EMNLP-CoNLL 2007

Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning

June 28–30, 2007
Prague, Czech Republic

The Association for Computational Linguistics

Production and Manufacturing by
Omnipress
2600 Anderson Street
Madison, WI 53704
USA

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

Preface

Welcome to EMNLP-CoNLL 2007, an unprecedented joint meeting of the Conference on Empirical Methods in Natural Language Processing (EMNLP) and the Conference on Computational Natural Language Learning (CoNLL).

The conference is a joint effort of SIGDAT and SIGNLL, the ACL Special Interest Groups that usually organize the annual EMNLP and CoNLL conferences, respectively.

Our field is growing rapidly. This year, EMNLP-CoNLL considered a remarkable 398 submissions,[1] accepting 109 of them (for an acceptance rate of 27%). It is startling to realize that even the ACL conferences were not this large until two years ago.

Only 66 of the accepted papers were scheduled for presentation as talks, and 43 more as posters. We took pains to ensure that the poster sessions would be leisurely and interactive.

In addition, two sessions of the conference and 22 specially designated short papers in this volume are devoted to the CoNLL Shared Task competition, an annual tradition. The 2007 competition concerns dependency parsing, with both a multilingual track and a domain adaptation track.

Several innovations this year have received positive feedback and are worth mentioning:

- To encourage thorough citation of related work, a paper's References section was *not* counted against the 8-page limit for submitted papers or the 9-page limit for camera-ready papers.

  (Note that authors were allowed an extra page in the camera-ready version to help them effectively address reviewers' comments, following an innovation at EMNLP 2006.)

- The review form was redesigned (starting from the fine review form of EACL 2006) to provide clearer and more consistent guidance to reviewers, area chairs, and authors. Authors were directed to consult the review form, which was posted at the conference website, while preparing their submissions and when interpreting their review scores.

- Some of our submissions (fewer than 1/3) appeared to be revisions of rejected ACL 2007 submissions. Where possible, we tried to conserve valuable information and effort from the ACL 2007 reviewing process by re-assigning one, though only one, of the ACL reviewers to such a paper.

  Such re-reviewers were instructed to give the new, revised submission the fresh reading that it deserved, but they were also encouraged to bring up points that still applied from any of the ACL 2007 reviews or discussion.

[1] Of the original 419 submissions, 17 were withdrawn (usually upon acceptance elsewhere), and 4 more were rejected without review (for violating the conference's standards on length, anonymization, or plagiarism).

- By accepting many posters and presenting them all *simultaneously*, we hoped to accommodate a large audience without overcrowding at each poster.

  The large number of posters in turn required a long period for poster viewing. With a total of 5 hours spanning two receptions, a conferencegoer can engage with nearly half of the posters for 15 minutes of personalized discussion each. This makes the posters roughly as visible as the talks, which are split into parallel sessions.

- In addition to the Best Paper Award (see Session 1), we are considering organizing—if logistically feasible—an "Audience Choice" award for the most worthwhile *presentation* at the conference. Such a prize would reward authors who not only produced outstanding research but also communicated it clearly and enjoyably at the conference meeting.

It is my privilege to thank the many individuals—most of them listed on the following pages—whose generous efforts have made this conference possible. Foremost are the 16 dedicated area chairs and 370 reviewers, who worked together hard and thoughtfully to select this excellent program and provide valuable feedback to the authors. Also as part of the technical program, Joakim Nivre chaired the organization of the CoNLL Shared Task and the resulting short papers; Taku Kudo ably identified ACL 2007 resubmissions (see above); and Hal Daumé III kindly chaired the best paper award committee. Eric Ringger put a great deal of effort into producing this fine proceedings volume, with support from Su Jian, the ACL publications chair. Jan Hajic coordinated the many local arrangements, along with Priscilla Rasmussen, Anna Kotesovcova, Jiri Mirovsky, Pavel Stranak, Zdenek Zabokrtsky, and no doubt others; we are very grateful to them for making everything run smoothly in Prague. Antal van den Bosch, Dan Jurafsky, Eric Gaussier, and Ken Church provided much valuable advice over the past months based on their experience. Finally, let us not forget the hundreds of authors who actually produced the excellent research in this volume, and the invited speakers who graciously traveled a long way to enlighten us.

Enjoy the conference!

Jason Eisner
EMNLP-CoNLL Chair
May 2007

Organizers

Program Chair:

Jason Eisner, Johns Hopkins University

Area Chairs:

David Chiang, USC Information Sciences Institute
Alexander Clark, Royal Holloway University of London
Michael Collins, Massachusetts Institute of Technology
Hal Daumé III, University of Utah
Sanjeev Khudanpur, Johns Hopkins University
Katrin Kirchhoff, University of Washington
Dekang Lin, Google Inc.
Manabu Okumura, Tokyo Institute of Technology
Anoop Sarkar, Simon Fraser University
Suzanne Stevenson, University of Toronto
Tomek Strzalkowski, SUNY Albany
Carlo Strapparava, Istituto Trentino di Cultura
Hans Uszkoreit, Saarland University DFKI
Menno van Zaanen, Macquarie University
Marilyn Walker, University of Sheffield
Janyce Wiebe, University of Pittsburgh

CoNLL Shared Task Organizers:

Joakim Nivre, Växjö University and Uppsala University (chair)
Johan Hall, Växjö University
Sandra Kübler, Indiana University
Ryan McDonald, Google Inc.
Jens Nilsson, Växjö University
Sebastian Riedel, University of Edinburgh
Deniz Yuret, Koç University

Local Arrangements Chair:

Jan Hajic, Charles University

Publications Chair:

Eric Ringger, Brigham Young University

Reviewers:

John Aberdeen, Pieter Adriaans, Eugene Agichtein, Eneko Agirre, Gregory Aist, Enrique Alfonseca, Rie Ando, Necip Fazil Ayan, Leif Azzopardi;

Srinivas Bangalore, Marco Baroni, Regina Barzilay, Roberto Basili, Sugato Basu, Ron Bekkerman, Anja Belz, Sabine Bergler, Shane Bergsma, Dan Bikel, Misha Bilenko, Dave Blei, John Blitzer, Rens Bod, Bernd Bohnet, Johan Bos, Thorsten Brants, Eric Breck, Chris Brew, David Brooks, Charles Brown, Wray Buntine, John Burger;

Janet Cahn, Giuseppe Carenini, Xavier Carreras Pérez, Neus Català, Damir Ćavar, Joyce Chai, Yee Seng Chan, Jason Chang, Ciprian Chelba, Hsin-Hsi Chen, Stanley Chen, Colin Cherry, Yejin Choi, Jennifer Chu-Carroll, Grace Chung, Ken Church, Massimiliano Ciaramita, Alexander Clark, Stephen Clark, John Coleman, Paul Cook, Christophe Costa Florêncio, Mathias Creutz, Dan Cristea, András Csomai, Elsa Cubel, Silviu Cucerzan, Aron Culotta, James Curran;

Walter Daelemans, Robert Dale, R. I. Damper, Hal Daumé III, Eric Villemonte de la Clergerie, Maarten de Rijke, Christy Doran, Mark Dras, Amit Dubey, Kevin Duh;

Phil Edmonds, Noémie Elhadad, T. Mark Ellison, Ahmad Emami, Katrin Erk, David Evans, Rémi Eyraud;

Afsaneh Fazly, Marcello Federico, Karim Filali, Jenny Finkel, Kate Forbes-Riley, Eric Fosler-Lussier, George Foster, Mary Ellen Foster, Alexander Fraser, Dayne Freitag, Atsushi Fujii, Sean Fulop, Pascale Fung;

Tamas Gaal, Evgeniy Gabrilovich, Michel Galley, Michael Gamon, Claire Gardent, Jeroen Geertzen, Dale Gerdemann, Ulrich Germann, Daniel Gildea, Roxana Girju, John Goldsmith, Jade Goldstein, Sharon Goldwater, Cyril Goutte, Mark Greenwood, Gregory Grefenstette, Tom Griffiths, Iryna Gurevych;

Kadri Hacioglu, Aria Haghighi, Udo Hahn, Jan Hajic, Dilek Hakkani-Tür, Keith Hall, Susan Haller, Hilda Hardy, Mary Harper, Mary Hearne, Marti Hearst, Peter Heeman, James Henderson, John Henderson, Mark Hepple, Ryuichiro Higashinaka, Tsutomu Hirao, Graeme Hirst, Julia Hockenmaier, Chu-Ren Huang, Liang Huang, Annette Hulth, Rebecca Hwa;

Diana Inkpen, Kentaro Inui, Abe Ittycheriah;

Martin Jansche, Valentin Jijkoun, Mark Johnson, Kristiina Jokinen;

Min-Yen Kan, Hiroshi Kanayama, Damianos Karakos, Lauri Karttunen, Rohit Kate, Frank Keller, André Kempe, Adam Kilgarriff, Soo-Min Kim, George Kiraz, Katrin Kirchhoff, Chunyu Kit, Kevin Knight, Alistair Knott, Philipp Koehn, Rob Koeling, Alexander Koller, Grzegorz Kondrak, Stasinos Konstantopoulos, Terry Koo, Moshe Koppel, Anna Korhonen, András Kornai, Kimmo Koskenniemi, Sandra Kübler, Roland Kuhn, Shankar Kumar, Hong-Kwang Kuo, Sadao Kurohashi;

Philippe Langlais, Guy Lapalme, Mirella Lapata, Eric Laporte, Staffan Larsson, Geunbae Lee, Lillian Lee, Oliver Lemon, Lori Levin, Roger Levy, Hang Li, Wei Li, Chin-Yew Lin, Ying Lin, Bing Liu, Yang Liu, Adam Lopez, Saturnino Luz, Caroline Lyon;

Bernardo Magnini, Milind Mahajan, Fran cois Mairesse, Suresh Manandhar, Lidia Mangu, Gideon Mann, Christopher Manning, Daniel Marcu, Mitchell Marcus, Katja Markert, David Martinez, Yuji Matsumoto, Takuya Matsuzaki, Irina Matveeva, John Maxwell, David McClosky, Ryan McDonald, Susan McRoy, Helen Meng, Wolfgang Menzel, Detmar Meurers, Rada Mihalcea, Eleni Miltsakaki, Gilad Mishne, Yusuke Miyao, Marie-Francine Moens, Saif Mohammad, Mehryar Mohri, Christof Monz, Robert Moore, Tatsunori Mori, Alessandro Moschitti, Karin Müller, Dragos Stefan Munteanu;

Vivi Nastase, Roberto Navigli, Mark-Jan Nederhof, Ani Nenkova, Hwee Tou Ng, Grace Ngai, Patrick Nguyen, Vincent Ng, Malvina Nissim, Cheng Niu, Joakim Nivre, Tadashi Nomoto, David Novick;

Tim Oates, Franz Och, Kemal Oflazer, Paul Ogilvie, Miles Osborne;

Sebastian Padó, Tim Paek, Chris Pal, David Palmer, Martha Palmer, Bo Pang, Patrick Pantel, Marius Pasca, Rebecca Passonneau, Slav Petrov, Fabio Pianesi, Paul Piwek, Ferran Pla, Massimo Poesio, Richard Power, David Powers, Sameer Pradhan, Rashmi Prasad, Adam Przepiórkowski, Stephen Pulman, Vasin Punyakanok, Matthew Purver;

Yan Qu, Chris Quirk;

Owen Rambow, Deepak Ravichandran, Giuseppe Riccardi, Sebastian Riedel, Stefan Riezler, German Rigau, Ellen Riloff, Sophie Rosset;

Kenji Sagae, Magnus Sahlgren, Helmut Schmid, Patrick Schone, Sabine Schulte im Walde, Tanja Schultz, Dale Schuurmans, Holger Schwenk, Frédérique Segond, Satoshi Sekine, Izhak Shafran, Libin Shen, Khalil Sima'an, Michel Simard, Wojciech Skut, David A. Smith, Noah A. Smith, Rion Snow, Stephen Soderland, Swapna Somasundaran, Radu Soricut, Caroline Sporleder, Richard Sproat, Padmini Srinivasan, Brad Starkie, Manfred Stede, Mark Steedman, Mark Stevenson, Matthew Stone, Veselin Stoyanov, Carlo Strapparava, Michael Strube, Jian Su, Maosong Sun, Mihai Surdeanu, Charles Sutton, Hisami Suzuki, Jun Suzuki, Marc Swerts, Stan Szpakowicz;

Maite Taboada, Hiroya Takamura, Isabelle Tellier, Joel Tetreault, Simone Teufel, Mariët Theune, Franck Thollard, Christoph Tillmann, Kristina Toutanova, Vivian Tsang, Jun'ichi Tsujii, Dan Tufis;

Nicola Ueffing, Ozlem Uzuner;

Antal van den Bosch, Ashish Venugopal, Renata Vieira, Marc Vilain, Begoña Villada Moirón, Aline Villavicencio, Carl Vogel, Stephan Vogel, Piek Vossen;

Qin Wang, Shaojun Wang, Wei Wang, Wen Wang, Ye-Yi Wang, Wayne Ward, Taro Watanabe, Andy Way, Nick Webb, Bonnie Webber, Ralph Weischedel, Edward Whittaker, Richard Wicentowski, Yorick Wilks, Theresa Wilson, Shuly Wintner, Dekai Wu, Xiaoyun Wu;

Aris Xanthos, Fei Xia, Eric Xing, Feiyu Xu, Peng Xu;

Scott Wen-tau Yih;

Fabio Zanzotto, Richard Zens, ChengXiang Zhai, Hao Zhang, Tong Zhang, Ying Zhang, Shubin Zhao, Guoding Zhou, Liang Zhou, Jerry Zhu, Michael Zock, Jelle Zuidema

CoNLL Shared Task Reviewers:

Jason Baldridge, Sabine Buchholz, James Clarke, Gülşen Eryiğit, Kilian Evang, Johan Hall, Julia Hockenmaier, Yuval Krymolowski, Sandra Kübler, Erwin Marsi, Ryan McDonald, Beáta Megyesi, Jens Nilsson, Joakim Nivre, Sebastian Riedel, Yannick Versley, Alexander Yeh, Deniz Yuret

Table of Contents

Conference Program Overview

Thursday, June 28, 2007

9:00–10:45	Session 1: Plenary Session
10:45–11:15	Morning Break
11:15–12:30	Sessions 2a and 2b
12:30–14:00	Lunch
14:00–15:40	Sessions 3a and 3b
15:40–16:00	Afternoon Break
16:00–18:30	Session 4: All Posters

Friday, June 29, 2007

9:00–10:40	Sessions 5a and 5b
10:40–11:15	Morning Break
11:15–12:30	Sessions 6a and 6b
12:30–14:00	Lunch
14:00–15:40	Sessions 7a and 7b
15:40–16:00	Afternoon Break
16:00–18:30	Session 8: All Posters

Saturday, June 30, 2007

9:00–10:00	Session 9: Plenary Session
10:00–10:50	Sessions 10a, 10b, and 10c
10:50–11:15	Morning Break
11:15–12:30	Sessions 11a, 11b, and 11c
12:30–14:00	Lunch
14:00–15:40	Sessions 12a, 12b, and 12c
15:40–16:15	Afternoon Break
16:15–17:30	Sessions 13a, 13b, and 13c
17:30	Concluding Session

Conference Program

Thursday, June 28, 2007

Thursday, June 28, 2007 (continued)

Session 3a: Generation, Summarization, and Discourse

Session 3b: Parsing

Thursday, June 28, 2007 (continued)

Session 4: All Posters (16:00–18:30)

Thursday, June 28, 2007 (continued)

Session 4: All Posters (16:00–18:30) (continued)

Thursday, June 28, 2007 (continued)

Session 4: All Posters (16:00–18:30) (continued)

Thursday, June 28, 2007 (continued)

Session 4: All Posters (16:00–18:30) (continued)

Friday, June 29, 2007

Session 5a: Semantics

Session 5b: Parsing

Session 6a: Document Analysis

Session 6b: Grammar Learning

Friday, June 29, 2007 (continued)

Session 7a: Information Extraction

Session 7b: Machine Translation

Session 8: All Posters (16:00–18:30)

Consult the list of poster titles under Session 4.

Saturday, June 30, 2007

Session 9: Plenary Session

9:00–10:00 Invited Talk: *Hashing, Sketching, and Other Approximate Algorithms for High-Dimensional Data*
Piotr Indyk, Massachusetts Institute of Technology

Session 10a: Machine Learning (supervised classifiers)

10:00–10:25 *Scalable Term Selection for Text Categorization*
Jingyang Li and Maosong Sun

10:25–10:50 *Active Learning for Word Sense Disambiguation with Methods for Addressing the Class Imbalance Problem*
Jingbo Zhu and Eduard Hovy

Session 10b: Machine Learning (sequential models)

10:00–10:25 *Semi-Supervised Structured Output Learning Based on a Hybrid Generative and Discriminative Approach*
Jun Suzuki, Akinori Fujino and Hideki Isozaki

10:25–10:50 *Finding Good Sequential Model Structures using Output Transformations*
Edward Loper

Session 10c: Information Retrieval

10:00–10:25 *A Statistical Language Modeling Approach to Lattice-Based Spoken Document Retrieval*
Tee Kiah Chia, Haizhou Li and Hwee Tou Ng

10:25–10:50 *Learning Noun Phrase Query Segmentation*
Shane Bergsma and Qin Iris Wang

Saturday, June 30, 2007 (continued)

Saturday, June 30, 2007 (continued)

Session 12a: CoNLL Shared Task Session (dependency parsing)

Saturday, June 30, 2007 (continued)

Session 12b: Machine Translation

Session 12c: Word Senses

Saturday, June 30, 2007 (continued)

Session 13a: CoNLL Shared Task Session (dependency parsing)

16:15–16:30 *Dependency Parsing and Domain Adaptation with LR Models and Parser Ensembles*
Kenji Sagae and Jun'ichi Tsujii

16:30–16:45 *Frustratingly Hard Domain Adaptation for Dependency Parsing*
Mark Dredze, John Blitzer, Partha Pratim Talukdar, Kuzman Ganchev, João Graca and Fernando Pereira

16:45–17:15 Analysis: Sandra Kübler, Ryan McDonald

17:15–17:30 Discussion

Session 13b: Sentiment

16:15–16:40 *Crystal: Analyzing Predictive Opinions on the Web*
Soo-Min Kim and Eduard Hovy

16:40–17:05 *Extracting Aspect-Evaluation and Aspect-Of Relations in Opinion Mining*
Nozomi Kobayashi, Kentaro Inui and Yuji Matsumoto

17:05–17:30 *Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents*
Nobuhiro Kaji and Masaru Kitsuregawa

Session 13c: Tagging

16:15–16:40 *Determining Case in Arabic: Learning Complex Linguistic Behavior Requires Complex Linguistic Features*
Nizar Habash, Ryan Gabbard, Owen Rambow, Seth Kulick and Mitch Marcus

16:40–17:05 *Mandarin Part-of-Speech Tagging and Discriminative Reranking*
Zhongqiang Huang, Mary Harper and Wen Wang

17:05–17:30 *Building Domain-Specific Taggers without Annotated (Domain) Data*
John Miller, Manabu Torii and K. Vijay-Shanker

Concluding Session

17:30 Closing Remarks

Additional CoNLL Shared Task Papers (dependency parsing)