emnlp2017

Copenhagen, Denmark
September 7-11, 2017

Conference on Empirical Methods in Natural Language Processing

Conference Proceedings

www.emnlp2017.net

The EMNLP organizers gratefully acknowledge support from the following sponsors:

Platinum

Gold

Silver

Bronze

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-945626-83-8

Preface by the General Chair

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

We have an exciting program lined up for you, with three Invited talks, fifteen workshops, seven tutorials, nine TACL presentations, 322 reviewed papers presented as both oral talks and posters, and twenty-one demos. I am especially grateful to our Program Chairs, Rebecca Hwa and Sebastian Riedel, who did a fantastic job managing a backbreaking 1,500 paper submissions (1466 reviewed papers). This involved 51 Area chairs and 980 reviewers. We tried some new things this year (never conducive to a smooth process) including a more careful handling of the COIs that result from Area Chair submissions, and the addition of a meta-review step to encourage more thoughtful reviewing. We are soliciting feedback on the meta-review process, from both reviewers and authors. Despite the additional time involvement, many of the Area Chairs embraced this new approach, and would like to repeat it. However, there are clearly a few dissenters, since Rebecca and Sebastian ended up writing around 200 meta-reviews themselves at the last minute! We are also trying to raise the visibility and status of the poster sessions by integrating them as parallel sessions alongside oral talks, with poster session chairs. This is in response to the survey results from EMNLP 2015 that indicated a decided preference for smaller, more frequent poster sessions during the day rather than evening mega-sessions. Finally, Rebecca and Sebastian are bringing you three outstanding invited speakers, Dan Jurafsky, Sharon Goldwater, and Nando de Freitas. No program chairs ever worked harder to bring you a superb set of presentations in an attendee friendly setting.

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

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

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

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

Best Regards,
Martha Palmer
EMNLP 2017 General Chair

Preface by the Program Committee Co-Chairs

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

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

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

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

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

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

- Martha Palmer, whose encouragement and advice as the general chair has been invaluable every step of the way;

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

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

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

Organizing Committee

General Chair
Martha Palmer, University of Colorado

Local Arrangements Chair
Priscilla Rasmussen, ACL Business Manager

Program Committee Co-chairs
Rebecca Hwa, University of Pittsburgh
Sebastian Riedel, University College London

Local Arrangements Co-chairs
Dirk Hovy, University of Copenhagen
Anders Søgaard, University of Copenhagen

Local Sponsorship Chair
Daniel Hardt, Copenhagen Business School

Workshop Co-chairs
Victoria Fossum, Google
Karl Moritz Hermann, DeepMind

Tutorial Co-chairs
Alexandra Birch, University of Edinburgh
Nathan Schneider, Georgetown University

Demos Co-chairs
Lucia Specia, University of Sheffield
Matt Post, Johns Hopkins University
Michael Paul, University of Colorado

Publications Sr Chair
Siddharth Patwardhan, Apple

Publications Jr Chair
Preethi Raghavan, TJ Watson Lab, IBM

Publicity Chair
Isabelle Augenstein, University of Copenhagen

Web Chair
Anders Johannsen, Apple

Conference Handbook Chair
Joachim Bingel, University of Copenhagen

Conference Handbook Advisor
Matt Post, Johns Hopkins University

Handbook Proofreader
Pontus Stenetorp, University College London

Conference App Chair
Chloé Braud, University of Copenhagen

Student Scholarship Co-chair and Student Volunteer Coordinator
Željko Agić, IT University of Copenhagen
Yonatan Bisk, University of Southern California, ISI

SIGDAT Liason
Chris Callison-Birch, University of Pennsylvania

Program Committee Co-chairs
Rebecca Hwa, University of Pittsburgh
Sebastian Riedel, University College London

Area Chairs
Information Extraction, Information Retrieval, and Question Answering
Mihai Surdeanu, University of Arizona
Jing Jiang, Singapore Management University
Hinrich Schütze, LMU Munich
Sameer Singh, UC Irvine
Scott Wen-tau Yih, MSR
Tomek Strzalkowski, SUNY Albany

Language and Vision
Sanja Fidler, University of Toronto
Hannaneh Hajishirzi, University of Washington

Linguistic Theories and Psycholinguistics
William Schuler, The Ohio State University

Machine Learning
Mohit Bansal, UNC Chapel Hill
Jordan Boyd-Graber, University of Colorado
Trevor Cohn, University of Melbourne
Hal Daumé, University of Maryland
Alona Fyshe, University of Victoria
Anoop Sarkar, Simon Fraser University

Machine Translation and Multilinguality
Marine Carpuat, University of Maryland
David Chiang, University of Notre Dame
Mona Diab, George Washington University
Dan Gildea, University of Rochester
Philipp Koehn, Johns Hopkins University

Segmentation, Tagging, and Parsing
Jinho Choi, Emory University
Julia Hockenmaier, University of Illinois at Urbana-Champaign
Alexander Rush, Harvard University
Zhang Yue, Singapore University of Technology and Design

Semantics
Roberto Basili, University of Roma, Tor Vergata
Chris Biemann, University of Hamburg
Ed Grefenstette, DeepMind
Tom Kwiatkowski, Google
Sameer Pradhan, cemantix.org and Boulder Learning, Inc
Vivek Srikumar, University of Utah

Sentiment Analysis and Opinion Mining
Bing Liu, University of Illinois at Chicago
Rada Mihalcea, University of Michigan
Saif M. Mohammad, National Research Council Canada

Social Media and Computational Social Science
David Bamman, University of California, Berkeley
Cristian Danescu-Niculescu-Mizil, Cornell University
Mark Dredze, Johns Hopkins University
Jacob Eisenstein, Georgia Tech

Spoken Language Processing
Mari Ostendorf, University of Washington

Summarization, Generation, Discourse, Dialogue
Joyce Chai, Michigan State University
Jennifer Chu-Carroll, Elemental Cognition
Kentaro Inui, Tohoku University
Gina Levow, University of Washington
Amanda Stent, Bloomberg LP

Text Mining and NLP Applications
Dipanjan Das, Google
Mamoru Komachi, Tokyo Metropolitan University
Anna Korhonen, University of Cambridge
Nitin Madnani, Educational Testing Service (ETS)
Marie-Francine Moens, KU Leuven
Thamar Solorio, University of Houston
Andreas Vlachos, University of Sheffield

Primary Reviewers

Muhammad Abdul-Mageed; Amjad Abu-Jbara; Heike Adel; Željko Agić; Eneko Agirre; Salah Ait-Mokhtar; Ahmet Aker; Cem Akkaya; Afra Alishahi; Alexandre Allauzen; Tim Althoff; Carlos Alzate; Bharat Ram Ambati; Antonios Anastasopoulos; Daniel Andor; Jacob Andreas; Nicholas Andrews;

Anietie Andy; Gabor Angeli; Marianna Apidianaki; Yuki Arase; Arturo Argueta; Ron Artstein; Yoav Artzi; Ehsaneddin Asgari; Nicholas Asher; Ramón Astudillo; Isabelle Augenstein; Michael Auli; Eleftherios Avramidis; amittai axelrod; Wilker Aziz; Yoram Bachrach; Hessam Bagherinezhad; Collin Baker; Niranjan Balasubramanian; Jason Baldridge; Timothy Baldwin; Tyler Baldwin; Kalika Bali; Miguel Ballesteros; Siddhartha Banerjee; Chitta Baral; Marco Baroni; Alberto Barrón-Cedeño; Pierpaolo Basile; Fernando Batista; Riza Theresa Batista-Navarro; Timo Baumann; Daniel Beck; Beata Beigman Klebanov; Núria Bel; Yonatan Belinkov; Dane Bell; Eric Bell; Kedar Bellare; Islam Beltagy; Anja Belz; Emily M. Bender; Darina Benikova; Luciana Benotti; Jonathan Berant; Taylor Berg-Kirkpatrick; Nicola Bertoldi; Laurent Besacier; Steven Bethard; Chandra Bhagavatula; Suma Bhat; Archna Bhatia; Joachim Bingel; Alexandra Birch; Arianna Bisazza; Yonatan Bisk; Johannes Bjerva; Anders Björkelund; Alan W Black; Eduardo Blanco; John Blitzer; Ivan Bogatyy; Bernd Bohnet; Gemma Boleda; Kalina Bontcheva; Stefan Bordag; Johan Bos; Houda Bouamor; Guillaume Bouchard; Samuel Bowman; Kristy Boyer; S.R.K. Branavan; Chloé Braud; Fabienne Braune; Felipe Bravo-Marquez; Chris Brew; Chris Brockett; Julian Brooke; Caroline Brun; Dominique Brunato; William Bryce; Christian Buck; Paul Buitelaar; Florin Bulgarov; Wray Buntine; Jill Burstein; Bill Byrne; Donna Byron; Elena Cabrio; Aoife Cahill; Nicoletta Calzolari; Jose Camacho-Collados; Erik Cambria; Nicola Cancedda; Marie Candito; Hailong Cao; Cornelia Caragea; Dallas Card; Xavier Carreras; Francisco Casacuberta; Tommaso Caselli; Taylor Cassidy; Vittorio Castelli; Asli Celikyilmaz; Daniel Cer; Özlem Çetinoğlu; Mauro Cettolo; Nathanael Chambers; Yee Seng Chan; Angel Chang; Baobao Chang; Kai-Wei Chang; Snigdha Chaturvedi; Wanxiang Che; Ciprian Chelba; Boxing Chen; Chen Chen; Danqi Chen; Hsin-Hsi Chen; John Chen; Lei Chen; Tao Chen; Wenliang Chen; Yidong Chen; Yun-Nung Chen; Zhiyuan Chen; Jianpeng Cheng; Colin Cherry; Jackie Chi Kit Cheung; Hai Leong Chieu; laura chiticariu; Eunsol Choi; Yejin Choi; Christos Christodoulopoulos; Grzegorz Chrupała; Tagyoung Chung; Philipp Cimiano; Kevin Clark; Stephen Clark; Ann Clifton; Maximin Coavoux; Anne Cocos; Nigel Collier; Michael Collins; Miriam Connor; John Conroy; Matthieu Constant; Danish Contractor; Ryan Cotterell; Benoit Crabbé; Danilo Croce; Montse Cuadros; Iria da Cunha; Andrew Dai; Bhavana Dalvi; Lena Dankin; Amitava Das; Pradeep Dasigi; Munmun De Choudhury; Adrià de Gispert; Daniël de Kok; Eric De La Clergerie; Gerard de Melo; Vera Demberg; Thomas Demeester; Dina Demner-Fushman; Steve DeNeefe; John DeNero; Lingjia Deng; Pascal Denis; Michael Denkowski; Tejaswini Deoskar; Valeria dePaiva; Leon Derczynski; Nina Dethlefs; Ann devitt; Jacob Devlin; Giuseppe Di Fabbrizio; Georgiana Dinu; Ellen Dodge; Jesse Dodge; A. Seza Doğruöz; Li Dong; Qing Dou; Doug Downey; Gabriel Doyle; Eduard Dragut; Mark Dras; Markus Dreyer; Lan Du; Loic Dugast; Kevin Duh; Greg Durrett; Chris Dyer; Marc Dymetman; Richard Eckart de Castilho; Judith Eckle-Kohler; Maud Ehrmann; Andreas Eisele; Jason Eisner; Asif Ekbal; Michael Elhadad; Messina Enza; Gülşen Eryiğit; Hugo Jair Escalante; Keelan Evanini; James Fan; Federico Fancellu; Hao Fang; Stefano Faralli; Richárd Farkas; Noura Farra; Manaal Faruqui; Benoit Favre; Afsaneh Fazly; Marcello Federico; Christian Federmann; Geli Fei; Anna Feldman; Yang Feng; Yansong Feng; Raquel Fernandez; Daniel Fernández-González; Olivier Ferret; Katja Filippova; Andrew Finch; Nicholas FitzGerald; Jeffrey Flanigan; Lucie Flekova; Michael Flor; Radu Florian; Antske Fokkens; José A. R. Fonollosa; Mikel L. Forcada; Eric Fosler-Lussier; George Foster; Jennifer Foster; Stefan L. Frank; Stella Frank; Alexander Fraser; Lea Fremmann; Daniel Fried; Mario Fritz; Xiao Fu; Atsushi Fujii; Matthias Gallé; Michel Galley; Michael Gamon; Kuzman Ganchev; Juri Ganitkevitch; Wei Gao; Claire Gardent; Matt Gardner; Ekaterina Garmash; Dan Garrette; Milica Gasic; Tao Ge; Michaela Geierhos; Spandana Gella; Georgi Georgiev; Kallirroi Georgila; Ulrich Germann; George Giannakopoulos; Kevin Gimpel; Filip Ginter; Filip Ginter; Goran Glavaš; Alfio Gliozzo; Koldo Gojenola; Yoav Goldberg; Dan Goldwasser; Juan Carlos Gomez; Carlos Gómez-Rodríguez; Graciela

Gonzalez; Matthew R. Gormley; Cyril Goutte; Amit Goyal; Pawan Goyal; Joao Graca; David Grangier; Spence Green; Eleni Gregoromichelaki; Cyril Grouin; Adam Grycner; Curry Guinn; Hongyu GUO; Jiafeng Guo; Jiang Guo; Weiwei Guo; Nitish Gupta; Pankaj Gupta; Sonal Gupta; Francisco Guzmán; Nizar Habash; Barry Haddow; Gholamreza Haffari; Masato Hagiwara; Udo Hahn; Gus Hahn-Powell; Dilek Hakkani-Tur; David Hall; Keith Hall; Na-Rae Han; Oul Han; Shuguang Han; Xianpei Han; Sanda Harabagiu; Daniel Hardt; Mark Hasegawa-Johnson; Kazuma Hashimoto; Eva Hasler; Helen Hastie; Katsuhiko Hayashi; He He; Hua He; Luheng He; Wenqi He; Xiaodong He; Yifan He; Yulan He; Kenneth Heafield; Michael Heilman; James Henderson; John Henderson; Aron Henriksson; Aurélie Herbelot; Teresa Herrmann; Daniel Hershcovich; Jack Hessel; Ryuichiro Higashinaka; Derrick Higgins; Felix Hill; Erhard Hinrichs; Gerold Hintz; Tsutomu Hirao; Julia Hirschberg; Graeme Hirst; Hieu Hoang; Nathan Hodas; Kristy Hollingshead; Ales Horak; Chiori Hori; Julian Hough; Yifan Hu; Fei Huang; Heyan Huang; Liang Huang; Lifu Huang; Minlie Huang; Ruihong Huang; Shujian Huang; Xuanjing Huang; Zhongqiang Huang; Luwen Huangfu; Samar Husain; Young-Sook Hwang; Gonzalo Iglesias; Ryu Iida; Diana Inkpen; Naoya Inoue; Radu Tudor Ionescu; Ozan Irsoy; Alexei V. Ivanov; Mohit Iyyer; Guillaume Jacquet; Peter Jansen; Yangfeng Ji; Ping Jian; Wenbin Jiang; Anders Johannsen; Michael Johnston; Kristiina Jokinen; Doug Jones; Shafiq Joty; Marcin Junczys-Dowmunt; Dan Jurafsky; David Jurgens; Nobuhiro Kaji; Pallika Kanani; Hiroshi Kanayama; Dimitri Kartsaklis; Arzoo Katiyar; Daisuke Kawahara; Aniruddha Kembhavi; Ruth Kempson; Casey Kennington; Mitesh M. Khapra; Tushar Khot; Bernd Kiefer; Douwe Kiela; Yuta Kikuchi; Jin-Dong Kim; Seokhwan Kim; Tracy Holloway King; Brian Kingsbury; Svetlana Kiritchenko; Chunyu Kit; Roman Klinger; Julien Kloetzer; Kevin Knight; Simon Kocbek; Ekaterina Kochmar; Thomas Kollar; Kazunori Komatani; Rik Koncel-Kedziorski; Lingpeng Kong; Ioannis Konstas; Parisa Kordjamshidi; Alexander Kotov; Zornitsa Kozareva; Mikhail Kozhevnikov; Martin Krallinger; Jayant Krishnamurthy; Canasai Kruengkrai; Lun-Wei Ku; Sandra Kübler; Marco Kuhlmann; Roland Kuhn; Shankar Kumar; Jonathan K. Kummerfeld; Tsung-Ting Kuo; Sadao Kurohashi; Nate Kushman; Polina Kuznetsova; Igor Labutov; Mathias Lambert; Patrik Lambert; Vasileios Lampos; Ian Lane; Ni Lao; Mirella Lapata; Jey Han Lau; Alon Lavie; Joseph Le Roux; John Lee; Kenton Lee; Sungjin Lee; Els Lefever; Tao Lei; Alessandro Lenci; Omer Levy; Roger Levy; Mike Lewis; Fangtao Li; Haibo Li; Jing Li; Junyi Jessy Li; Qi Li; Sujian Li; Wenjie Li; Yanen Li; Zhenghua Li; Maria Liakata; Constantine Lignos; Chuan-Jie Lin; Victoria Lin; Wang Ling; Xiao Ling; Tal Linzen; Pierre Lison; Diane Litman; Changsong Liu; Fei Liu; Fei Liu; Jiangming Liu; Jing Liu; Kang Liu; Qian Liu; Qun Liu; Ting Liu; Yang Liu; Yang Liu; Yiqun Liu; Zhanyi Liu; Zhiyuan Liu; Adam Lopez; Oier Lopez de Lacalle; Adrian Pastor López Monroy; Annie Louis; Ryan Lowe; Bin Lu; Wei Lu; Yi Luan; Marco Lui; Minh-Thang Luong; Franco M. Luque; Anh Tuan Luu; Teresa Lynn; Ji Ma; Xuezhe Ma; Klaus Macherey; Wolfgang Macherey; Pierre Magistry; Suraj Maharjan; Wolfgang Maier; Igor Malioutov; Shervin Malmasi; Suresh Manandhar; Gideon Mann; Christopher D. Manning; Saab Mansour; Amin Mantrach; Diego Marcheggiani; Daniel Marcu; Anna Margolis; Alex Marin; Héctor Martínez Alonso; André F. T. Martins; Bruno Martins; Yuichiroh Matsubayashi; Yuji Matsumoto; Takuya Matsuzaki; Austin Matthews; Arne Mauser; Jonathan May; Diana Maynard; Andrew McCallum; Diana McCarthy; David McClosky; Yashar Mehdad; Yelena Mejova; Pablo Mendes; Helen Meng; Haitao Mi; Yishu Miao; Claudiu Mihăilă; Timothy Miller; Tristan Miller; Bonan Min; Paramita Mirza; Dipendra Misra; Dipendra Misra; Makoto Miwa; Daichi Mochihashi; Ashutosh Modi; Karo Moilanen; Manuel Montes; Christof Monz; Taesun Moon; Raymond Mooney; Roser Morante; Véronique MORICEAU; Alessandro Moschitti; Nasrin Mostafazadeh; Roozbeh Mottaghi; Animesh Mukherjee; Dragos Munteanu; Yugo Murawaki; Smaranda Muresan; Kenton Murray; Maria Nadejde; Ajay Nagesh; Mikio Nakano; Ndapandula Nakashole; Preslav Nakov; Courtney Napoles; Jason Naradowsky; Karthik Narasimhan; Shashi Narayan; Alexis Nasr; Vivi Nastase; Borja

Navarro; Roberto Navigli; Adeline Nazarenko; Mark-Jan Nederhof; Arvind Neelakantan; Sapna Negi; Aida Nematzadeh; Graham Neubig; Hwee Tou Ng; Jun-Ping Ng; Vincent Ng; Axel-Cyrille Ngonga Ngomo; Dong Nguyen; Thien Huu Nguyen; Toan Nguyen; Jian-Yun NIE; Nobal Bikram Niraula; Joakim Nivre; Hiroshi Noji; Joel Nothman; Pierre Nugues; Diarmuid Ó Séaghdha; Brendan O'Connor; Stephan Oepen; Kemal Oflazer; Alice Oh; Jong-Hoon Oh; Naoaki Okazaki; Manabu Okumura; Constantin Orasan; Vicente Ordonez; Myle Ott; Muntsa Padró; Alexis Palmer; Martha Palmer; Shimei Pan; Sinno Jialin Pan; Alexander Panchenko; Bo Pang; Denis Paperno; Aasish Pappu; Natalie Parde; Ankur Parikh; Niki Parmar; Patrick Paroubek; Rebecca J. Passonneau; Panupong Pasupat; John K Pate; Siddharth Patwardhan; Michael J. Paul; Adam Pauls; Umashanthi Pavalanathan; Ellie Pavlick; Adam Pease; Viktor Pekar; Anselmo Peñas; Xiaochang Peng; Sergio Penkale; Gerald Penn; Julien PEREZ; Verónica Pérez-Rosas; Johann Petrak; Slav Petrov; Nghia The Pham; Peter Phandi; Olivier Pietquin; Daniele Pighin; Mohammad Taher Pilehvar; Joelle Pineau; Yuval Pinter; Emily Pitler; Barbara Plank; Julien Plu; Massimo Poesio; Tamara Polajnar; Heather Pon-Barry; Simone Paolo Ponzetto; Ana-Maria Popescu; Andrei Popescu-Belis; Maja Popović; Fred Popowich; Soujanya Poria; Matt Post; Christopher Potts; Vinodkumar Prabhakaran; Daniel Preoţiuc-Pietro; Prokopis Prokopidis; Emily Prud'hommeaux; Jay Pujara; Matthew Purver; Ashequl Qadir; Longhua Qian; Xian Qian; Long Qiu; Minghui Qiu; Ella Rabinovich; Will Radford; Alessandro Raganato; Preethi Raghavan; Altaf Rahman; Nazneen Fatema Rajani; Rafal Rak; Bhuvana Ramabhadran; Carlos Ramisch; Delip Rao; Ari Rappoport; Mohammad Sadegh Rasooli; Sravana Reddy; Ines Rehbein; Roi Reichart; Nils Reimers; David Reitter; Steffen Remus; Xiang Ren; Adithya Renduchintala; Corentin Ribeyre; Giuseppe Riccardi; Matthew Richardson; Martin Riedl; Verena Rieser; Stefan Riezler; German Rigau; Ellen Riloff; Laura Rimell; Fabio Rinaldi; Eric Ringger; Brian Riordan; Alan Ritter; Brian Roark; Kirk Roberts; Molly Roberts; Tim Rocktäschel; Anna Rohrbach; Marcus Rohrbach; Lina M. Rojas Barahona; Stephen Roller; Salvatore Romeo; Paolo Rosso; Mihai Rotaru; Benjamin Roth; Michael Roth; Alla Rozovskaya; Vasile Rus; Attapol Rutherford; Derek Ruths; Fatiha Sadat; Mehrnoosh Sadrzadeh; Markus Saers; Kenji Sagae; Horacio Saggion; Rishiraj Saha Roy; Magnus Sahlgren; Keisuke Sakaguchi; Mohammad Salameh; Shimi Salant; Yunita Sari; Ruhi Sarikaya; Ryohei Sasano; Hassan Sawaf; Asad Sayeed; David Schlangen; Jonathan Schler; Natalie Schluter; Helmut Schmid; Sebastian Schuster; H. Andrew Schwartz; Lane Schwartz; Roy Schwartz; Stephanie Schwartz; Holger Schwenk; Djamé Seddah; Satoshi Sekine; Ethan Selfridge; Jean Senellart; Rico Sennrich; Minjoon Seo; BURR SETTLES; Izhak Shafran; Kashif Shah; Samira Shaikh; Amr Sharaf; Rebecca Sharp; Wade Shen; xiaodong shi; Chaitanya Shivade; Prasha Shrestha; Avirup Sil; Fabrizio Silvestri; Yanchuan Sim; Dan Simonson; Kiril Simov; Steve Skiena; Kevin Small; Sharon Small; Noah A. Smith; Jan Šnajder; Parinaz Sobhani; Richard Socher; Anders Søgaard; Swapna Somasundaran; Hyun-Je Song; Linfeng Song; Sandeep Soni; Alessandro Sordoni; Aitor Soroa; Victor Soto; Rohini Srihari; Somayajulu Sripada; Christian Stab; Sanja Štajner; Miloš Stanojević; Gabriel Stanovsky; Manfred Stede; Mark Steedman; Pontus Stenetorp; Lucas Sterckx; Mark Stevenson; Brandon Stewart; Ian Stewart; Svetlana Stoyanchev; Veselin Stoyanov; Carlo Strapparava; Karl Stratos; Keh-Yih Su; Qi Su; Yu Su; Fabian Suchanek; Alane Suhr; Fei Sun; Huan Sun; Le Sun; Ming Sun; Jun Suzuki; Swabha Swayamdipta; Idan Szpektor; Oscar Täckström; Shabnam Tafreshi; Hiroya Takamura; Partha Talukdar; Chenhao Tan; Hristo Tanev; Duyu Tang; Jian Tang; Xavier Tannier; Makarand Tapaswi; Christoph Teichmann; TestSeb TestRiedel; Joel Tetreault; Simone Teufel; Kapil Thadani; Sam Thomson; Jörg Tiedemann; Christoph Tillmann; Ivan Titov; Takenobu Tokunaga; Gaurav Singh Tomar; Nadi Tomeh; Marc Tomlinson; Sara Tonelli; Kentaro Torisawa; Lamia Tounsi; Isabel Trancoso; Adam Trischler; Yuta Tsuboi; Oren Tsur; Yoshimasa Tsuruoka; Yulia Tsvetkov; Marco Turchi; Ferhan Ture; Kateryna Tymoshenko; Raghavendra Udupa; Stefan Ultes; Lyle Ungar; Shyam Upadhyay; L. Alfonso Urena Lopez; Dmitry Ustalov; Jakob Uszkoreit; Masao Utiyama; Naushad UzZaman; Sowmya Vajjala;

Marco A. Valenzuela-Escárcega; Tim Van de Cruys; Kees van Deemter; Ielka van der Sluis; Benjamin Van Durme; Gertjan van Noord; Marten van Schijndel; David Vandyke; Tony Veale; Eva Maria Vecchi; sriram venkatapathy; Giulia Venturi; Ashish Venugopal; Patrick Verga; Marc Verhagen; Yannick Versley; Guido Vetere; Paul Vicol; David Vilar; Aline Villavicencio; Rob Voigt; Svitlana Volkova; Ivan Vulić; Yogarshi Vyas; V.G.Vinod Vydiswaran; Henning Wachsmuth; Joachim Wagner; Byron Wallace; Matthew Walter; Stephen Wan; Xiaojun Wan; Chuan Wang; Dingquan Wang; Hai Wang; Houfeng WANG; Lu Wang; Qin Iris Wang; Shuai Wang; Shuohang Wang; William Yang Wang; Zhiguo Wang; Zhongqing Wang; Leo Wanner; Taro Watanabe; Nick Webb; Bonnie Webber; Ingmar Weber; Kellie Webster; Julie Weeds; Furu Wei; Gerhard Weikum; David Weir; Ralph Weischedel; David Weiss; Dirk Weissenborn; Robert West; Michael White; Michael Wick; Michael Wiegand; John Wieting; Jason D Williams; Steven Wilson; Guillaume Wisniewski; Michael Wojatzki; Kam-Fai Wong; Hua Wu; Stephen Wu; Joern Wuebker; Rui Xia; Deyi Xiong; Wei Xu; Wenduan Xu; Yadollah Yaghoobzadeh; Adam Yala; Rui Yan; Bishan Yang; Diyi Yang; Yi Yang; Roman Yangarber; Berrin Yanikoglu; Helen Yannakoudakis; tae yano; Mark Yatskar; Seid Muhie Yimam; Wenpeng Yin; Dani Yogatama; Naoki Yoshinaga; Jianfei Yu; Mo Yu; Ning Yu; Zhuoran Yu; François Yvon; Taras Zagibalov; Marcos Zampieri; Fabio Massimo Zanzotto; Sina Zarrieß; Victoria Zayats; Richard Zens; Torsten Zesch; Luke Zettlemoyer; Congle Zhang; Hao Zhang; Justine Zhang; Lei Zhang; Meishan Zhang; Min Zhang; Qi Zhang; Yuan Zhang; Hai Zhao; Jun Zhao; Shiqi Zhao; Wayne Xin Zhao; Bowen Zhou; Muhua Zhu; Xiaodan Zhu; Yukun Zhu; Michael Zock; Chengqing Zong; Ingrid Zukerman; Pierre Zweigenbaum

Invited Speaker: Dan Jurafsky, Stanford University
"Does This Vehicle Belong to You”? Processing the Language of Policing for Improving Police-Community Relations"

Abstract: Police body-cameras have the potential to play an important role in understanding and improving police-community relations. In this talk I describe a series of studies conducted by our large interdisciplinary team at Stanford that use speech and natural language processing on body-camera recordings to model the interactions between police officers and community members in traffic stops. We use text and speech features to automatically measure linguistic aspects of the interaction, from discourse factors like conversational structure to social factors like respect. I describe the differences we find in the language directed toward black versus white community members, and offer suggestions for how these findings can be used to help improve the fraught relations between police officers and the communities they serve.

Bio: Dan Jurafsky is Professor and Chair of Linguistics and Professor of Computer Science, at Stanford University. His research has focused on the extraction of meaning, intention, and affect from text and speech, on the processing of Chinese, and on applying natural language processing to the cognitive and social sciences. Dan’s deep interest in NLP education led him to co-write with Jim Martin the widely-used textbook "Speech and Language Processing” (whose 3rd edition is in (slow) progress) and co-teach with Chris Manning the first massive open online class on natural language processing. Dan was the recipient of the 2002 MacArthur Fellowship and is a 2015 James Beard Award Nominee for his book, "The Language of Food: A Linguist Reads the Menu".

Invited Speaker: Sharon Goldwater, University of Edinburgh
Towards more universal language technology: unsupervised learning from speech

Abstract: Speech and language processing has advanced enormously in the last decade, with successful applications in machine translation, voice-activated search, and even language-enabled personal assistants. Yet these systems typically still rely on learning from very large quantities of human-annotated data. These resource-intensive methods mean that effective technology is available for only a tiny fraction of the world’s 7000 or so languages, mainly those spoken in large rich countries.

This talk describes our recent work on developing unsupervised speech technology, where transcripts and pronunciation dictionaries are not used. The work is inspired by considering both how young infants may begin to acquire the sounds and words of their language, and how we might develop systems to help linguists analyze and document endangered languages. I will first present work on learning from speech audio alone, where the system must learn to segment the speech stream into word tokens and cluster repeated instances of the same word together to learn a lexicon of vocabulary items. The approach combines Bayesian and neural network methods to address learning at the word and sub-word levels.

Bio: Sharon Goldwater is a Reader at the University of Edinburgh’s School of Informatics, where she is a member of the Institute for Language, Cognition and Computation. She received her PhD in 2007 from Brown University and spent two years as a postdoctoral researcher at Stanford University before moving to Edinburgh. Her research interests include unsupervised learning for speech and language processing, computer modelling of language acquisition in children, and computational studies of language use. Dr. Goldwater co-chaired the 2014 Conference of the European Chapter of the Association for Computational Linguistics and is Chair-Elect of EACL. She has served on the editorial boards of the Transactions of the Association for Computational Linguistics, the Computational Linguistics journal, and OPEN MIND: Advances in Cognitive Science (a new open-access journal). In 2016, she received the Roger Needham Award from the British Computer Society, awarded for "distinguished research contribution in computer science by a UK-based researcher who has completed up to 10 years of post-doctoral research."

Invited Speaker: Nando de Freitas, Google Deepmind
Physical simulation, learning and language

Abstract: Simulated physical environments, with common physical laws, objects and agents with bodies, provide us with consistency to facilitate transfer and continual learning. In such environments, research topics such as learning to experiment, learning to learn and emergent communication can be easily explored. Given the relevance of these topics to language, it is natural to ask ourselves whether research in language would benefit from the development of such environments, and whether language can contribute toward improving such environments and agents within them. This talk will provide an overview of some of these environments, discuss learning to learn and its potential relevance to language, and present some deep reinforcement learning agents that capitalize on formal language instructions to develop disentangled interpretable representations that allow them to generalize to a wide variety of zero-shot semantic tasks. The talk will pose more questions than answers in the hope of stimulating discussion.

Bio: I was born in Zimbabwe, with malaria. I was a refugee from the war in Mocambique and thanks to my parents getting in debt to buy me a passport from a corrupt official, I grew up in Portugal without water and electricity, before the EU got there, and without my parents who were busy making money to pay their debt. At 8, I joined my parents in Venezuela and began school in the hood; see City of God. I moved to South Africa after high-school and sold beer illegally in black-townships for a living until 1991. Apartheid was the worst thing I ever experienced. I did my BSc in electrical engineering and MSc in control at the University of the Witwatersrand, where I strived to be the best student to prove to racists that anyone can do it. I did my PhD on Bayesian methods for neural networks at Trinity College, Cambridge University. I did a postdoc in Artificial Intelligence at UC Berkeley. I became a Full Professor at the University of British Columbia, before joining the University of Oxford in 2013. I quit Oxford in 2017 to join DeepMind full-time, where I lead the Machine Learning team. I aim to solve intelligence so that future generations have a better life. I have been a Senior Fellow of the Canadian Institute for Advanced Research for a long time. Some of my recent awards, mostly thanks to my collaborators, include: Best Paper Award at the International Conference on Machine Learning (2016), Best Paper Award at the International Conference on Learning Representations (2016), Winner of round 5 of the Yelp Dataset Challenge (2015), Distinguished Paper Award at the International Joint Conference on Artificial Intelligence (2013), Charles A. McDowell Award for Excellence in Research (2012), and Mathematics of Information Technology and Complex Systems Young Researcher Award (2010).

Table of Contents

Conference Program

Friday, September 8, 2017

19:00–22:00 Welcome Reception

Saturday, September 9, 2017

07:30–17:30 Registration Day 1

08:00–08:30 ***Morning Coffee***

08:30–09:00 Plenary Session. Opening Remarks

08:30–09:00 *Opening Remarks*
General Chair, PC Co-Chairs

09:00–10:00 Plenary Session. Invited Talk by Nando de Freitas

09:00–10:00 *Physical simulation, learning and language*
Nando de Freitas

10:00–10:30 ***Coffee Break***

10:30–12:10 Session 1A: Syntax 1

10:30–10:55 *Monolingual Phrase Alignment on Parse Forests*
Yuki Arase and Jun'ichi Tsujii

10:55–11:20 *Fast(er) Exact Decoding and Global Training for Transition-Based Dependency Parsing via a Minimal Feature Set*
Tianze Shi, Liang Huang and Lillian Lee

11:20–11:45 *Parsing with Traces: An O(n^4) Algorithm and a Structural Representation*
Jonathan K. Kummerfeld and Dan Klein

11:45–12:10 *Quasi-Second-Order Parsing for 1-Endpoint-Crossing, Pagenumber-2 Graphs*
Junjie Cao, Sheng Huang, Weiwei Sun and Xiaojun Wan

10:30–12:10 Session 1B: Information Extraction 1

10:30–10:55 *Position-aware Attention and Supervised Data Improve Slot Filling*
Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli and Christopher D. Manning

10:55–11:20 *Heterogeneous Supervision for Relation Extraction: A Representation Learning Approach*
Liyuan Liu, Xiang Ren, Qi Zhu, Shi Zhi, Huan Gui, Heng Ji and Jiawei Han

11:20–11:45 *Integrating Order Information and Event Relation for Script Event Prediction*
Zhongqing Wang, Yue Zhang and Ching-Yun Chang

11:45–12:10 *Entity Linking for Queries by Searching Wikipedia Sentences*
Chuanqi Tan, Furu Wei, Pengjie Ren, Weifeng Lv and Ming Zhou

Saturday, September 9, 2017 (continued)

10:30–12:10 Session 1C: Multilingual NLP

10:30–10:55 *Train-O-Matic: Large-Scale Supervised Word Sense Disambiguation in Multiple Languages without Manual Training Data*
Tommaso Pasini and Roberto Navigli

10:55–11:20 *Universal Semantic Parsing*
Siva Reddy, Oscar Täckström, Slav Petrov, Mark Steedman and Mirella Lapata

11:20–11:45 *Mimicking Word Embeddings using Subword RNNs*
Yuval Pinter, Robert Guthrie and Jacob Eisenstein

11:45–12:10 *Past, Present, Future: A Computational Investigation of the Typology of Tense in 1000 Languages*
Ehsaneddin Asgari and Hinrich Schütze

10:30–12:10 Session 1D: Poster Session: Demo

12:10–13:40 ***Lunch***

13:40–15:20 Session 2A: Machine Translation 1

13:40–14:05 *Neural Machine Translation with Source-Side Latent Graph Parsing*
Kazuma Hashimoto and Yoshimasa Tsuruoka

14:05–14:30 *Neural Machine Translation with Word Predictions*
Rongxiang Weng, Shujian Huang, Zaixiang Zheng, XIN-YU DAI and Jiajun CHEN

14:30–14:55 *Towards Decoding as Continuous Optimisation in Neural Machine Translation*
Cong Duy Vu Hoang, Gholamreza Haffari and Trevor Cohn

14:55–15:20 *Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation*
Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes and Jeffrey Dean

Saturday, September 9, 2017 (continued)

13:40–15:20 Session 2D: Poster Session. Embeddings

Hierarchical Embeddings for Hypernymy Detection and Directionality
Kim Anh Nguyen, Maximilian Köper, Sabine Schulte im Walde and Ngoc Thang Vu

Ngram2vec: Learning Improved Word Representations from Ngram Co-occurrence Statistics
Zhe Zhao, Tao Liu, Shen Li, Bofang Li and Xiaoyong Du

Dict2vec : Learning Word Embeddings using Lexical Dictionaries
Julien Tissier, Christopher Gravier and Amaury Habrard

Learning Chinese Word Representations From Glyphs Of Characters
Tzu-ray Su and Hung-yi Lee

Learning Paraphrastic Sentence Embeddings from Back-Translated Bitext
John Wieting, Jonathan Mallinson and Kevin Gimpel

Joint Embeddings of Chinese Words, Characters, and Fine-grained Subcharacter Components
Jinxing Yu, Xun Jian, Hao Xin and Yangqiu Song

Exploiting Morphological Regularities in Distributional Word Representations
Arihant Gupta, Syed Sarfaraz Akhtar, Avijit Vajpayee, Arjit Srivastava, Madan Gopal Jhanwar and Manish Shrivastava

Exploiting Word Internal Structures for Generic Chinese Sentence Representation
Shaonan Wang, Jiajun Zhang and Chengqing Zong

High-risk learning: acquiring new word vectors from tiny data
Aurélie Herbelot and Marco Baroni

Word Embeddings based on Fixed-Size Ordinally Forgetting Encoding
Joseph Sanu, Mingbin Xu, Hui Jiang and Quan Liu

VecShare: A Framework for Sharing Word Representation Vectors
Jared Fernandez, Zhaocheng Yu and Doug Downey

Learning the Structure of Variable-Order CRFs: a finite-state perspective
Thomas Lavergne and François Yvon

Sparse Communication for Distributed Gradient Descent
Alham Fikri Aji and Kenneth Heafield

Why ADAGRAD Fails for Online Topic Modeling
You Lu, Jeffrey Lund and Jordan Boyd-Graber

13:40–15:20 Session 2F: Poster Session. Sentiment Analysis 1

Recurrent Attention Network on Memory for Aspect Sentiment Analysis
Peng Chen, Zhongqian Sun, Lidong Bing and Wei Yang

A Cognition Based Attention Model for Sentiment Analysis
Yunfei Long, Lu Qin, Rong Xiang, Minglei Li and Chu-Ren Huang

Author-aware Aspect Topic Sentiment Model to Retrieve Supporting Opinions from Reviews
Lahari Poddar, Wynne Hsu and Mong Li Lee

Magnets for Sarcasm: Making Sarcasm Detection Timely, Contextual and Very Personal
Aniruddha Ghosh and Tony Veale

Identifying Humor in Reviews using Background Text Sources
Alex Morales and Chengxiang Zhai

Sentiment Lexicon Construction with Representation Learning Based on Hierarchical Sentiment Supervision
Leyi Wang and Rui Xia

Towards a Universal Sentiment Classifier in Multiple languages
Kui Xu and Xiaojun Wan

Capturing User and Product Information for Document Level Sentiment Analysis with Deep Memory Network
Zi-Yi Dou

Identifying and Tracking Sentiments and Topics from Social Media Texts during Natural Disasters
Min Yang, Jincheng Mei, Heng Ji, zhao wei, Zhou Zhao and Xiaojun Chen

Refining Word Embeddings for Sentiment Analysis
Liang-Chih Yu, Jin Wang, K. Robert Lai and Xuejie Zhang

A Multilayer Perceptron based Ensemble Technique for Fine-grained Financial Sentiment Analysis
Md Shad Akhtar, Abhishek Kumar, Deepanway Ghosal, Asif Ekbal and Pushpak Bhattacharyya

Sentiment Intensity Ranking among Adjectives Using Sentiment Bearing Word Embeddings
Raksha Sharma, Arpan Somani, Lakshya Kumar and Pushpak Bhattacharyya

Sentiment Lexicon Expansion Based on Neural PU Learning, Double Dictionary Lookup, and Polarity Association
Yasheng Wang, Yang Zhang and Bing Liu

15:20–15:50 ***Coffee Break***

15:50–17:30 **Session 3A: Machine Learning 2**

15:50–16:15 *DeepPath: A Reinforcement Learning Method for Knowledge Graph Reasoning*
Wenhan Xiong, Thien Hoang and William Yang Wang

16:15–16:40 *Task-Oriented Query Reformulation with Reinforcement Learning*
Rodrigo Nogueira and Kyunghyun Cho

16:40–17:05 *Sentence Simplification with Deep Reinforcement Learning*
Xingxing Zhang and Mirella Lapata

17:05–17:30 *Learning how to Active Learn: A Deep Reinforcement Learning Approach*
Meng Fang, Yuan Li and Trevor Cohn

Saturday, September 9, 2017 (continued)

15:50–17:30 Session 3B: Generation

15:50–16:15 *Split and Rephrase*
Shashi Narayan, Claire Gardent, Shay B. Cohen and Anastasia Shimorina

16:15–16:40 *Neural Response Generation via GAN with an Approximate Embedding Layer*
Zhen Xu, Bingquan Liu, Baoxun Wang, Chengjie SUN, Xiaolong Wang, Zhuoran Wang and Chao Qi

16:40–17:05 *A Hybrid Convolutional Variational Autoencoder for Text Generation*
Stanislau Semeniuta, Aliaksei Severyn and Erhardt Barth

17:05–17:30 *Filling the Blanks (hint: plural noun) for Mad Libs Humor*
Nabil Hossain, John Krumm, Lucy Vanderwende, Eric Horvitz and Henry Kautz

15:50–17:30 Session 3C: Semantics 1

15:50–16:15 *Measuring Thematic Fit with Distributional Feature Overlap*
Enrico Santus, Emmanuele Chersoni, Alessandro Lenci and Philippe Blache

16:15–16:40 *SCDV : Sparse Composite Document Vectors using soft clustering over distributional representations*
Dheeraj Mekala, Vivek Gupta, Bhargavi Paranjape and Harish Karnick

16:40–17:05 *Supervised Learning of Universal Sentence Representations from Natural Language Inference Data*
Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault and Antoine Bordes

17:05–17:30 *Determining Semantic Textual Similarity using Natural Deduction Proofs*
Hitomi Yanaka, Koji Mineshima, Pascual Martínez-Gómez and Daisuke Bekki

Saturday, September 9, 2017 (continued)

15:50–17:30 Session 3D: Poster Session. Syntax 2

Multi-Grained Chinese Word Segmentation
Chen Gong, Zhenghua Li, Min Zhang and Xinzhou Jiang

Don't Throw Those Morphological Analyzers Away Just Yet: Neural Morphological Disambiguation for Arabic
Nasser Zalmout and Nizar Habash

Paradigm Completion for Derivational Morphology
Ryan Cotterell, Ekaterina Vylomova, Huda Khayrallah, Christo Kirov and David Yarowsky

A Sub-Character Architecture for Korean Language Processing
Karl Stratos

Do LSTMs really work so well for PoS tagging? – A replication study
Tobias Horsmann and Torsten Zesch

The Labeled Segmentation of Printed Books
Lara McConnaughey, Jennifer Dai and David Bamman

Cross-lingual Character-Level Neural Morphological Tagging
Ryan Cotterell and Georg Heigold

Word-Context Character Embeddings for Chinese Word Segmentation
Hao Zhou, Zhenting Yu, Yue Zhang, Shujian Huang, XIN-YU DAI and Jiajun Chen

Segmentation-Free Word Embedding for Unsegmented Languages
Takamasa Oshikiri

15:50–17:30 Session 3E: Poster Session. Question Answering and Machine Comprehension

From Textbooks to Knowledge: A Case Study in Harvesting Axiomatic Knowledge from Textbooks to Solve Geometry Problems
Mrinmaya Sachan, Kumar Dubey and Eric Xing

RACE: Large-scale ReAding Comprehension Dataset From Examinations
Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang and Eduard Hovy

Beyond Sentential Semantic Parsing: Tackling the Math SAT with a Cascade of Tree Transducers
Mark Hopkins, Cristian Petrescu-Prahova, Roie Levin, Ronan Le Bras, Alvaro Herrasti and Vidur Joshi

Learning Fine-Grained Expressions to Solve Math Word Problems
Danqing Huang, Shuming Shi, Chin-Yew Lin and Jian Yin

Structural Embedding of Syntactic Trees for Machine Comprehension
Rui Liu, Junjie Hu, Wei Wei, Zi Yang and Eric Nyberg

World Knowledge for Reading Comprehension: Rare Entity Prediction with Hierarchical LSTMs Using External Descriptions
Teng Long, Emmanuel Bengio, Ryan Lowe, Jackie Chi Kit Cheung and Doina Precup

Two-Stage Synthesis Networks for Transfer Learning in Machine Comprehension
David Golub, Po-Sen Huang, Xiaodong He and Li Deng

Deep Neural Solver for Math Word Problems
Yan Wang, Xiaojiang Liu and Shuming Shi

Latent Space Embedding for Retrieval in Question-Answer Archives
Deepak P, Dinesh Garg and Shirish Shevade

Question Generation for Question Answering
Nan Duan, Duyu Tang, Peng Chen and Ming Zhou

Learning to Paraphrase for Question Answering
Li Dong, Jonathan Mallinson, Siva Reddy and Mirella Lapata

Temporal Information Extraction for Question Answering Using Syntactic Dependencies in an LSTM-based Architecture
Yuanliang Meng, Anna Rumshisky and Alexey Romanov

Ranking Kernels for Structures and Embeddings: A Hybrid Preference and Classification Model
Kateryna Tymoshenko, Daniele Bonadiman and Alessandro Moschitti

Recovering Question Answering Errors via Query Revision
Semih Yavuz, Izzeddin Gur, Yu Su and Xifeng Yan

15:50–17:30 Session 3F: Poster Session. Multimodal NLP 1

An empirical study on the effectiveness of images in Multimodal Neural Machine Translation
Jean-Benoit Delbrouck and Stéphane Dupont

Sound-Word2Vec: Learning Word Representations Grounded in Sounds
Ashwin Vijayakumar, Ramakrishna Vedantam and Devi Parikh

The Promise of Premise: Harnessing Question Premises in Visual Question Answering
Aroma Mahendru, Viraj Prabhu, Akrit Mohapatra, Dhruv Batra and Stefan Lee

Guided Open Vocabulary Image Captioning with Constrained Beam Search
Peter Anderson, Basura Fernando, Mark Johnson and Stephen Gould

Zero-Shot Activity Recognition with Verb Attribute Induction
Rowan Zellers and Yejin Choi

Deriving continous grounded meaning representations from referentially structured multimodal contexts
Sina Zarrieß and David Schlangen

Hierarchically-Attentive RNN for Album Summarization and Storytelling
Licheng Yu, Mohit Bansal and Tamara Berg

Video Highlight Prediction Using Audience Chat Reactions
Cheng-Yang Fu, Joon Lee, Mohit Bansal and Alexander Berg

Saturday, September 9, 2017 (continued)

Reinforced Video Captioning with Entailment Rewards
Ramakanth Pasunuru and Mohit Bansal

Evaluating Hierarchies of Verb Argument Structure with Hierarchical Clustering
Jesse Mu, Joshua K. Hartshorne and Timothy O'Donnell

Incorporating Global Visual Features into Attention-based Neural Machine Translation.
Iacer Calixto and Qun Liu

Mapping Instructions and Visual Observations to Actions with Reinforcement Learning
Dipendra Misra, John Langford and Yoav Artzi

An analysis of eye-movements during reading for the detection of mild cognitive impairment
Kathleen C. Fraser, Kristina Lundholm Fors, Dimitrios Kokkinakis and Arto Nordlund

Evaluating Low-Level Speech Features Against Human Perceptual Data
Caitlin Richter, Naomi H Feldman, Harini Salgado and Aren Jansen

Sunday, September 10, 2017

07:30–17:30 Registration Day 2

08:00–09:00 *Morning Coffee*

09:00–10:00 Plenary Session. Invited Talk by Sharon Goldwater

09:00–10:00 *Towards more universal language technology: unsupervised learning from speech*
Sharon Goldwater

10:00–10:30 *Coffee Break*

10:30–12:10 Session 4A: Reading and Retrieving

10:30–10:55 *A Structured Learning Approach to Temporal Relation Extraction*
Qiang Ning, Zhili Feng and Dan Roth

10:55–11:20 *Importance sampling for unbiased on-demand evaluation of knowledge base population*
Arun Chaganty, Ashwin Paranjape, Percy Liang and Christopher D. Manning

11:20–11:45 *PACRR: A Position-Aware Neural IR Model for Relevance Matching*
Kai Hui, Andrew Yates, Klaus Berberich and Gerard de Melo

11:45–12:10 *Globally Normalized Reader*
Jonathan Raiman and John Miller

10:30–12:10 Session 4B: Multimodal NLP 2

10:30–10:55 *Speech segmentation with a neural encoder model of working memory*
Micha Elsner and Cory Shain

10:55–11:20 *Speaking, Seeing, Understanding: Correlating semantic models with conceptual representation in the brain*
Luana Bulat, Stephen Clark and Ekaterina Shutova

11:20–11:45 *Multi-modal Summarization for Asynchronous Collection of Text, Image, Audio and Video*
Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang and Chengqing Zong

11:45–12:10 *Tensor Fusion Network for Multimodal Sentiment Analysis*
Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria and Louis-Philippe Morency

10:30–12:10 Session 4C: Human Centered NLP and Linguistic Theory

10:30–12:10 Session 4D: Poster Session. Semantics 2

A Continuously Growing Dataset of Sentential Paraphrases
Wuwei Lan, Siyu Qiu, Hua He and Wei Xu

Cross-domain Semantic Parsing via Paraphrasing
Yu Su and Xifeng Yan

A Joint Sequential and Relational Model for Frame-Semantic Parsing
Bishan Yang and Tom Mitchell

Getting the Most out of AMR Parsing
Chuan Wang and Nianwen Xue

AMR Parsing using Stack-LSTMs
Miguel Ballesteros and Yaser Al-Onaizan

An End-to-End Deep Framework for Answer Triggering with a Novel Group-Level Objective
Jie Zhao, Yu Su, Ziyu Guan and Huan Sun

Predicting Word Association Strengths
Andrew Cattle and Xiaojuan Ma

10:30–12:10 Session 4E: Poster Session. Discourse

Learning Contextually Informed Representations for Linear-Time Discourse Parsing
Yang Liu and Mirella Lapata

Multi-task Attention-based Neural Networks for Implicit Discourse Relationship Representation and Identification
Man Lan, Jianxiang Wang, Yuanbin Wu, Zheng-Yu Niu and Haifeng Wang

Chinese Zero Pronoun Resolution with Deep Memory Network
Qingyu Yin, Yu Zhang, Weinan Zhang and Ting Liu

How much progress have we made on RST discourse parsing? A replication study of recent results on the RST-DT
Mathieu Morey, Philippe Muller and Nicholas Asher

10:30–12:10 Session 4F: Poster Session. Machine Translation and Multilingual NLP 1

Neural Lattice-to-Sequence Models for Uncertain Inputs
Matthias Sperber, Graham Neubig, Jan Niehues and Alex Waibel

Memory-augmented Neural Machine Translation
Yang Feng, Shiyue Zhang, Andi Zhang, Dong Wang and Andrew Abel

Dynamic Data Selection for Neural Machine Translation
Marlies van der Wees, Arianna Bisazza and Christof Monz

Neural Machine Translation Leveraging Phrase-based Models in a Hybrid Search
Leonard Dahlmann, Evgeny Matusov, Pavel Petrushkov and Shahram Khadivi

Translating Phrases in Neural Machine Translation
Xing Wang, Zhaopeng Tu, Deyi Xiong and Min Zhang

Towards Bidirectional Hierarchical Representations for Attention-based Neural Machine Translation
Baosong Yang, Derek F. Wong, Tong Xiao, Lidia S. Chao and Jingbo Zhu

Massive Exploration of Neural Machine Translation Architectures
Denny Britz, Anna Goldie, Minh-Thang Luong and Quoc Le

Learning Translations via Matrix Completion
Derry Tanti Wijaya, Brendan Callahan, John Hewitt, Jie Gao, Xiao Ling, Marianna Apidianaki and Chris Callison-Burch

Reinforcement Learning for Bandit Neural Machine Translation with Simulated Human Feedback
Khanh Nguyen, Hal Daumé III and Jordan Boyd-Graber

Towards Compact and Fast Neural Machine Translation Using a Combined Method
Xiaowei Zhang, Wei Chen, Feng Wang, Shuang Xu and Bo Xu

Instance Weighting for Neural Machine Translation Domain Adaptation
Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen and Eiichiro Sumita

Sunday, September 10, 2017 (continued)

Regularization techniques for fine-tuning in neural machine translation
Antonio Valerio Miceli Barone, Barry Haddow, Ulrich Germann and Rico Sennrich

Source-Side Left-to-Right or Target-Side Left-to-Right? An Empirical Comparison of Two Phrase-Based Decoding Algorithms
Yin-Wen Chang and Michael Collins

Using Target-side Monolingual Data for Neural Machine Translation through Multi-task Learning
Tobias Domhan and Felix Hieber

12:10–13:40 ***Lunch***

12:40–13:40 **SIGDAT Business Meeting**

13:40–15:20 **Session 5A: Semantics 3**

13:40–14:05 *Encoding Sentences with Graph Convolutional Networks for Semantic Role Labeling*
Diego Marcheggiani and Ivan Titov

14:05–14:30 *Neural Semantic Parsing with Type Constraints for Semi-Structured Tables*
Jayant Krishnamurthy, Pradeep Dasigi and Matt Gardner

14:30–14:55 *Joint Concept Learning and Semantic Parsing from Natural Language Explanations*
Shashank Srivastava, Igor Labutov and Tom Mitchell

14:55–15:20 *Grasping the Finer Point: A Supervised Similarity Network for Metaphor Detection*
Marek Rei, Luana Bulat, Douwe Kiela and Ekaterina Shutova

Sunday, September 10, 2017 (continued)

13:40–15:20 Session 5D: Poster Session. Syntax 3

13:40–15:20 Session 5E: Poster Session. Relations

Global Normalization of Convolutional Neural Networks for Joint Entity and Relation Classification
Heike Adel and Hinrich Schütze

End-to-End Neural Relation Extraction with Global Optimization
Meishan Zhang, Yue Zhang and Guohong Fu

KGEval: Accuracy Estimation of Automatically Constructed Knowledge Graphs
Prakhar Ojha and Partha Talukdar

Sparsity and Noise: Where Knowledge Graph Embeddings Fall Short
Jay Pujara, Eriq Augustine and Lise Getoor

Dual Tensor Model for Detecting Asymmetric Lexico-Semantic Relations
Goran Glavaš and Simone Paolo Ponzetto

Incorporating Relation Paths in Neural Relation Extraction
Wenyuan Zeng, Yankai Lin, Zhiyuan Liu and Maosong Sun

Adversarial Training for Relation Extraction
Yi Wu, David Bamman and Stuart Russell

Context-Aware Representations for Knowledge Base Relation Extraction
Daniil Sorokin and Iryna Gurevych

A Soft-label Method for Noise-tolerant Distantly Supervised Relation Extraction
Tianyu Liu, Kexiang Wang, Baobao Chang and Zhifang Sui

A Sequential Model for Classifying Temporal Relations between Intra-Sentence Events
Prafulla Kumar Choubey and Ruihong Huang

Deep Residual Learning for Weakly-Supervised Relation Extraction
YiYao Huang and William Yang Wang

Noise-Clustered Distant Supervision for Relation Extraction: A Nonparametric Bayesian Perspective
Qing Zhang and Houfeng Wang

Exploring Vector Spaces for Semantic Relations
Kata Gábor, Haifa Zargayouna, Isabelle Tellier, Davide Buscaldi and Thierry Charnois

Temporal dynamics of semantic relations in word embeddings: an application to predicting armed conflict participants
Andrey Kutuzov, Erik Velldal and Lilja Øvrelid

13:40–15:20 Session 5F: Poster Session. Language Models, Text Mining, and Crowd Sourcing

Dynamic Entity Representations in Neural Language Models
Yangfeng Ji, Chenhao Tan, Sebastian Martschat, Yejin Choi and Noah A. Smith

Towards Quantum Language Models
Ivano Basile and Fabio Tamburini

Reference-Aware Language Models
Zichao Yang, Phil Blunsom, Chris Dyer and Wang Ling

A Simple Language Model based on PMI Matrix Approximations
Oren Melamud, Ido Dagan and Jacob Goldberger

Syllable-aware Neural Language Models: A Failure to Beat Character-aware Ones
Zhenisbek Assylbekov, Rustem Takhanov, Bagdat Myrzakhmetov and Jonathan N. Washington

Inducing Semantic Micro-Clusters from Deep Multi-View Representations of Novels
Lea Frermann and György Szarvas

Initializing Convolutional Filters with Semantic Features for Text Classification
Shen Li, Zhe Zhao, Tao Liu, Renfen Hu and Xiaoyong Du

Shortest-Path Graph Kernels for Document Similarity
Giannis Nikolentzos, Polykarpos Meladianos, Francois Rousseau, Yannis Stavrakas and Michalis Vazirgiannis

Adapting Topic Models using Lexical Associations with Tree Priors
Weiwei Yang, Jordan Boyd-Graber and Philip Resnik

Finding Patterns in Noisy Crowds: Regression-based Annotation Aggregation for Crowdsourced Data
Natalie Parde and Rodney Nielsen

CROWD-IN-THE-LOOP: A Hybrid Approach for Annotating Semantic Roles
Chenguang Wang, Alan Akbik, laura chiticariu, Yunyao Li, Fei Xia and Anbang Xu

A Joint Many-Task Model: Growing a Neural Network for Multiple NLP Tasks
Kazuma Hashimoto, caiming xiong, Yoshimasa Tsuruoka and Richard Socher

15:20–15:50 ***Coffee Break***

15:50–17:30 **Session 6A: Machine Translation 2**

15:50–16:15 *Earth Mover's Distance Minimization for Unsupervised Bilingual Lexicon Induction*
Meng Zhang, Yang Liu, Huanbo Luan and Maosong Sun

16:15–16:40 *Unfolding and Shrinking Neural Machine Translation Ensembles*
Felix Stahlberg and Bill Byrne

16:40–17:05 *Graph Convolutional Encoders for Syntax-aware Neural Machine Translation*
Joost Bastings, Ivan Titov, Wilker Aziz, Diego Marcheggiani and Khalil Simaan

17:05–17:30 *Trainable Greedy Decoding for Neural Machine Translation*
Jiatao Gu, Kyunghyun Cho and Victor O.K. Li

Sunday, September 10, 2017 (continued)

15:50–17:30 Session 6B: Text Mining and NLP applications

15:50–17:30 Session 6C: Machine Comprehension

15:50–17:30 **Session 6E: Poster Session. Summarization, Generation, Dialog, and Discourse 2**

15:50–17:30 Session 6F: Poster Session. Computational Social Science 2

Never Abandon Minorities: Exhaustive Extraction of Bursty Phrases on Microblogs Using Set Cover Problem
Masumi Shirakawa, Takahiro Hara and Takuya Maekawa

18:00–22:00 Social Event

Monday, September 11, 2017

07:30–17:30 Registration Day 3

08:00–09:00 *Morning Coffee*

09:00–10:00 Plenary Session. Invited Talk by Dan Jurafsky

09:00–10:00 *"Does This Vehicle Belong to You"? Processing the Language of Policing for Improving Police-Community Relations*
Dan Jurafsky

10:00–10:30 *Coffee Break*

10:30–12:10 Session 7A: Machine Learning 3

10:30–10:55 *Maximum Margin Reward Networks for Learning from Explicit and Implicit Supervision*
Haoruo Peng, Ming-Wei Chang and Wen-tau Yih

10:55–11:20 *The Impact of Modeling Overall Argumentation with Tree Kernels*
Henning Wachsmuth, Giovanni Da San Martino, Dora Kiesel and Benno Stein

11:20–11:45 *Learning Generic Sentence Representations Using Convolutional Neural Networks*
Zhe Gan, Yunchen Pu, Ricardo Henao, Chunyuan Li, Xiaodong He and Lawrence Carin

11:45–12:10 *Repeat before Forgetting: Spaced Repetition for Efficient and Effective Training of Neural Networks*
Hadi Amiri, Timothy Miller and Guergana Savova

10:30–12:10 Session 7D: Poster Session. Machine Translation and Multilingual NLP 2

Semantic Specialisation of Distributional Word Vector Spaces using Monolingual and Cross-Lingual Constraints
Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen and Steve Young

Counterfactual Learning from Bandit Feedback under Deterministic Logging : A Case Study in Statistical Machine Translation
Carolin Lawrence, Artem Sokolov and Stefan Riezler

10:30–12:10 Session 7E: Poster Session. Information Extraction 2

Learning Fine-grained Relations from Chinese User Generated Categories
Chengyu Wang, Yan Fan, Xiaofeng He and Aoying Zhou

Improving Slot Filling Performance with Attentive Neural Networks on Dependency Structures
Lifu Huang, Avirup Sil, Heng Ji and Radu Florian

Identifying Products in Online Cybercrime Marketplaces: A Dataset for Fine-grained Domain Adaptation
Greg Durrett, Jonathan K. Kummerfeld, Taylor Berg-Kirkpatrick, Rebecca Portnoff, Sadia Afroz, Damon McCoy, Kirill Levchenko and Vern Paxson

Labeling Gaps Between Words: Recognizing Overlapping Mentions with Mention Separators
Aldrian Obaja Muis and Wei Lu

Deep Joint Entity Disambiguation with Local Neural Attention
Octavian-Eugen Ganea and Thomas Hofmann

MinIE: Minimizing Facts in Open Information Extraction
Kiril Gashteovski, Rainer Gemulla and Luciano Del Corro

Scientific Information Extraction with Semi-supervised Neural Tagging
Yi Luan, Mari Ostendorf and Hannaneh Hajishirzi

NITE: A Neural Inductive Teaching Framework for Domain Specific NER
Siliang Tang, Ning Zhang, Jinjiang Zhang, Fei Wu and Yueting Zhuang

Speeding up Reinforcement Learning-based Information Extraction Training using Asynchronous Methods
Aditya Sharma, Zarana Parekh and Partha Talukdar

Leveraging Linguistic Structures for Named Entity Recognition with Bidirectional Recursive Neural Networks
Peng-Hsuan Li, Ruo-Ping Dong, Yu-Siang Wang, Ju-Chieh Chou and Wei-Yun Ma

Fast and Accurate Entity Recognition with Iterated Dilated Convolutions
Emma Strubell, Patrick Verga, David Belanger and Andrew McCallum

Entity Linking via Joint Encoding of Types, Descriptions, and Context
Nitish Gupta, Sameer Singh and Dan Roth

An Insight Extraction System on BioMedical Literature with Deep Neural Networks
Hua He, Kris Ganjam, Navendu Jain, Jessica Lundin, Ryen White and Jimmy Lin

10:30–12:10 Session 7F: Poster Session. NLP Applications

Word Etymology as Native Language Interference
Vivi Nastase and Carlo Strapparava

A Simpler and More Generalizable Story Detector using Verb and Character Features
Joshua Eisenberg and Mark Finlayson

Multi-modular domain-tailored OCR post-correction
Sarah Schulz and Jonas Kuhn

Learning to Predict Charges for Criminal Cases with Legal Basis
Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang and Dongyan Zhao

Quantifying the Effects of Text Duplication on Semantic Models
Alexandra Schofield, Laure Thompson and David Mimno

Identifying Semantically Deviating Outlier Documents
Honglei Zhuang, Chi Wang, Fangbo Tao, Lance Kaplan and Jiawei Han

Detecting and Explaining Causes From Text For a Time Series Event
Dongyeop Kang, Varun Gangal, Ang Lu, Zheng Chen and Eduard Hovy

A Novel Cascade Model for Learning Latent Similarity from Heterogeneous Sequential Data of MOOC
Zhuoxuan Jiang, Shanshan Feng, Gao Cong, Chunyan Miao and Xiaoming Li

Identifying the Provision of Choices in Privacy Policy Text
Kanthashree Mysore Sathyendra, Shomir Wilson, Florian Schaub, Sebastian Zimmeck and Norman Sadeh

An Empirical Analysis of Edit Importance between Document Versions
Tanya Goyal, Sachin Kelkar, Manas Agarwal and Jeenu Grover

Transition-Based Disfluency Detection using LSTMs
Shaolei Wang, Wanxiang Che, Yue Zhang, Meishan Zhang and Ting Liu

Neural Sequence-Labelling Models for Grammatical Error Correction
Helen Yannakoudakis, Marek Rei, Øistein E. Andersen and Zheng Yuan

Adapting Sequence Models for Sentence Correction
Allen Schmaltz, Yoon Kim, Alexander Rush and Stuart Shieber

12:10–13:40 ***Lunch***

13:40–15:25 **Session 8A: Machine Translation and Multilingual/Multimodal NLP (Short)**

13:40–13:55 *A Study of Style in Machine Translation: Controlling the Formality of Machine Translation Output*
Xing Niu, Marianna Martindale and Marine Carpuat

13:55–14:10 *Sharp Models on Dull Hardware: Fast and Accurate Neural Machine Translation Decoding on the CPU*
Jacob Devlin

14:10–14:25 *Exploiting Cross-Sentence Context for Neural Machine Translation*
Longyue Wang, Zhaopeng Tu, Andy Way and Qun Liu

14:25–14:40 *Cross-Lingual Transfer Learning for POS Tagging without Cross-Lingual Resources*
Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya and Eric Fosler-Lussier

14:40–14:55 *Image Pivoting for Learning Multilingual Multimodal Representations*
Spandana Gella, Rico Sennrich, Frank Keller and Mirella Lapata

14:55–15:10 *Neural Machine Translation with Source Dependency Representation*
Kehai Chen, Rui Wang, Masao Utiyama, Lemao Liu, Akihiro Tamura, Eiichiro Sumita and Tiejun Zhao

15:10–15:25 *Visual Denotations for Recognizing Textual Entailment*
Dan Han, Pascual Martínez-Gómez and Koji Mineshima

13:40–15:25 Session 8B: Machine Learning (Short)

13:40–13:55 *Sequence Effects in Crowdsourced Annotations*
Nitika Mathur, Timothy Baldwin and Trevor Cohn

13:55–14:10 *No Need to Pay Attention: Simple Recurrent Neural Networks Work!*
Ferhan Ture and Oliver Jojic

14:10–14:25 *The strange geometry of skip-gram with negative sampling*
David Mimno and Laure Thompson

14:25–14:40 *Natural Language Processing with Small Feed-Forward Networks*
Jan A. Botha, Emily Pitler, Ji Ma, Anton Bakalov, Alex Salcianu, David Weiss, Ryan McDonald and Slav Petrov

14:40–14:55 *Deep Multi-Task Learning for Aspect Term Extraction with Memory Interaction*
Xin Li and Wai Lam

14:55–15:10 *Analogs of Linguistic Structure in Deep Representations*
Jacob Andreas and Dan Klein

15:10–15:25 *A Simple Regularization-based Algorithm for Learning Cross-Domain Word Embeddings*
Wei Yang, Wei Lu and Vincent Zheng

Monday, September 11, 2017 (continued)

13:40–15:25 Session 8C: NLP Applications (Short)

15:25–15:50 *Coffee Break*

Monday, September 11, 2017 (continued)

15:50–17:25 Plenary Session. Best Paper

17:00–17:25 *Bringing Structure into Summaries: Crowdsourcing a Benchmark Corpus of Concept Maps*
Tobias Falke and Iryna Gurevych

16:20–16:35 *Natural Language Does Not Emerge 'Naturally' in Multi-Agent Dialog*
Satwik Kottur, José Moura, Stefan Lee and Dhruv Batra

16:35–17:00 *Depression and Self-Harm Risk Assessment in Online Forums*
Andrew Yates, Arman Cohan and Nazli Goharian

15:55–16:20 *Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints*
Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez and Kai-Wei Chang

17:25–17:45 Plenary Session. Closing Remarks

17:25–17:45 *Closing Remarks*
General Chair

Neural Sequence Learning Models for Word Sense Disambiguation

Alessandro Raganato, Claudio Delli Bovi and **Roberto Navigli**
Department of Computer Science
Sapienza University of Rome
{raganato,dellibovi,navigli}@di.uniroma1.it

Abstract

Word Sense Disambiguation models exist in many flavors. Even though supervised ones tend to perform best in terms of accuracy, they often lose ground to more flexible knowledge-based solutions, which do not require training by a word expert for every disambiguation target. To bridge this gap we adopt a different perspective and rely on sequence learning to frame the disambiguation problem: we propose and study in depth a series of end-to-end neural architectures directly tailored to the task, from bidirectional Long Short-Term Memory to encoder-decoder models. Our extensive evaluation over standard benchmarks and in multiple languages shows that sequence learning enables more versatile all-words models that consistently lead to state-of-the-art results, even against word experts with engineered features.

1 Introduction

As one of the long-standing challenges in Natural Language Processing (NLP), Word Sense Disambiguation (Navigli, 2009, WSD) has received considerable attention over recent years. Indeed, by dealing with lexical ambiguity an effective WSD model brings numerous benefits to a variety of downstream tasks and applications, from Information Retrieval and Extraction (Zhong and Ng, 2012; Delli Bovi et al., 2015) to Machine Translation (Carpuat and Wu, 2007; Xiong and Zhang, 2014; Neale et al., 2016). Recently, WSD has also been leveraged to build continuous vector representations for word senses (Chen et al., 2014; Iacobacci et al., 2015; Flekova and Gurevych, 2016).

Inasmuch as WSD is described as the task of associating words in context with the most suitable entries in a pre-defined sense inventory, the majority of WSD approaches to date can be grouped into two main categories: supervised (or semi-supervised) and knowledge-based. Supervised models have been shown to consistently outperform knowledge-based ones in all standard benchmarks (Raganato et al., 2017), at the expense, however, of harder training and limited flexibility. First of all, obtaining reliable sense-annotated corpora is highly expensive and especially difficult when non-expert annotators are involved (de Lacalle and Agirre, 2015), and as a consequence approaches based on unlabeled data and semi-supervised learning are emerging (Taghipour and Ng, 2015b; Başkaya and Jurgens, 2016; Yuan et al., 2016; Pasini and Navigli, 2017).

Apart from the shortage of training data, a crucial limitation of current supervised approaches is that a dedicated classifier (*word expert*) needs to be trained for every target lemma, making them less flexible and hampering their use within end-to-end applications. In contrast, knowledge-based systems do not require sense-annotated data and often draw upon the structural properties of lexico-semantic resources (Agirre et al., 2014; Moro et al., 2014; Weissenborn et al., 2015). Such systems construct a model based only on the underlying resource, which is then able to handle multiple target words at the same time and disambiguate them jointly, whereas word experts are forced to treat each disambiguation target in isolation.

In this paper our focus is on supervised WSD, but we depart from previous approaches and adopt a different perspective on the task: instead of framing a separate classification problem for each given word, we aim at modeling the joint disambiguation of the target text as a whole in terms of a sequence labeling problem. From this standpoint, WSD amounts to translating a sequence of words into a sequence of potentially sense-tagged tokens.

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1156–1167
Copenhagen, Denmark, September 7–11, 2017.

With this in mind, we design, analyze and compare experimentally various neural architectures of different complexities, ranging from a single bidirectional Long Short-Term Memory (Graves and Schmidhuber, 2005, LSTM) to a sequence-to-sequence approach (Sutskever et al., 2014). Each architecture reflects a particular way of modeling the disambiguation problem, but they all share some key features that set them apart from previous supervised approaches to WSD: they are trained end-to-end from sense-annotated text to sense labels, and learn a single all-words model from the training data, without fine tuning or explicit engineering of local features.

The contributions of this paper are twofold. First, we show that neural sequence learning represents a novel and effective alternative to the traditional way of modeling supervised WSD, enabling a single all-words model to compete with a pool of word experts and achieve state-of-the-art results, while also being easier to train, arguably more versatile to use within downstream applications, and directly adaptable to different languages without requiring additional sense-annotated data (as we show in Section 6.2); second, we carry out an extensive experimental evaluation where we compare various neural architectures designed for the task (and somehow left underinvestigated in previous literature), exploring different configurations and training procedures, and analyzing their strengths and weaknesses on all the standard benchmarks for all-words WSD.

2 Related Work

The literature on WSD is broad and comprehensive (Agirre and Edmonds, 2007; Navigli, 2009): new models are continuously being developed (Yuan et al., 2016; Tripodi and Pelillo, 2017; Butnaru et al., 2017) and tested over a wide variety of standard benchmarks (Edmonds and Cotton, 2001; Snyder and Palmer, 2004; Pradhan et al., 2007; Navigli et al., 2007, 2013; Moro and Navigli, 2015). Moreover, the field has been explored in depth from different angles by means of extensive empirical studies and evaluation frameworks (Pilehvar and Navigli, 2014; Iacobacci et al., 2016; McCarthy et al., 2016; Raganato et al., 2017).

As regards supervised WSD, traditional approaches are generally based on extracting local features from the words surrounding the target, and then training a classifier (Zhong and Ng, 2010; Shen et al., 2013) for each target lemma. In their latest developments, these models include more complex features based on word embeddings (Taghipour and Ng, 2015b; Rothe and Schütze, 2015; Iacobacci et al., 2016).

The recent upsurge of neural networks has also contributed to fueling WSD research: Yuan et al. (2016) rely on a powerful neural language model to obtain a latent representation for the whole sentence containing a target word w; their instance-based system then compares that representation with those of example sentences annotated with the candidate meanings of w. Similarly, Context2Vec (Melamud et al., 2016) makes use of a bidirectional LSTM architecture trained on an unlabeled corpus and learns a context vector for each sense annotation in the training data. Finally, Kågebäck and Salomonsson (2016) present a supervised classifier based on bidirectional LSTM for the lexical sample task (Kilgarriff, 2001; Mihalcea et al., 2004). All these contributions have shown that supervised neural models can achieve state-of-the-art performances without taking advantage of external resources or language-specific features. However, they all consider each target word as a separate classification problem and, to the best of our knowledge, very few attempts have been made to disambiguate a text jointly using sequence learning (Ciaramita and Altun, 2006).

Sequence learning, especially using LSTM (Hochreiter and Schmidhuber, 1997; Graves and Schmidhuber, 2005; Graves, 2013), has become a well-established standard in numerous NLP tasks (Zhou and Xu, 2015; Ma and Hovy, 2016; Wang and Chang, 2016). In particular, sequence-to-sequence models (Sutskever et al., 2014) have grown increasingly popular and are used extensively in, e.g., Machine Translation (Cho et al., 2014; Bahdanau et al., 2015), Sentence Representation (Kiros et al., 2015), Syntactic Parsing (Vinyals et al., 2015), Conversation Modeling (Vinyals and Le, 2015), Morphological Inflection (Faruqui et al., 2016) and Text Summarization (Gu et al., 2016). In line with this trend, we focus on the (so far unexplored) context of supervised WSD, and investigate state-of-the-art all-words approaches that are based on neural sequence learning and capable of disambiguating all target content words within an input text, a key feature in several knowledge-based approaches.

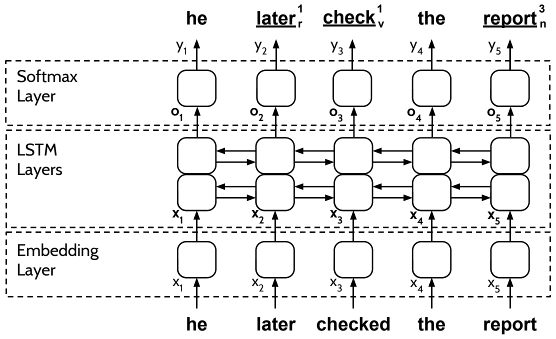

Figure 1: Bidirectional LSTM sequence labeling architecture for WSD (2 hidden layers). We use the notation of Navigli (2009) for word senses: w_p^i is the i-th sense of w with part of speech p.

3 Sequence Learning for Word Sense Disambiguation

In this section we define WSD in terms of a sequence learning problem. While in its classical formulation (Navigli, 2009) WSD is viewed as a classification problem for a given word w in context, with word senses of w being the class labels, here we consider a variable-length sequence of input symbols $\vec{x} = \langle x_1, ..., x_T \rangle$ and we aim at predicting a sequence of output symbols $\vec{y} = \langle y_1, ..., y_{T'} \rangle$.[1] Input symbols are word tokens drawn from a given vocabulary V.[2] Output symbols are either drawn from a pre-defined sense inventory S (if the corresponding input symbols are open-class content words, i.e., nouns, verbs, adjectives or adverbs), or from the same input vocabulary V (e.g., if the corresponding input symbols are function words, like prepositions or determiners). Hence, we can define a WSD model in terms of a function that maps sequences of symbols $x_i \in V$ into sequences of symbols $y_j \in O = S \cup V$.

Here all-words WSD is no longer broken down into a series of distinct and separate classification tasks (one per target word) but rather treated directly at the sequence level, with a single model handling all disambiguation decisions. In what follows, we describe three different models for accomplishing this: a traditional LSTM-based model (Section 3.1), a variant that incorporates an attention mechanism (Section 3.2), and an encoder-decoder architecture (Section 3.3).

[1] In general $\vec{x}$ and $\vec{y}$ might have different lengths, e.g., if $\vec{x}$ contains a multi-word expression (*European Union*) which is mapped to a unique sense identifier (`European Union`$_n^1$).

[2] V generalizes traditional vocabularies used in WSD and includes both word lemmas and inflected forms.

3.1 Bidirectional LSTM Tagger

The most straightforward way of modeling WSD as formulated in Section 3 is that of considering a sequence labeling architecture that tags each symbol $x_i \in V$ in the input sequence with a label $y_j \in O$. Even though the formulation is rather general, previous contributions (Melamud et al., 2016; Kågebäck and Salomonsson, 2016) have already shown the effectiveness of recurrent neural networks for WSD. We follow the same line and employ a bidirectional LSTM architecture: in fact, important clues for disambiguating a target word could be located anywhere in the context (not necessarily before the target) and for a model to be effective it is crucial that it exploits information from the whole input sequence at every time step.

Architecture. A sketch of our bidirectional LSTM tagger is shown in Figure 1. It consists of:

- An embedding layer that converts each word $x_i \in \vec{x}$ into a real-valued d-dimensional vector $\mathbf{x}_i$ via the embedding matrix $\mathbf{W} \in \mathbb{R}^{d \times |V|}$;
- One or more stacked layers of bidirectional LSTM (Graves and Schmidhuber, 2005). The hidden state vectors $\mathbf{h}_i$ and output vectors $\mathbf{o}_i$ at the i^{th} time step are then obtained as the concatenations of the forward and backward pass vectors $\overrightarrow{\mathbf{h}}_i$, $\overrightarrow{\mathbf{o}}_i$ and $\overleftarrow{\mathbf{h}}_i$, $\overleftarrow{\mathbf{o}}_i$;
- A fully-connected layer with softmax activation that turns the output vector $\mathbf{o}_i$ at the i^{th} time step into a probability distribution over the output vocabulary O.

Training. The tagger is trained on a dataset of N labeled sequences $\{(\vec{x}_k, \vec{y}_k)\}_{k=1}^N$ directly obtained from the sentences of a sense-annotated corpus, where each $\vec{x}_k$ is a sequence of word tokens, and each $\vec{y}_k$ is a sequence containing both word tokens and sense labels. Ideally $\vec{y}_k$ is a copy of $\vec{x}_k$ where each content word is sense-tagged. This is, however, not the case in many real-world datasets, where only a subset of the content words is annotated; hence the architecture is designed to deal with both fully and partially annotated sentences. Apart from sentence splitting and tokenization, no preprocessing is required on the training data.

3.2 Attentive Bidirectional LSTM Tagger

The bidirectional LSTM tagger of Section 3.1 exploits information from the whole input sequence $\vec{x}$, which is encoded in the hidden state $\mathbf{h}_i$. However, certain elements of $\vec{x}$ might be more discriminative than others in predicting the output label at a given time step (e.g., the syntactic subject and object when predicting the sense label of a verb).

We model this hunch by introducing an attention mechanism, already proven to be effective in other NLP tasks (Bahdanau et al., 2015; Vinyals et al., 2015), into the sequence labeling architecture of Section 3.1. The resulting *attentive* bidirectional LSTM tagger augments the original architecture with an attention layer, where a context vector $\mathbf{c}$ is computed from all the hidden states $\mathbf{h}_1, ..., \mathbf{h}_T$ of the bidirectional LSTM. The attentive tagger first reads the entire input sequence $\vec{x}$ to construct $\mathbf{c}$, and then exploits $\mathbf{c}$ to predict the output label y_j at each time step, by concatenating it with the output vector $\mathbf{o}_j$ of the bidirectional LSTM (Figure 2).

We follow previous work (Vinyals et al., 2015; Zhou et al., 2016) and compute $\mathbf{c}$ as the weighted sum of the hidden state vectors $\mathbf{h}_1, ..., \mathbf{h}_T$. Formally, let $H \in \mathbb{R}^{n \times T}$ be the matrix of hidden state vectors $[\,\mathbf{h}_1, ..., \mathbf{h}_T\,]$, where n is the hidden state dimension and T is the input sequence length (cf. Section 3). $\mathbf{c}$ is obtained as follows:

$$\begin{aligned} \mathbf{u} &= \omega^T \tanh(H) \\ \mathbf{a} &= softmax(\mathbf{u}) \\ \mathbf{c} &= H\mathbf{a}^T \end{aligned} \qquad (1)$$

where $\omega \in \mathbb{R}^n$ is a parameter vector, and $\mathbf{a} \in \mathbb{R}^T$ is the vector of normalized attention weights.

3.3 Sequence-to-Sequence Model

The attentive tagger of Section 3.2 performs a two-pass procedure by first reading the input sequence $\vec{x}$ to construct the context vector $\mathbf{c}$, and then predicting an output label y_j for each element in $\vec{x}$. In this respect, the attentive architecture can effectively be viewed as an encoder for $\vec{x}$. A further generalization of this model would then be a complete encoder-decoder architecture (Sutskever et al., 2014) where WSD is treated as a sequence-to-sequence mapping (*sequence-to-sequence WSD*), i.e., as the "translation" of word sequences into sequences of potentially sense-tagged tokens.

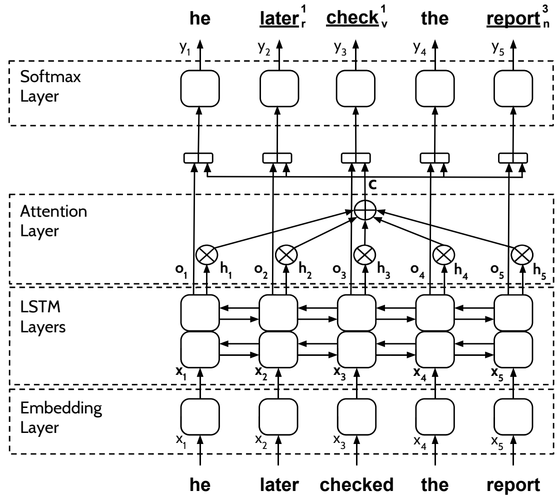

Figure 2: Attentive bidirectional LSTM sequence labeling architecture for WSD (2 hidden layers).

In the sequence-to-sequence framework, a variable-length sequence of input symbols $\vec{x}$ is represented as a sequence of vectors $\vec{\mathbf{x}} = \langle \mathbf{x}_1, ..., \mathbf{x}_T \rangle$ by converting each symbol $x_i \in \vec{x}$ into a real-valued vector $\mathbf{x}_i$ via an embedding layer, and then fed to an encoder, which generates a fixed-dimensional vector representation of the sequence. Traditionally, the encoder function is a Recurrent Neural Network (RNN) such that:

$$\begin{aligned} \mathbf{h}_t &= f(\mathbf{h}_{t-1}, \mathbf{x}_t) \\ \mathbf{c} &= q(\{\mathbf{h}_1, ..., \mathbf{h}_T\}) \end{aligned} \qquad (2)$$

where $\mathbf{h}_t \in \mathbb{R}^n$ is the n-dimensional hidden state vector at time t, $\mathbf{c} \in \mathbb{R}^n$ is a vector generated from the whole sequence of input states, and f and q are non-linear functions.[3] A decoder is then trained to predict the next output symbol y_t given the encoded input vector $\mathbf{c}$ and all the previously predicted output symbols $\langle y_1, ..., y_{t-1} \rangle$. More formally, the decoder defines a probability over the output sequence $\vec{y} = \langle y_1, ..., y_{T'} \rangle$ by decomposing the joint probability into ordered conditionals:

$$p(\vec{y} \,|\, \vec{x}) = \prod_{t=1}^{T'} p(y_t \,|\, \mathbf{c}, \langle y_1, ..., y_{t-1} \rangle) \qquad (3)$$

Typically a decoder RNN defines the hidden state at time t as $\mathbf{s}_t = g(\mathbf{s}_{t-1}, \{\mathbf{c},\, y_{t-1}\})$ and then feeds $\mathbf{s}_t$ to a softmax layer in order to obtain a conditional probability over output symbols.

[3] For instance, Sutskever et al. (2014) used an LSTM as f, and $q(\{\mathbf{h}_1, ..., \mathbf{h}_T\}) = \mathbf{h}_T$.

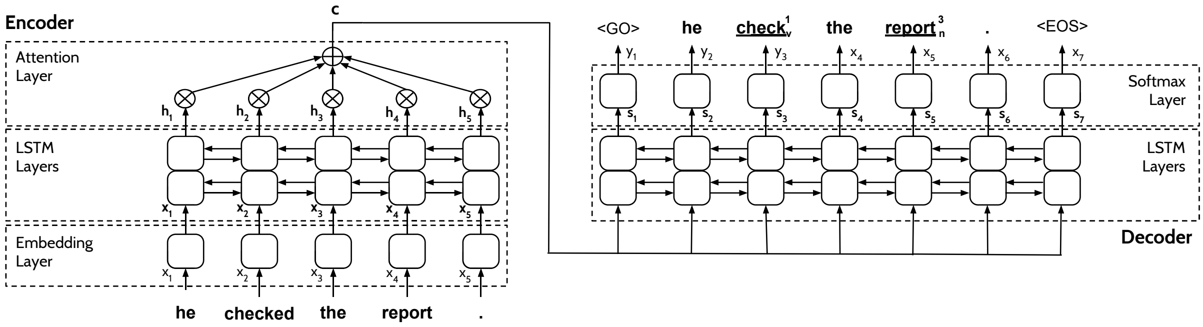

Figure 3: Encoder-decoder architecture for sequence-to-sequence WSD, with 2 bidirectional LSTM layers and an attention layer.

In the context of WSD framed as a sequence learning problem, a sequence-to-sequence model takes as input a training set of labeled sequences (cf. Section 3.1) and learns to replicate an input sequence $\vec{x}$ while replacing each content word with its most suitable word sense from S. In other words, sequence-to-sequence WSD can be viewed as the combination of two sub-tasks:

- A *memorization* task, where the model learns to replicate the input sequence token by token at decoding time;
- The actual *disambiguation* task where the model learns to replace content words across the input sequence with their most suitable senses from the sense inventory S.

In the latter stage, multi-word expressions (such as nominal entity mentions or phrasal verbs) are replaced by their sense identifiers, hence yielding an output sequence that might have a different length than $\vec{x}$.

Architecture. The encoder-decoder architecture generalizes over both the models in Sections 3.1 and 3.2. In particular, we include one or more bidirectional LSTM layers at the core of both the encoder and the decoder modules. The encoder utilizes an embedding layer (cf. Section 3.1) to convert input symbols into embedded representations, feeds it to the bidirectional LSTM layer, and then constructs the context vector $\mathbf{c}$, either by simply letting $\mathbf{c} = \mathbf{h}_T$ (i.e., the hidden state of the bidirectional LSTM layer after reading the whole input sequence), or by computing the weighted sum described in Section 3.2 (if an attention mechanism is employed). In either case, the context vector $\mathbf{c}$ is passed over to the decoder, which generates the output symbols sequentially based on $\mathbf{c}$ and the current hidden state $\mathbf{s}_t$, using one or more bidirectional LSTM layers as in the encoder module. Instead of feeding $\mathbf{c}$ to the decoder only at the first time step (Sutskever et al., 2014; Vinyals and Le, 2015), we condition each output symbol y_t on $\mathbf{c}$, allowing the decoder to peek into the input at every step, as in Cho et al. (2014). Finally, a fully-connected layer with softmax activation converts the current output vector of the last LSTM layer into a probability distribution over the output vocabulary O. The complete encoder-decoder architecture (including the attention mechanism) is shown in Figure 3.

4 Multitask Learning with Multiple Auxiliary Losses

Several recent contributions (Søgaard and Goldberg, 2016; Bjerva et al., 2016; Plank et al., 2016; Luong et al., 2016) have shown the effectiveness of *multitask learning* (Caruana, 1997, MTL) in a sequence learning scenario. In MTL the idea is that of improving generalization performance by leveraging training signals contained in related tasks, in order to exploit their commonalities and differences. MTL is typically carried out by training a single architecture using multiple loss functions and a shared representation, with the underlying intention of improving a main task by incorporating joint learning of one or more related auxiliary tasks. From a practical point of view, MTL works by including one task-specific output layer per additional task, usually at the outermost level of the architecture, while keeping the remaining hidden layers common across all tasks.

In line with previous approaches, and guided by the intuition that WSD is strongly linked to other NLP tasks at various levels, we also design and study experimentally a multitask augmentation of the models described in Section 3. In particular, we consider two auxiliary tasks:

- **Part-of-speech (POS) tagging**, a standard auxiliary task extensively studied in previous work (Søgaard and Goldberg, 2016; Plank et al., 2016). Predicting the part-of-speech tag for a given token can also be informative for word senses, and help in dealing with cross-POS lexical ambiguities (e.g., *book a flight* vs. *reading a good book*);
- **Coarse-grained semantic labels (LEX)** based on the WordNet (Miller et al., 1990) lexicographer files,[4] i.e., 45 coarse-grained semantic categories manually associated with all the synsets in WordNet on the basis of both syntactic and logical groupings (e.g., *noun.location*, or *verb.motion*). These very coarse semantic labels, recently employed in a multitask setting by Alonso and Plank (2017), group together related senses and help the model to generalize, especially over senses less covered at training time.

We follow previous work (Plank et al., 2016; Alonso and Plank, 2017) and define an auxiliary loss function for each additional task. The overall loss is then computed by summing the main loss (i.e., the one associated with word sense labels) and all the auxiliary losses taken into account.

As regards the architecture, we consider both the models described in Sections 3.2 and 3.3 and modify them by adding two softmax layers in addition to the one in the original architecture. Figure 4 illustrates this for the attentive tagger of Section 3.2, considering both POS and LEX as auxiliary tasks. At the j^{th} time step the model predicts a sense label y_j together with a part-of-speech tag POS_j and a coarse semantic label LEX_j.[5]

5 Experimental Setup

In this section we detail the setup of our experimental evaluation. We first describe the training corpus and all the standard benchmarks for all-words WSD; we then report technical details on the architecture and on the training process for all the models described throughout Section 3 and their multitask augmentations (Section 4).

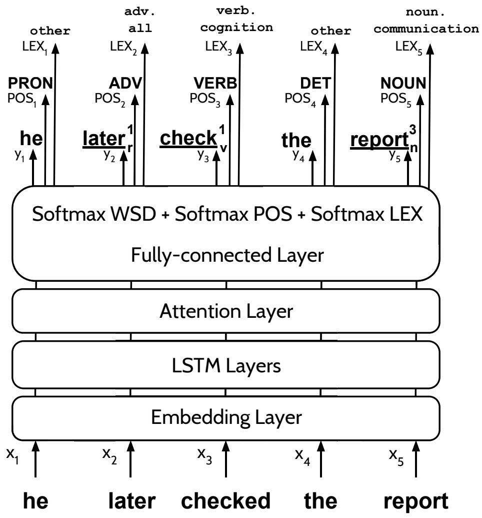

Figure 4: Multitask augmentation (with both POS and LEX as auxiliary tasks) for the attentive bidirectional LSTM tagger of Section 3.2.

Evaluation Benchmarks. We evaluated our models on the English all-words WSD task, considering both the fine-grained and coarse-grained benchmarks (Section 6.1). As regards fine-grained WSD, we relied on the evaluation framework of Raganato et al. (2017), which includes five standardized test sets from the Senseval/SemEval series: Senseval-2 (Edmonds and Cotton, 2001, **SE2**), Senseval-3 (Snyder and Palmer, 2004, **SE3**), SemEval-2007 (Pradhan et al., 2007, **SE07**), SemEval-2013 (Navigli et al., 2013, **SE13**) and SemEval-2015 (Moro and Navigli, 2015, **SE15**). Due to the lack of a reasonably large development set for our setup, we considered the smallest among these test sets, i.e., **SE07**, as development set and excluded it from the evaluation of Section 6.1. As for coarse-grained WSD, we used the SemEval-2007 task 7 test set (Navigli et al., 2007), which is not included in the standardized framework, and mapped the original sense inventory from WordNet 2.1 to WordNet 3.0.[6] Finally, we carried out an experiment on multilingual WSD using the Italian, German, French and Spanish data of **SE13**. For these benchmarks we relied on BabelNet (Navigli and Ponzetto, 2012)[7] as unified sense inventory.

[4] `https://wordnet.princeton.edu/man/lexnames.5WN.html`

[5] We use a dummy LEX label (`other`) for punctuation and function words.

[6] We utilized the original sense-key mappings available at `http://wordnetcode.princeton.edu/3.0` for nouns and verbs, and the automatic mappings by Daudé et al. (2003) for the remaining parts of speech (not available in the original mappings).

[7] `http://babelnet.org`

	Dev	Test Datasets				Concatenation of All Test Datasets				
	SE07	SE2	SE3	SE13	SE15	Nouns	Verbs	Adj.	Adv.	All
BLSTM	61.8	71.4	68.8	65.6	69.2	70.2	56.3	75.2	**84.4**	68.9
BLSTM + att.	62.4	71.4	**70.2**	66.4	70.8	71.0	**58.4**	75.2	83.5	69.7
BLSTM + att. + LEX	63.7	**72.0**	69.4	66.4	**72.4**	**71.6**	57.1	**75.6**	83.2	**69.9**
BLSTM + att. + LEX + POS	**64.8**	**72.0**	69.1	**66.9**	71.5	71.5	57.5	75.0	83.8	**69.9**
Seq2Seq	60.9	68.5	67.9	65.3	67.0	68.7	54.5	74.0	81.2	67.3
Seq2Seq + att.	62.9	69.9	69.6	65.6	67.7	69.5	57.2	74.5	81.8	68.4
Seq2Seq + att. + LEX	64.6	70.6	67.8	66.5	68.7	70.4	55.7	73.3	82.9	68.5
Seq2Seq + att. + LEX + POS	63.1	70.1	68.5	66.5	69.2	70.1	55.2	75.1	84.4	68.6
IMS	61.3	70.9	69.3	65.3	69.5	70.5	55.8	75.6	82.9	68.9
IMS+emb	**62.6**	**72.2**	**70.4**	65.9	71.5	**71.9**	56.6	**75.9**	**84.7**	**70.1**
Context2Vec	61.3	71.8	69.1	65.6	**71.9**	71.2	**57.4**	75.2	82.7	69.6
Lesk$_{ext}$+emb	⋆56.7	63.0	63.7	66.2	64.6	70.0	51.1	51.7	80.6	64.2
UKB$_{gloss}$ w2w	42.9	63.5	55.4	⋆62.9	63.3	64.9	41.4	69.5	69.7	61.1
Babelfy	51.6	⋆67.0	63.5	**66.4**	70.3	68.9	50.7	73.2	79.8	66.4
MFS	54.5	65.6	⋆66.0	63.8	⋆67.1	67.7	49.8	73.1	80.5	65.5

Table 1: F-scores (%) for English all-words fine-grained WSD on the test sets in the framework of Raganato et al. (2017) (including the development set **SE07**). The first system with a statistically significant difference from our best models is marked with ⋆ (unpaired t-test, $p < 0.05$).

At testing time, given a target word w, our models used the probability distribution over O, computed by the softmax layer at the corresponding time step, to rank the candidate senses of w; we then simply selected the top ranking candidate as output of the model.

Architecture Details. To set a level playing field with comparison systems on English all-words WSD, we followed Raganato et al. (2017) and, for all our models, we used a layer of word embeddings pre-trained[8] on the English ukWaC corpus (Baroni et al., 2009) as initialization, and kept them fixed during the training process. For all architectures we then employed 2 layers of bidirectional LSTM with 2048 hidden units (1024 units per direction).

As regards multilingual all-words WSD (Section 6.2), we experimented, instead, with two different configurations of the embedding layer: the pre-trained bilingual embeddings by Mrkšić et al. (2017) for all the language pairs of interest (EN-IT, EN-FR, EN-DE, and EN-ES), and the pre-trained multilingual 512-dimensional embeddings for 12 languages by Ammar et al. (2016).

Training. We used SemCor 3.0 (Miller et al., 1993) as training corpus for all our experiments. Widely known and utilized in the WSD literature, SemCor is one of the largest corpora annotated manually with word senses from the sense inventory of WordNet (Miller et al., 1990) for all open-class parts of speech. We used the standardized version of SemCor as provided in the evaluation framework[9] which also includes coarse-grained POS tags from the universal tagset. All models were trained for a fixed number of epochs $E = 40$ using Adadelta (Zeiler, 2012) with learning rate 1.0 and batch size 32. After each epoch we evaluated our models on the development set, and then compared the best iterations (E^*) on the development set with the reported state of the art in each benchmark.

6 Experimental Results

Throughout this section we identify the models based on the LSTM tagger (Sections 3.1-3.2) by the label **BLSTM**, and the sequence-to-sequence models (Section 3.3) by the label **Seq2Seq**.

6.1 English All-words WSD

Table 1 shows the performance of our models on the standardized benchmarks for all-words fine-grained WSD. We report the F1-score on each in-

[8]We followed Iacobacci et al. (2016) and used the Word2Vec (Mikolov et al., 2013) skip-gram model with 400 dimensions, 10 negative samples and a window size of 10.

[9]`http://lcl.uniroma1.it/wsdeval`

SemEval-2007 task 7			
BLSTM + att. + LEX	83.0	IMS	81.9
BLSTM + att. + LEX + POS	**83.1**	Chen et al. (2014)	82.6
Seq2Seq + att. + LEX	82.3	Yuan et al. (2016)	**82.8**
Seq2Seq + att. + LEX + POS	81.6	UKB w2w	80.1

Table 2: F-scores (%) for coarse-grained WSD.

	SemEval-2013 task 12			
	IT	**FR**	**DE**	**ES**
BLSTM (bilingual)	61.6	55.2	**69.2**	65.0
BLSTM (multilingual)	62.0	55.5	**69.2**	66.4
UMCC-DLSI	**65.8**	**60.5**	62.1	**71.0**
DAEBAK!	61.3	53.8	59.1	60.0
MFS	57.5	45.3	67.4	64.5

Table 3: F-scores (%) for multilingual WSD.

dividual test set, as well as the F1-score obtained on the concatenation of all four test sets, divided by part-of-speech tag.

We compared against the best supervised and knowledge-based systems evaluated on the same framework. As supervised systems, we considered **Context2Vec** (Melamud et al., 2016) and It Makes Sense (Zhong and Ng, 2010, **IMS**), both the original implementation and the best configuration reported by Iacobacci et al. (2016, **IMS**+**emb**), which also integrates word embeddings using exponential decay.[10] All these supervised systems were trained on the standardized version of SemCor. As knowledge-based systems we considered the embeddings-enhanced version of Lesk by Basile et al. (2014, **Lesk**$_{ext}$+**emb**), UKB (Agirre et al., 2014) (**UKB**$_{gloss}$ **w2w**) , and **Babelfy** (Moro et al., 2014). All these systems relied on the Most Frequent Sense (**MFS**) baseline as back-off strategy.[11] Overall, both **BLSTM** and **Seq2Seq** achieved results that are either state-of-the-art or statistically equivalent (unpaired t-test, $p < 0.05$) to the best supervised system in each benchmark, performing on par with word experts tuned over explicitly engineered features (Iacobacci et al., 2016). Interestingly enough, **BLSTM** models tended consistently to outperform their **Seq2Seq** counterparts, suggesting that an encoder-decoder architecture, despite being more powerful, might be suboptimal for WSD. Furthermore, introducing LEX (cf. Section 4) as auxiliary task was generally helpful; on the other hand, POS did not seem to help, corroborating previous findings (Alonso and Plank, 2017; Bingel and Søgaard, 2017).

The overall performance by part of speech was consistent with the above analysis, showing that our models outperformed all knowledge-based systems, while obtaining results that are superior or equivalent to the best supervised models. It is worth noting that RNN-based architectures outperformed classical supervised approaches (Zhong and Ng, 2010; Iacobacci et al., 2016) when dealing with verbs, which are shown to be highly ambiguous (Raganato et al., 2017).

The performance on coarse-grained WSD followed the same trend (Table 2). Both **BLSTM** and **Seq2Seq** outperformed UKB (Agirre et al., 2014) and IMS trained on SemCor (Taghipour and Ng, 2015a), as well as recent supervised approaches based on distributional semantics and neural architectures (Chen et al., 2014; Yuan et al., 2016).

6.2 Multilingual All-words WSD

All the neural architectures described in this paper can be readily adapted to work with different languages without adding sense-annotated data in the target language. In fact, as long as the first layer (cf. Figures 1-3) is equipped with *bilingual* or *multilingual* embeddings where word vectors in the training and target language are defined in the same space, the training process can be left unchanged, even if based only on English data. The underlying assumption is that words that are translations of each other (e.g., *house* in English and *casa* in Italian) are mapped to word embeddings that are as close as possible in the vector space.

In order to assess this, we considered one of our best models (**BLSTM+att.+LEX**) and replaced the monolingual embeddings with bilingual and multilingual embeddings (as specified in Section 5), leaving the rest of the architecture unchanged. We then trained these architectures on the same English training data, and ran the resulting models on the multilingual benchmarks of SemEval-2013 for Italian, French, German and Spanish. While doing this, we exploited BabelNet's inter-resource mappings to convert WordNet sense labels (used at training time) into BabelNet synsets compliant with the sense inventory of the task.

F-score figures (Table 3) show that bilingual and multilingual models, despite being trained only on English data, consistently outperformed the MFS

[10] We are not including Yuan et al. (2016), as their models are not available and not replicable on the standardized test sets, being based on proprietary data.

[11] Since each system always outputs an answer, F-score equals both precision and recall, and statistical significance can be expressed with respect to any of these measures.

baseline and achieved results that are competitive with the best participating systems in the task. We also note that the overall F-score performance did not change substantially (and slightly improved) when moving from bilingual to multilingual models, despite the increase in the number of target languages treated simultaneously.

6.3 Discussion and Error Analysis

All the neural models evaluated in Section 6.1 utilized the MFS back-off strategy for instances unseen at training time, which amounted to 9.4% overall for fine-grained WSD and 10.5% for coarse-grained WSD. Back-off strategy aside, 85% of the times the top candidate sense for a target instance lay within the 10 most probable entries in the probability distribution over O computed by the softmax layer.[12] In fact, our sequence models learned, on the one hand, to associate a target word with its candidate senses (something word experts are not required to learn, as they only deal with a single word type at a time); on the other, they tended to generate softmax distributions reflecting the semantics of the surrounding context. For example, in the sentence:

(a) The two *justices* have been attending federalist society events for years,

our model correctly disambiguated *justices* with the WordNet sense $\texttt{justice}_n^3$ (public official) rather than $\texttt{justice}_n^1$ (the quality of being just), and the corresponding softmax distribution was heavily biased towards words and senses related to persons or groups (*commissioners*, *defendants*, *jury*, *cabinet*, *directors*). On the other hand, in the sentence:

(b) Xavi Hernandez, the player of Barcelona, has 106 *matches*,

the same model disambiguated *matches* with the wrong WordNet sense $\texttt{match}_n^1$ (tool for starting a fire). This suggests that the signal carried by discriminative words like *player* vanishes rather quickly. In order to enforce global coherence further, recent contributions have proposed more sophisticated models where recurrent architectures are combined with Conditional Random Fields (Huang et al., 2015; Ma and Hovy, 2016). Finally, a number of errors were connected to shorter sentences with limited context for disambiguation: in fact, we noted that the average precision of our model, without MFS back-off, increased by 6.2% (from 74.6% to 80.8%) on sentences with more than 20 word tokens.

[12] We refer here to the same model considered in Section 6.2 (i.e., **BLSTM+att.+LEX**).

7 Conclusion

In this paper we adopted a new perspective on supervised WSD, so far typically viewed as a classification problem at the word level, and framed it using neural sequence learning. To this aim we defined, analyzed and compared experimentally different end-to-end models of varying complexities, including augmentations based on an attention mechanism and multitask learning.

Unlike previous supervised approaches, where a dedicated model needs to be trained for every content word and each disambiguation target is treated in isolation, sequence learning approaches learn a single model in one pass from the training data, and then disambiguate jointly all target words within an input text. The resulting models consistently achieved state-of-the-art (or statistically equivalent) figures in all benchmarks for all-words WSD, both fine-grained and coarse-grained, effectively demonstrating that we can overcome the so far undisputed and long-standing word-expert assumption of supervised WSD, while retaining the accuracy of supervised word experts.

Furthermore, these models are sufficiently flexible to allow them, for the first time in WSD, to be readily adapted to languages different from the one used at training time, and still achieve competitive results (as shown in Section 6.2). This crucial feature could potentially pave the way for cross-lingual supervised WSD, and overcome the shortage of sense-annotated data in multiple languages that, to date, has prevented the development of supervised models for languages other than English.

As future work, we plan to extend our evaluation to larger sense-annotated corpora (Raganato et al., 2016) as well as to different sense inventories and different languages. We also plan to exploit the flexibility of our models by integrating them into downstream applications, such as Machine Translation and Information Extraction.

Acknowledgments

The authors gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE No. 726487.

References

Eneko Agirre and Philip Edmonds. 2007. *Word sense disambiguation: Algorithms and applications*, volume 33. Springer Science & Business Media.

Eneko Agirre, Oier Lopez de Lacalle, and Aitor Soroa. 2014. Random Walks for Knowledge-Based Word Sense Disambiguation. *Comput. Linguist.*, 40(1):57–84.

Héctor Martínez Alonso and Barbara Plank. 2017. When is Multitask Learning Effective? Semantic Sequence Prediction under Varying Data Conditions. In *Proc. of ACL*, pages 44–53.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively Multilingual Word Embeddings. *CoRR*, abs/1602.01925.

Osman Başkaya and David Jurgens. 2016. Semi-supervised Learning with Induced Word Senses for State of the Art Word Sense Disambiguation. *J. Artif. Int. Res.*, 55(1):1025–1058.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR Workshop*.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2014. An Enhanced Lesk Word Sense Disambiguation Algorithm through a Distributional Semantic Model. In *Proc. of COLING*, pages 1591–1600.

Joachim Bingel and Anders Søgaard. 2017. Identifying Beneficial Task Relations for Multi-task Learning in Deep Neural Networks. In *Proc. of EACL*, pages 164–169.

Johannes Bjerva, Barbara Plank, and Johan Bos. 2016. Semantic Tagging with Deep Residual Networks. In *Proc. of COLING*, pages 3531–3541.

Andrei Butnaru, Radu Tudor Ionescu, and Florentina Hristea. 2017. ShotgunWSD: an Unsupervised Algorithm for Global Word Sense Disambiguation Inspired by DNA Sequencing. In *Proc. of EACL*, pages 916–926.

Marine Carpuat and Dekai Wu. 2007. Improving Statistical Machine Translation using Word Sense Disambiguation. In *Proc. of EMNLP*, pages 61–72.

Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28(1):41–75.

Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. 2014. A Unified Model for Word Sense Representation and Disambiguation. In *Proc. of EMNLP*, pages 1025–1035.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proc. of EMNLP*, pages 1724–1734.

Massimiliano Ciaramita and Yasemin Altun. 2006. Broad-coverage Sense Disambiguation and Information Extraction with a Supersense Sequence Tagger. In *Proc. of EMNLP*, pages 594–602.

Jordi Daudé, Lluís Padró, and German Rigau. 2003. Validation and Tuning of WordNet Mapping Techniques. In *Proc. of RANLP*, pages 117–123.

Oier Lopez de Lacalle and Eneko Agirre. 2015. A Methodology for Word Sense Disambiguation at 90% based on large-scale Crowd Sourcing. In *Proc. of SEM*, pages 61–70.

Claudio Delli Bovi, Luca Telesca, and Roberto Navigli. 2015. Large-Scale Information Extraction from Textual Definitions through Deep Syntactic and Semantic Analysis. *TACL*, 3:529–543.

Philip Edmonds and Scott Cotton. 2001. Senseval-2: Overview. In *Proc. of SENSEVAL*, pages 1–5.

Manaal Faruqui, Yulia Tsvetkov, Graham Neubig, and Chris Dyer. 2016. Morphological Inflection Generation Using Character Sequence to Sequence Learning. In *Proc. of NAACL-HLT*, pages 634–643.

Lucie Flekova and Iryna Gurevych. 2016. Supersense embeddings: A unified model for supersense interpretation, prediction, and utilization. In *Proc. of ACL*, pages 2029–2041.

Alex Graves. 2013. Generating Sequences With Recurrent Neural Networks. *CoRR*, abs/1308.0850.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6):602–610.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating Copying Mechanism in Sequence-to-Sequence Learning. In *Proc. of ACL*, pages 1631–1640.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.*, 9(8):1735–1780.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. *CoRR*, abs/1508.01991.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. SensEmbed: Learning Sense Embeddings for Word and Relational Similarity. In *Proc. of ACL*, pages 95–105.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for Word Sense Disambiguation: An Evaluation Study. In *Proc. of ACL*, pages 897–907.

Mikael Kågebäck and Hans Salomonsson. 2016. Word Sense Disambiguation using a Bidirectional LSTM. In *Proceedings of CogALex*, pages 51–56.

Adam Kilgarriff. 2001. English Lexical Sample Task Description. In *Proc. of SENSEVAL*, pages 17–20.

Ryan Kiros, Yukun Zhu, Ruslan R. Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Proc. of NIPS*, pages 3294–3302.

Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In *ICLR Workshop*.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In *Proc. of ACL*, pages 1064–1074.

Diana McCarthy, Marianna Apidianaki, and Katrin Erk. 2016. Word Sense Clustering and Clusterability. *Comput. Linguist.*, 42(2):245–275.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning Generic Context Embedding with Bidirectional LSTM. In *Proc. of CoNLL*, pages 51–61.

Rada Mihalcea, Timothy Chklovski, and Adam Kilgarriff. 2004. The Senseval-3 English Lexical Sample Task. In *Proc. of SENSEVAL*, pages 25–28.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *ICLR Workshop*.

George A. Miller, R.T. Beckwith, Christiane D. Fellbaum, D. Gross, and K. Miller. 1990. WordNet: an online lexical database. *International Journal of Lexicography*, 3(4):235–244.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Proc. of HLT*, pages 303–308.

Andrea Moro and Roberto Navigli. 2015. SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking. In *Proc. of SemEval*, pages 288–297.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *TACL*, 2:231–244.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Roi Reichart, Ira Leviant, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic Specialisation of Distributional Word Vector Spaces using Monolingual and Cross-Lingual Constraints. *TACL*, 5.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):10.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. SemEval-2013 Task 12: Multilingual Word Sense Disambiguation. In *Proc. of SemEval*, volume 2, pages 222–231.

Roberto Navigli, Kenneth C. Litkowski, and Orin Hargraves. 2007. SemEval-2007 Task 07: Coarse-grained English All-words Task. In *Proc. of SemEval*, pages 30–35.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, 193:217–250.

Steven Neale, Lus Gomes, Eneko Agirre, Oier Lopez de Lacalle, and Antnio Branco. 2016. Word Sense-Aware Machine Translation: Including Senses as Contextual Features for Improved Translation Models. In *Proc. of LREC*, pages 2777–2783.

Tommaso Pasini and Roberto Navigli. 2017. Train-O-Matic: Large-Scale Supervised Word Sense Disambiguation in Multiple Languages without Manual Training Data. In *Proc. of EMNLP*.

Mohammad Taher Pilehvar and Roberto Navigli. 2014. A Large-scale Pseudoword-based Evaluation Framework for State-of-the-art Word Sense Disambiguation. *Comput. Linguist.*, 40(4):837–881.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss. In *Proc. of ACL*, pages 412–418.

Sameer S. Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. SemEval-2007 Task 17: English Lexical Sample, SRL and All Words. In *Proc. of SemEval-2007*, pages 87–92.

Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. 2016. Automatic Construction and Evaluation of a Large Semantically Enriched Wikipedia. In *Proc. of IJCAI*, pages 2894–2900.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison. In *Proc. of EACL*, pages 99–110.

Sascha Rothe and Hinrich Schütze. 2015. AutoExtend: Extending Word Embeddings to Embeddings for Synsets and Lexemes. In *Proc. of ACL*, pages 1793–1803.

Hui Shen, Razvan Bunescu, and Rada Mihalcea. 2013. Coarse to Fine Grained Sense Disambiguation in Wikipedia. In *Proc. of SEM*, pages 22–31.

Benjamin Snyder and Martha Palmer. 2004. The english all-words task. In *Proc. of Senseval-3*, pages 41–43.

Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *Proc. of ACL*, pages 231–235.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Proc. of NIPS*, pages 3104–3112.

Kaveh Taghipour and Hwee Tou Ng. 2015a. One Million Sense-Tagged Instances for Word Sense Disambiguation and Induction. In *Proc. of CoNLL*, pages 338–344.

Kaveh Taghipour and Hwee Tou Ng. 2015b. Semi-Supervised Word Sense Disambiguation Using Word Embeddings in General and Specific Domains. In *Proc. of NAACL-HLT*, pages 314–323.

Rocco Tripodi and Marcello Pelillo. 2017. A Game-Theoretic Approach to Word Sense Disambiguation. *Comput. Linguist.*, 43(1):31–70.

Oriol Vinyals, Ł ukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton. 2015. Grammar as a foreign language. In *Proc. of NIPS*, pages 2773–2781.

Oriol Vinyals and Quoc V. Le. 2015. A Neural Conversational Model. In *Proc. of ICML, JMLR: W&CP*, volume 37.

Wenhui Wang and Baobao Chang. 2016. Graph-based Dependency Parsing with Bidirectional LSTM. In *Proc. of ACL*, pages 2306–2315.

Dirk Weissenborn, Leonhard Hennig, Feiyu Xu, and Hans Uszkoreit. 2015. Multi-Objective Optimization for the Joint Disambiguation of Nouns and Named Entities. In *Proc. of ACL*, pages 596–605.

Deyi Xiong and Min Zhang. 2014. A Sense-Based Translation Model for Statistical Machine Translation. In *Proc. of ACL*, pages 1459–1469.

Dayu Yuan, Ryan Doherty, Julian Richardson, Colin Evans, and Eric Altendorf. 2016. Semi-supervised Word Sense Disambiguation with Neural Models. In *Proc. of COLING*, pages 1374–1385.

Matthew D. Zeiler. 2012. ADADELTA: An Adaptive Learning Rate Method. *CoRR*, abs/1212.5701.

Zhi Zhong and Hwee Tou Ng. 2010. It Makes Sense: A Wide-Coverage Word Sense Disambiguation System for Free Text. In *Proc. of ACL: System Demonstrations*, pages 78–83.

Zhi Zhong and Hwee Tou Ng. 2012. Word Sense Disambiguation Improves Information Retrieval. In *Proc. of ACL*, pages 273–282.

Jie Zhou and Wei Xu. 2015. End-to-End Learning of Semantic Role Labeling using Recurrent Neural Networks. In *Proc. of ACL*, pages 1127–1137.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. In *Proc. of ACL*, pages 207–212.