in this space as its type. By leveraging manual annotations available for a small set of existing event types, our framework can be applied to new unseen event types without additional manual annotations. When tested on 23 unseen event types, our zero-shot framework, without manual annotations, achieved performance comparable to a supervised model trained from 3,000 sentences annotated with 500 event mentions.

Recursive Neural Structural Correspondence Network for Cross-domain Aspect and Opinion Co-Extraction
Wenya Wang and Sinno Jialin Pan 11:45–12:10

Fine-grained opinion analysis aims to extract aspect and opinion terms from each sentence for opinion summarization. Supervised learning methods have proven to be effective for this task. However, in many domains, the lack of labeled data hinders the learning of a precise extraction model. In this case, unsupervised domain adaptation methods are desired to transfer knowledge from the source domain to any unlabeled target domain. In this paper, we develop a novel recursive neural network that could reduce domain shift effectively in word level through syntactic relations. We treat these relations as invariant "pivot information" across domains to build structural correspondences and generate an auxiliary task to predict the relation between any two adjacent words in the dependency tree. In the end, we demonstrate state-of-the-art results on three benchmark datasets.

Session 7D: Dialog System 3

212–213 *Chair: Bernardo Magnini*

Deep Dyna-Q: Integrating Planning for Task-Completion Dialogue Policy Learning
Baolin Peng, Xiujun Li, Jianfeng Gao, Jingjing Liu, and Kam-Fai Wong 10:30–10:55

Training a task-completion dialogue agent via reinforcement learning (RL) is costly because it requires many interactions with real users. One common alternative is to use a user simulator. However, a user simulator usually lacks the language complexity of human interlocutors and the biases in its design may tend to degrade the agent. To address these issues, we present Deep Dyna-Q, which to our knowledge is the first deep RL framework that integrates planning for task-completion dialogue policy learning. We incorporate into the dialogue agent a model of the environment, referred to as the world model, to mimic real user response and generate simulated experience. During dialogue policy learning, the world model is constantly updated with real user experience to approach real user behavior, and in turn, the dialogue agent is optimized using both real experience and simulated experience. The effectiveness of our approach is demonstrated on a movie-ticket booking task in both simulated and human-in-the-loop settings.

Learning to Ask Questions in Open-domain Conversational Systems with Typed Decoders
Yansen Wang, Chenyi Liu, Minlie Huang, and Liqiang Nie 10:55–11:20

Asking good questions in open-domain conversational systems is quite significant but rather untouched. This task, substantially different from traditional question generation, requires to question not only with various patterns but also on diverse and relevant topics. We observe that a good question is a natural composition of interrogatives, topic words, and ordinary words. Interrogatives lexicalize the pattern of questioning, topic words address the key information for topic transition in dialogue, and ordinary words play syntactical and grammatical roles in making a natural sentence. We devise two typed decoders (soft typed decoder and hard typed decoder) in which a type distribution over the three types is estimated and the type distribution is used to modulate the final generation distribution. Extensive experiments show that the typed decoders outperform state-of-the-art baselines and can generate more meaningful questions.

Personalizing Dialogue Agents: I have a dog, do you have pets too?
Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston 11:20–11:45

Chit-chat models are known to have several problems: they lack specificity, do not display a consistent personality and are often not very captivating. In this work we present the task of making chit-chat more engaging by conditioning on profile information. We collect data and train models to (i)condition on their given profile information; and (ii) information about the person they are talking to, resulting in improved dialogues, as measured by next utterance prediction. Since (ii) is initially unknown our model is trained to engage its partner with personal topics, and we show the resulting dialogue can be used to predict profile information about the interlocutors.

Efficient Large-Scale Neural Domain Classification with Personalized Attention
Young-Bum Kim, Dongchan Kim, Anjishnu Kumar, and Ruhi Sarikaya 11:45–12:10

In this paper, we explore the task of mapping spoken language utterances to one of thousands of natural language understanding domains in intelligent personal digital assistants (IPDAs). This scenario is observed in mainstream IPDAs in industry that allow third parties to develop thousands of new domains to augment built-in first party domains to rapidly increase domain coverage and overall IPDA capabilities. We propose a scalable neural model architecture with a shared encoder, a novel attention mechanism that incorporates personalization information and domain-specific classifiers that solves the problem efficiently. Our architecture is designed to efficiently accommodate incremental domain additions achieving two orders of magnitude speed up compared to full model retraining. We consider the practical constraints of real-time production systems, and design to minimize memory footprint and runtime latency. We demonstrate that incorporating personalization significantly improves domain classification accuracy in a setting with thousands of overlapping domains.

Session 7E: Multimodal

219 *Chair: Julia Hockenmaier*

Multimodal Affective Analysis Using Hierarchical Attention Strategy with Word-Level Alignment
Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic 10:30–10:55

Multimodal affective computing, learning to recognize and interpret human affect and subjective information from multiple data sources, is still a challenge because: (i) it is hard to extract informative features to represent human affects from heterogeneous inputs; (ii) current fusion strategies only fuse different modalities at abstract levels, ignoring time-dependent interactions between modalities. Addressing such issues, we introduce a hierarchical multimodal architecture with attention and word-level fusion to classify utterance-level sentiment and emotion from text and audio data. Our introduced model outperforms state-of-the-art approaches on published datasets, and we demonstrate that our model is able to visualize and interpret synchronized attention over modalities.

Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph
AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency 10:55–11:20

Analyzing human multimodal language is an emerging area of research in NLP. Intrinsically this language is multimodal (heterogeneous), sequential and asynchronous; it consists of the language (words), visual (expressions) and acoustic (paralinguistic) modalities all in the form of asynchronous coordinated sequences. From a resource perspective, there is a genuine need for large scale datasets that allow for in-depth studies of this form of language. In this paper we introduce CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI), the largest dataset of sentiment analysis and emotion recognition to date. Using data from CMU-MOSEI and a novel multimodal fusion technique called the Dynamic Fusion Graph (DFG), we conduct experimentation to exploit how modalities interact with each other in human multimodal language. Unlike previously proposed fusion techniques, DFG is highly interpretable and achieves competative performance when compared to the previous state of the art.

Efficient Low-rank Multimodal Fusion With Modality-Specific Factors
Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency 11:20–11:45

Multimodal research is an emerging field of artificial intelligence, and one of the main research problems in this field is multimodal fusion. The fusion of multimodal data is the process of integrating multiple unimodal representations into one compact multimodal representation. Previous research in this field has exploited the expressiveness of tensors for multimodal representation. However, these methods often suffer from exponential increase in dimensions and in computational complexity introduced by transformation of input into tensor. In this paper, we propose the Low- rank Multimodal Fusion method, which performs multimodal fusion using low-rank tensors to improve efficiency. We evaluate our model on three different tasks: multimodal sentiment analysis, speaker trait analysis, and emotion recognition. Our model achieves competitive results on all these tasks while drastically reducing computational complexity. Additional experiments also show that our model can perform robustly for a wide range of low-rank settings, and is indeed much more efficient in both training and inference compared to other methods that utilize tensor representations.

[TACL] Whodunnit? Crime Drama as a Case for Natural Language Understanding
Lea Frermann, Shay Cohen, and Mirella Lapata 11:45–12:10

In this paper we argue that crime drama exemplified in television programs such as CSI:Crime Scene Investigation is an ideal testbed for approximating real-world natural language understanding and the complex inferences associated with it. We propose to treat crime drama as a new inference task, capitalizing on the fact that each episode poses the same basic question (i.e., who committed the crime) and naturally provides the answer when the perpetrator is revealed. We develop a new dataset based on CSI episodes, formalize perpetrator identification as a sequence labeling problem, and develop an LSTM-based model which learns from multi-modal data. Experimental results show that an incremental inference strategy is key to making accurate guesses as well as learning from representations fusing textual, visual, and acoustic input.

Session 7F: Discourse

220 *Chair: Anette Frank*

Discourse Coherence: Concurrent Explicit and Implicit Relations
Hannah Rohde, Alexander Johnson, Nathan Schneider, and Bonnie Webber 10:30–10:55

Theories of discourse coherence posit relations between discourse segments as a key feature of coherent text. Our prior work suggests that multiple discourse relations can be simultaneously operative between two segments for reasons not predicted by the literature. Here we test how this joint presence can lead participants to endorse seemingly divergent conjunctions (e.g., BUT and SO) to express the link they see between two segments. These apparent divergences are not symptomatic of participant naivety or bias, but arise reliably from the concurrent availability of multiple relations between segments – some available through explicit signals and some via inference. We believe that these new results can both inform future progress in theoretical work on discourse coherence and lead to higher levels of performance in discourse parsing.

A Spatial Model for Extracting and Visualizing Latent Discourse Structure in Text
Shashank Srivastava and Nebojsa Jojic 10:55–11:20

We present a generative probabilistic model of documents as sequences of sentences, and show that inference in it can lead to extraction of long-range latent discourse structure from a collection of documents. The approach is based on embedding sequences of sentences from longer texts into a 2- or 3-D spatial grids, in which one or two coordinates model smooth topic transitions, while the third captures the sequential nature of the modeled text. A significant advantage of our approach is that the learned models are naturally visualizable and interpretable, as semantic similarity and sequential structure are modeled along orthogonal directions in the grid. We show that the method is effective in capturing discourse structures in narrative text across multiple genres, including biographies, stories, and newswire reports. In particular, our method outperforms or is competitive with state-of-the-art generative approaches on tasks such as predicting the outcome of a story, and sentence ordering.

Joint Reasoning for Temporal and Causal Relations
Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth 11:20–11:45

Understanding temporal and causal relations between events is a fundamental natural language understanding task. Because a cause must occur earlier than its effect, temporal and causal relations are closely related and one relation often dictates the value of the other. However, limited attention has been paid to studying these two relations jointly. This paper presents a joint inference framework for them using constrained conditional models (CCMs). Specifically, we formulate the joint problem as an integer linear programming (ILP) problem, enforcing constraints that are inherent in the nature of time and causality. We show that the joint inference framework results in statistically significant improvement in the extraction of both temporal and causal relations from text.

Modeling Naive Psychology of Characters in Simple Commonsense Stories
Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi 11:45–12:10

Understanding a narrative requires reading between the lines and reasoning about the unspoken but obvious implications about events and people's mental states — a capability that is trivial for humans but remarkably hard for machines. To facilitate research addressing this challenge, we introduce a new annotation framework to explain naive psychology of story characters as fully-specified chains of mental states with respect to motivations and emotional reactions. Our work presents a new large-scale dataset with rich low-level annotations and establishes baseline performance on several new tasks, suggesting avenues for future research.

Poster Session 3

Time: *12:30–14:00* Location: Melbourne Room 1 & 2

Poster Session 3A: Student Research Workshop

Biomedical Document Retrieval for Clinical Decision Support System
Jainisha Sankhavara

The availability of huge amount of biomedical literature have opened up new possibilities to apply Information Retrieval and NLP for mining documents from them. In this work, we are focusing on biomedical document retrieval from literature for clinical decision support systems. We compare statistical and NLP based approaches of query reformulation for biomedical document retrieval. Also, we have modeled the biomedical document retrieval as a learning to rank problem. We report initial results for statistical and NLP based query reformulation approaches and learning to rank approach with future direction of research.

A Computational Approach to Feature Extraction for Identification of Suicidal Ideation in Tweets
Ramit Sawhney, Prachi Manchanda, Raj Singh, and Swati Aggarwal

Technological advancements in the World Wide Web and social networks in particular coupled with an increase in social media usage has led to a positive correlation between the exhibition of Suicidal ideation on websites such as Twitter and cases of suicide. This paper proposes a novel supervised approach for detecting suicidal ideation in content on Twitter. A set of features is proposed for training both linear and ensemble classifiers over a dataset of manually annotated tweets. The performance of the proposed methodology is compared against four baselines that utilize varying approaches to validate its utility. The results are finally summarized by reflecting on the effect of the inclusion of the proposed features one by one for suicidal ideation detection.

BCSAT : A Benchmark Corpus for Sentiment Analysis in Telugu Using Word-level Annotations
Sreekavitha Parupalli, Vijjini Anvesh Rao, and Radhika Mamidi

The presented work aims at generating a systematically annotated corpus that can support the enhancement of sentiment analysis tasks in Telugu using word-level sentiment annotations. From OntoSenseNet, we extracted 11,000 adjectives, 253 adverbs, 8483 verbs and sentiment annotation is done by language experts. We discuss the methodology followed for the polarity annotations and validate the developed resource. This work aims at developing a benchmark corpus, as an extension to SentiWordNet, and baseline accuracy for a model where lexeme annotations are applied for sentiment predictions. The fundamental aim of this paper is to validate and study the possibility of utilizing machine learning algorithms, word-level sentiment annotations in the task of automated sentiment identification. Furthermore, accuracy is improved by annotating the bi-grams extracted from the target corpus.

Reinforced Extractive Summarization with Question-Focused Rewards
Kristjan Arumae and Fei Liu

We investigate a new training paradigm for extractive summarization. Traditionally, human abstracts are used to derive goldstandard labels for extraction units. However, the labels are often inaccurate, because human abstracts and source documents cannot be easily aligned at the word level. In this paper we convert human abstracts to a set of Cloze-style comprehension questions. System summaries are encouraged to preserve salient source content useful for answering questions and share common words with the abstracts. We use reinforcement learning to explore the space of possible extractive summaries and introduce a question-focused reward function to promote concise, fluent, and informative summaries. Our experiments show that the proposed method is effective. It surpasses state-of-the-art systems on the standard summarization dataset.

Graph-based Filtering of Out-of-Vocabulary Words for Encoder-Decoder Models
Satoru Katsumata, Yukio Matsumura, Hayahide Yamagishi, and Mamoru Komachi

Encoder-decoder models typically only employ words that are frequently used in the training corpus because of the computational costs and/or to exclude noisy words. However, this vocabulary set may still include words that interfere with learning in encoder-decoder models. This paper proposes a method for selecting more suitable words for learning encoders by utilizing not only frequency, but also co-occurrence information, which we capture using the HITS algorithm. The proposed method is applied to two tasks: machine translation and grammatical error correction. For Japanese-to-English translation, this method achieved a BLEU score that was 0.56 points more than that of a baseline. It also outperformed the baseline method for English grammatical error correction, with an F-measure that

was 1.48 points higher.

Exploring Chunk Based Templates for Generating a subset of English Text
Nikhilesh Bhatnagar, Manish Shrivastava, and Radhika Mamidi

Natural Language Generation (NLG) is a research task which addresses the automatic generation of natural language text representative of an input non-linguistic collection of knowledge. In this paper, we address the task of the generation of grammatical sentences in an isolated context given a partial bag-of-words which the generated sentence must contain. We view the task as a search problem (a problem of choice) involving combinations of smaller chunk based templates extracted from a training corpus to construct a complete sentence. To achieve that, we propose a fitness function which we use in conjunction with an evolutionary algorithm as the search procedure to arrive at a potentially grammatical sentence (modeled by the fitness score) which satisfies the input constraints.

Trick Me If You Can: Adversarial Writing of Trivia Challenge Questions
Eric Wallace and Jordan Boyd-Graber

Modern question answering systems have been touted as approaching human performance. However, existing question answering datasets are imperfect tests. Questions are written with humans in mind, not computers, and often do not properly expose model limitations. To address this, we develop an adversarial writing setting, where humans interact with trained models and try to break them. This annotation process yields a challenge set, which despite being easy for trivia players to answer, systematically stumps automated question answering systems. Diagnosing model errors on the evaluation data provides actionable insights to explore in developing robust and generalizable question answering systems.

Alignment Analysis of Sequential Segmentation of Lexicons to Improve Automatic Cognate Detection
Pranav A

Ranking functions in information retrieval are often used in search engines to extract the relevant answers to the query. This paper makes use of this notion of information retrieval and applies onto the problem domain of cognate detection. The main contributions of this paper are: (1) positional tokenization, which incorporates the sequential notion; (2) graphical error modelling, which calculates the morphological shifts. The current research work only distinguishes whether a pair of words are cognates or not. However, we also study if we could predict a possible cognate from the given input. Our study shows that language modelling based retrieval functions with positional tokenization and error modelling tend to give better results than competing baselines.

Mixed Feelings: Natural Text Generation with Variable, Coexistent Affective Categories
Lee Kezar

Conversational agents, having the goal of natural language generation, must rely on language models which can integrate emotion into their responses. Recent projects outline models which can produce emotional sentences, but unlike human language, they tend to be restricted to one affective category out of a few. To my knowledge, none allow for the intentional coexistence of multiple emotions on the word or sentence level. Building on prior research which allows for variation in the intensity of a singular emotion, this research proposal outlines an LSTM (Long Short-Term Memory) language model which allows for variation in multiple emotions simultaneously.

Automatic Spelling Correction for Resource-Scarce Languages using Deep Learning
Pravallika Etoori, Manoj Chinnakotla, and Radhika Mamidi

Spelling correction is a well-known task in Natural Language Processing (NLP). Automatic spelling correction is important for many NLP applications like web search engines, text summarization, sentiment analysis etc. Most approaches use parallel data of noisy and correct word mappings from different sources as training data for automatic spelling correction. Indic languages are resource-scarce and do not have such parallel data due to low volume of queries and non-existence of such prior implementations. In this paper, we show how to build an automatic spelling corrector for resource-scarce languages. We propose a sequence-to-sequence deep learning model which trains end-to-end. We perform experiments on synthetic datasets created for Indic languages, Hindi and Telugu, by incorporating the spelling mistakes committed at character level. A comparative evaluation shows that our model is competitive with the existing spell checking and correction techniques for Indic languages.

Automatic Question Generation using Relative Pronouns and Adverbs
Payal Khullar, Konigari Rachna, Mukul Hase, and Manish Shrivastava

This paper presents a system that automatically generates multiple, natural language questions using relative pronouns and relative adverbs from complex English sentences. Our system is syntax-based, runs on dependency parse information of a single-sentence input, and achieves high accuracy in terms of syntactic correctness, semantic adequacy, fluency and uniqueness. One of the key advantages of our system, in comparison with other rule-based ap-

proaches, is that we nearly eliminate the chances of getting a wrong wh- word in the generated question, by fetching the requisite wh-word from the input sentence itself. Depending upon the input, we generate both factoid and descriptive type questions. To the best of our information, the exploitation of wh-pronouns and wh-adverbs to generate questions is novel in the Automatic Question Generation task.

Poster Session 3B: Document Analysis

A Deep Relevance Model for Zero-Shot Document Filtering
Chenliang Li, Wei Zhou, Feng Ji, Yu Duan, and Haiqing Chen

In the era of big data, focused analysis for diverse topics with a short response time becomes an urgent demand. As a fundamental task, information filtering therefore becomes a critical necessity. In this paper, we propose a novel deep relevance model for zero-shot document filtering, named DAZER. DAZER estimates the relevance between a document and a category by taking a small set of seed words relevant to the category. With pre-trained word embeddings from a large external corpus, DAZER is devised to extract the relevance signals by modeling the hidden feature interactions in the word embedding space. The relevance signals are extracted through a gated convolutional process. The gate mechanism controls which convolution filters output the relevance signals in a category dependent manner. Experiments on two document collections of two different tasks (i.e., topic categorization and sentiment analysis) demonstrate that DAZER significantly outperforms the existing alternative solutions, including the state-of-the-art deep relevance ranking models.

Disconnected Recurrent Neural Networks for Text Categorization
Baoxin Wang

Recurrent neural network (RNN) has achieved remarkable performance in text categorization. RNN can model the entire sequence and capture long-term dependencies, but it does not do well in extracting key patterns. In contrast, convolutional neural network (CNN) is good at extracting local and position-invariant features. In this paper, we present a novel model named disconnected recurrent neural network (DRNN), which incorporates position-invariance into RNN. By limiting the distance of information flow in RNN, the hidden state at each time step is restricted to represent words near the current position. The proposed model makes great improvements over RNN and CNN models and achieves the best performance on several benchmark datasets for text categorization.

Joint Embedding of Words and Labels for Text Classification
Guoyin Wang, Chunyuan Li, Wenlin Wang, Yizhe Zhang, Dinghan Shen, Xinyuan Zhang, Ricardo Henao, and Lawrence Carin

Word embeddings are effective intermediate representations for capturing semantic regularities between words, when learning the representations of text sequences. We propose to view text classification as a label-word joint embedding problem: each label is embedded in the same space with the word vectors. We introduce an attention framework that measures the compatibility of embeddings between text sequences and labels. The attention is learned on a training set of labeled samples to ensure that, given a text sequence, the relevant words are weighted higher than the irrelevant ones. Our method maintains the interpretability of word embeddings, and enjoys a built-in ability to leverage alternative sources of information, in addition to input text sequences. Extensive results on the several large text datasets show that the proposed framework outperforms the state-of-the-art methods by a large margin, in terms of both accuracy and speed.

Neural Sparse Topical Coding
Min Peng, Qianqian Xie, Yanchun Zhang, Hua Wang, Xiuzhen Zhang, Jimin Huang, and Gang Tian

Topic models with sparsity enhancement have been proven to be effective at learning discriminative and coherent latent topics of short texts, which is critical to many scientific and engineering applications. However, the extensions of these models require carefully tailored graphical models and re-deduced inference algorithms, limiting their variations and applications. We propose a novel sparsity-enhanced topic model, Neural Sparse Topical Coding (NSTC) base on a sparsity-enhanced topic model called Sparse Topical Coding (STC). It focuses on replacing the complex inference process with the back propagation, which makes the model easy to explore extensions. Moreover, the external semantic information of words in word embeddings is incorporated to improve the representation of short texts. To illustrate the flexibility offered by the neural network based framework, we present three extensions base on NSTC without re-deduced inference algorithms. Experiments on Web Snippet and 20Newsgroups datasets demonstrate that our models outperform existing methods.

Document Similarity for Texts of Varying Lengths via Hidden Topics
Hongyu Gong, Tarek Sakakini, Suma Bhat, and JinJun Xiong

Measuring similarity between texts is an important task for several applications. Available approaches to measure document similarity are inadequate for document pairs that have non-comparable lengths, such as a long document and its summary. This is because of the lexical, contextual and the abstraction gaps between a long document of rich details and its concise summary of abstract information. In this paper, we present a document matching approach to bridge this gap, by comparing the texts in a common space of hidden topics. We evaluate the matching algorithm

on two matching tasks and find that it consistently and widely outperforms strong baselines. We also highlight the benefits of the incorporation of domain knowledge to text matching.

Eyes are the Windows to the Soul: Predicting the Rating of Text Quality Using Gaze Behaviour
Sandeep Mathias, Diptesh Kanojia, Kevin Patel, Samarth Agrawal, Abhijit Mishra, and Pushpak Bhattacharyya

Predicting a reader's rating of text quality is a challenging task that involves estimating different subjective aspects of the text, like structure, clarity, etc. Such subjective aspects are better handled using cognitive information. One such source of cognitive information is gaze behaviour. In this paper, we show that gaze behaviour does indeed help in effectively predicting the rating of text quality. To do this, we first we model text quality as a function of three properties - organization, coherence and cohesion. Then, we demonstrate how capturing gaze behaviour helps in predicting each of these properties, and hence the overall quality, by reporting improvements obtained by adding gaze features to traditional textual features for score prediction. We also hypothesize that if a reader has fully understood the text, the corresponding gaze behaviour would give a better indication of the assigned rating, as opposed to partial understanding. Our experiments validate this hypothesis by showing greater agreement between the given rating and the predicted rating when the reader has a full understanding of the text.

Multi-Input Attention for Unsupervised OCR Correction
Rui Dong and David Smith

We propose a novel approach to OCR post-correction that exploits repeated texts in large corpora both as a source of noisy target outputs for unsupervised training and as a source of evidence when decoding. A sequence-to-sequence model with attention is applied for single-input correction, and a new decoder with multi-input attention averaging is developed to search for consensus among multiple sequences. We design two ways of training the correction model without human annotation, either training to match noisily observed textual variants or bootstrapping from a uniform error model. On two corpora of historical newspapers and books, we show that these unsupervised techniques cut the character and word error rates nearly in half on single inputs and, with the addition of multi-input decoding, can rival supervised methods.

Building Language Models for Text with Named Entities
Md Rizwan Parvez, Saikat Chakraborty, Baishakhi Ray, and Kai-Wei Chang

Text in many domains involves a significant amount of named entities. Predicting the entity names is often challenging for a language model as they appear less frequent on the training corpus. In this paper, we propose a novel and effective approach to building a language model which can learn the entity names by leveraging their entity type information. We also introduce two benchmark datasets based on recipes and Java programming codes, on which we evaluate the proposed model. Experimental results show that our model achieves 52.2% better perplexity in recipe generation and 22.06% on code generation than state-of-the-art language models.

hyperdoc2vec: Distributed Representations of Hypertext Documents
Jialong Han, Yan Song, Wayne Xin Zhao, Shuming Shi, and Haisong Zhang

Hypertext documents, such as web pages and academic papers, are of great importance in delivering information in our daily life. Although being effective on plain documents, conventional text embedding methods suffer from information loss if directly adapted to hyper-documents. In this paper, we propose a general embedding approach for hyper-documents, namely, hyperdoc2vec, along with four criteria characterizing necessary information that hyper-document embedding models should preserve. Systematic comparisons are conducted between hyperdoc2vec and several competitors on two tasks, i.e., paper classification and citation recommendation, in the academic paper domain. Analyses and experiments both validate the superiority of hyperdoc2vec to other models w.r.t. the four criteria.

Entity-Duet Neural Ranking: Understanding the Role of Knowledge Graph Semantics in Neural Information Retrieval
Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu

This paper presents the Entity-Duet Neural Ranking Model (EDRM), which introduces knowledge graphs to neural search systems. EDRM represents queries and documents by their words and entity annotations. The semantics from knowledge graphs are integrated in the distributed representations of their entities, while the ranking is conducted by interaction-based neural ranking networks. The two components are learned end-to-end, making EDRM a natural combination of entity-oriented search and neural information retrieval. Our experiments on a commercial search log demonstrate the effectiveness of EDRM. Our analyses reveal that knowledge graph semantics significantly improve the generalization ability of neural ranking models.

Automatic Academic Paper Rating Based on Modularized Hierarchical Convolutional Neural Network
Pengcheng Yang, Xu Sun, Wei Li, and Shuming Ma

As more and more academic papers are being submitted to conferences and journals, evaluating all these papers by professionals is time-consuming and can cause inequality due to the personal factors of the reviewers. In this paper, in order to assist professionals in evaluating academic papers, we propose a novel task: automatic academic paper rating (AAPR), which automatically determine whether to accept academic papers. We build a new dataset for this task and propose a novel modularized hierarchical convolutional neural network to achieve automatic academic paper rating. Evaluation results show that the proposed model outperforms the baselines by a large margin. The dataset and code are available at `https://github.com/lancopku/AAPR`

Automated essay scoring with string kernels and word embeddings
Madalina Cozma, Andrei Butnaru, and Radu Tudor Ionescu

In this work, we present an approach based on combining string kernels and word embeddings for automatic essay scoring. String kernels capture the similarity among strings based on counting common character n-grams, which are a low-level yet powerful type of feature, demonstrating state-of-the-art results in various text classification tasks such as Arabic dialect identification or native language identification. To our best knowledge, we are the first to apply string kernels to automatically score essays. We are also the first to combine them with a high-level semantic feature representation, namely the bag-of-super-word-embeddings. We report the best performance on the Automated Student Assessment Prize data set, in both in-domain and cross-domain settings, surpassing recent state-of-the-art deep learning approaches.

Party Matters: Enhancing Legislative Embeddings with Author Attributes for Vote Prediction
Anastassia Kornilova, Daniel Argyle, and Vladimir Eidelman

Predicting how Congressional legislators will vote is important for understanding their past and future behavior. However, previous work on roll-call prediction has been limited to single session settings, thus not allowing for generalization across sessions. In this paper, we show that text alone is insufficient for modeling voting outcomes in new contexts, as session changes lead to changes in the underlying data generation process. We propose a novel neural method for encoding documents alongside additional metadata, achieving an average of a 4% boost in accuracy over the previous state-of-the-art.

Dynamic and Static Topic Model for Analyzing Time-Series Document Collections
Rem Hida, Naoya Takeishi, Takehisa Yairi, and Koichi Hori

For extracting meaningful topics from texts, their structures should be considered properly. In this paper, we aim to analyze structured time-series documents such as a collection of news articles and a series of scientific papers, wherein topics evolve along time depending on multiple topics in the past and are also related to each other at each time. To this end, we propose a dynamic and static topic model, which simultaneously considers the dynamic structures of the temporal topic evolution and the static structures of the topic hierarchy at each time. We show the results of experiments on collections of scientific papers, in which the proposed method outperformed conventional models. Moreover, we show an example of extracted topic structures, which we found helpful for analyzing research activities.

PhraseCTM: Correlated Topic Modeling on Phrases within Markov Random Fields
Weijing Huang

Recent emerged phrase-level topic models are able to provide topics of phrases, which are easy to read for humans. But these models are lack of the ability to capture the correlation structure among the discovered numerous topics. We propose a novel topic model PhraseCTM and a two-stage method to find out the correlated topics at phrase level. In the first stage, we train PhraseCTM, which models the generation of words and phrases simultaneously by linking the phrases and component words within Markov Random Fields when they are semantically coherent. In the second stage, we generate the correlation of topics from PhraseCTM. We evaluate our method by a quantitative experiment and a human study, showing the correlated topic modeling on phrases is a good and practical way to interpret the underlying themes of a corpus.

A Document Descriptor using Covariance of Word Vectors
Marwan Torki

In this paper, we address the problem of finding a novel document descriptor based on the covariance matrix of the word vectors of a document. Our descriptor has a fixed length, which makes it easy to use in many supervised and unsupervised applications. We tested our novel descriptor in different tasks including supervised and unsupervised settings. Our evaluation shows that our document covariance descriptor fits different tasks with competitive performance against state-of-the-art methods.

Poster Session 3C: Semantics

Neural Natural Language Inference Models Enhanced with External Knowledge
Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Diana Inkpen, and Si Wei

Modeling natural language inference is a very challenging task. With the availability of large annotated data, it has recently become feasible to train complex models such as neural-network-based inference models, which have shown to achieve the state-of-the-art performance. Although there exist relatively large annotated data, can machines learn all knowledge needed to perform natural language inference (NLI) from these data? If not, how can neural-network-based NLI models benefit from external knowledge and how to build NLI models to leverage it? In this paper, we enrich the state-of-the-art neural natural language inference models with external knowledge. We demonstrate that the proposed models improve neural NLI models to achieve the state-of-the-art performance on the SNLI and MultiNLI datasets.

AdvEntuRe: Adversarial Training for Textual Entailment with Knowledge-Guided Examples
Dongyeop Kang, Tushar Khot, Ashish Sabharwal, and Eduard Hovy

We consider the problem of learning textual entailment models with limited supervision (5K-10K training examples), and present two complementary approaches for it. First, we propose knowledge-guided adversarial example generators for incorporating large lexical resources in entailment models via only a handful of rule templates. Second, to make the entailment model—a discriminator—more robust, we propose the first GAN-style approach for training it using a natural language example generator that iteratively adjusts to the discriminator's weaknesses. We demonstrate effectiveness using two entailment datasets, where the proposed methods increase accuracy by 4.7% on SciTail and by 2.8% on a 1% sub-sample of SNLI. Notably, even a single hand-written rule, negate, improves the accuracy of negation examples in SNLI by 6.1%.

Learning with Structured Representations for Negation Scope Extraction
Hao Li and Wei Lu

We report an empirical study on the task of negation scope extraction given the negation cue. Our key observation is that certain useful information such as features related to negation cue, long-distance dependencies as well as some latent structural information can be exploited for such a task. We design approaches based on conditional random fields (CRF), semi-Markov CRF, as well as latent-variable CRF models to capture such information. Extensive experiments on several standard datasets demonstrate that our approaches are able to achieve better results than existing approaches reported in the literature.

End-Task Oriented Textual Entailment via Deep Explorations of Inter-Sentence Interactions
Wenpeng Yin, Dan Roth, and Hinrich Schütze

This work deals with SciTail, a natural entailment challenge derived from a multi-choice question answering problem. The premises and hypotheses in SciTail were generated with no awareness of each other, and did not specifically aim at the entailment task. This makes it more challenging than other entailment data sets and more directly useful to the end-task – question answering. We propose DEISTE (deep explorations of inter-sentence interactions for textual entailment) for this entailment task. Given word-to-word interactions between the premise-hypothesis pair (P, H), DEISTE consists of: (i) a parameter-dynamic convolution to make important words in P and H play a dominant role in learnt representations; and (ii) a position-aware attentive convolution to encode the representation and position information of the aligned word pairs. Experiments show that DEISTE gets ~ 5% improvement over prior state of the art and that the pretrained DEISTE on SciTail generalizes well on RTE-5.

Sense-Aware Neural Models for Pun Location in Texts
Yitao Cai, Yin Li, and Xiaojun Wan

A homographic pun is a form of wordplay in which one signifier (usually a word) suggests two or more meanings by exploiting polysemy for an intended humorous or rhetorical effect. In this paper, we focus on the task of pun location, which aims to identify the pun word in a given short text. We propose a sense-aware neural model to address this challenging task. Our model first obtains several WSD results for the text, and then leverages a bidirectional LSTM network to model each sequence of word senses. The outputs at each time step for different LSTM networks are then concatenated for prediction. Evaluation results on the benchmark SemEval 2017 dataset demonstrate the efficacy of our proposed model.

Subword-level Word Vector Representations for Korean
Sungjoon Park, Jeongmin Byun, Sion Baek, Yongseok Cho, and Alice Oh

Research on distributed word representations is focused on widely-used languages such as English. Although the same methods can be used for other languages, language-specific knowledge can enhance the accuracy and richness

of word vector representations. In this paper, we look at improving distributed word representations for Korean using knowledge about the unique linguistic structure of Korean. Specifically, we decompose Korean words into the jamo-level, beyond the character-level, allowing a systematic use of subword information. To evaluate the vectors, we develop Korean test sets for word similarity and analogy and make them publicly available. The results show that our simple method outperforms word2vec and character-level Skip-Grams on semantic and syntactic similarity and analogy tasks and contributes positively toward downstream NLP tasks such as sentiment analysis.

Incorporating Chinese Characters of Words for Lexical Sememe Prediction
Huiming Jin, Hao Zhu, Zhiyuan Liu, Ruobing Xie, Maosong Sun, Fen Lin, and Leyu Lin

Sememes are minimum semantic units of concepts in human languages, such that each word sense is composed of one or multiple sememes. Words are usually manually annotated with their sememes by linguists, and form linguistic common-sense knowledge bases widely used in various NLP tasks. Recently, the lexical sememe prediction task has been introduced. It consists of automatically recommending sememes for words, which is expected to improve annotation efficiency and consistency. However, existing methods of lexical sememe prediction typically rely on the external context of words to represent the meaning, which usually fails to deal with low-frequency and out-of-vocabulary words. To address this issue for Chinese, we propose a novel framework to take advantage of both internal character information and external context information of words. We experiment on HowNet, a Chinese sememe knowledge base, and demonstrate that our framework outperforms state-of-the-art baselines by a large margin, and maintains a robust performance even for low-frequency words.

SemAxis: A Lightweight Framework to Characterize Domain-Specific Word Semantics Beyond Sentiment
Jisun An, Haewoon Kwak, and Yong-Yeol Ahn

Because word semantics can substantially change across communities and contexts, capturing domain-specific word semantics is an important challenge. Here, we propose SemAxis, a simple yet powerful framework to characterize word semantics using many semantic axes in word-vector spaces beyond sentiment. We demonstrate that SemAxis can capture nuanced semantic representations in multiple online communities. We also show that, when the sentiment axis is examined, SemAxis outperforms the state-of-the-art approaches in building domain-specific sentiment lexicons.

End-to-End Reinforcement Learning for Automatic Taxonomy Induction
Yuning Mao, Xiang Ren, Jiaming Shen, Xiaotao Gu, and Jiawei Han

We present a novel end-to-end reinforcement learning approach to automatic taxonomy induction from a set of terms. While prior methods treat the problem as a two-phase task (i.e., detecting hypernymy pairs followed by organizing these pairs into a tree-structured hierarchy), we argue that such two-phase methods may suffer from error propagation, and cannot effectively optimize metrics that capture the holistic structure of a taxonomy. In our approach, the representations of term pairs are learned using multiple sources of information and used to determine *which* term to select and *where* to place it on the taxonomy via a policy network. All components are trained in an end-to-end manner with cumulative rewards, measured by a holistic tree metric over the training taxonomies. Experiments on two public datasets of different domains show that our approach outperforms prior state-of-the-art taxonomy induction methods up to 19.6% on ancestor F1.

Incorporating Glosses into Neural Word Sense Disambiguation
Fuli Luo, Tianyu Liu, Qiaolin Xia, Baobao Chang, and Zhifang Sui

Word Sense Disambiguation (WSD) aims to identify the correct meaning of polysemous words in the particular context. Lexical resources like WordNet which are proved to be of great help for WSD in the knowledge-based methods. However, previous neural networks for WSD always rely on massive labeled data (context), ignoring lexical resources like glosses (sense definitions). In this paper, we integrate the context and glosses of the target word into a unified framework in order to make full use of both labeled data and lexical knowledge. Therefore, we propose GAS: a gloss-augmented WSD neural network which jointly encodes the context and glosses of the target word. GAS models the semantic relationship between the context and the gloss in an improved memory network framework, which breaks the barriers of the previous supervised methods and knowledge-based methods. We further extend the original gloss of word sense via its semantic relations in WordNet to enrich the gloss information. The experimental results show that our model outperforms the state-of-the-art systems on several English all-words WSD datasets.

A Rank-Based Similarity Metric for Word Embeddings
Enrico Santus, Hongmin Wang, Emmanuele Chersoni, and Yue Zhang

Word Embeddings have recently imposed themselves as a standard for representing word meaning in NLP. Semantic similarity between word pairs has become the most common evaluation benchmark for these representations, with

vector cosine being typically used as the only similarity metric. In this paper, we report experiments with a rank-based metric for WE, which performs comparably to vector cosine in similarity estimation and outperforms it in the recently-introduced and challenging task of outlier detection, thus suggesting that rank-based measures can improve clustering quality.

Addressing Noise in Multidialectal Word Embeddings
Alexander Erdmann, Nasser Zalmout, and Nizar Habash

Word embeddings are crucial to many natural language processing tasks. The quality of embeddings relies on large non-noisy corpora. Arabic dialects lack large corpora and are noisy, being linguistically disparate with no standardized spelling. We make three contributions to address this noise. First, we describe simple but effective adaptations to word embedding tools to maximize the informative content leveraged in each training sentence. Second, we analyze methods for representing disparate dialects in one embedding space, either by mapping individual dialects into a shared space or learning a joint model of all dialects. Finally, we evaluate via dictionary induction, showing that two metrics not typically reported in the task enable us to analyze our contributions' effects on low and high frequency words. In addition to boosting performance between 2-53%, we specifically improve on noisy, low frequency forms without compromising accuracy on high frequency forms.

GNEG: Graph-Based Negative Sampling for word2vec
Zheng Zhang and Pierre Zweigenbaum

Negative sampling is an important component in word2vec for distributed word representation learning. We hypothesize that taking into account global, corpus-level information and generating a different noise distribution for each target word better satisfies the requirements of negative examples for each training word than the original frequency-based distribution. In this purpose we pre-compute word co-occurrence statistics from the corpus and apply to it network algorithms such as random walk. We test this hypothesis through a set of experiments whose results show that our approach boosts the word analogy task by about 5% and improves the performance on word similarity tasks by about 1% compared to the skip-gram negative sampling baseline.

Unsupervised Learning of Style-sensitive Word Vectors
Reina Akama, Kento Watanabe, Sho Yokoi, Sosuke Kobayashi, and Kentaro Inui

This paper presents the first study aimed at capturing stylistic similarity between words in an unsupervised manner. We propose extending the continuous bag of words (CBOW) embedding model (Mikolov et al., 2013b) to learn style-sensitive word vectors using a wider context window under the assumption that the style of all the words in an utterance is consistent. In addition, we introduce a novel task to predict lexical stylistic similarity and to create a benchmark dataset for this task. Our experiment with this dataset supports our assumption and demonstrates that the proposed extensions contribute to the acquisition of style-sensitive word embeddings.

Poster Session 3D: Sentiment Analysis and Argument Mining

Bilingual Sentiment Embeddings: Joint Projection of Sentiment Across Languages
Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde

Sentiment analysis in low-resource languages suffers from a lack of annotated corpora to estimate high-performing models. Machine translation and bilingual word embeddings provide some relief through cross-lingual sentiment approaches. However, they either require large amounts of parallel data or do not sufficiently capture sentiment information. We introduce Bilingual Sentiment Embeddings (BLSE), which jointly represent sentiment information in a source and target language. This model only requires a small bilingual lexicon, a source-language corpus annotated for sentiment, and monolingual word embeddings for each language. We perform experiments on three language combinations (Spanish, Catalan, Basque) for sentence-level cross-lingual sentiment classification and find that our model significantly out- performs state-of-the-art methods on four out of six experimental setups, as well as capturing complementary information to machine translation. Our analysis of the resulting embedding space provides evidence that it represents sentiment information in the resource-poor target language without any annotated data in that language.

Learning Domain-Sensitive and Sentiment-Aware Word Embeddings
Bei Shi, Zihao Fu, Lidong Bing, and Wai Lam

Word embeddings have been widely used in sentiment classification because of their efficacy for semantic representations of words. Given reviews from different domains, some existing methods for word embeddings exploit sentiment information, but they cannot produce domain-sensitive embeddings. On the other hand, some other existing methods can generate domain-sensitive word embeddings, but they cannot distinguish words with similar contexts but opposite sentiment polarity. We propose a new method for learning domain-sensitive and sentiment-aware embeddings that simultaneously capture the information of sentiment semantics and domain sensitivity of individual words. Our method can automatically determine and produce domain-common embeddings and domain-specific embeddings. The differentiation of domain-common and domain-specific words enables the advantage of data augmentation of common semantics from multiple domains and capture the varied semantics of specific words from different domains at the same time. Experimental results show that our model provides an effective way to learn domain-sensitive and sentiment-aware word embeddings which benefit sentiment classification at both sentence level and lexicon term level.

Cross-Domain Sentiment Classification with Target Domain Specific Information
Minlong Peng, Qi Zhang, Yu-gang Jiang, and Xuanjing Huang

The task of adopting a model with good performance to a target domain that is different from the source domain used for training has received considerable attention in sentiment analysis. Most existing approaches mainly focus on learning representations that are domain-invariant in both the source and target domains. Few of them pay attention to domain-specific information, which should also be informative. In this work, we propose a method to simultaneously extract domain specific and invariant representations and train a classifier on each of the representation, respectively. And we introduce a few target domain labeled data for learning domain-specific information. To effectively utilize the target domain labeled data, we train the domain invariant representation based classifier with both the source and target domain labeled data and train the domain-specific representation based classifier with only the target domain labeled data. These two classifiers then boost each other in a co-training style. Extensive sentiment analysis experiments demonstrated that the proposed method could achieve better performance than state-of-the-art methods.

Aspect Based Sentiment Analysis with Gated Convolutional Networks
Wei Xue and Tao Li

Aspect based sentiment analysis (ABSA) can provide more detailed information than general sentiment analysis, because it aims to predict the sentiment polarities of the given aspects or entities in text. We summarize previous approaches into two subtasks: aspect-category sentiment analysis (ACSA) and aspect-term sentiment analysis (ATSA). Most previous approaches employ long short-term memory and attention mechanisms to predict the sentiment polarity of the concerned targets, which are often complicated and need more training time. We propose a model based on convolutional neural networks and gating mechanisms, which is more accurate and efficient. First, the novel Gated Tanh-ReLU Units can selectively output the sentiment features according to the given aspect or entity. The architecture is much simpler than attention layer used in the existing models. Second, the computations of our model could be easily parallelized during training, because convolutional layers do not have time dependency as in LSTM layers, and gating units also work independently. The experiments on SemEval datasets demonstrate the efficiency and effectiveness of our models.

A Helping Hand: Transfer Learning for Deep Sentiment Analysis
Xin Dong and Gerard de Melo

Deep convolutional neural networks excel at sentiment polarity classification, but tend to require substantial amounts of training data, which moreover differs quite significantly between domains. In this work, we present an approach to feed generic cues into the training process of such networks, leading to better generalization abilities given limited training data. We propose to induce sentiment embeddings via supervision on extrinsic data, which are then fed into the model via a dedicated memory-based component. We observe significant gains in effectiveness on a range of different datasets in seven different languages.

Cold-Start Aware User and Product Attention for Sentiment Classification
Reinald Kim Amplayo, Jihyeok Kim, Sua Sung, and Seung-won Hwang

The use of user/product information in sentiment analysis is important, especially for cold-start users/products, whose number of reviews are very limited. However, current models do not deal with the cold-start problem which is typical in review websites. In this paper, we present Hybrid Contextualized Sentiment Classifier (HCSC), which contains two modules: (1) a fast word encoder that returns word vectors embedded with short and long range dependency features; and (2) Cold-Start Aware Attention (CSAA), an attention mechanism that considers the existence of cold-start problem when attentively pooling the encoded word vectors. HCSC introduces shared vectors that are constructed from similar users/products, and are used when the original distinct vectors do not have sufficient information (i.e. cold-start). This is decided by a frequency-guided selective gate vector. Our experiments show that in terms of RMSE, HCSC performs significantly better when compared with on famous datasets, despite having less complexity, and thus can be trained much faster. More importantly, our model performs significantly better than previous models when the training data is sparse and has cold-start problems.

Modeling Deliberative Argumentation Strategies on Wikipedia
Khalid Al Khatib, Henning Wachsmuth, Kevin Lang, Jakob Herpel, Matthias Hagen, and Benno Stein

This paper studies how the argumentation strategies of participants in deliberative discussions can be supported computationally. Our ultimate goal is to predict the best next deliberative move of each participant. In this paper, we present a model for deliberative discussions and we illustrate its operationalization. Previous models have been built manually based on a small set of discussions, resulting in a level of abstraction that is not suitable for move recommendation. In contrast, we derive our model statistically from several types of metadata that can be used for move description. Applied to six million discussions from Wikipedia talk pages, our approach results in a model with 13 categories along three dimensions: discourse acts, argumentative relations, and frames. On this basis, we automatically generate a corpus with about 200,000 turns, labeled for the 13 categories. We then operationalize the model with three supervised classifiers and provide evidence that the proposed categories can be predicted.

[TACL] Bootstrap Domain-Specific Sentiment Classifiers from Unlabelled Corpora
Andrius Mudinas, Dell Zhang, and Mark Levene

There is often the need to perform sentiment classification in a particular domain where no labelled document is available. Although we could make use of a general-purpose off-the-shelf sentiment classifier or a pre-built one for a different domain, the effectiveness would be inferior. In this paper, we explore the possibility of building domain-specific sentiment classifiers with unlabelled documents only. Our investigation indicates that in the word embeddings learnt from the unlabelled corpus of a given domain, the distributed word representations (vectors) for opposite sentiments form distinct clusters, though those clusters are not transferable across domains. Exploiting such a clustering structure, we are able to utilise machine learning algorithms to induce a quality domain-specific sentiment lexicon from just a few typical sentiment words ("seeds"). An important finding is that simple linear model based supervised learning algorithms (such as linear SVM) can actually work better than more sophisticated semi-supervised/transductive learning algorithms which represent the state-of-the-art technique for sentiment lexicon induction. The induced lexicon could be applied directly in a lexicon-based method for sentiment classification, but a higher performance could be achieved through a two-phase bootstrapping method which uses the induced lexicon to assign positive/negative sentiment scores to unlabelled documents first, and then uses those documents found to have clear sentiment signals as pseudo-labelled examples to train a document sentiment classifier via supervised learning algorithms (such as LSTM). On several benchmark datasets for document sentiment classification, our end-to-end pipelined approach which is overall unsupervised (except for a tiny set of seed words) outperforms existing unsupervised approaches and achieves an accuracy comparable to that of fully supervised approaches.

Exploiting Document Knowledge for Aspect-level Sentiment Classification
Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier

Attention-based long short-term memory (LSTM) networks have proven to be useful in aspect-level sentiment clas-

sification. However, due to the difficulties in annotating aspect-level data, existing public datasets for this task are all relatively small, which largely limits the effectiveness of those neural models. In this paper, we explore two approaches that transfer knowledge from document-level data, which is much less expensive to obtain, to improve the performance of aspect-level sentiment classification. We demonstrate the effectiveness of our approaches on 4 public datasets from SemEval 2014, 2015, and 2016, and we show that attention-based LSTM benefits from document-level knowledge in multiple ways.

Modeling Sentiment Association in Discourse for Humor Recognition
Lizhen Liu, Donghai Zhang, and Wei Song

Humor is one of the most attractive parts in human communication. However, automatically recognizing humor in text is challenging due to the complex characteristics of humor. This paper proposes to model sentiment association between discourse units to indicate how the punchline breaks the expectation of the setup. We found that discourse relation, sentiment conflict and sentiment transition are effective indicators for humor recognition. On the perspective of using sentiment related features, sentiment association in discourse is more useful than counting the number of emotional words.

Double Embeddings and CNN-based Sequence Labeling for Aspect Extraction
Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu

One key task of fine-grained sentiment analysis of product reviews is to extract product aspects or features that users have expressed opinions on. This paper focuses on supervised aspect extraction using deep learning. Unlike other highly sophisticated supervised deep learning models, this paper proposes a novel and yet simple CNN model employing two types of pre-trained embeddings for aspect extraction: general-purpose embeddings and domain-specific embeddings. Without using any additional supervision, this model achieves surprisingly good results, outperforming state-of-the-art sophisticated existing methods. To our knowledge, this paper is the first to report such double embeddings based CNN model for aspect extraction and achieve very good results.

Will it Blend? Blending Weak and Strong Labeled Data in a Neural Network for Argumentation Mining
Eyal Shnarch, Carlos Alzate, Lena Dankin, Martin Gleize, Yufang Hou, Leshem Choshen, Ranit Aharonov, and Noam Slonim

The process of obtaining high quality labeled data for natural language understanding tasks is often slow, error-prone, complicated and expensive. With the vast usage of neural networks, this issue becomes more notorious since these networks require a large amount of labeled data to produce satisfactory results. We propose a methodology to blend high quality but scarce strong labeled data with noisy but abundant weak labeled data during the training of neural networks. Experiments in the context of topic-dependent evidence detection with two forms of weak labeled data show the advantages of the blending scheme. In addition, we provide a manually annotated data set for the task of topic-dependent evidence detection. We believe that blending weak and strong labeled data is a general notion that may be applicable to many language understanding tasks, and can especially assist researchers who wish to train a network but have a small amount of high quality labeled data for their task of interest.

Poster Session 3E: Vision, Multimodal, Grounding, Speech

Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning
Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut

We present a new dataset of image caption annotations, Conceptual Captions, which contains an order of magnitude more images than the MS-COCO dataset (Lin et al., 2014) and represents a wider variety of both images and image caption styles. We achieve this by extracting and filtering image caption annotations from billions of webpages. We also present quantitative evaluations of a number of image captioning models and show that a model architecture based on Inception-ResNetv2 (Szegedy et al., 2016) for image-feature extraction and Transformer (Vaswani et al., 2017) for sequence modeling achieves the best performance when trained on the Conceptual Captions dataset.

Learning Translations via Images with a Massively Multilingual Image Dataset
John Hewitt, Daphne Ippolito, Brendan Callahan, Reno Kriz, Derry Tanti Wijaya, and Chris Callison-Burch

We conduct the most comprehensive study to date into translating words via images. To facilitate research on the task, we introduce a large-scale multilingual corpus of images, each labeled with the word it represents. Past datasets have been limited to only a few high-resource languages and unrealistically easy translation settings. In contrast, we have collected by far the largest available dataset for this task, with images for approximately 10,000 words in each of 100 languages. We run experiments on a dozen high resource languages and 20 low resources languages, demonstrating the effect of word concreteness and part-of-speech on translation quality. %We find that while image features work best for concrete nouns, they are sometimes effective on other parts of speech. To improve image-based translation, we introduce a novel method of predicting word concreteness from images, which improves on a previous state-of-the-art unsupervised technique. This allows us to predict when image-based translation may be effective, enabling consistent improvements to a state-of-the-art text-based word translation system. Our code and the Massively Multilingual Image Dataset (MMID) are available at http://multilingual-images.org/.

On the Automatic Generation of Medical Imaging Reports
Baoyu Jing, Pengtao Xie, and Eric Xing

Medical imaging is widely used in clinical practice for diagnosis and treatment. Report-writing can be error-prone for unexperienced physicians, and time-consuming and tedious for experienced physicians. To address these issues, we study the automatic generation of medical imaging reports. This task presents several challenges. First, a complete report contains multiple heterogeneous forms of information, including findings and tags. Second, abnormal regions in medical images are difficult to identify. Third, the reports are typically long, containing multiple sentences. To cope with these challenges, we (1) build a multi-task learning framework which jointly performs the prediction of tags and the generation of paragraphs, (2) propose a co-attention mechanism to localize regions containing abnormalities and generate narrations for them, (3) develop a hierarchical LSTM model to generate long paragraphs. We demonstrate the effectiveness of the proposed methods on two publicly available dataset.

Attacking Visual Language Grounding with Adversarial Examples: A Case Study on Neural Image Captioning
Hongge Chen, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, and Cho-Jui Hsieh

Visual language grounding is widely studied in modern neural image captioning systems, which typically adopts an encoder-decoder framework consisting of two principal components: a convolutional neural network (CNN) for image feature extraction and a recurrent neural network (RNN) for language caption generation. To study the robustness of language grounding to adversarial perturbations in machine vision and perception, we propose Show-and-Fool, a novel algorithm for crafting adversarial examples in neural image captioning. The proposed algorithm provides two evaluation approaches, which check if we can mislead neural image captioning systems to output some randomly chosen captions or keywords. Our extensive experiments show that our algorithm can successfully craft visually-similar adversarial examples with randomly targeted captions or keywords, and the adversarial examples can be made highly transferable to other image captioning systems. Consequently, our approach leads to new robustness implications of neural image captioning and novel insights in visual language grounding.

Think Visually: Question Answering through Virtual Imagery
Ankit Goyal, Jian Wang, and Jia Deng

In this paper, we study the problem of geometric reasoning (a form of visual reasoning) in the context of question-answering. We introduce Dynamic Spatial Memory Network (DSMN), a new deep network architecture that specializes in answering questions that admit latent visual representations, and learns to generate and reason over such

representations. Further, we propose two synthetic benchmarks, FloorPlanQA and ShapeIntersection, to evaluate the geometric reasoning capability of QA systems. Experimental results validate the effectiveness of our proposed DSMN for visual thinking tasks.

Interactive Language Acquisition with One-shot Visual Concept Learning through a Conversational Game
Haichao Zhang, Haonan Yu, and Wei Xu

Building intelligent agents that can communicate with and learn from humans in natural language is of great value. Supervised language learning is limited by the ability of capturing mainly the statistics of training data, and is hardly adaptive to new scenarios or flexible for acquiring new knowledge without inefficient retraining or catastrophic forgetting. We highlight the perspective that conversational interaction serves as a natural interface both for language learning and for novel knowledge acquisition and propose a joint imitation and reinforcement approach for grounded language learning through an interactive conversational game. The agent trained with this approach is able to actively acquire information by asking questions about novel objects and use the just-learned knowledge in subsequent conversations in a one-shot fashion. Results compared with other methods verified the effectiveness of the proposed approach.

A Purely End-to-End System for Multi-speaker Speech Recognition
Hiroshi Seki, Takaaki Hori, Shinji Watanabe, Jonathan Le Roux, and John R Hershey

Recently, there has been growing interest in multi-speaker speech recognition, where the utterances of multiple speakers are recognized from their mixture. Promising techniques have been proposed for this task, but earlier works have required additional training data such as isolated source signals or senone alignments for effective learning. In this paper, we propose a new sequence-to-sequence framework to directly decode multiple label sequences from a single speech sequence by unifying source separation and speech recognition functions in an end-to-end manner. We further propose a new objective function to improve the contrast between the hidden vectors to avoid generating similar hypotheses. Experimental results show that the model is directly able to learn a mapping from a speech mixture to multiple label sequences, achieving 83.1% relative improvement compared to a model trained without the proposed objective. Interestingly, the results are comparable to those produced by previous end-to-end works featuring explicit separation and recognition modules.

[TACL] Learning Representations Specialized in Spatial Knowledge: Leveraging Language and Vision
Guillem Collell and Marie-Francine Moens

Spatial understanding is crucial in many real-world problems, yet little progress has been made towards building representations that capture spatial knowledge. Here, we move one step forward in this direction and learn such representations by leveraging a task consisting in predicting continuous 2D spatial arrangements of objects given object-relationship-object instances (e.g., “cat under chair") and a simple neural network model that learns the task from annotated images. We show that the model succeeds in this task and that it is furthermore capable of predicting correct spatial arrangements for unseen objects if either CNN features or word embeddings of the objects are provided. The differences between visual and linguistic features are discussed. Next, to evaluate the spatial representations learned in the previous task, we introduce a task and a dataset consisting in a set of crowdsourced human ratings of spatial similarity for object pairs. We find that both CNN features and word embeddings predict well human judgments of similarity and that these vectors can be further specialized in spatial knowledge if we update them when training the model that predicts spatial arrangements of objects. Overall, this paper paves the way towards building distributed spatial representations, contributing to the understanding of spatial expressions in language.

Investigating Audio, Video, and Text Fusion Methods for End-to-End Automatic Personality Prediction
Onno Kampman, Elham J. Barezi, Dario Bertero, and Pascale Fung

We propose a tri-modal architecture to predict Big Five personality trait scores from video clips with different channels for audio, text, and video data. For each channel, stacked Convolutional Neural Networks are employed. The channels are fused both on decision-level and by concatenating their respective fully connected layers. It is shown that a multimodal fusion approach outperforms each single modality channel, with an improvement of 9.4% over the best individual modality (video). Full backpropagation is also shown to be better than a linear combination of modalities, meaning complex interactions between modalities can be leveraged to build better models. Furthermore, we can see the prediction relevance of each modality for each trait. The described model can be used to increase the emotional intelligence of virtual agents.

Poster Session 3F: Morphology, Tagging, Parsing

A Structured Variational Autoencoder for Contextual Morphological Inflection
Ryan Cotterell, Jason Naradowsky, Sebastian J. Mielke, and Lawrence Wolf-Sonkin

Statistical morphological inflectors are typically trained on fully supervised, type-level data. One remaining open research question is the following: How can we effectively exploit raw, token-level data to improve their performance? To this end, we introduce a novel generative latent-variable model for the semi-supervised learning of inflection generation. To enable posterior inference over the latent variables, we derive an efficient variational inference procedure based on the wake-sleep algorithm. We experiment on 23 languages, using the Universal Dependencies corpora in a simulated low-resource setting, and find improvements of over 10% absolute accuracy in some cases.

Morphosyntactic Tagging with a Meta-BiLSTM Model over Context Sensitive Token Encodings
Bernd Bohnet, Ryan McDonald, Gonçalo Simões, Daniel Andor, Emily Pitler, and Joshua Maynez

The rise of neural networks, and particularly recurrent neural networks, has produced significant advances in part-of-speech tagging accuracy. One characteristic common among these models is the presence of rich initial word encodings. These encodings typically are composed of a recurrent character-based representation with dynamically and pre-trained word embeddings. However, these encodings do not consider a context wider than a single word and it is only through subsequent recurrent layers that word or sub-word information interacts. In this paper, we investigate models that use recurrent neural networks with sentence-level context for initial character and word-based representations. In particular we show that optimal results are obtained by integrating these context sensitive representations through synchronized training with a meta-model that learns to combine their states.

Neural Factor Graph Models for Cross-lingual Morphological Tagging
Chaitanya Malaviya, Matthew R. Gormley, and Graham Neubig

Morphological analysis involves predicting the syntactic traits of a word (e.g. POS: Noun, Case: Acc, Gender: Fem). Previous work in morphological tagging improves performance for low-resource languages (LRLs) through cross-lingual training with a high-resource language (HRL) from the same family, but is limited by the strict, often false, assumption that tag sets exactly overlap between the HRL and LRL. In this paper we propose a method for cross-lingual morphological tagging that aims to improve information sharing between languages by relaxing this assumption. The proposed model uses factorial conditional random fields with neural network potentials, making it possible to (1) utilize the expressive power of neural network representations to smooth over superficial differences in the surface forms, (2) model pairwise and transitive relationships between tags, and (3) accurately generate tag sets that are unseen or rare in the training data. Experiments on four languages from the Universal Dependencies Treebank demonstrate superior tagging accuracies over existing cross-lingual approaches.

Global Transition-based Non-projective Dependency Parsing
Carlos Gómez-Rodríguez, Tianze Shi, and Lillian Lee

Shi, Huang, and Lee (2017a) obtained state-of-the-art results for English and Chinese dependency parsing by combining dynamic-programming implementations of transition-based dependency parsers with a minimal set of bidirectional LSTM features. However, their results were limited to projective parsing. In this paper, we extend their approach to support non-projectivity by providing the first practical implementation of the MH_4 algorithm, an $O(n^4)$ mildly nonprojective dynamic-programming parser with very high coverage on non-projective treebanks. To make MH_4 compatible with minimal transition-based feature sets, we introduce a transition-based interpretation of it in which parser items are mapped to sequences of transitions. We thus obtain the first implementation of global decoding for non-projective transition-based parsing, and demonstrate empirically that it is effective than its projective counterpart in parsing a number of highly non-projective languages.

Constituency Parsing with a Self-Attentive Encoder
Nikita Kitaev and Dan Klein

We demonstrate that replacing an LSTM encoder with a self-attentive architecture can lead to improvements to a state-of-the-art discriminative constituency parser. The use of attention makes explicit the manner in which information is propagated between different locations in the sentence, which we use to both analyze our model and propose potential improvements. For example, we find that separating positional and content information in the encoder can lead to improved parsing accuracy. Additionally, we evaluate different approaches for lexical representation. Our parser achieves new state-of-the-art results for single models trained on the Penn Treebank: 93.55 F1 without the use of any external data, and 95.13 F1 when using pre-trained word representations. Our parser also outperforms the previous best-published accuracy figures on 8 of the 9 languages in the SPMRL dataset.

Pre- and In-Parsing Models for Neural Empty Category Detection
Yufei Chen, Yuanyuan Zhao, Weiwei Sun, and Xiaojun Wan

Motivated by the positive impact of empty category on syntactic parsing, we study neural models for pre- and in-parsing detection of empty category, which has not previously been investigated. We find several non-obvious facts: (a) BiLSTM can capture non-local contextual information which is essential for detecting empty categories, (b) even with a BiLSTM, syntactic information is still able to enhance the detection, and (c) automatic detection of empty categories improves parsing quality for overt words. Our neural ECD models outperform the prior state-of-the-art by significant margins.

Composing Finite State Transducers on GPUs
Arturo Argueta and David Chiang

Weighted finite state transducers (FSTs) are frequently used in language processing to handle tasks such as part-of-speech tagging and speech recognition. There has been previous work using multiple CPU cores to accelerate finite state algorithms, but limited attention has been given to parallel graphics processing unit (GPU) implementations. In this paper, we introduce the first (to our knowledge) GPU implementation of the FST composition operation, and we also discuss the optimizations used to achieve the best performance on this architecture. We show that our approach obtains speedups of up to 6 times over our serial implementation and 4.5 times over OpenFST.

Supervised Treebank Conversion: Data and Approaches
Xinzhou Jiang, Zhenghua Li, Bo Zhang, Min Zhang, Sheng Li, and Luo Si

Treebank conversion is a straightforward and effective way to exploit various heterogeneous treebanks for boosting parsing performance. However, previous work mainly focuses on unsupervised treebank conversion and has made little progress due to the lack of manually labeled data where each sentence has two syntactic trees complying with two different guidelines at the same time, referred as bi-tree aligned data. In this work, we for the first time propose the task of supervised treebank conversion. First, we manually construct a bi-tree aligned dataset containing over ten thousand sentences. Then, we propose two simple yet effective conversion approaches (pattern embedding and treeLSTM) based on the state-of-the-art deep biaffine parser. Experimental results show that 1) the two conversion approaches achieve comparable conversion accuracy, and 2) treebank conversion is superior to the widely used multi-task learning framework in multi-treebank exploitation and leads to significantly higher parsing accuracy.

Object-oriented Neural Programming (OONP) for Document Understanding
Zhengdong Lu, Xianggen Liu, Haotian Cui, Yukun Yan, and Daqi Zheng

We propose Object-oriented Neural Programming (OONP), a framework for semantically parsing documents in specific domains. Basically, OONP reads a document and parses it into a predesigned object-oriented data structure that reflects the domain-specific semantics of the document. An OONP parser models semantic parsing as a decision process: a neural net-based Reader sequentially goes through the document, and builds and updates an intermediate ontology during the process to summarize its partial understanding of the text. OONP supports a big variety of forms (both symbolic and differentiable) for representing the state and the document, and a rich family of operations to compose the representation. An OONP parser can be trained with supervision of different forms and strength, including supervised learning (SL), reinforcement learning (RL) and hybrid of the two. Our experiments on both synthetic and real-world document parsing tasks have shown that OONP can learn to handle fairly complicated ontology with training data of modest sizes.

An Empirical Study of Building a Strong Baseline for Constituency Parsing
Jun Suzuki, Sho Takase, Hidetaka Kamigaito, Makoto Morishita, and Masaaki Nagata

This paper investigates the construction of a strong baseline based on general purpose sequence-to-sequence models for constituency parsing. We incorporate several techniques that were mainly developed in natural language generation tasks, e.g., machine translation and summarization, and demonstrate that the sequence-to-sequence model achieves the current top-notch parsers' performance (almost) without requiring any explicit task-specific knowledge or architecture of constituent parsing.

Parser Training with Heterogeneous Treebanks
Sara Stymne, Miryam de Lhoneux, Aaron Smith, and Joakim Nivre

How to make the most of multiple heterogeneous treebanks when training a monolingual dependency parser is an open question. We start by investigating previously suggested, but little evaluated, strategies for exploiting multiple treebanks based on concatenating training sets, with or without fine-tuning. We go on to propose a new method based on treebank embeddings. We perform experiments for several languages and show that in many cases fine-tuning and treebank embeddings lead to substantial improvements over single treebanks or concatenation, with average gains of 2.0–3.5 LAS points. We argue that treebank embeddings should be preferred due to their conceptual simplicity,

flexibility and extensibility.

Generalized chart constraints for efficient PCFG and TAG parsing
Stefan Grünewald, Sophie Henning, and Alexander Koller

Chart constraints, which specify at which string positions a constituent may begin or end, have been shown to speed up chart parsers for PCFGs. We generalize chart constraints to more expressive grammar formalisms and describe a neural tagger which predicts chart constraints at very high precision. Our constraints accelerate both PCFG and TAG parsing, and combine effectively with other pruning techniques (coarse-to-fine and supertagging) for an overall speedup of two orders of magnitude, while improving accuracy.

Demo Poster Session 3

Time: *12:30–14:00* Location: Melbourne Room 1 & 2

The SUMMA Platform: A Scalable Infrastructure for Multi-lingual Multi-media Monitoring
Ulrich Germann, Renars Liepins, Guntis Barzdins, Didzis Gosko, Sebastião Miranda, and David Nogueira

The open-source SUMMA Platform is a highly scalable distributed architecture for monitoring a large number of media broadcasts in parallel, with a lag behind actual broadcast time of at most a few minutes. The Platform offers a fully automated media ingestion pipeline capable of recording live broadcasts, detection and transcription of spoken content, translation of all text (original or transcribed) into English, recognition and linking of Named Entities, topic detection, clustering and cross-lingual multi-document summarization of related media items, and last but not least, extraction and storage of factual claims in these news items. Browser-based graphical user interfaces provide humans with aggregated information as well as structured access to individual news items stored in the Platform's database. This paper describes the intended use cases and provides an overview over the system's implementation.

CRUISE: Cold-Start New Skill Development via Iterative Utterance Generation
Yilin Shen, Avik Ray, Abhishek Patel, and Hongxia Jin

We present a system, CRUISE, that guides ordinary software developers to build a high quality natural language understanding (NLU) engine from scratch. This is the fundamental step of building a new skill in personal assistants. Unlike existing solutions that require either developers or crowdsourcing to manually generate and annotate a large number of utterances, we design a hybrid rule-based and data-driven approach with the capability to iteratively generate more and more utterances. Our system only requires light human workload to iteratively prune incorrect utterances. CRUISE outputs a well trained NLU engine and a large scale annotated utterance corpus that third parties can use to develop their custom skills. Using both benchmark dataset and custom datasets we collected in real-world settings, we validate the high quality of CRUISE generated utterances via both competitive NLU performance and human evaluation. We also show the largely reduced human workload in terms of both cognitive load and human pruning time consumption.

Praaline: An Open-Source System for Managing, Annotating, Visualising and Analysing Speech Corpora
George Christodoulides

In this system demonstration we present the latest developments of Praaline, an open-source software system for constituting and managing, manually and automatically annotating, visualising and analysing spoken language and multimodal corpora. We review the system's functionality and design architecture, present current use cases and directions for future development.

Marian: Fast Neural Machine Translation in C++
Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch

We present Marian, an efficient and self-contained Neural Machine Translation framework with an integrated automatic differentiation engine based on dynamic computation graphs. Marian is written entirely in C++. We describe the design of the encoder-decoder framework and demonstrate that a research-friendly toolkit can achieve high training and translation speed.

DeepPavlov: Open-Source Library for Dialogue Systems
Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, Alexey Litinsky, Varvara Logacheva, Alexey Lymar, Valentin Malykh, Maxim Petrov, Vadim Polulyakh, Leonid Pugachev, Alexey Sorokin, Maria Vikhreva, and Marat Zaynutdinov

Adoption of messaging communication and voice assistants has grown rapidly in the last years. This creates a demand for tools that speed up prototyping of feature-rich dialogue systems. An open-source library DeepPavlov is tailored for development of conversational agents. The library prioritises efficiency, modularity, and extensibility with the goal to make it easier to develop dialogue systems from scratch and with limited data available. It supports modular as well as end-to-end approaches to implementation of conversational agents. Conversational agent consists of skills and every skill can be decomposed into components. Components are usually models which solve typical NLP tasks such as intent classification, named entity recognition or pre-trained word vectors. Sequence-to-sequence chit-chat skill, question answering skill or task-oriented skill can be assembled from components provided in the library.

RETURNN as a Generic Flexible Neural Toolkit with Application to Translation and Speech Recognition
Albert Zeyer, Tamer Alkhouli, and Hermann Ney

We compare the fast training and decoding speed of RETURNN of attention mod- els for translation, due to fast CUDA LSTM kernels, and a fast pure TensorFlow beam search decoder. We show that a layer-wise pretraining scheme for recurrent attention models gives over 1% BLEU improvement absolute and it allows to train deeper recurrent encoder networks. Promising preliminary results on max. expected BLEU training are presented. We are able to train state-of-the-art models for translation and end-to-end models for speech recognition and show results on WMT 2017 and Switchboard. The flexibility of RETURNN allows a fast research feedback loop to experiment with alternative architectures, and its generality allows to use it on a wide range of applications.

A Flexible, Efficient and Accurate Framework for Community Question Answering Pipelines
Salvatore Romeo, Giovanni Da San Martino, Alberto Barrón-Cedeño, and Alessandro Moschitti

Although deep neural networks have been proving to be excellent tools to deliver state-of-the-art results, when data is scarce and the tackled tasks involve complex semantic inference, deep linguistic processing and traditional structure-based approaches, such as tree kernel methods, are an alternative solution. Community Question Answering is a research area that benefits from deep linguistic analysis to improve the experience of the community of forum users. In this paper, we present a UIMA framework to distribute the computation of cQA tasks over computer clusters such that traditional systems can scale to large datasets and deliver fast processing.

Moon IME: Neural-based Chinese Pinyin Aided Input Method with Customizable Association
Yafang Huang, Zuchao Li, Zhuosheng Zhang, and Hai Zhao

Chinese pinyin input method engine (IME) lets user conveniently input Chinese into a computer by typing pinyin through the common keyboard. In addition to offering high conversion quality, modern pinyin IME is supposed to aid user input with extended association function. However, existing solutions for such functions are roughly based on oversimplified matching algorithms at word-level, whose resulting products provide limited extension associated with user inputs. This work presents the Moon IME, a pinyin IME that integrates the attention-based neural machine translation (NMT) model and Information Retrieval (IR) to offer amusive and customizable association ability. The released IME is implemented on Windows via text services framework.

Session 8 Overview – Wednesday, July 18, 2018

	Track A *Semantics 2 (Short)*	**Track B** *Machine Translation, Multilinguality 2 (Short)*	**Track C** *Information Extraction 2 (Short)*	**Track D** *Generation, Summarization (Short)*	**Track E** *Machine Learning, Question Answering (Short)*	**Track F** *Sentiment (Short)*
	Plenary	203–204	210–211	212–213	219	220
14:00	Exploring Semantic Properties of Sentence Embeddings *Zhu, Li, and Melo*	Adaptive Knowledge Sharing in Multi-Task Learning: Improving Low-Resource Neural Machine Translation *Zaremoodi, Buntine, and Haffari*	Enhancing Drug-Drug Interaction Extraction from Texts by Molecular Structure Information *Asada, Miwa, and Sasaki*	Personalized Review Generation By Expanding Phrases and Attending on Aspect-Aware Representations *Ni and McAuley*	Long Short-Term Memory as a Dynamically Computed Element-wise Weighted Sum *Levy, Lee, FitzGerald, and Zettlemoyer*	A Multi-sentiment-resource Enhanced Attention Network for Sentiment Classification *Lei, Yang, Yang, and Liu*
14:15	Scoring Lexical Entailment with a Supervised Directional Similarity Network *Rei, Gerz, and Vulić*	Automatic Estimation of Simultaneous Interpreter Performance *Stewart, Vogler, Hu, Boyd-Graber, and Neubig*	diaNED: Time-Aware Named Entity Disambiguation for Diachronic Corpora *Agarwal, Strötgen, Del Corro, Hoffart, and Weikum*	Learning Simplifications for Specific Target Audiences *Scarton and Specia*	On the Practical Computational Power of Finite Precision RNNs for Language Recognition *Weiss, Goldberg, and Yahav*	Pretraining Sentiment Classifiers with Unlabeled Dialog Data *Shimizu, Shimizu, and Kobayashi*
14:30	Extracting Commonsense Properties from Embeddings with Limited Human Guidance *Yang, Birnbaum, Wang, and Downey*	Polyglot Semantic Role Labeling *Mulcaire, Swayamdipta, and Smith*	Examining Temporality in Document Classification *Huang and Paul*	Split and Rephrase: Better Evaluation and Stronger Baselines *Aharoni and Goldberg*	A Co-Matching Model for Multi-choice Reading Comprehension *Wang, Yu, Jiang, and Chang*	Disambiguating False-Alarm Hashtag Usages in Tweets for Irony Detection *Huang, Chen, and Chen*
14:45	Breaking NLI Systems with Sentences that Require Simple Lexical Inferences *Glockner, Shwartz, and Goldberg*	Learning Cross-lingual Distributed Logical Representations for Semantic Parsing *Zou and Lu*	Personalized Language Model for Query Auto-Completion *Jaech and Ostendorf*	Autoencoder as Assistant Supervisor: Improving Text Representation for Chinese Social Media Text Summarization *Ma, Sun, Lin, and Wang*	Tackling the Story Ending Biases in The Story Cloze Test *Sharma, Allen, Bakhshandeh, and Mostafazadeh*	Cross-Target Stance Classification with Self-Attention Networks *Xu, Paris, Nepal, and Sparks*

Parallel Session 8

Session 8A: Semantics 2 (Short)

Plenary *Chair: Omri Abend*

Exploring Semantic Properties of Sentence Embeddings
Xunjie Zhu, Tingfeng Li, and Gerard de Melo 14:00–14:15

Neural vector representations are ubiquitous throughout all subfields of NLP. While word vectors have been studied in much detail, thus far only little light has been shed on the properties of sentence embeddings. In this paper, we assess to what extent prominent sentence embedding methods exhibit select semantic properties. We propose a framework that generate triplets of sentences to explore how changes in the syntactic structure or semantics of a given sentence affect the similarities obtained between their sentence embeddings.

Scoring Lexical Entailment with a Supervised Directional Similarity Network
Marek Rei, Daniela Gerz, and Ivan Vulić 14:15–14:30

We present the Supervised Directional Similarity Network, a novel neural architecture for learning task-specific transformation functions on top of general-purpose word embeddings. Relying on only a limited amount of supervision from task-specific scores on a subset of the vocabulary, our architecture is able to generalise and transform a general-purpose distributional vector space to model the relation of lexical entailment. Experiments show excellent performance on scoring graded lexical entailment, raising the state-of-the-art on the HyperLex dataset by approximately 25%.

Extracting Commonsense Properties from Embeddings with Limited Human Guidance
Yiben Yang, Larry Birnbaum, Ji-Ping Wang, and Doug Downey 14:30–14:45

Intelligent systems require common sense, but automatically extracting this knowledge from text can be difficult. We propose and assess methods for extracting one type of commonsense knowledge, object-property comparisons, from pre-trained embeddings. In experiments, we show that our approach exceeds the accuracy of previous work but requires substantially less hand-annotated knowledge. Further, we show that an active learning approach that synthesizes common-sense queries can boost accuracy.

Breaking NLI Systems with Sentences that Require Simple Lexical Inferences
Max Glockner, Vered Shwartz, and Yoav Goldberg 14:45–15:00

We create a new NLI test set that shows the deficiency of state-of-the-art models in inferences that require lexical and world knowledge. The new examples are simpler than the SNLI test set, containing sentences that differ by at most one word from sentences in the training set. Yet, the performance on the new test set is substantially worse across systems trained on SNLI, demonstrating that these systems are limited in their generalization ability, failing to capture many simple inferences.

Session 8B: Machine Translation, Multilinguality 2 (Short)

203–204 *Chair: Yulia Tsvetkov*

Adaptive Knowledge Sharing in Multi-Task Learning: Improving Low-Resource Neural Machine Translation
Poorya Zaremoodi, Wray Buntine, and Gholamreza Haffari 14:00–14:15

Neural Machine Translation (NMT) is notorious for its need for large amounts of bilingual data. An effective approach to compensate for this requirement is Multi-Task Learning (MTL) to leverage different linguistic resources as a source of inductive bias. Current MTL architectures are based on the Seq2Seq transduction, and (partially) share different components of the models among the tasks. However, this MTL approach often suffers from task interference and is not able to fully capture commonalities among subsets of tasks. We address this issue by extending the recurrent units with multiple "blocks" along with a trainable "routing network". The routing network enables adaptive collaboration by dynamic sharing of blocks conditioned on the task at hand, input, and model state. Empirical evaluation of two low-resource translation tasks, English to Vietnamese and Farsi, show +1 BLEU score improvements compared to strong baselines.

Automatic Estimation of Simultaneous Interpreter Performance
Craig Stewart, Nikolai Vogler, Junjie Hu, Jordan Boyd-Graber, and Graham Neubig 14:15–14:30

Simultaneous interpretation, translation of the spoken word in real-time, is both highly challenging and physically demanding. Methods to predict interpreter confidence and the adequacy of the interpreted message have a number of potential applications, such as in computer-assisted interpretation interfaces or pedagogical tools. We propose the task of predicting simultaneous interpreter performance by building on existing methodology for quality estimation (QE) of machine translation output. In experiments over five settings in three language pairs, we extend a QE pipeline to estimate interpreter performance (as approximated by the METEOR evaluation metric) and propose novel features reflecting interpretation strategy and evaluation measures that further improve prediction accuracy.

Polyglot Semantic Role Labeling
Phoebe Mulcaire, Swabha Swayamdipta, and Noah A. Smith 14:30–14:45

Previous approaches to multilingual semantic dependency parsing treat languages independently, without exploiting the similarities between semantic structures across languages. We experiment with a new approach where we combine resources from different languages in the CoNLL 2009 shared task to build a single polyglot semantic dependency parser. Notwithstanding the absence of parallel data, and the dissimilarity in annotations between languages, our approach results in improvement in parsing performance on several languages over a monolingual baseline. Analysis of the polyglot models' performance provides a new understanding of the similarities and differences between languages in the shared task.

Learning Cross-lingual Distributed Logical Representations for Semantic Parsing
Yanyan Zou and Wei Lu 14:45–15:00

With the development of several multilingual datasets used for semantic parsing, recent research efforts have looked into the problem of learning semantic parsers in a multilingual setup. However, how to improve the performance of a monolingual semantic parser for a specific language by leveraging data annotated in different languages remains a research question that is under-explored. In this work, we present a study to show how learning distributed representations of the logical forms from data annotated in different languages can be used for improving the performance of a monolingual semantic parser. We extend two existing monolingual semantic parsers to incorporate such cross-lingual distributed logical representations as features. Experiments show that our proposed approach is able to yield improved semantic parsing results on the standard multilingual GeoQuery dataset.

Session 8C: Information Extraction 2 (Short)

210–211 *Chair: Radu Florian*

Enhancing Drug-Drug Interaction Extraction from Texts by Molecular Structure Information
Masaki Asada, Makoto Miwa, and Yutaka Sasaki 14:00–14:15

We propose a novel neural method to extract drug-drug interactions (DDIs) from texts using external drug molecular structure information. We encode textual drug pairs with convolutional neural networks and their molecular pairs with graph convolutional networks (GCNs), and then we concatenate the outputs of these two networks. In the experiments, we show that GCNs can predict DDIs from the molecular structures of drugs in high accuracy and the molecular information can enhance text-based DDI extraction by 2.39 percent points in the F-score on the DDIExtraction 2013 shared task data set.

diaNED: Time-Aware Named Entity Disambiguation for Diachronic Corpora
Prabal Agarwal, Jannik Strötgen, Luciano Del Corro, Johannes Hoffart, and Gerhard Weikum 14:15–14:30

Named Entity Disambiguation (NED) systems perform well on news articles and other texts covering a specific time interval. However, NED quality drops when inputs span long time periods like in archives or historic corpora. This paper presents the first time-aware method for NED that resolves ambiguities even when mention contexts give only few cues. The method is based on computing temporal signatures for entities and comparing these to the temporal contexts of input mentions. Our experiments show superior quality on a newly created diachronic corpus.

Examining Temporality in Document Classification
Xiaolei Huang and Michael J. Paul 14:30–14:45

Many corpora span broad periods of time. Language processing models trained during one time period may not work well in future time periods, and the best model may depend on specific times of year (e.g., people might describe hotels differently in reviews during the winter versus the summer). This study investigates how document classifiers trained on documents from certain time intervals perform on documents from other time intervals, considering both seasonal intervals (intervals that repeat across years, e.g., winter) and non-seasonal intervals (e.g., specific years). We show experimentally that classification performance varies over time, and that performance can be improved by using a standard domain adaptation approach to adjust for changes in time.

Personalized Language Model for Query Auto-Completion
Aaron Jaech and Mari Ostendorf 14:45–15:00

Query auto-completion is a search engine feature whereby the system suggests completed queries as the user types. Recently, the use of a recurrent neural network language model was suggested as a method of generating query completions. We show how an adaptable language model can be used to generate personalized completions and how the model can use online updating to make predictions for users not seen during training. The personalized predictions are significantly better than a baseline that uses no user information.

Session 8D: Generation, Summarization (Short)

212–213 *Chair: Hiroya Takamura*

Personalized Review Generation By Expanding Phrases and Attending on Aspect-Aware Representations

Jianmo Ni and Julian McAuley 14:00–14:15

In this paper, we focus on the problem of building assistive systems that can help users to write reviews. We cast this problem using an encoder-decoder framework that generates personalized reviews by expanding short phrases (e.g. review summaries, product titles) provided as input to the system. We incorporate aspect-level information via an aspect encoder that learns aspect-aware user and item representations. An attention fusion layer is applied to control generation by attending on the outputs of multiple encoders. Experimental results show that our model successfully learns representations capable of generating coherent and diverse reviews. In addition, the learned aspect-aware representations discover those aspects that users are more inclined to discuss and bias the generated text toward their personalized aspect preferences.

Learning Simplifications for Specific Target Audiences

Carolina Scarton and Lucia Specia 14:15–14:30

Text simplification (TS) is a monolingual text-to-text transformation task where an original (complex) text is transformed into a target (simpler) text. Most recent work is based on sequence-to-sequence neural models similar to those used for machine translation (MT). Different from MT, TS data comprises more elaborate transformations, such as sentence splitting. It can also contain multiple simplifications of the same original text targeting different audiences, such as school grade levels. We explore these two features of TS to build models tailored for specific grade levels. Our approach uses a standard sequence-to-sequence architecture where the original sequence is annotated with information about the target audience and/or the (predicted) type of simplification operation. We show that it outperforms state-of-the-art TS approaches (up to 3 and 12 BLEU and SARI points, respectively), including when training data for the specific complex-simple combination of grade levels is not available, i.e. zero-shot learning.

Split and Rephrase: Better Evaluation and Stronger Baselines

Roee Aharoni and Yoav Goldberg 14:30–14:45

Splitting and rephrasing a complex sentence into several shorter sentences that convey the same meaning is a challenging problem in NLP. We show that while vanilla seq2seq models can reach high scores on the proposed benchmark (Narayan et al., 2017), they suffer from memorization of the training set which contains more than 89% of the unique simple sentences from the validation and test sets. To aid this, we present a new train-development-test data split and neural models augmented with a copy-mechanism, outperforming the best reported baseline by 8.68 BLEU and fostering further progress on the task.

Autoencoder as Assistant Supervisor: Improving Text Representation for Chinese Social Media Text Summarization

Shuming Ma, Xu Sun, Junyang Lin, and Houfeng Wang 14:45–15:00

Most of the current abstractive text summarization models are based on the sequence-to-sequence model (Seq2Seq). The source content of social media is long and noisy, so it is difficult for Seq2Seq to learn an accurate semantic representation. Compared with the source content, the annotated summary is short and well written. Moreover, it shares the same meaning as the source content. In this work, we supervise the learning of the representation of the source content with that of the summary. In implementation, we regard a summary autoencoder as an assistant supervisor of Seq2Seq. Following previous work, we evaluate our model on a popular Chinese social media dataset. Experimental results show that our model achieves the state-of-the-art performances on the benchmark dataset.

Session 8E: Machine Learning, Question Answering (Short)

219 *Chair: Massimo Piccardi*

Long Short-Term Memory as a Dynamically Computed Element-wise Weighted Sum
Omer Levy, Kenton Lee, Nicholas FitzGerald, and Luke Zettlemoyer 14:00–14:15

LSTMs were introduced to combat vanishing gradients in simple RNNs by augmenting them with gated additive recurrent connections. We present an alternative view to explain the success of LSTMs: the gates themselves are versatile recurrent models that provide more representational power than previously appreciated. We do this by decoupling the LSTM's gates from the embedded simple RNN, producing a new class of RNNs where the recurrence computes an element-wise weighted sum of context-independent functions of the input. Ablations on a range of problems demonstrate that the gating mechanism alone performs as well as an LSTM in most settings, strongly suggesting that the gates are doing much more in practice than just alleviating vanishing gradients.

On the Practical Computational Power of Finite Precision RNNs for Language Recognition
Gail Weiss, Yoav Goldberg, and Eran Yahav 14:15–14:30

While Recurrent Neural Networks (RNNs) are famously known to be Turing complete, this relies on infinite precision in the states and unbounded computation time. We consider the case of RNNs with finite precision whose computation time is linear in the input length. Under these limitations, we show that different RNN variants have different computational power. In particular, we show that the LSTM and the Elman-RNN with ReLU activation are strictly stronger than the RNN with a squashing activation and the GRU. This is achieved because LSTMs and ReLU-RNNs can easily implement counting behavior. We show empirically that the LSTM does indeed learn to effectively use the counting mechanism.

A Co-Matching Model for Multi-choice Reading Comprehension
Shuohang Wang, Mo Yu, Jing Jiang, and Shiyu Chang 14:30–14:45

Multi-choice reading comprehension is a challenging task, which involves the matching between a passage and a question-answer pair. This paper proposes a new co-matching approach to this problem, which jointly models whether a passage can match both a question and a candidate answer. Experimental results on the RACE dataset demonstrate that our approach achieves state-of-the-art performance.

Tackling the Story Ending Biases in The Story Cloze Test
Rishi Sharma, James Allen, Omid Bakhshandeh, and Nasrin Mostafazadeh 14:45–15:00

The Story Cloze Test (SCT) is a recent framework for evaluating story comprehension and script learning. There have been a variety of models tackling the SCT so far. Although the original goal behind the SCT was to require systems to perform deep language understanding and commonsense reasoning for successful narrative understanding, some recent models could perform significantly better than the initial baselines by leveraging human-authorship biases discovered in the SCT dataset. In order to shed some light on this issue, we have performed various data analysis and analyzed a variety of top performing models presented for this task. Given the statistics we have aggregated, we have designed a new crowdsourcing scheme that creates a new SCT dataset, which overcomes some of the biases. We benchmark a few models on the new dataset and show that the top-performing model on the original SCT dataset fails to keep up its performance. Our findings further signify the importance of benchmarking NLP systems on various evolving test sets.

Session 8F: Sentiment (Short)

220 *Chair: Ion Androutsopoulos*

A Multi-sentiment-resource Enhanced Attention Network for Sentiment Classification
Zeyang Lei, Yujiu Yang, Min Yang, and Yi Liu 14:00–14:15

Deep learning approaches for sentiment classification do not fully exploit sentiment linguistic knowledge. In this paper, we propose a Multi-sentiment-resource Enhanced Attention Network (MEAN) to alleviate the problem by integrating three kinds of sentiment linguistic knowledge (e.g., sentiment lexicon, negation words, intensity words) into the deep neural network via attention mechanisms. By using various types of sentiment resources, MEAN utilizes sentiment-relevant information from different representation sub-spaces, which makes it more effective to capture the overall semantics of the sentiment, negation and intensity words for sentiment prediction. The experimental results demonstrate that MEAN has robust superiority over strong competitors.

Pretraining Sentiment Classifiers with Unlabeled Dialog Data
Toru Shimizu, Nobuyuki Shimizu, and Hayato Kobayashi 14:15–14:30

The huge cost of creating labeled training data is a common problem for supervised learning tasks such as sentiment classification. Recent studies showed that pretraining with unlabeled data via a language model can improve the performance of classification models. In this paper, we take the concept a step further by using a conditional language model, instead of a language model. Specifically, we address a sentiment classification task for a tweet analysis service as a case study and propose a pretraining strategy with unlabeled dialog data (tweet-reply pairs) via an encoder-decoder model. Experimental results show that our strategy can improve the performance of sentiment classifiers and outperform several state-of-the-art strategies including language model pretraining.

Disambiguating False-Alarm Hashtag Usages in Tweets for Irony Detection
Hen-Hsen Huang, Chiao-Chen Chen, and Hsin-Hsi Chen 14:30–14:45

The reliability of self-labeled data is an important issue when the data are regarded as ground-truth for training and testing learning-based models. This paper addresses the issue of false-alarm hashtags in the self-labeled data for irony detection. We analyze the ambiguity of hashtag usages and propose a novel neural network-based model, which incorporates linguistic information from different aspects, to disambiguate the usage of three hashtags that are widely used to collect the training data for irony detection. Furthermore, we apply our model to prune the self-labeled training data. Experimental results show that the irony detection model trained on the less but cleaner training instances outperforms the models trained on all data.

Cross-Target Stance Classification with Self-Attention Networks
Chang Xu, Cecile Paris, Surya Nepal, and Ross Sparks 14:45–15:00

In stance classification, the target on which the stance is made defines the boundary of the task, and a classifier is usually trained for prediction on the same target. In this work, we explore the potential for generalizing classifiers between different targets, and propose a neural model that can apply what has been learned from a source target to a destination target. We show that our model can find useful information shared between relevant targets which improves generalization in certain scenarios.

Session 9: Best Paper Session (Sponsored by Apple & Amazon)

The process for selecting best papers and honourable mentions

The Program Committee Co-Chairs (PCs) have defined a multi-step process. Area Chairs (ACs) were asked to select a number of top papers in their areas satisfying as many as possible of the following criteria:

- high quality
- nominated for the award by at least one primary reviewer
- bringing disruptive ground-breaking innovation as compared to the current mainstream

ACs re-read their finalists and discussed among themselves the merits of the nominee's work with the help of the primary reviews. ACs then submitted the papers to the PCs along with their selection decisions. PCs balanced ACs' nominations for diversity and representativeness among areas and the review consistency. They prepared the papers in Softconf for best-paper reviewing and selection. There were 52 best paper candidates.

In parallel, PCs formed the best paper selection committee (BPC) from 22 experts in the field with a mix of expertise and backgrounds and at a good seniority level. In case of COIs, the BPC member was excluded from the further evaluation process. BPC members reviewed 6-8 papers each and provided a short review with respect to the best paper criteria.

Based on BPC recommendations, there were about 20 papers left in the pool. PCs then re-read those papers and discussed their particular merits. Finally, 6 long papers and 2 short papers were selected as honourable mentions. For the best papers, 3 long papers and 2 short papers were selected for presentation in the closing conference session.

The selected honourable mentions and best papers emphasize the diversity of the ACL in terms of research questions, methods, and interdisciplinarity.

Best Long Papers

- *Finding syntax in human encephalography with beam search.* John Hale, Chris Dyer, Adhiguna Kuncoro and Jonathan Brennan.
- *Learning to Ask Good Questions: Ranking Clarification Questions using Neural Expected Value of Perfect Information.* Sudha Rao and Hal Daumé III.
- *Let's do it "again": A First Computational Approach to Detecting Adverbial Presupposition Triggers.* Andre Cianflone, Yulan Feng, Jad Kabbara and Jackie Chi Kit Cheung.

Best Short Papers

- *Know What You Don't Know: Unanswerable Questions for SQuAD.* Pranav Rajpurkar, Robin Jia and Percy Liang.
- *'Lighter' Can Still Be Dark: Modeling Comparative Color Descriptions.* Olivia Winn and Smaranda Muresan.

Session 9: Best Paper Session

Plenary *Chair: Iryna Gurevych, Yusuke Miyao*

Know What You Don't Know: Unanswerable Questions for SQuAD
Pranav Rajpurkar, Robin Jia, and Percy Liang 15:30–15:45

Extractive reading comprehension systems can often locate the correct answer to a question in a context document, but they also tend to make unreliable guesses on questions for which the correct answer is not stated in the context. Existing datasets either focus exclusively on answerable questions, or use automatically generated unanswerable questions that are easy to identify. To address these weaknesses, we present SQuADRUn, a new dataset that combines the existing Stanford Question Answering Dataset (SQuAD) with over 50,000 unanswerable questions written adversarially by crowdworkers to look similar to answerable ones. To do well on SQuADRUn, systems must not only answer questions when possible, but also determine when no answer is supported by the paragraph and abstain from answering. SQuADRUn is a challenging natural language understanding task for existing models: a strong neural system that gets 86% F1 on SQuAD achieves only 66% F1 on SQuADRUn. We release SQuADRUn to the community as the successor to SQuAD.

'Lighter' Can Still Be Dark: Modeling Comparative Color Descriptions
Olivia Winn and Smaranda Muresan 15:45–16:00

We propose a novel paradigm of grounding comparative adjectives within the realm of color descriptions. Given a reference RGB color and a comparative term (e.g., lighter, darker), our model learns to ground the comparative as a direction in the RGB space such that the colors along the vector, rooted at the reference color, satisfy the comparison. Our model generates grounded representations of comparative adjectives with an average accuracy of 0.65 cosine similarity to the desired direction of change. These vectors approach colors with Delta-E scores of under 7 compared to the target colors, indicating the differences are very small with respect to human perception. Our approach makes use of a newly created dataset for this task derived from existing labeled color data.

Finding syntax in human encephalography with beam search
John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan Brennan 16:00–16:25

Recurrent neural network grammars (RNNGs) are generative models of (tree , string) pairs that rely on neural networks to evaluate derivational choices. Parsing with them using beam search yields a variety of incremental complexity metrics such as word surprisal and parser action count. When used as regressors against human electrophysiological responses to naturalistic text, they derive two amplitude effects: an early peak and a P600-like later peak. By contrast, a non-syntactic neural language model yields no reliable effects. Model comparisons attribute the early peak to syntactic composition within the RNNG. This pattern of results recommends the RNNG+beam search combination as a mechanistic model of the syntactic processing that occurs during normal human language comprehension.

Learning to Ask Good Questions: Ranking Clarification Questions using Neural Expected Value of Perfect Information
Sudha Rao and Hal Daumé III 16:25–16:50

Inquiry is fundamental to communication, and machines cannot effectively collaborate with humans unless they can ask questions. In this work, we build a neural network model for the task of ranking clarification questions. Our model is inspired by the idea of expected value of perfect information: a good question is one whose expected answer will be useful. We study this problem using data from StackExchange, a plentiful online resource in which people routinely ask clarifying questions to posts so that they can better offer assistance to the original poster. We create a dataset of clarification questions consisting of 77K posts paired with a clarification question (and answer) from three domains of StackExchange: askubuntu, unix and superuser. We evaluate our model on 500 samples of this dataset against expert human judgments and demonstrate significant improvements over controlled baselines.

Let's do it "again": A First Computational Approach to Detecting Adverbial Presupposition Triggers
Andre Cianflone, Yulan Feng, Jad Kabbara, and Jackie Chi Kit Cheung 16:50–17:15

We introduce the novel task of predicting adverbial presupposition triggers, which is useful for natural language generation tasks such as summarization and dialogue systems. We introduce two new corpora, derived from the Penn Treebank and the Annotated English Gigaword dataset and investigate the use of a novel attention mechanism tailored to this task. Our attention mechanism augments a baseline recurrent neural network without the need for additional trainable parameters, minimizing the added computational cost of our mechanism. We demonstrate that this model statistically outperforms our baselines.

6 Workshops

Thursday

Friday

BioNLP: BioNLP 2018 Workshop

Organizers: *Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii*
Venue: 207

Thursday, July 19, 2018

09:00–09:15 **Opening remarks**

09:15–10:30 **Session 1: Clinical NLP**

09:15–09:30 Embedding Transfer for Low-Resource Medical Named Entity Recognition: A Case Study on Patient Mobility
Denis Newman-Griffis and Ayah Zirikly

09:30–09:45 Multi-task learning for interpretable cause of death classification using key phrase prediction
Serena Jeblee, Mireille Gomes, and Graeme Hirst

09:45–10:00 Identifying Risk Factors For Heart Disease in Electronic Medical Records: A Deep Learning Approach
Thanat Chokwijitkul, Anthony Nguyen, Hamed Hassanzadeh, and Siegfried Perez

10:00–10:15 Keyphrases Extraction from User-Generated Contents in Healthcare Domain Using Long Short-Term Memory Networks
Ilham Fathy Saputra, Rahmad Mahendra, and Alfan Farizki Wicaksono

10:15–10:30 Identifying Key Sentences for Precision Oncology Using Semi-Supervised Learning
Jurica Ševa, Martin Wackerbauer, and Ulf Leser

10:30–11:00 **Coffee Break**

11:00–12:00 **Session 2: Foundations**

11:00–11:15 Ontology alignment in the biomedical domain using entity definitions and context
Lucy Wang, Chandra Bhagavatula, Mark Neumann, Kyle Lo, Chris Wilhelm, and Waleed Ammar

11:15–11:30 Sub-word information in pre-trained biomedical word representations: evaluation and hyper-parameter optimization
Dieter Galea, Ivan Laponogov, and Kirill Veselkov

11:30–11:45 PICO Element Detection in Medical Text via Long Short-Term Memory Neural Networks
Di Jin and Peter Szolovits

11:45–12:00 Coding Structures and Actions with the COSTA Scheme in Medical Conversations
Nan Wang, Yan Song, and Fei Xia

12:00–13:30 **Lunch break**

13:30–14:30 **Invited Talk: "Automating systematic reviews: progress and challenges" – Paul Glasziou**

14:30–15:30 **Session 3 Literature mining and retrieval; Question Answering**

14:30–14:45 A Neural Autoencoder Approach for Document Ranking and Query Refinement in Pharmacogenomic Information Retrieval
Jonas Pfeiffer, Samuel Broscheit, Rainer Gemulla, and Mathias Göschl

14:45–15:00 Biomedical Event Extraction Using Convolutional Neural Networks and Dependency Parsing
Jari Björne and Tapio Salakoski

15:00–15:15 BioAMA: Towards an End to End BioMedical Question Answering System
Vasu Sharma, Nitish Kulkarni, Srividya Pranavi, Gabriel Bayomi, Eric Nyberg, and Teruko Mitamura

15:15–15:30 Phrase2VecGLM: Neural generalized language model–based semantic tagging for complex query reformulation in medical IR
Manirupa Das, Eric Fosler-Lussier, Simon Lin, Soheil Moosavinasab, David Chen, Steve Rust, Yungui Huang, and Rajiv Ramnath

15:30–16:00 **Coffee Break**

16:00–16:15 **Invited Presentation: "A Corpus with Multi-Level Annotations of Patients, Interventions and Outcomes to Support Language Processing for Medical Literature"–Ben Nye**

16:15–18:00 **Poster Session**

- Convolutional neural networks for chemical-disease relation extraction are improved with character-based word embeddings
Dat Quoc Nguyen and Karin Verspoor
- Domain Adaptation for Disease Phrase Matching with Adversarial Networks
Miaofeng Liu, Jialong Han, Haisong Zhang, and Yan Song
- Predicting Discharge Disposition Using Patient Complaint Notes in Electronic Medical Records
Mohamad Salimi and Alla Rozovskaya
- Bacteria and Biotope Entity Recognition Using A Dictionary-Enhanced Neural Network Model
Qiuyue Wang and Xiaofeng Meng
- SingleCite: Towards an improved Single Citation Search in PubMed
Lana Yeganova, Donald C Comeau, Won Kim, W John Wilbur, and Zhiyong Lu
- A Framework for Developing and Evaluating Word Embeddings of Drug-named Entity
Mengnan Zhao, Aaron J. Masino, and Christopher C. Yang
- MeSH-based dataset for measuring the relevance of text retrieval
Won Gyu Kim, Lana Yeganova, Donald C Comeau, W John Wilbur, and Zhiyong Lu
- CRF-LSTM Text Mining Method Unveiling the Pharmacological Mechanism of Off-target Side Effect of Anti-Multiple Myeloma Drugs
Kaiyin Zhou, Sheng Zhang, Xiangyu Meng, Qi Luo, Yuxing Wang, Ke Ding, Yukun Feng, Mo Chen, Kevin Bretonnel Cohen, and Jingbo Xia
- Prediction Models for Risk of Type-2 Diabetes Using Health Claims
Masatoshi Nagata, Kohichi Takai, Keiji Yasuda, Panikos Heracleous, and Akio Yoneyama
- On Learning Better Embeddings from Chinese Clinical Records: Study on Combining In-Domain and Out-Domain Data
Yaqiang Wang, Yunhui Chen, Hongping Shu, and Yongguang Jiang
- Investigating Domain-Specific Information for Neural Coreference Resolution on Biomedical Texts
Long Trieu, Nhung Nguyen, Makoto Miwa, and Sophia Ananiadou
- Toward Cross-Domain Engagement Analysis in Medical Notes
Sara Rosenthal and Adam Faulkner

CogCL: The Eighth Workshop on Cognitive Aspects of Computational Language Learning and Processing

Organizers: *Marco Idiart, Alessandro Lenci, Thierry Poibeau, and Aline Villavicencio*

Venue: 218

Thursday, July 19, 2018

09:00–09:10 **Welcome and Opening Session**

09:10–09:30 **Session I - Semantics**

09:10–09:30 Predicting Brain Activation with WordNet Embeddings
João António Rodrigues, Ruben Branco, João Silva, Chakaveh Saedi, and António Branco

09:30–10:30 **Invited Talk I**

10:30–11:00 **Coffee Break**

11:00–11:50 **Session II - Production**

11:00–11:20 Do Speakers Produce Discourse Connectives Rationally?
Frances Yung and Vera Demberg

11:20–11:50 Language Production Dynamics with Recurrent Neural Networks
Jesús Calvillo and Matthew Crocker

11:50–12:30 **Poster Session**

11:50–12:30 Multi-glance Reading Model for Text Understanding
Pengcheng Zhu, Yujiu Yang, Wenqiang Gao, and Yi Liu

11:50–12:30 Predicting Japanese Word Order in Double Object Constructions
Masayuki Asahara, Satoshi Nambu, and Shin-Ichiro Sano

11:50–12:30 Affordances in Grounded Language Learning
Stephen McGregor and KyungTae Lim

12:30–14:00 **Lunch**

14:00–15:00 **Invited Talk II**

15:00–15:30 **Session III - Processing**

15:00–15:30 Rating Distributions and Bayesian Inference: Enhancing Cognitive Models of Spatial Language Use
Thomas Kluth and Holger Schultheis

15:30–16:00 **Coffee Break**

16:00–17:00 **Session IV - Syntax and Parsing**

16:00–16:30 The Role of Syntax During Pronoun Resolution: Evidence from fMRI
Jixing Li, Murielle Fabre, Wen-Ming Luh, and John Hale

16:30–17:00 A Sound and Complete Left-Corner Parsing for Minimalist Grammars
Miloš Stanojević and Edward Stabler

17:00–17:30 **Panel, Business Meeting and Closing Session**

DeepLo: Workshop on Deep Learning Approaches for Low Resource NLP

Organizers: *Reza Haffari, Colin Cherry, George Foster, Shahram Khadivi, and Bahar Salehi*
Venue: 211

Thursday, July 19, 2018

09:00–09:10 **Opening Remarks**

09:10–09:50 **Invited Talk (Stefan Riezler)**

09:50–10:30 **Invited Talk (Sujith Ravi)**

10:30–11:00 **Coffee Break**

11:00–12:40 **Oral Presentations**

11:00–11:25 Phrase-Based & Neural Unsupervised Machine Translation
Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato

11:25–11:50 Character-level Supervision for Low-resource POS Tagging
Katharina Kann, Johannes Bjerva, Isabelle Augenstein, Barbara Plank, and Anders Søgaard

11:50–12:15 Training a Neural Network in a Low-Resource Setting on Automatically Annotated Noisy Data
Michael A. Hedderich and Dietrich Klakow

12:15–12:40 Exploiting Cross-Lingual Subword Similarities in Low-Resource Document Classification
Mozhi Zhang, Yoshinari Fujinuma, and Jordan Boyd-Graber

12:40–14:00 **Lunch Break**

14:00–15:30 **Poster Session**

- Multi-task learning for historical text normalization: Size matters
Marcel Bollmann, Anders Søgaard, and Joachim Bingel
- Compositional Language Modeling for Icon-Based Augmentative and Alternative Communication
Shiran Dudy and Steven Bedrick
- Multimodal Neural Machine Translation for Low-resource Language Pairs using Synthetic Data
Koel Dutta Chowdhury, Mohammed Hasanuzzaman, and Qun Liu
- Morphological neighbors beat word2vec on the long tail
Clayton Greenberg, Mittul Singh, and Dietrich Klakow
- Multi-Task Active Learning for Neural Semantic Role Labeling on Low Resource Conversational Corpus
Fariz Ikhwantri, Samuel Louvan, Kemal Kurniawan, Bagas Abisena, Valdi Rachman, Alfan Farizki Wicaksono, and Rahmad Mahendra
- Domain Adapted Word Embeddings for Improved Sentiment Classification
Prathusha K Sarma, Yingyu Liang, and Bill Sethares
- Investigating Effective Parameters for Fine-tuning of Word Embeddings Using Only a Small Corpus
Kanako Komiya and Hiroyuki Shinnou

- Dependency Parsing of Code-Switching Data with Cross-Lingual Feature Representations
  KyungTae Lim, Niko Partanen, Michael Rießler, and Thierry Poibeau
- Semi-Supervised Learning with Auxiliary Evaluation Component for Large Scale e-Commerce Text Classification
  Mingkuan Liu, Musen Wen, Selcuk Kopru, Xianjing Liu, and Alan Lu
- Low-rank passthrough neural networks
  Antonio Valerio Miceli Barone
- Embedding Transfer for Low-Resource Medical Named Entity Recognition: A Case Study on Patient Mobility
  Denis Newman-Griffis and Ayah Zirikly

15:30–16:00 **Coffee Break**

16:00–16:40 **Invited Talk (Trevor Cohn)**

16:40–17:40 **Panel Discussion**

17:40–17:55 **Closing Remarks**

MSR: The First Workshop on Multilingual Surface Realization

Organizers: *Simon Mille, Anja Belz, Bernd Bohnet, Emily Pitler, and Leo Wanner*
Venue: 216

Thursday, July 19, 2018

08:45–09:00 **Opening Remarks**

09:00–10:00 **Invited Talk (Hadar Shemtov)**

10:00–10:30 **Shared Task overview and results**

- The First Multilingual Surface Realisation Shared Task (SR'18): Overview and Evaluation Results
  Simon Mille, Anja Belz, Bernd Bohnet, Yvette Graham, Emily Pitler, and Leo Wanner

10:30–11:00 **Coffee Break**

11:00–12:30 **Oral Presentations**

- BinLin: A Simple Method of Dependency Tree Linearization
  Yevgeniy Puzikov and Iryna Gurevych
- IIT (BHU) Varanasi at MSR-SRST 2018: A Language Model Based Approach for Natural Language Generation
  Shreyansh Singh, Ayush Sharma, Avi Chawla, and A.K. Singh
- Surface Realization Shared Task 2018 (SR18): The Tilburg University Approach
  Thiago Castro Ferreira, Sander Wubben, and Emiel Krahmer

12:30–13:45 **Lunch**

13:45–14:15 **Oral Presentation**

- The OSU Realizer for SRST '18: Neural Sequence-to-Sequence Inflection and Incremental Locality-Based Linearization
  David King and Michael White

14:15–15:30 **Poster Session**

- AX Semantics' Submission to the Surface Realization Shared Task 2018
  Andreas Madsack, Johanna Heininger, Nyamsuren Davaasambuu, Vitaliia Voronik, Michael Käufl, and Robert Weißgraeber
- NILC-SWORNEMO at the Surface Realization Shared Task: Exploring Syntax-Based Word Ordering using Neural Models
  Marco Antonio Sobrevilla Cabezudo and Thiago Pardo
- The DipInfo-UniTo system for SRST 2018
  Valerio Basile and Alessandro Mazzei
- Generating High-Quality Surface Realizations Using Data Augmentation and Factored Sequence Models
  Henry Elder and Chris Hokamp

15:30–16:00 **Coffee Break**

16:00–17:30 **Panel/Discussions**

NLPTEA: The 5th Workshop on Natural Language Processing Techniques for Educational Applications

Organizers: *Yuen-Hsien Tseng, Hsin-Hsi Chen, Vincent Ng, and Mamoru Komachi*

Venue: 217

Thursday, July 19, 2018

09:20–09:30 **Opening Remarks**

09:30–10:30 **Invited Talk**

09:30–10:30 **Multi-word Expressions in Second Language Learning (Yuji Matsumoto)**

10:30–11:00 **Coffee Break**

11:00–12:40 **Regular Paper Session**

11:00–11:20 Generating Questions for Reading Comprehension using Coherence Relations
Takshak Desai, Parag Dakle, and Dan Moldovan

11:20–11:40 Syntactic and Lexical Approaches to Reading Comprehension
Henry Lin

11:40–12:00 Feature Optimization for Predicting Readability of Arabic L1 and L2
Hind Saddiki, Nizar Habash, Violetta Cavalli-Sforza, and Muhamed Al Khalil

12:00–12:20 A Tutorial Markov Analysis of Effective Human Tutorial Sessions
Nabin Maharjan and Vasile Rus

12:20–12:40 Thank "Goodness"! A Way to Measure Style in Student Essays
Sandeep Mathias and Pushpak Bhattacharyya

12:40–14:10 **Lunch**

14:10–15:30 **Shared Task Session**

14:10–14:30 Overview of NLPTEA-2018 Share Task Chinese Grammatical Error Diagnosis
Gaoqi Rao, Qi Gong, Baolin Zhang, and Endong Xun

14:30–14:45 Chinese Grammatical Error Diagnosis using Statistical and Prior Knowledge driven Features with Probabilistic Ensemble Enhancement
Ruiji Fu, Zhengqi Pei, Jiefu Gong, Wei Song, Dechuan Teng, Wanxiang Che, Shijin Wang, Guoping Hu, and Ting Liu

14:45–15:00 A Hybrid System for Chinese Grammatical Error Diagnosis and Correction
Chen Li, Junpei Zhou, Zuyi Bao, Hengyou Liu, Guangwei Xu, and Linlin Li

15:00–15:15 LingCASS Solution to the NLP-TEA CGED Shared Task 2018
Qinan Hu, Yongwei Zhang, Fang Liu, and Yueguo Gu

15:15–15:30 Chinese Grammatical Error Diagnosis Based on Policy Gradient LSTM Model
Changliang Li and Ji Qi

15:30–16:00 **Coffee Break**

16:00–17:00 **Poster Session**

- The Importance of Recommender and Feedback Features in a Pronunciation Learning Aid
  Dzikri Fudholi and Hanna Suominen
- Selecting NLP Techniques to Evaluate Learning Design Objectives in Collaborative Multi-perspective Elaboration Activities
  Aneesha Bakharia

- Augmenting Textual Qualitative Features in Deep Convolution Recurrent Neural Network for Automatic Essay Scoring
  Tirthankar Dasgupta, Abir Naskar, Lipika Dey, and Rupsa Saha
- Joint learning of frequency and word embeddings for multilingual readability assessment
  Dieu-Thu Le, Cam-Tu Nguyen, and Xiaoliang Wang
- MULLE: A grammar-based Latin language learning tool to supplement the classroom setting
  Herbert Lange and Peter Ljunglöf
- Textual Features Indicative of Writing Proficiency in Elementary School Spanish Documents
  Gemma Bel-Enguix, Diana Dueñas Chavez, and Arturo Curiel Díaz
- Assessment of an Index for Measuring Pronunciation Difficulty
  Katsunori Kotani and Takehiko Yoshimi
- A Short Answer Grading System in Chinese by Support Vector Approach
  Shih-Hung Wu and Wen-Feng Shih
- From Fidelity to Fluency: Natural Language Processing for Translator Training
  Oi Yee Kwong
- Countering Position Bias in Instructor Interventions in MOOC Discussion Forums
  Muthu Kumar Chandrasekaran and Min-Yen Kan
- Measuring Beginner Friendliness of Japanese Web Pages explaining Academic Concepts by Integrating Neural Image Feature and Text Features
  Hayato Shiokawa, Kota Kawaguchi, Bingcai Han, Takehito Utsuro, Yasuhide Kawada, Masaharu Yoshioka, and Noriko Kando
- Learning to Automatically Generate Fill-In-The-Blank Quizzes
  Edison Marrese-Taylor, Ai Nakajima, Yutaka Matsuo, and Ono Yuichi
- Multilingual Short Text Responses Clustering for Mobile Educational Activities: a Preliminary Exploration
  Yuen-Hsien Tseng, Lung-Hao Lee, Yu-Ta Chien, Chun-Yen Chang, and Tsung-Yen Li
- Chinese Grammatical Error Diagnosis Based on CRF and LSTM-CRF model
  Yujie Zhou, Yinan Shao, and Yong Zhou
- Contextualized Character Representation for Chinese Grammatical Error Diagnosis
  Jianbo Zhao, Si Li, and Zhiqing Lin
- CMMC-BDRC Solution to the NLP-TEA-2018 Chinese Grammatical Error Diagnosis Task
  Zhang Yongwei, Hu Qinan, Liu Fang, and Gu Yueguo
- Detecting Simultaneously Chinese Grammar Errors Based on a BiLSTM-CRF Model
  Yajun Liu, Hongying Zan, Mengjie Zhong, and Hongchao Ma
- A Hybrid Approach Combining Statistical Knowledge with Conditional Random Fields for Chinese Grammatical Error Detection
  Yiyi Wang and Chilin Shih
- CYUT-III Team Chinese Grammatical Error Diagnosis System Report in NLPTEA-2018 CGED Shared Task
  Shih-Hung Wu, JUN-WEI Wang, Liang-Pu Chen, and Ping-Che Yang
- Detecting Grammatical Errors in the NTOU CGED System by Identifying Frequent Subsentences
  Chuan-Jie Lin and Shao-Heng Chen

17:00–17:10 **Closing Remarks**

CALCS: The Third Workshop on Computational Approaches to Linguistic Code-Switching

Organizers: *Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Thamar Solorio, Mona Diab, and Julia Hirschberg*

Venue: 209

Thursday, July 19, 2018

09:00–10:30 **Session 1 Invited Talk and Oral Presentations**

09:00–09:05 **Opening Remarks (Thamar Solorio)**

09:05–09:50 **Invited Talk: Learning to Codeswitch (Pascale Fung)**

09:50–10:10 Joint Part-of-Speech and Language ID Tagging for Code-Switched Data
Victor Soto and Julia Hirschberg

10:10–10:30 Phone Merging For Code-Switched Speech Recognition
Sunit Sivasankaran, Brij Mohan Lal Srivastava, Sunayana Sitaram, Kalika Bali, and Monojit Choudhury

10:30–11:00 **Coffee Break**

11:00–12:00 **Session 2 Oral Presentations**

11:00–11:20 Improving Neural Network Performance by Injecting Background Knowledge: Detecting Code-switching and Borrowing in Algerian texts
Wafia Adouane, Jean-Philippe Bernardy, and Simon Dobnik

11:20–11:40 Code-Mixed Question Answering Challenge: Crowd-sourcing Data and Techniques
Khyathi Chandu, Ekaterina Loginova, Vishal Gupta, Josef van Genabith, Günter Neuman, Manoj Chinnakotla, Eric Nyberg, and Alan W. Black

11:40–12:00 Transliteration Better than Translation? Answering Code-mixed Questions over a Knowledge Base
Vishal Gupta, Manoj Chinnakotla, and Manish Shrivastava

12:00–13:30 **Lunch Break**

13:30–14:15 **Session 3 Invited Talk**

13:30–14:15 **Invited Talk: Variation in Codeswitched Language: a Psycholinguistic Approach to What, When, and Why (Melinda Fricke)**

14:15–15:30 **Session 4 Poster Session**

- Language Identification and Analysis of Code-Switched Social Media Text
Deepthi Mave, Suraj Maharjan, and Thamar Solorio
- Code-Switching Language Modeling using Syntax-Aware Multi-Task Learning
Genta Indra Winata, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung
- Predicting the presence of a Matrix Language in code-switching
Barbara Bullock, Wally Guzman, Jacqueline Serigos, Vivek Sharath, and Almeida Jacqueline Toribio
- Automatic Detection of Code-switching Style from Acoustics
SaiKrishna Rallabandi, Sunayana Sitaram, and Alan W. Black
- Accommodation of Conversational Code-Choice
Anshul Bawa, Monojit Choudhury, and Kalika Bali

- Language Informed Modeling of Code-Switched Text
  Khyathi Chandu, Thomas Manzini, Sumeet Singh, and Alan W. Black
- GHHT at CALCS 2018: Named Entity Recognition for Dialectal Arabic Using Neural Networks
  Mohammed Attia, Younes Samih, and Wolfgang Maier
- Simple Features for Strong Performance on Named Entity Recognition in Code-Switched Twitter Data
  Devanshu Jain, Maria Kustikova, Mayank Darbari, Rishabh Gupta, and Stephen Mayhew
- Bilingual Character Representation for Efficiently Addressing Out-of-Vocabulary Words in Code-Switching Named Entity Recognition
  Genta Indra Winata, Chien-Sheng Wu, Andrea Madotto, and Pascale Fung
- Named Entity Recognition on Code-Switched Data Using Conditional Random Fields
  Utpal Kumar Sikdar, Biswanath Barik, and Björn Gambäck
- The University of Texas System Submission for the Code-Switching Workshop Shared Task 2018
  Florian Janke, Tongrui Li, Eric Rincón, Gualberto Guzmán, Barbara Bullock, and Almeida Jacqueline Toribio
- Tackling Code-Switched NER: Participation of CMU
  Parvathy Geetha, Khyathi Chandu, and Alan W. Black
- Multilingual Named Entity Recognition on Spanish-English Code-switched Tweets using Support Vector Machines
  Daniel Claeser, Samantha Kent, and Dennis Felske

15:30–16:00 **Coffee Break**

16:00–17:00 **Session 5 Shared Task Talks**

16:00–16:10 Named Entity Recognition on Code-Switched Data: Overview of the CALCS 2018 Shared Task
Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Mona Diab, Julia Hirschberg, and Thamar Solorio

16:10–16:30 IIT (BHU) Submission for the ACL Shared Task on Named Entity Recognition on Code-switched Data
Shashwat Trivedi, Harsh Rangwani, and Anil Kumar Singh

16:30–16:50 Code-Switched Named Entity Recognition with Embedding Attention
Changhan Wang, Kyunghyun Cho, and Douwe Kiela

16:50–17:00 **Closing Remarks (Victor Soto)**

MRQA: Workshop on Machine Reading for Question Answering

Organizers: *Eunsol Choi, Minjoon Seo, Danqi Chen, Robin Jia, and Jonathan Berant*
Venue: 210

Thursday, July 19, 2018

08:45–09:00 **Opening Remarks**

Session 1

09:00–09:35 **Invited Talk: Phil Blunsom, University of Oxford / Deepmind**

09:35–10:10 **Invited Talk: Sebastian Riedel, University College London**

10:10–10:30 **Best Paper Talk: A Systematic Classification of Knowledge, Reasoning, and Context within the ARC Dataset**

10:30–11:00 **Morning coffee break**

Session 2

11:00–11:35 **Invited Talk: Richard Socher, Salesforce Research**

11:35–12:10 **Invited Talk: Jianfeng Gao, Microsoft Research**

12:10–13:45 **Lunch**

Session 3

13:45–14:20 **Invited Talk: Antoine Bordes, Facebook AI Research**

14:20–15:30 **Poster Session (with a spotlight presentation)**

14:20–15:30 Ruminating Reader: Reasoning with Gated Multi-hop Attention
Yichen Gong and Samuel Bowman

14:20–15:30 Systematic Error Analysis of the Stanford Question Answering Dataset
Marc-Antoine Rondeau and T. J. Hazen

14:20–15:30 A Multi-Stage Memory Augmented Neural Network for Machine Reading Comprehension
Seunghak Yu, Sathish Reddy Indurthi, Seohyun Back, and Haejun Lee

14:20–15:30 Tackling Adversarial Examples in QA via Answer Sentence Selection
Yuanhang Ren, Ye Du, and Di Wang

14:20–15:30 DuReader: a Chinese Machine Reading Comprehension Dataset from Real-world Applications
Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang

14:20–15:30 Robust and Scalable Differentiable Neural Computer for Question Answering
Jörg Franke, Jan Niehues, and Alexander Waibel

14:20–15:30 A Systematic Classification of Knowledge, Reasoning, and Context within the ARC Dataset
Michael Boratko, Harshit Padigela, Divyendra Mikkilineni, Pritish Yuvraj, Rajarshi Das, Andrew McCallum, Maria Chang, Achille Fokoue-Nkoutche, Pavan Kapanipathi, Nicholas Mattei, Ryan Musa, Kartik Talamadupula, and Michael Witbrock

14:20–15:30 RECIPE: Applying Open Domain Question Answering to Privacy Policies
Yan Shvartzshanider, Ananth Balashankar, Thomas Wies, and Lakshminarayanan Subramanian

14:20–15:30 Neural Models for Key Phrase Extraction and Question Generation
Sandeep Subramanian, Tong Wang, Xingdi Yuan, Saizheng Zhang, Adam Trischler, and Yoshua Bengio

14:20–15:30 Comparative Analysis of Neural QA models on SQuAD
Soumya Wadhwa, Khyathi Chandu, and Eric Nyberg

14:20–15:30 Adaptations of ROUGE and BLEU to Better Evaluate Machine Reading Comprehension Task
An Yang, Kai Liu, Jing Liu, Yajuan Lyu, and Sujian Li

15:30–16:00 **Afternoon coffee break**

16:00–17:00 **Panel Discussion**

RELNLP: Workshop on Relevance of Linguistic Structure in Neural Architectures for NLP

Organizers: *Georgiana Dinu, Miguel Ballesteros, Avirup Sil, Sam Bowman, Wael Hamza, Anders Søgaard, Tahira Naseem, and Yoav Goldberg*

Venue: 208

Thursday, June 19, 2018

08:50–09:00 **Opening Remarks**

Session 1

09:00–10:00 **Invited Talk: Jason Eisner**

10:00–10:20 Compositional Morpheme Embeddings with Affixes as Functions and Stems as Arguments
Daniel Edmiston and Karl Stratos

10:20–11:00 **Break**

Session 2

11:00–12:00 **Invited Talk: Mark Johnson**

12:00–12:20 Unsupervised Source Hierarchies for Low-Resource Neural Machine Translation
Anna Currey and Kenneth Heafield

12:20–13:30 **Lunch**

13:30–14:30 **Session 3 (Poster)**

- Latent Tree Learning with Differentiable Parsers: Shift-Reduce Parsing and Chart Parsing
Jean Maillard and Stephen Clark
- Syntax Helps ELMo Understand Semantics: Is Syntax Still Relevant in a Deep Neural Architecture for SRL?
Emma Strubell and Andrew McCallum
- Subcharacter Information in Japanese Embeddings: When Is It Worth It?
Marzena Karpinska, Bofang Li, Anna Rogers, and Aleksandr Drozd
- A neural parser as a direct classifier for head-final languages
Hiroshi Kanayama, Masayasu Muraoka, and Ryosuke Kohita
- Syntactic Dependency Representations in Neural Relation Classification
Farhad Nooralahzadeh and Lilja Øvrelid

14:30–15:30 **Invited Talk: Emily Bender**

15:30–16:00 **Break**

Session 4

16:00–17:00 **Invited Talk: Chris Dyer**

17:00–18:00 **Panel discussion**

ECONLP: The First Workshop on Economics and Natural Language Processing

Organizers: *Udo Hahn, Véronique Hoste, and Ming-Feng Tsai*
Venue: 207

Friday, July 20, 2018

09:00–09:30 **Introduction to the ECONLP Workshop (Udo Hahn)**

09:30–10:00 Economic Event Detection in Company-Specific News Text
Gilles Jacobs, Els Lefever, and Véronique Hoste

10:00–10:30 Causality Analysis of Twitter Sentiments and Stock Market Returns
Narges Tabari, Piyusha Biswas, Bhanu Praneeth, Armin Seyeditabari, Mirsad Hadzikadic, and Wlodek Zadrozny

10:30–11:00 **Morning Coffee Break**

11:00–11:20 A Corpus of Corporate Annual and Social Responsibility Reports: 280 Million Tokens of Balanced Organizational Writing
Sebastian G.M. Händschke, Sven Buechel, Jan Goldenstein, Philipp Poschmann, Tinghui Duan, Peter Walgenbach, and Udo Hahn

11:20–11:40 Word Embeddings-Based Uncertainty Detection in Financial Disclosures
Christoph Kilian Theil, Sanja Stajner, and Heiner Stuckenschmidt

11:40–12:00 A Simple End-to-End Question Answering Model for Product Information
Tuan Lai, Trung Bui, Sheng Li, and Nedim Lipka

12:00–14:00 **Lunch Break**

14:00–14:20 Sentence Classification for Investment Rules Detection
Youness Mansar and Sira Ferradans

14:20–14:40 Leveraging News Sentiment to Improve Microblog Sentiment Classification in the Financial Domain
Tobias Daudert, Paul Buitelaar, and Sapna Negi

14:40–15:00 Implicit and Explicit Aspect Extraction in Financial Microblogs
Thomas Gaillat, Bernardo Stearns, Gopal Sridhar, Ross McDermott, Manel Zarrouk, and Brian Davis

15:00–15:20 Unsupervised Word Influencer Networks from News Streams
Ananth Balashankar, Sunandan Chakraborty, and Lakshminarayanan Subramanian

15:30–16:00 **Afternoon Coffee Break**

16:00–16:30 **Discussion and Wrap-up**

16:30–17:30 **Poster Session - All of the Papers Presented at the Workshop**

RepL4NLP: The Third Workshop on Representation Learning for NLP

Organizers: *Isabelle Augenstein, Kris Cao, He He, Felix Hill, Spandana Gella, Jamie Kiros, Hongyuan Mei, and Dipendra Misra*

Venue: 210

Friday, July 20, 2018

09:30–09:45 **Welcome and Opening Remarks**

09:45–14:45 **Keynote Session**

09:45–10:30 **Invited Talk 1 (Yejin Choi)**

10:30–11:00 **Coffee Break**

11:00–11:45 **Invited Talk 2 (Trevor Cohn)**

11:45–12:30 **Invited Talk 3 (Margaret Mitchell)**

12:30–14:00 **Lunch**

14:00–14:45 **Invited Talk 4 (Yoav Goldberg)**

14:45–15:00 **Outstanding Papers Spotlight Presentations**

15:00–16:30 **Poster Session (including Coffee Break from 15:30-16:00) + Drinks Reception**

- Corpus Specificity in LSA and Word2vec: The Role of Out-of-Domain Documents
  Edgar Altszyler, Mariano Sigman, and Diego Fernandez Slezak
- Hierarchical Convolutional Attention Networks for Text Classification
  Shang Gao, Arvind Ramanathan, and Georgia Tourassi
- Extrofitting: Enriching Word Representation and its Vector Space with Semantic Lexicons
  Hwiyeol Jo and Stanley Jungkyu Choi
- Chat Discrimination for Intelligent Conversational Agents with a Hybrid CNN-LMTGRU Network
  Dennis Singh Moirangthem and Minho Lee
- Text Completion using Context-Integrated Dependency Parsing
  Amr Rekaby Salama, Özge Alacam, and Wolfgang Menzel
- Quantum-Inspired Complex Word Embedding
  Qiuchi Li, Sagar Uprety, Benyou Wang, and Dawei Song
- Natural Language Inference with Definition Embedding Considering Context On the Fly
  Kosuke Nishida, Kyosuke Nishida, Hisako Asano, and Junji Tomita
- Comparison of Representations of Named Entities for Document Classification
  Lidia Pivovarova and Roman Yangarber
- Speeding up Context-based Sentence Representation Learning with Non-autoregressive Convolutional Decoding
  Shuai Tang, Hailin Jin, Chen Fang, Zhaowen Wang, and Virginia de Sa
- Connecting Supervised and Unsupervised Sentence Embeddings
  Gil Levi
- A Hybrid Learning Scheme for Chinese Word Embedding
  Weiguo Sheng and Weiguo Sheng

- Unsupervised Random Walk Sentence Embeddings: A Strong but Simple Baseline
  Kawin Ethayarajh and Graeme Hirst
- Evaluating Word Embeddings in Multi-label Classification Using Fine-Grained Name Typing
  Yadollah Yaghoobzadeh, Katharina Kann, and Hinrich Schütze
- A Dense Vector Representation for Open-Domain Relation Tuples
  Ade Romadhony, Alfan Farizki Wicaksono, Ayu Purwarianti, and Dwi Hendratmo Widyantoro
- Exploiting Common Characters in Chinese and Japanese to Learn Cross-Lingual Word Embeddings via Matrix Factorization
  Jilei Wang, Shiying Luo, Weiyan Shi, Tao Dai, and Shu-Tao Xia
- WordNet Embeddings
  Chakaveh Saedi, António Branco, João António Rodrigues, and João Silva
- Knowledge Graph Embedding with Numeric Attributes of Entities
  Yanrong Wu and Zhichun Wang
- Injecting Lexical Contrast into Word Vectors by Guiding Vector Space Specialisation
  Ivan Vulić
- Characters or Morphemes: How to Represent Words?
  Ahmet Üstün, Murathan Kurfalı, and Burcu Can
- Learning Hierarchical Structures On-The-Fly with a Recurrent-Recursive Model for Sequences
  Athul Paul Jacob, Zhouhan Lin, Alessandro Sordoni, and Yoshua Bengio
- Limitations of Cross-Lingual Learning from Image Search
  Mareike Hartmann and Anders Søgaard
- Learning Semantic Textual Similarity from Conversations
  Yinfei Yang, Steve Yuan, Daniel Cer, Sheng-Yi Kong, Noah Constant, Petr Pilar, Heming Ge, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil
- Multilingual Seq2seq Training with Similarity Loss for Cross-Lingual Document Classification
  Katherine Yu, Haoran Li, and Barlas Oguz
- LSTMs Exploit Linguistic Attributes of Data
  Nelson F. Liu, Omer Levy, Roy Schwartz, Chenhao Tan, and Noah A. Smith
- Learning Distributional Token Representations from Visual Features
  Samuel Broscheit
- Jointly Embedding Entities and Text with Distant Supervision
  Denis Newman-Griffis, Albert M. Lai, and Eric Fosler-Lussier
- A Sequence-to-Sequence Model for Semantic Role Labeling
  Angel Daza and Anette Frank
- Predicting Concreteness and Imageability of Words Within and Across Languages via Word Embeddings
  Nikola Ljubešić, Darja Fišer, and Anita Peti-Stantić

16:30–17:30 **Panel Discussion**

17:30–17:40 **Closing Remarks + Best Paper Awards Announcement**

Challenge-MML: The First Grand Challenge and Workshop on Human Multimodal Language

Organizers: *Amir Zadeh, Paul Pu Liang, Louis-Philippe Morency, Soujanya Poria, Erik Cambria, and Stefan Scherer*

Venue: 217

Friday, July 20, 2018

09:00–10:30 **Session 1**

09:00–09:10 **Opening Remarks**

09:10–10:00 **Keynote (Bing Liu)**

10:00–10:10 Getting the subtext without the text: Scalable multimodal sentiment classification from visual and acoustic modalities
Nathaniel Blanchard, Daniel Moreira, Aparna Bharati, and Walter Scheirer

10:10–10:20 Recognizing Emotions in Video Using Multimodal DNN Feature Fusion
Jennifer Williams, Steven Kleinegesse, Ramona Comanescu, and Oana Radu

10:20–10:30 Multimodal Relational Tensor Network for Sentiment and Emotion Classification
Saurav Sahay, Shachi H Kumar, Rui Xia, Jonathan Huang, and Lama Nachman

10:30–11:00 **Coffee Break**

11:00–12:30 **Session 2**

11:00–11:50 **Keynote (Sharon Oviatt)**

11:50–12:00 **Advances in Multimodal Datasets (Paul Pu Liang)**

12:00–12:10 Convolutional Attention Networks for Multimodal Emotion Recognition from Speech and Text Data
Woo Yong Choi, Kyu Ye Song, and Chan Woo Lee

12:10–12:20 Sentiment Analysis using Imperfect Views from Spoken Language and Acoustic Modalities
Imran Sheikh, Sri Harsha Dumpala, Rupayan Chakraborty, and Sunil Kumar Kopparapu

12:20–12:30 Polarity and Intensity: the Two Aspects of Sentiment Analysis
Leimin Tian, Catherine Lai, and Johanna Moore

12:30–13:30 **Lunch Break**

13:30–15:00 **Session 3**

13:30–14:20 **Keynote (Roland Goecke)**

14:20–14:30 ASR-based Features for Emotion Recognition: A Transfer Learning Approach
Noé Tits, Kevin El Haddad, and Thierry Dutoit

14:30–14:40 Seq2Seq2Sentiment: Multimodal Sequence to Sequence Models for Sentiment Analysis
Hai Pham, Thomas Manzini, Paul Pu Liang, and Barnabas Poczos

14:40–14:50 DNN Multimodal Fusion Techniques for Predicting Video Sentiment
Jennifer Williams, Ramona Comanescu, Oana Radu, and Leimin Tian

14:50–15:00 **Grand Challenge Results**

15:00–15:05 **Workshop End**

SocialNLP: The Sixth Workshop on Natural Language Processing for Social Media

Organizers: *Lun-Wei Ku and Cheng-Te Li*
Venue: 216

Friday, July 20, 2018

09:20–10:30 **Keynote Speech (I):The Search for Emotions, Creativity, and Fairness in Language (Dr. Saif Mohammad (NSF))**

10:30–11:00 **Coffee Break**

11:00–12:20 **Technical Session 1**

- Sociolinguistic Corpus of WhatsApp Chats in Spanish among College Students
  Alejandro Dorantes, Gerardo Sierra, Tlauhlia Yamín Donohue Pérez, Gemma Bel-Enguix, and Mónica Jasso Rosales
- A Crowd-Annotated Spanish Corpus for Humor Analysis
  Santiago Castro, Luis Chiruzzo, Aiala Rosá, Diego Garat, and Guillermo Moncecchi
- A Twitter Corpus for Hindi-English Code Mixed POS Tagging
  Kushagra Singh, Indira Sen, and Ponnurangam Kumaraguru
- Detecting Offensive Tweets in Hindi-English Code-Switched Language
  Puneet Mathur, Rajiv Shah, Ramit Sawhney, and Debanjan Mahata

12:20–13:20 **Lunch**

13:20–14:30 **Keynote Speech (II): Understanding Online Social Behaviors through Automatic Language Analysis (Dr. Yi-Chia Wang (Uber))**

14:30–15:30 **EmotionX Challenge Session**

- SocialNLP 2018 EmotionX Challenge Overview: Recognizing Emotions in Dialogues
  Chao-Chun Hsu and Lun-Wei Ku
- EmotionX-DLC: Self-Attentive BiLSTM for Detecting Sequential Emotions in Dialogues
  Linkai Luo, Haiqin Yang, and Francis Y. L. Chin
- EmotionX-AR: CNN-DCNN autoencoder based Emotion Classifier
  Sopan Khosla
- EmotionX-SmartDubai_NLP: Detecting User Emotions In Social Media Text
  Hessa AlBalooshi, Shahram Rahmanian, and Rahul Venkatesh Kumar
- EmotionX-Area66: Predicting Emotions in Dialogues using Hierarchical Attention Network with Sequence Labeling
  Rohit Saxena, Savita Bhat, and Niranjan Pedanekar
- EmotionX-JTML: Detecting emotions with Attention
  Johnny Torres

15:30–16:00 **Coffee Break**

16:00–17:00 **Technical Session 2**

- Towards Automation of Sense-type Identification of Verbs in OntoSenseNet
  Sreekavitha Parupalli, Vijjini Anvesh Rao, and Radhika Mamidi
- Improving Classification of Twitter Behavior During Hurricane Events
  Kevin Stowe, Jennings Anderson, Martha Palmer, Leysia Palen, and Ken Anderson

- Political discourse classification in social networks using context sensitive convolutional neural networks
  Aritz Bilbao-Jayo and Aitor Almeida

17:00–17:10 **Closing**

NMT: The 2nd Workshop on Neural Machine Translation and Generation

Organizers: *Alexandra Birch, Andrew Finch, Thang Luong, Graham Neubig, and Yusuke Oda*
Venue: 211

Friday, July 20, 2018

09:00–09:10 **Welcome and Opening Remarks**

- Findings of the Second Workshop on Neural Machine Translation and Generation
  Alexandra Birch, Andrew Finch, Minh-Thang Luong, Graham Neubig, and Yusuke Oda

09:10–10:00 **Keynote 1 (Jacob Devlin)**

10:00–10:30 **Shared Task Overview**

10:30–11:00 **Coffee Break**

11:00–11:30 **Marian: Fast Neural Machine Translation in C++**

11:30–12:20 **Keynote 2 (Rico Sennrich)**

12:20–13:20 **Lunch Break**

13:20–13:50 **Best Paper Session**

13:50–14:40 **Keynote 3 (Jason Weston)**

14:40–15:30 **Keynote 4 (Yulia Tsvetkov)**

15:30–16:00 **Coffee Break**

16:00–17:30 **Poster Session**

- A Shared Attention Mechanism for Interpretation of Neural Automatic Post-Editing Systems
  Inigo Jauregi Unanue, Ehsan Zare Borzeshi, and Massimo Piccardi
- Iterative Back-Translation for Neural Machine Translation
  Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn
- Inducing Grammars with and for Neural Machine Translation
  Yonatan Bisk and Ke Tran
- Regularized Training Objective for Continued Training for Domain Adaptation in Neural Machine Translation
  Huda Khayrallah, Brian Thompson, Kevin Duh, and Philipp Koehn
- Controllable Abstractive Summarization
  Angela Fan, David Grangier, and Michael Auli
- Enhancement of Encoder and Attention Using Target Monolingual Corpora in Neural Machine Translation
  Kenji Imamura, Atsushi Fujita, and Eiichiro Sumita
- Document-Level Adaptation for Neural Machine Translation
  Sachith Sri Ram Kothur, Rebecca Knowles, and Philipp Koehn
- On the Impact of Various Types of Noise on Neural Machine Translation
  Huda Khayrallah and Philipp Koehn
- Bi-Directional Neural Machine Translation with Synthetic Parallel Data
  Xing Niu, Michael Denkowski, and Marine Carpuat

- Multi-Source Neural Machine Translation with Missing Data
  Yuta Nishimura, Katsuhito Sudoh, Graham Neubig, and Satoshi Nakamura
- Towards one-shot learning for rare-word translation with external experts
  Ngoc-Quan Pham, Jan Niehues, and Alexander Waibel
- NICT Self-Training Approach to Neural Machine Translation at NMT-2018
  Kenji Imamura and Eiichiro Sumita
- Fast Neural Machine Translation Implementation
  Hieu Hoang, Tomasz Dwojak, Rihards Krislauks, Daniel Torregrosa, and Kenneth Heafield
- OpenNMT System Description for WNMT 2018: 800 words/sec on a single-core CPU
  Jean Senellart, Dakun Zhang, Bo Wang, Guillaume Klein, Jean-Pierre Ramatchandirin, Josep Crego, and Alexander Rush
- Marian: Cost-effective High-Quality Neural Machine Translation in C++
  Marcin Junczys-Dowmunt, Kenneth Heafield, Hieu Hoang, Roman Grundkiewicz, and Anthony Aue
- On Individual Neurons in Neural Machine Translation
  D. Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass
- Parameter Sharing Strategies in Neural Machine Translation
  Sébastien Jean, Stanislas Lauly, and Kyunghyun Cho
- Modeling Latent Sentence Structure in Neural Machine Translation
  Joost Bastings, Wilker Aziz, Ivan Titov, and Khalil Simaan
- Extreme Adaptation for Personalized Neural Machine Translation
  Paul Michel and Graham Neubig
- Exploiting Semantics in Neural Machine Translation with Graph Convolutional Networks
  Diego Marcheggiani, Joost Bastings, and Ivan Titov

17:30–17:40 **Closing Remarks**

NEWS: The Seventh Named Entities Workshop

Organizers: *Nancy Chen, Rafael E. Banchs, Xiangyu Duan, Min Zhang, and Haizhou Li*
Venue: 208

Friday, July 20, 2018

08:30–08:40 **Opening Remarks**

08:40–09:00 Automatic Extraction of Entities and Relation from Legal Documents
Judith Jeyafreeda Andrew

09:00–09:20 Connecting Distant Entities with Induction through Conditional Random Fields for Named Entity Recognition: Precursor-Induced CRF
Wangjin Lee and Jinwook Choi

09:20–09:40 A Sequence Learning Method for Domain-Specific Entity Linking
Emrah Inan and Oguz Dikenelli

09:40–10:00 Attention-based Semantic Priming for Slot-filling
Jiewen Wu, Rafael E. Banchs, Luis Fernando D'Haro, Pavitra Krishnaswamy, and Nancy Chen

10:00–10:20 Named Entity Recognition for Hindi-English Code-Mixed Social Media Text
Vinay Singh, Deepanshu Vijay, Syed Sarfaraz Akhtar, and Manish Shrivastava

10:30–11:00 **Coffee Break**

11:00–11:20 Forms of Anaphoric Reference to Organisational Named Entities: Hoping to widen appeal, they diversified
Christian Hardmeier, Luca Bevacqua, Sharid Loáiciga, and Hannah Rohde

11:20–11:40 Named-Entity Tagging and Domain adaptation for Better Customized Translation
Zhongwei Li, Xuancong Wang, AiTi Aw, Eng Siong Chng, and Haizhou Li

12:00–14:00 **Lunch**

14:10–14:20 NEWS 2018 Whitepaper
Nancy Chen, Xiangyu Duan, Min Zhang, Rafael E. Banchs, and Haizhou Li

14:20–14:40 Report of NEWS 2018 Named Entity Transliteration Shared Task
Nancy Chen, Rafael E. Banchs, Min Zhang, Xiangyu Duan, and Haizhou Li

14:40–15:00 Statistical Machine Transliteration Baselines for NEWS 2018
Snigdha Singhania, Minh Nguyen, Gia H Ngo, and Nancy Chen

15:00–15:20 A Deep Learning Based Approach to Transliteration
Soumyadeep Kundu, Sayantan Paul, and Santanu Pal

15:30–16:00 **Coffee Break**

16:00–16:20 Comparison of Assorted Models for Transliteration
Saeed Najafi, Bradley Hauer, Rashed Rubby Riyadh, Leyuan Yu, and Grzegorz Kondrak

16:20–16:40 Neural Machine Translation Techniques for Named Entity Transliteration
Roman Grundkiewicz and Kenneth Heafield

16:40–17:00 Low-Resource Machine Transliteration Using Recurrent Neural Networks of Asian Languages
Ngoc Tan Le and Fatiha Sadat

NLPOSS: Workshop for NLP Open Source Software

Organizers: *Eunjeong L. Park, Masato Hagiwara, Dmitrijs Milajevs, and Liling Tan*
Venue: 209

Friday, July 20, 2018

08:45–09:00 **Loading Presentations to Computer**

09:00–09:05 **Opening Remarks**

09:05–09:50 **Invited Talk 1 (Joel Nothman)**

09:50–10:30 **Lightning Presentation for Posters Session 1**

10:30–11:00 **Coffee Break**

11:00–11:45 **Poster Session 1**

- AllenNLP: A Deep Semantic Natural Language Processing Platform
  Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer
- Stop Word Lists in Free Open-source Software Packages
  Joel Nothman, Hanmin Qin, and Roman Yurchak
- Texar: A Modularized, Versatile, and Extensible Toolbox for Text Generation
  Zhiting Hu, Zichao Yang, Tiancheng Zhao, Haoran Shi, Junxian He, Di Wang, Xuezhe Ma, Zhengzhong Liu, Xiaodan Liang, Lianhui Qin, Devendra Singh Chaplot, Bowen Tan, Xingjiang Yu, and Eric Xing
- The ACL Anthology: Current State and Future Directions
  Daniel Gildea, Min-Yen Kan, Nitin Madnani, Christoph Teichmann, and Martin Villalba
- The risk of sub-optimal use of Open Source NLP Software: UKB is inadvertently state-of-the-art in knowledge-based WSD
  Eneko Agirre, Oier Lopez de Lacalle, and Aitor Soroa

12:00–14:00 **Lunch**

14:00–14:45 **Invited Talk 2 (Christopher Manning)**

14:45–15:30 **Lightning Presentation for Posters 2**

15:30–16:00 **Break**

16:00–16:45 **Poster Session 2**

- Baseline: A Library for Rapid Modeling, Experimentation and Development of Deep Learning Algorithms targeting NLP
  Daniel Pressel, Sagnik Ray Choudhury, Brian Lester, Yanjie Zhao, and Matt Barta
- OpenSeq2Seq: Extensible Toolkit for Distributed and Mixed Precision Training of Sequence-to-Sequence Models
  Oleksii Kuchaiev, Boris Ginsburg, Igor Gitman, Vitaly Lavrukhin, Carl Case, and Paulius Micikevicius
- Integrating Multiple NLP Technologies into an Open-source Platform for Multilingual Media Monitoring
  Ulrich Germann, Renars Liepins, Didzis Gosko, and Guntis Barzdins
- The Annotated Transformer
  Alexander Rush

16:45–17:30 **Invited Talk 3 (Matthew Honnibal and Ines Montani)**

17:30–17:35 **Closing Remarks**

7

Anti-harassment Policy

The open exchange of ideas, the freedom of thought and expression, and respectful scientific debate are central to the aims and goals of a ACL conference. These require a community and an environment that recognizes the inherent worth of every person and group, that fosters dignity, understanding, and mutual respect, and that embraces diversity. For these reasons, ACL is dedicated to providing a harassment-free experience for participants at our events and in our programs.

Harassment and hostile behavior are unwelcome at any ACL conference. This includes: speech or behavior (including in public presentations and on-line discourse) that intimidates, creates discomfort, or interferes with a person's participation or opportunity for participation in the conference. We aim for ACL conferences to be an environment where harassment in any form does not happen, including but not limited to: harassment based on race, gender, religion, age, color, national origin, ancestry, disability, sexual orientation, or gender identity. Harassment includes degrading verbal comments, deliberate intimidation, stalking, harassing photography or recording, inappropriate physical contact, and unwelcome sexual attention.

It is the responsibility of the community as a whole to promote an inclusive and positive environment for our scholarly activities. In addition, any participant who experiences harassment or hostile behavior may contact any current member of the ACL Board or contact Priscilla Rasmussen, who is usually available at the registration desk of the conference. Please be assured that if you approach us, your concerns will be kept in strict confidence, and we will consult with you on any actions taken.

The ACL board members are listed at:

`https://www.aclweb.org/portal/about`

The full policy and its implementation is defined at:

`https://www.aclweb.org/adminwiki/index.php?title=Anti-Harassment_Policy`

8

ACL Author Guidelines

Preserving Double Blind Review

The following rules and guidelines are meant to protect the integrity of double-blind review and ensure that submissions are reviewed fairly. The rules make reference to the anonymity period, which runs from 1 month before the submission deadline up to the date when your paper is either accepted, rejected, or withdrawn.

- You may not make a non-anonymized version of your paper available online to the general community (for example, via a preprint server) during the anonymity period. By a version of a paper we understand another paper having essentially the same scientific content but possibly differing in minor details (including title and structure) and/or in length (e.g., an abstract is a version of the paper that it summarizes).

- If you have posted a non-anonymized version of your paper online before the start of the anonymity period, you may submit an anonymized version to the conference. The submitted version must not refer to the non-anonymized version, and you must inform the program chair(s) that a non-anonymized version exists. You may not update the non-anonymized version during the anonymity period, and we ask you not to advertise it on social media or take other actions that would further compromise double-blind reviewing during the anonymity period.

- Note that, while you are not prohibited from making a non-anonymous version available online before the start of the anonymity period, this does make double-blind reviewing more difficult to maintain, and we therefore encourage you to wait until the end of the anonymity period if possible. Alternatively, you may consider submitting your work to the Computational Linguistics journal, which does not require anonymization and has a track for "short" (i.e., conference-length) papers.

Citation and Comparison

If you are aware of previous research that appears sound and is relevant to your work, you should cite it even if it has not been peer-reviewed, and certainly if it influenced your own work. However, refereed publications take priority over unpublished work reported in preprints. Specifically:

- You are expected to cite all refereed publications relevant to your submission, but you may be excused for not knowing about all unpublished work (especially work that has been recently posted and/or is not widely cited).
- In cases where a preprint has been superseded by a refereed publication, the refereed publication should be cited in addition to or instead of the preprint version.

Papers (whether refereed or not) appearing less than 3 months before the submission deadline are considered contemporaneous to your submission, and you are therefore not obliged to make detailed comparisons that require additional experimentation and/or in-depth analysis.

9 Local Guide

This guide was originally written by Tony Wirth, with additions by Jey Han Lau, Trevor Cohn and Timothy Baldwin.

Introduction

Melburnians are only too happy to tell you that their city is the cultural capital of Australia, or the sporting capital of Australia. Though true, Melbourne's real international claim is that its food is world class and its coffee is even better, with outstanding cafés and coffee roasters scattered across the city. A combination of strong competition and a multicultural population that frequently dines out has kindled a vibrant café and restaurant scene.

The CBD, as we like to call it, aka "the city" or Melbourne 3000 – or Downtown if you want to sound like a tourist – has the highest density of restaurants in greater Melbourne. In general, the streets to the east of Elizabeth Street, especially, east of Russell Street, are best for evenings. For an outstanding gelato experience, with experimental flavours, drop into Gelateria Primavera. Chinatown has its spine along Little Bourke Street, including the long-standing Golden Orchids.

Closest to the conference venue, the most vibrant area is near the corner of Katherine Place and Flinders Lane. Further east along Flinders Lane, you can find Dukes coffee roasters. Degraves Street, either side of Flinders Lane, is packed with cafés, and is very popular with visitors and locals alike.

There are several quality *rooftop bars* in the CBD, including Bomba, Siglo, Campari House, Rooftop Bar (Curtin House), Madame Brussels, and Red Hummingbird. For a view of the city, and of local commuters rushing for their trains, head to the delightful Arbory. Either side, along the Yarra River, Ponyfish Island and Riverland are gems.

Explore Melbourne's *laneways*: some of them have amazing places to eat (Hardware Lane, Centre Place, Degraves Street), some have fantastic artwork (Hosier Lane), some are just a little scary looking.

To browse a list of curated cafés, restaurants and other notable places interactively on your smartphone, please check out the following Google Maps Places: `https://goo.gl/maps/6Q96QDCHbJD2`.

Useful Links

Food `https://www.broadsheet.com.au/melbourne`

Cafés `https://www.beanhunter.com/melbourne`
`https://www.broadsheet.com.au/melbourne/guides/best-coffee`

Things To Do `http://www.visitvictoria.com/Regions/Melbourne`

Getting Around

Melbourne is famous for its trams, and has one of the most extensive tram networks in the world. The CBD is designated a Free Tram Zone, meaning that you can ride any tram for free, although beware, as the Free Tram Zone finishes one stop short of the conference venue and ticket inspectors frequent the fringes of the zone trying to catch out those who ride without a ticket, and are infamous for their intolerance (including tourists). If you wish to catch a tram beyond the Free Tram Zone, you will need to purchase a Myki ticket from a newsagent or one of the many vending machines at tram stops, and "touch on" each time you get on a tram. Note that if you touch on within the Free Tram Zone, you will be charged. No, not the most user-friendly system in the world, but there are clear announcements in the trams of whether you are in the Free Tram Zone or not. The same Myki ticket will work on local trains and busses, should you need to travel further afield. Google maps gives a good overview of public transport options, as does `https://www.ptv.vic.gov.au/` and the PTV phone app.

Uber operates in Melbourne, and tends to be cheaper than taxis for short trips.

Surrounding Suburbs

South Melbourne Some of the great cafés of Melbourne, St Ali and Chez Dré, are just a few blocks south of the conference venue.

Fitzroy Taking the Number 96 tram north for about 25 minutes, you'll arrive in Fitzroy. Apart from the CBD, this inner suburb has the highest concentration of restaurants. There are classic atmospheric pubs, such as The Napier, Labour in Vain, and The Standard, as well as the newer Naked for Satan with its rooftop Naked in the Sky. The cafés are top notch, with Industry Beans featuring in an article by *The Huffington Post* on the most hipster neighborhoods in the world. In general, Brunswick Street and Smith Street reward aimless wandering, and can be reached on the 59 tram.

Bayside Port Melbourne is a short ride down the 109 tram. Bay Street has several good eateries, including bakery Noisette. St Kilda is the southern terminus of the Number 96 tram. The Sunday esplanade market is a Melbourne classic, the scenic railway at Luna Park is the oldest continuously running roller coaster in the world, and the continental cakes along Acland Street have kept the area buzzing for decades.

Guide to be a Tourist in Melbourne

- Buying coffee at Starbucks (Melburnians do pride themselves on their local café culture!)
- Meaning anything other than "Aussie rules football" when referring to "football" (Melburnians are, in large part, famously one-eyed when it comes to football codes)... in fact using the term "football" at all, as Australians love to abbreviate everything, including "footie"

- Not having an immediate response and breaking into impassioned dialogue/song when asked "who do you barrack for" (referring, of course, to the footie team you support)

- Tipping — tipping culture is very limited in Melbourne, and it is only at high-end restaurants where there is really any expectation of a tip, and even here it is optional. When eating out in large groups, high-end restaurants will sometimes charge a group surcharge, meaning even less reason to tip. Certainly there is no need to tip in taxis or at cafés (other than in the form of loose change in the tipping jar).

- Not having a humorous/whimsical come-back at the ready at all times — Melburnians are generally a very friendly, laid-back bunch who try not to take themselves too seriously (except when it comes to footie, of course), and like to light-heartedly "take the piss" when the opportunity arises

- Comparing Sydney with Melbourne favourably in any way — Melburnians are very proud of their city, and fiercely territorial when it comes to comparisons with Sydney

- Asking about "Australian" eating options — Melbourne is proudly multicultural and very proud of its "foodie" culture, and "Australian" cuisine is representative of that: a melting pot of the myriad of different cuisines of the many migrant groups who make up its population; if there is an "Australian" cuisine, it is in the blending/fusion of different cuisines. Particular cuisines where Melbourne excels include Chinese, Japanese, Korean, Vietnamese, Malaysian, Indonesian, Italian, and Greek, with many fantastic options within easy access of the conference venue.

- Being suspicious of off-laneway, hidden-by-dumpsters cafés — in general, the more obscure the location, the more hipster the café and more authentically local the clientele.

- Walking/riding/driving on the right — yes, we get that much of the rest of the world drives on the left, but we don't.

- Not holding the door for the person behind you — not particular to Melbourne necessarily, and people aren't being chauvinistic or patronising in any way, it's just a politeness thing here.

- Getting annoyed when people approach you to ask if you need directions when you have a map out or are otherwise looking lost — don't feel patronised or defensive, it's just the locals being friendly.

- Asking where the kangaroos or koalas are in central Melbourne — kangaroos (or "roos" as the locals call them) outnumber people in Australia (koalas are less common), yes, but you won't find them hopping down Swanston St. Where you will spot them is in places where there is open farmland, where they thrive.

- Being offended at the standards of "colourful" language in Australia — it's not uncommon for Australians to pepper their speech with swear words, as a way of colouring their speech, with no intention of offending anyone's sensibilities.

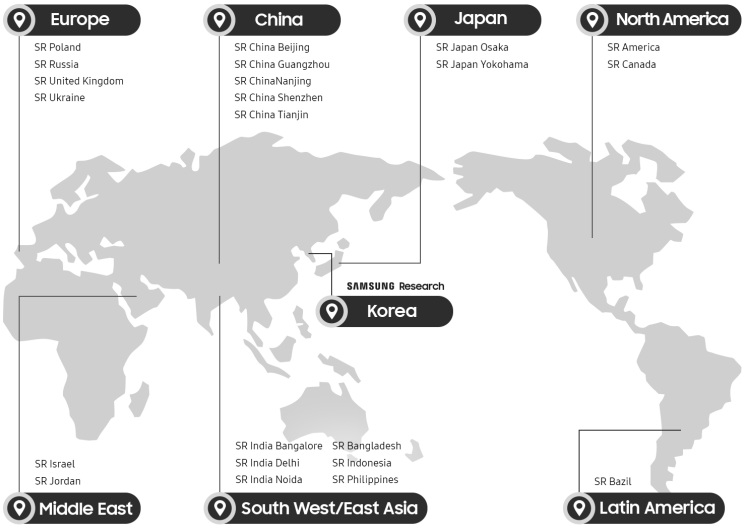
Index

SAMSUNG Research

Shape the Future with Innovation and Intelligence

To create a valuable future through innovationn and intelligence

Global R&D Network



RESEARCH AREAS

Artificial Intelligence | Data Intelligence | IoT | Tizen | Smart Machine | Security | Next Generation Media

HOW TO APPLY

- career.sr@samsung.com
- http://research.samsung.com

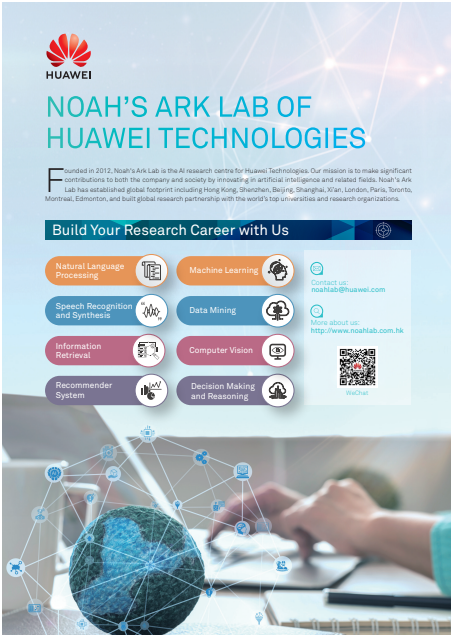
Megagon Labs

Megagon Labs (formerly Recruit Institute of Technology) is the technology hub and research lab for Recruit Holdings, a company that provides over 200 online services in the areas of human resources, travel, housing, education, restaurants and many other areas in which people make daily lifestyle decisions.

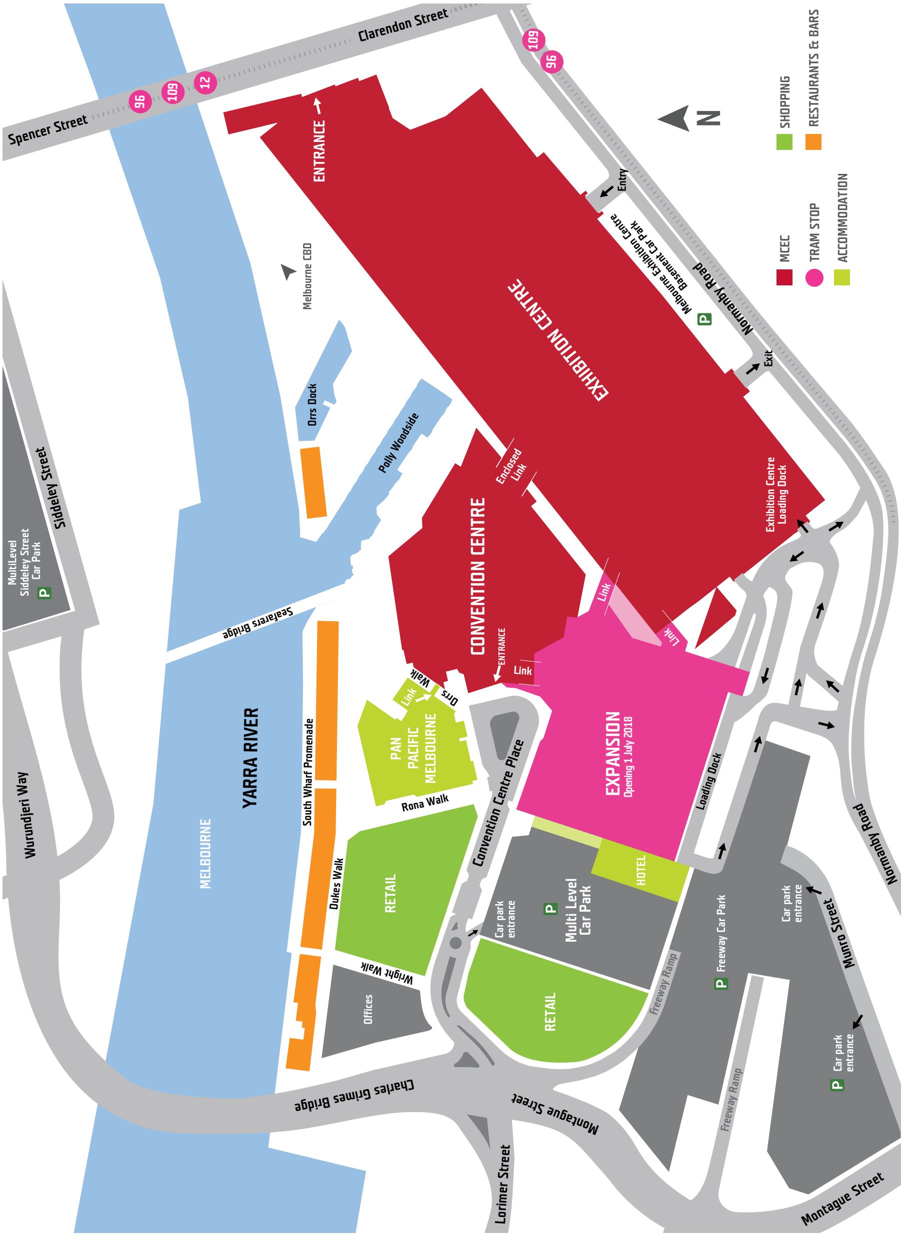
We conduct world-class research in the several areas, including natural language processing, machine learning, data management, data integration, and artificial intelligence. We collaborate with research in universities and publish in top-notch conferences.

Megagon Labs is hiring Research Scientists and Software Engineers. Visit our website at www.megagon.ai for more information.

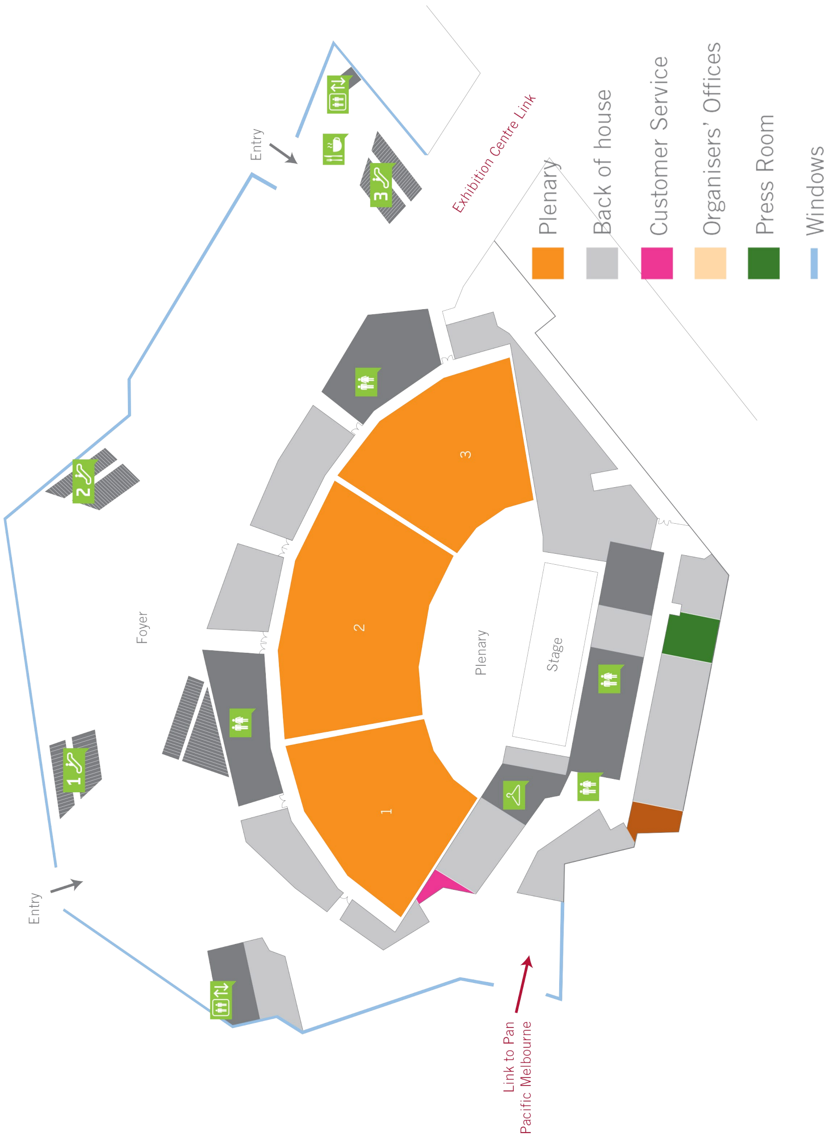

Tencent
AI
Tencent is applying its technology leadership, deep understanding of connection and user behavior and open platform/ecosystem to create AI breakthroughs that can solve the world's most pressing challenges. We are not innovating for its own sake or simply seeking to win AI contents. We are exploring how AI can enhance the quality of human life.
Tencent's mission is to enhance the quality of human life through Internet services and AI enables us to do this in powerful new ways. We are focused on solving key problems for our users, customers and society. Through fundamental and applied research, we seek to create products and services that make the benefits of AI accessible to everyone, everywhere.
Tencent AI Job List
Principle/Senior Researcher-Natural Language & Speech Processing, Tencent AI Lab
Location: Seattle Bellevue, United States
Research Scientist-Natural Language Processing, Tencent Lingo Lab
Location: Beijing / Shenzhen / Chengdu, China
Senior Researcher-Natural Language Processing , Conversational Bots, Tencent Cloud
Location: Shenzhen, China
(Senior) Researcher-Natural Language Processing, Wechat
Location: Beijing/Guangzhou/Shenzhen, China
Senior Researcher-Natural Language & Speech Processing, Tencent Games
Location: Shenzhen, China/Los Angeles, United States
Senior Researcher-Natural Language Processing, Financial Data Application Department
Location: Shenzhen, China
Connect with us
LinkedIn:@tencent
WeChat:@QQjobs
Tencent


HUAWEI
NOAH'S ARK LAB OF HUAWEI TECHNOLOGIES
Founded in 2012, Noah's Ark Lab is the AI research centre for Huawei Technologies. Our mission is to make significant contributions to both the company and society by innovating in artificial intelligence and related fields. Noah's Ark Lab has established global footprint including Hong Kong, Shenzhen, Beijing, Shanghai, Xi'an, London, Paris, Toronto, Montreal, Edmonton, and built global research partnership with the world's top universities and research organizations.
Build Your Research Career with Us
Natural Language Processing
Machine Learning
Speech Recognition and Synthesis
Data Mining
Information Retrieval
Computer Vision
Recommender System
Decision Making and Reasoning
Contact us:
noahlab@huawei.com
More about us:
http://www.noahlab.com.hk
WeChat

Getting from Conference to Student Recruitment and Social Event



Precinct Map



Convention Centre Ground Floor

Convention Centre Level Two

ACL gratefully acknowledges the following sponsors for their support!

Diamond

Google

ByteDance 字节跳动

facebook

SAMSUNG

Platinum

amazon

Tencent 腾讯

Gold

Microsoft

IBM Research AI

CVTE 视源股份

NAVER | LINE

digital health crc

Silver

NUANCE

HUAWEI

ELSEVIER

Bronze

Australian Government
Department of Defence
Science and Technology

Student Volunteer Sponsors

Supporter

ROAM

SINTELIX