CoNLL-X

Proceedings of the

Tenth Conference on Computational Natural Language Learning

8-9 June 2006
New York City, USA

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

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

Foreword

CoNLL has turned ten! With a mix of pride and amazement over how time flies, we now celebrate the tenth time that ACL's special interest group on natural language learning, SIGNLL, holds its yearly conference.

Having a yearly meeting was the major pillar of the design plan for SIGNLL, drawn up by a circle of enthusiastic like-minded people around 1995, headed by first president David Powers and first secretary Walter Daelemans. The first CoNLL was organized as a satellite event of ACL-97 in Madrid, in the capable hands of Mark Ellison. Since then, no single year has gone by without a CoNLL. The boards of SIGNLL (with consecutive presidents Michael Brent, Walter Daelemans, and Dan Roth) have made sure that CoNLL toured the world; twice it was held in the Asian-Pacific part of the world, four times in Europe, and four times in the North-American continent.

Over time, the field of computational linguistics got to know CoNLL for its particular take on empirical methods for NLP and the ties these methods have with areas outside the focus of the typical ACL conference. The image of CoNLL was furthermore boosted by the splendid concept of the shared task, the organized competition that tackles timely tasks in NLP and has produced both powerful and sobering scientific insights. The CoNLL shared tasks have produced benchmark data sets and results on which a significant body of work in computational linguistics is based nowadays. The first shared task was organized in 1999 on NP bracketing, by Erik Tjong Kim Sang and Miles Osborne. With the help of others, Erik continued the organization of shared tasks until 2003 (on syntactic chunking, clause identification, and named-entity recognition), after which Lluís Màrquez and Xavier Carreras organized two consecutive shared tasks on semantic role labeling (2004, 2005). This year's shared task on multi-lingual dependency parsing holds great promise in becoming a new landmark in NLP research.

With great gratitude we salute all past CoNLL programme chairs and reviewers who have made CoNLL possible, and who have contributed to this conference series, which we believe has a shining future ahead. We are still exploring unknown territory in the fields of language learning, where models of human learning and natural language processing may on one day be one. We hope we will see a long series of CoNLLs along that path.

1997 - Madrid, Spain (chair: T. Mark Ellison)
1998 - Sydney, Australia (chair: David Powers)
1999 - Bergen, Norway (chairs: Miles Osborne and Erik Tjong Kim Sang)
2000 - Lisbon, Portugal (chairs: Claire Cardie, Walter Daelemans, and Erik Tjong Kim Sang)
2001 - Toulouse, France (chairs: Walter Daelemans and Rémi Zajac)
2002 - Taipei, Taiwan (chairs: Dan Roth and Antal van den Bosch)
2003 - Edmonton, Canada (chairs: Walter Daelemans and Miles Osborne)
2004 - Boston, MA, USA (chairs: Hwee Tou Ng and Ellen Riloff)
2005 - Ann Arbor, MI, USA (chairs: Ido Dagan and Dan Gildea)
2006 - New York City, NY, USA (chairs: Lluís Màrquez and Dan Klein)

Antal van den Bosch, President
Hwee Tou Ng, Secretary

Preface

The 2006 Conference on Computational Natural Language Learning is the tenth in a series of yearly meetings organized by SIGNLL, the ACL special interest group on natural language learning. Due to the special occasion, we have brought out the celebratory Roman numerals: welcome to CoNLL-X! Presumably, next year we will return to CoNLL-2007 (until 2016, when perhaps we will see CoNLL-XX). CoNLL-X will be held in New York City on June 8-9, in conjunction with the HLT-NAACL 2006 conference.

A total of 52 papers were submitted to CoNLL's main session, from which only 18 were accepted. The 35% acceptance ratio maintains the high competitiveness of recent CoNLLs and is an indicator of this year's high-quality programme. We are very grateful to the CoNLL community for the large amount of exciting, diverse, and high-quality submissions we received. We are equally grateful to the program committee for their service in reviewing these submissions, on a very tight schedule. Your efforts made our job a pleasure.

As in previous years, we defined a topic of special interest for the conference. This year, we particularly encouraged submissions describing architectures, algorithms, methods, or models designed to improve the robustness of learning-based NLP systems. While the topic of interest was directly addressed by only a small number of the main session submissions, the shared task setting contributed significantly in this direction.

Also following CoNLL tradition, a centerpiece of the confernence is a shared task, this year on multilingual dependency parsing. The shared task was organized by Sabine Buchholz, Amit Dubey, Yuval Krymolwski, and Erwin Marsi, who worked very hard to make the shared task the success it has been. Up to 13 different languages were treated. 19 teams submitted results, from which 17 are presenting description papers in the proceedings. In our opinion, the current shared task constitutes a qualitative step ahead in the evolution of CoNLL shared tasks, and we hope that the resources created and the body of work presented will both serve as a benchmark and also have a substantial impact on future research on syntactic parsing.

Finally, we are delighted to announce that this year's invited speakers are Michael Collins and Walter Daelemans. In accordance with the tenth anniversary celebration, Walter Daelemans will look back at the 10 years of CoNLL conferences, presenting the state of the art in computational natural language learning, and suggesting a new "mission" for the future of field. Michael Collins, in turn, will talk about one of the important current research lines in the field: global learning architectures for structural and relational learning problems in natural language.

In addition to the program committee and shared task organizers, we are very indebted to the SIGNLL board members for very helpful discussion and advice, Erik Tjong Kim Sang, who acted as the information officer, and the HLT-NAACL 2006 conference organizers, in particular Robert Moore, Brian Roark, Sanjeev Khudanpur, Lucy Vanderwende, Roberto Pieraccini, and Liz Liddy for their help with local arrangements and the publication of the proceedings.

To all the attendees, enjoy the CoNLL-X conference!

Lluís Màrquez and Dan Klein
CoNLL-X Program Co-Chairs

Organizers:

Lluís Màrquez, Technical University of Catalonia, Spain
Dan Klein, University of California at Berkeley, USA

Shared Task Organizers:

Sabine Buchholz, Toshiba Research Europe Ltd, UK
Amit Dubey, University of Edinburgh, UK
Yuval Krymolowski, University of Haifa, Israel
Erwin Marsi, Tilburg University, The Netherlands

Information Officer:

Erik Tjong Kim Sang, University of Amsterdam, The Netherlands

Program Committee:

Eneko Agirre, University of the Basque Country, Spain
Regina Barzilay, Massachusetts Institute of Technology, USA
Thorsten Brants, Google Inc., USA
Xavier Carreras, Polytechnical University of Catalunya, Spain
Eugene Charniak, Brown University, USA
Alexander Clark, Royal Holloway University of London, UK
James Cussens, University of York, UK
Walter Daelemans, University of Antwerp, Belgium
Hal Daum, ISI, University of Southern California, USA
Radu Florian, IBM, USA
Dayne Freitag, Fair Isaac Corporation, USA
Daniel Gildea, University of Rochester, USA
Trond Grenager, Stanford University, USA
Marti Hearst, I-School, UC Berkeley, USA
Philipp Koehn, University of Edinburgh, UK
Roger Levy, University of Edinburgh, UK
Rob Malouf, San Diego State University, USA
Christopher Manning, Stanford University, USA
Yuji Matsumoto, Nara Institute of Science and Technology, Japan
Andrew McCallum, University of Massachusetts Amherst, USA
Rada Mihalcea, University of North Texas, USA
Alessandro Moschitti, University of Rome Tor Vergata, Italy
John Nerbonne, University of Groningen, The Netherlands
Hwee Tou Ng, National University of Singapore, Singapore
Franz Josef Och, Google Inc., USA
Miles Osborne, University of Edinburgh, UK

David Powers, Flinders University, Australia
Ellen Riloff, University of Utah, USA
Dan Roth, University of Illinois at Urbana-Champaign, USA
Anoop Sarkar, Simon Fraser University, Canada
Noah Smith, Johns Hopkins University, USA
Suzanne Stevenson, University of Toronto, Canada
Mihai Surdeanu, Polytechnical University of Catalunya, Spain
Charles Sutton, University of Massachusetts Amherst, USA
Kristina Toutanova, Microsoft Research, USA
Antal van den Bosch, Tilburg University, The Netherlands
Janyce Wiebe, University of Pittsburgh, USA
Dekai Wu, Hong Kong University of Science and Technology, Hong Kong

Additional Reviewers:

Sander Canisius, Michael Connor, Andras Csomai, Aron Culotta, Quang Do, Gholamreza Haffari, Yudong Liu, David Martinez, Vanessa Murdoch, Vasin Punyakanok, Lev Ravitov, Kevin Small,Dong Song, Adam Vogel

Invited Speakers:

Michael Collins, Massachusetts Institute of Technology, USA
Walter Daelemans, University of Antwerp, Belgium

Table of Contents

Conference Program

Thursday, June 8, 2006

8:45–8:50 Welcome

Session 1: Syntax and Statistical Parsing

8:50–9:15 *Porting Statistical Parsers with Data-Defined Kernels*
Ivan Titov and James Henderson

9:15–9:40 *Non-Local Modeling with a Mixture of PCFGs*
Slav Petrov, Leon Barrett and Dan Klein

9:40–10:05 *Improved Large Margin Dependency Parsing via Local Constraints and Laplacian Regularization*
Qin Iris Wang, Colin Cherry, Dan Lizotte and Dale Schuurmans

10:05–10:30 *What are the Productive Units of Natural Language Grammar? A DOP Approach to the Automatic Identification of Constructions.*
Willem Zuidema

10:30–11:00 coffee break

11:00–11:50 Invited Talk by Michael Collins

Session 2: Anaphora Resolution and Paraphrasing

11:50–12:15 *Resolving and Generating Definite Anaphora by Modeling Hypernymy using Unlabeled Corpora*
Nikesh Garera and David Yarowsky

12:15–12:40 *Investigating Lexical Substitution Scoring for Subtitle Generation*
Oren Glickman, Ido Dagan, Walter Daelemans, Mikaela Keller and Samy Bengio

12:40–14:00 lunch

Session 3: Shared Task on Dependency Parsing

14:00–15:30 Introduction and System presentation I

15:30–16:00 coffee break

16:00–18:00 System presentation II and Discussion

Friday, June 9, 2006

Session 4: Semantic Role Labeling and Semantics

8:50–9:15 *Semantic Role Recognition Using Kernels on Weighted Marked Ordered Labeled Trees*
Jun'ichi Kazama and Kentaro Torisawa

9:15–9:40 *Semantic Role Labeling via Tree Kernel Joint Inference*
Alessandro Moschitti, Daniele Pighin and Roberto Basili

9:40–10:05 *Can Human Verb Associations Help Identify Salient Features for Semantic Verb Classification?*
Sabine Schulte im Walde

10:05–10:30 *Applying Alternating Structure Optimization to Word Sense Disambiguation*
Rie Kubota Ando

10:30–11:00 coffee break

11:00–11:50 Invited Talk by Walter Daelemans

Session 5: Syntax and Unsupervised Learning

11:50–12:15 *Unsupervised Parsing with U-DOP*
Rens Bod

12:15–12:40 *A Lattice-Based Framework for Enhancing Statistical Parsers with Information from Unlabeled Corpora*
Michaela Atterer and Hinrich Schütze

12:40–14:00 lunch

13:30–14:00 SIGNLL business meeting

Session 6: Thematic Segmentation and Discourse Analysis

14:00–14:25 *Word Distributions for Thematic Segmentation in a Support Vector Machine Approach*
Maria Georgescul, Alexander Clark and Susan Armstrong

14:25–14:50 *Which Side are You on? Identifying Perspectives at the Document and Sentence Levels*
Wei-Hao Lin, Theresa Wilson, Janyce Wiebe and Alexander Hauptmann

Friday, June 9, 2006 (continued)

Session 7: Grammatical Inference

14:50–15:15 *Unsupervised Grammar Induction by Distribution and Attachment*
David J. Brooks

15:15–15:40 *Learning Auxiliary Fronting with Grammatical Inference*
Alexander Clark and Rémi Eyraud

15:40–16:00 coffee break

Session 8: Information Extraction and Named Entity Extraction

16:00–16:25 *Using Gazetteers in Discriminative Information Extraction*
Andrew Smith and Miles Osborne

16:25–16:50 *A Context Pattern Induction Method for Named Entity Extraction*
Partha Pratim Talukdar, Thorsten Brants, Mark Liberman and Fernando Pereira

16:50–17:00 Best Paper Award

17:00 Closing