

CoNLL-2009

**Proceedings of the
Thirteenth Conference on Computational
Natural Language Learning (CoNLL)**

Conference Chairs:
Suzanne Stevenson and Xavier Carreras

June 4–5, 2009
Boulder, Colorado

Production and Manufacturing by
Omnipress Inc.
2600 Anderson Street
Madison, WI 53707
USA



©2009 The Association for Computational Linguistics

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 978-1-932432-29-9

Introduction

The 2009 Conference on Computational Natural Language Learning is the thirteenth in the series of annual meetings organized by SIGNLL, the ACL special interest group on natural language learning. CoNLL-2009 will be held in Boulder, CO, 4–5 June 2009, in conjunction with NAACL HLT.

For our special focus this year in the main session of CoNLL, we invited papers on unsupervised, minimally supervised and semi-supervised methods in natural language learning, as well as on incremental learning methods. As with earlier CoNLLs, we encouraged papers that addressed these issues from the perspective both of human language acquisition and of NLP systems.

We received 70 submissions to the main session on these and other relevant topics, of which 11 were withdrawn. Of the remaining 59 papers, 15 were selected to appear in the conference program as oral presentations, and 10 were chosen as posters. All accepted papers appear here in the proceedings.

Our invited speakers reflect the state-of-the-art in human and machine learning of natural language, and we are grateful to Michael Frank and Andrew McCallum for agreeing to speak on their exciting new work in these areas.

As in previous years, CoNLL-2009 has a shared task, Syntactic and Semantic Dependencies in Multiple Languages. This is an extension of the CoNLL-2008 shared task to multiple languages (English plus Catalan, Chinese, Czech, German, Japanese and Spanish). Among the new features are compatible evaluation for several languages and their comparison, and learning curves for languages with large datasets. We expect that this major comparative exercise will lead to very enlightening results and discussion that will serve to move the field forward. The Shared Task papers are collected into an accompanying volume of CoNLL-2009. We thank Jan Hajic and the rest of the organizers for their great effort in running the Shared Task.

We would like to thank the members of the SIGNLL steering committee for useful discussion, especially Lluís Màrquez and Joakim Nivre, who helped us greatly with advice around the conference organization, and Erik Tjong Kim Sang, who acted as the information officer. We also appreciate the help we received from NAACL HLT organizers, including Martha Palmer, Mark Hasegawa-Johnson, Nizar Habash, Christy Doran, Eric Ringger, and Priscilla Rasmussen.

Finally, many thanks to Google for sponsoring the best paper award at CoNLL-2009.

We hope you find CoNLL-2009 a fruitful venue for discussion and interaction on the exciting topics covered by our program.

Suzanne Stevenson and Xavier Carreras

2009 Conference Chairs

Conference Chairs:

Suzanne Stevenson, University of Toronto
Xavier Carreras, MIT

Program Committee:

Steve Abney, University of Michigan
Afra Alishahi, Saarland University
Galen Andrew, Microsoft Research
Giuseppe Attardi, University of Pisa
Kirk Baker, Collexis Inc
Timothy Baldwin, University of Melbourne
Roberto Basili, University of Roma, Tor Vergata
Phil Blunsom, University of Edinburgh
S.R.K. Branavan, MIT
Chris Brew, Ohio State University
Sabine Buchholz, Toshiba Research Europe Ltd.
Paula Burity, Cambridge University
Nicola Cancedda, Xerox Research Center Europe
Sander Canisius, The Netherlands Cancer Institute
Claire Cardie, Cornell University
Ming-Wei Chang, University of Illinois at Urbana-Champaign
Ciprian Chelba, Google
Colin Cherry, Microsoft Research
Massimiliano Ciaramita, Google Research Zurich
Alexander Clark, Royal Holloway, University of London
Stephen Clark, University of Cambridge
James Clarke, University of Illinois at Urbana-Champaign
Michael Collins, MIT
Paul Cook, University of Toronto
James Cussens, University of York
Walter Daelemans, University of Antwerp
Hal Daume III, University of Utah
Jacob Eisenstein, University of Illinois at Urbana-Champaign
Katrín Erk, University of Texas at Austin
Anna Feldman, Montclair State University
Jenny Finkel, Stanford University
Radu Florian, IBM Research
Dayne Freitag, SRI International
Pascale Fung, Hong Kong University of Science and Technology
Michel Galley, Stanford University
Daniel Gildea, University of Rochester

Roxana Girju, University of Illinois at Urbana-Champaign
John Hale, Cornell University
James Henderson, University of Geneva
Julia Hockenmaier, University of Illinois at Urbana-Champaign
Liang Huang, Google Research
Richard Johansson, University of Trento
Rie Johnson (formerly, Ando), RJ Research Consulting
Rohit Kate, University of Texas at Austin
Philipp Koehn, University of Edinburgh
Rob Koeling, Sussex University
Anna Korhonen, Cambridge University
Shalom Lappin, King's College, London
Dekang Lin, Google Research
Xiaofei Lu, Pennsylvania State University
Rob Malouf, San Diego State University
Yuji Matsumoto, Nara Institute of Science and Technology
Diana McCarthy, University of Sussex
Ryan McDonald, Google Research
Paola Merlo, University of Geneva
Rada Mihalcea, University of North Texas
Saif Mohammad, University of Maryland
Alessandro Moschitti, University of Trento
Gabriele Musillo, University of Geneva
John Nerbonne, University of Groningen
Hwee Tou Ng, National University of Singapore
Vincent Ng, University of Texas at Dallas
Grace Ngai, The Hong Kong Polytechnic University
Joakim Nivre, Uppsala University and Växjö University
Franz Och, Google Research
Miles Osborne, University of Edinburgh
Chris Parisien, University of Toronto
Ted Pedersen, University of Minnesota, Duluth
Slav Petrov, UC Berkeley
David Powers, Flinders University
Chris Quirk, Microsoft Research
Ari Rappoport, Hebrew University
Lev Ratinov, University of Illinois at Urbana-Champaign
Sebastian Riedel, University of Tokyo
Brian Roark, Oregon Health & Science University
Dan Roth, University of Illinois at Urbana-Champaign
William Sakas, City University of New York
Sabine Schulte im Walde, University of Stuttgart
Libin Shen, BBN Technologies
David Smith, University of Massachusetts Amherst
Noah Smith, Carnegie Mellon University
Benjamin Snyder, MIT

Caroline Sporleder, Saarland University
Richard Sproat, Oregon Health & Science University
Mihai Surdeanu, Stanford University
Ivan Titov, University of Illinois at Urbana-Champaign
Erik Tjong Kim Sang, University of Groningen
Vivian Tsang, Bloorview Research Institute
Antal van den Bosch, Tilburg University
Aline Villavicencio, Federal University of Rio Grande do Sul
Charles Yang, University of Pennsylvania
Scott Wen-tau Yih, Microsoft Research
Deniz Yuret, Koç University
Luke Zettlemoyer, MIT

Additional Reviewers:

Yee Seng Chan, National University of Singapore
Michael Connor, University of Illinois at Urbana-Champaign
Dmitry Davidov, Hebrew University
Michael Demko, NextJ Systems
Ying Li, Hong Kong University of Science and Technology
Andrew MacKinlay, University of Melbourne
Preslav Nakov, National University of Singapore
Jeremy Nicholson, University of Melbourne
Jing Peng, Montclair State University
Jelena Prokic, University of Groningen
Kevin Small, University of Illinois at Urbana-Champaign
Oren Tsur, Hebrew University

Invited Speakers:

Michael C. Frank, MIT
Andrew McCallum, University of Massachusetts Amherst

Table of Contents

<i>Joint Inference for Natural Language Processing</i> Andrew McCallum	1
<i>Modeling Word Learning As Communicative Inference</i> Michael C. Frank	2
<i>Sample Selection for Statistical Parsers: Cognitively Driven Algorithms and Evaluation Measures</i> Roi Reichart and Ari Rappoport	3
<i>Data-Driven Dependency Parsing of New Languages Using Incomplete and Noisy Training Data</i> Kathrin Spreyer and Jonas Kuhn	12
<i>A Metalearning Approach to Processing the Scope of Negation</i> Roser Morante and Walter Daelemans	21
<i>Efficient Linearization of Tree Kernel Functions</i> Daniele Pighin and Alessandro Moschitti	30
<i>A Method for Stopping Active Learning Based on Stabilizing Predictions and the Need for User-Adjustable Stopping</i> Michael Bloodgood and Vijay Shanker	39
<i>Superior and Efficient Fully Unsupervised Pattern-based Concept Acquisition Using an Unsupervised Parser</i> Dmitry Davidov, Roi Reichart and Ari Rappoport	48
<i>Representing words as regions in vector space</i> Katrin Erk	57
<i>Interactive Feature Space Construction using Semantic Information</i> Dan Roth and Kevin Small	66
<i>Mining the Web for Reciprocal Relationships</i> Michael Paul, Roxana Girju and Chen Li	75
<i>Minimally Supervised Model of Early Language Acquisition</i> Michael Connor, Yael Gertner, Cynthia Fisher and Dan Roth	84
<i>Learning Where to Look: Modeling Eye Movements in Reading</i> Mattias Nilsson and Joakim Nivre	93
<i>Monte Carlo inference and maximization for phrase-based translation</i> Abhishek Arun, Chris Dyer, Barry Haddow, Phil Blunsom, Adam Lopez and Philipp Koehn .	102
<i>Investigating Automatic Alignment Methods for Slide Generation from Academic Papers</i> Brandon Beamer and Roxana Girju	111

<i>Glen, Glenda or Glendale: Unsupervised and Semi-supervised Learning of English Noun Gender</i> Shane Bergsma, Dekang Lin and Randy Goebel	120
<i>Improving Translation Lexicon Induction from Monolingual Corpora via Dependency Contexts and Part-of-Speech Equivalences</i> Nikesh Garera, Chris Callison-Burch and David Yarowsky	129
<i>An Intrinsic Stopping Criterion for Committee-Based Active Learning</i> Fredrik Olsson and Katrin Tomanek	138
<i>Design Challenges and Misconceptions in Named Entity Recognition</i> Lev Ratinov and Dan Roth	147
<i>Automatic Selection of High Quality Parses Created By a Fully Unsupervised Parser</i> Roi Reichart and Ari Rappoport	156
<i>The NVI Clustering Evaluation Measure</i> Roi Reichart and Ari Rappoport	165
<i>Lexical Patterns or Dependency Patterns: Which Is Better for Hypernym Extraction?</i> Erik Tjong Kim Sang and Katja Hofmann	174
<i>Improving Text Classification by a Sense Spectrum Approach to Term Expansion</i> Peter Wittek, Sándor Darányi and Chew Lim Tan	183
<i>A simple feature-copying approach for long-distance dependencies</i> Marc Vilain, Jonathan Huggins and Ben Wellner	192
<i>Fine-Grained Classification of Named Entities Exploiting Latent Semantic Kernels</i> Claudio Giuliano	201
<i>Using Encyclopedic Knowledge for Automatic Topic Identification</i> Kino Coursey, Rada Mihalcea and William Moen	210
<i>New Features for FrameNet - WordNet Mapping</i> Sara Tonelli and Daniele Pighin	219

Conference Program

Joint Inference for Natural Language Processing

Andrew McCallum

Modeling Word Learning As Communicative Inference

Michael C. Frank

Thursday, June 4, 2009

9:00–9:15 Opening Remarks

Session 1: Parsing and Tagging

9:15–9:40 *Sample Selection for Statistical Parsers: Cognitively Driven Algorithms and Evaluation Measures*

Roi Reichart and Ari Rappoport

9:40–10:05 *Data-Driven Dependency Parsing of New Languages Using Incomplete and Noisy Training Data*

Kathrin Spreyer and Jonas Kuhn

10:05–10:30 *A Metalearning Approach to Processing the Scope of Negation*

Roser Morante and Walter Daelemans

10:30–11:00 **Coffee Break**

Session 2: Learning Methods

11:00–10:25 *Efficient Linearization of Tree Kernel Functions*

Daniele Pighin and Alessandro Moschitti

11:25–11:40 *A Method for Stopping Active Learning Based on Stabilizing Predictions and the Need for User-Adjustable Stopping*

Michael Bloodgood and Vijay Shanker

11:40–12:40 **Invited Talk:** *Joint Inference for Natural Language Processing*

Andrew McCallum

12:40–14:00 **Lunch**

Thursday, June 4, 2009 (continued)

Shared Task: Overview and Oral Presentations

- 14:00–14:20 *The CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies in Multiple Languages*
Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue and Yi Zhang
- 14:20–14:30 *An Iterative Approach for Joint Dependency Parsing and Semantic Role Labeling*
Qifeng Dai, Enhong Chen and Liu Shi
- 14:30–14:40 *Joint Memory-Based Learning of Syntactic and Semantic Dependencies in Multiple Languages*
Roser Morante, Vincent Van Asch and Antal van den Bosch
- 14:40–14:50 *Hybrid Multilingual Parsing with HPSG for SRL*
Yi Zhang, Rui Wang and Stephan Oepen
- 14:50–15:00 *A Latent Variable Model of Synchronous Syntactic-Semantic Parsing for Multiple Languages*
Andrea Gesmundo, James Henderson, Paola Merlo and Ivan Titov
- 15:00–15:10 *Multilingual Semantic Role Labeling*
Anders Björkelund, Love Hafdell and Pierre Nugues
- 15:10–15:20 *Multilingual Dependency-based Syntactic and Semantic Parsing*
Wanxiang Che, Zhenghua Li, Yongqiang Li, Yuhang Guo, Bing Qin and Ting Liu
- 15:20–15:30 *Multilingual Dependency Learning: A Huge Feature Engineering Method to Semantic Dependency Parsing*
Hai Zhao, Wenliang Chen, Chunyu Kity and Guodong Zhou
- 15:30–15:40 *Multilingual Dependency Learning: Exploiting Rich Features for Tagging Syntactic and Semantic Dependencies*
Hai Zhao, Wenliang Chen, Jun'ichi Kazama, Kiyotaka Uchimoto and Kentaro Torisawa
- 15:40–16:10 **Coffee Break**

Thursday, June 4, 2009 (continued)

Shared Task: Poster Session (16:10–18:00)

Multilingual Semantic Role Labeling

Anders Björkelund, Love Hafdell and Pierre Nugues

Efficient Parsing of Syntactic and Semantic Dependency Structures

Berud Bohnet

Multilingual Dependency-based Syntactic and Semantic Parsing

Wanxiang Che, Zhenghua Li, Yongqiang Li, Yuhang Guo, Bing Qin and Ting Liu

An Iterative Approach for Joint Dependency Parsing and Semantic Role Labeling

Qifeng Dai, Enhong Chen and Liu Shi

A Latent Variable Model of Synchronous Syntactic-Semantic Parsing for Multiple Languages

Andrea Gesmundo, James Henderson, Paola Merlo and Ivan Titov

Exploring Multilingual Semantic Role Labeling

Baoli Li, Martin Emms, Saturnino Luz and Carl Vogel

A Second-Order Joint Eisner Model for Syntactic and Semantic Dependency Parsing

Xavier Lluís, Stefan Bott and Lluís Màrquez

Multilingual Semantic Role Labelling with Markov Logic

Ivan Meza-Ruiz and Sebastian Riedel

Joint Memory-Based Learning of Syntactic and Semantic Dependencies in Multiple Languages

Roser Morante, Vincent Van Asch and Antal van den Bosch

The Crotal SRL System : a Generic Tool Based on Tree-structured CRF

Erwan Moreau and Isabelle Tellier

Parsing Syntactic and Semantic Dependencies for Multiple Languages with A Pipeline Approach

Han Ren, Donghong Ji, Jing Wan and Mingyao Zhang

Multilingual Semantic Parsing with a Pipeline of Linear Classifiers

Oscar Täckström

Thursday, June 4, 2009 (continued)

A Joint Syntactic and Semantic Dependency Parsing System based on Maximum Entropy Models

Buzhou Tang, Lu Li, Xinxin Li, Xuan Wang and Xiaolong Wang

Multilingual Syntactic-Semantic Dependency Parsing with Three-Stage Approximate Max-Margin Linear Models

Yotaro Watanabe, Masayuki Asahara and Yuji Matsumoto

A Simple Generative Pipeline Approach to Dependency Parsing and Semantic Role Labeling

Daniel Zeman

Hybrid Multilingual Parsing with HPSG for SRL

Yi Zhang, Rui Wang and Stephan Oepen

Multilingual Dependency Learning: Exploiting Rich Features for Tagging Syntactic and Semantic Dependencies

Hai Zhao, Wenliang Chen, Jun'ichi Kazama, Kiyotaka Uchimoto and Kentaro Torisawa

Multilingual Dependency Learning: A Huge Feature Engineering Method to Semantic Dependency Parsing

Hai Zhao, Wenliang Chen, Chunyu Kity and Guodong Zhou

Friday, June 5, 2009

Session 3: Learning and Semantics

9:00–9:25 *Superior and Efficient Fully Unsupervised Pattern-based Concept Acquisition Using an Unsupervised Parser*

Dmitry Davidov, Roi Reichart and Ari Rappoport

9:25–9:50 *Representing words as regions in vector space*

Katrin Erk

9:50–10:15 *Interactive Feature Space Construction using Semantic Information*

Dan Roth and Kevin Small

10:15–10:40 *Mining the Web for Reciprocal Relationships*

Michael Paul, Roxana Girju and Chen Li

10:40–11:00 **Coffee Break**

Friday, June 5, 2009 (continued)

Session 4: Human Behavior

- 11:00–11:25 *Minimally Supervised Model of Early Language Acquisition*
Michael Connor, Yael Gertner, Cynthia Fisher and Dan Roth
- 11:25–11:40 *Learning Where to Look: Modeling Eye Movements in Reading*
Mattias Nilsson and Joakim Nivre
- 11:40–12:40 **Invited talk:** *Modeling Word Learning As Communicative Inference*
Michael C. Frank
- 12:40–14:00 **Lunch**
- 13:45–14:15 **SIGNLL Meeting**

Poster Session (14:00–15:30)

- Monte Carlo inference and maximization for phrase-based translation*
Abhishek Arun, Chris Dyer, Barry Haddow, Phil Blunsom, Adam Lopez and Philipp Koehn
- Investigating Automatic Alignment Methods for Slide Generation from Academic Papers*
Brandon Beamer and Roxana Girju
- Glen, Glenda or Glendale: Unsupervised and Semi-supervised Learning of English Noun Gender*
Shane Bergsma, Dekang Lin and Randy Goebel
- Improving Translation Lexicon Induction from Monolingual Corpora via Dependency Contexts and Part-of-Speech Equivalences*
Nikesh Garera, Chris Callison-Burch and David Yarowsky
- An Intrinsic Stopping Criterion for Committee-Based Active Learning*
Fredrik Olsson and Katrin Tomanek
- Design Challenges and Misconceptions in Named Entity Recognition*
Lev Ratinov and Dan Roth
- Automatic Selection of High Quality Parses Created By a Fully Unsupervised Parser*
Roi Reichart and Ari Rappoport

Friday, June 5, 2009 (continued)

The NVI Clustering Evaluation Measure

Roi Reichart and Ari Rappoport

Lexical Patterns or Dependency Patterns: Which Is Better for Hypernym Extraction?

Erik Tjong Kim Sang and Katja Hofmann

Improving Text Classification by a Sense Spectrum Approach to Term Expansion

Peter Wittek, Sándor Darányi and Chew Lim Tan

15:30–16:00 **Coffee Break**

Session 5: Semantic Annotation and Extraction

16:00–16:25 *A simple feature-copying approach for long-distance dependencies*

Marc Vilain, Jonathan Huggins and Ben Wellner

16:25–16:50 *Fine-Grained Classification of Named Entities Exploiting Latent Semantic Kernels*

Claudio Giuliano

16:50–17:15 *Using Encyclopedic Knowledge for Automatic Topic Identification*

Kino Coursey, Rada Mihalcea and William Moen

17:15–17:40 *New Features for FrameNet - WordNet Mapping*

Sara Tonelli and Daniele Pighin

17:40–18:00 Closing remarks and best paper award