

**Proceedings of the
Fourth Conference on
Computational Natural Language Learning
and of the
Second Learning Language in Logic Workshop**

Held in cooperation with ICGI-2000

**13–14 September 2000
Lisbon, Portugal**

**Proceedings of the
Fourth Conference on
Computational Natural Language Learning
and of the
Second Learning Language in Logic Workshop**

Held in cooperation with ICGI-2000

**13–14 September 2000
Lisbon, Portugal**

Order additional copies from:

Association for Computational Linguistics

75 Paterson Street

New Brunswick, NJ 08901 USA

+1-732-342-9100 phone

+1-732-342-9339 fax

acl@aclweb.org

Preface

The joint *Second Learning Language in Logic (LLL-2000) Workshop* and *Fourth Conference on Computational Natural Language Learning (CoNLL-2000)* took place September 13-14, 2000, at the Instituto Superior Técnico in Lisbon, Portugal and have been co-organized with the 5th International Colloquium on Grammatical Inference (ICGI-2000).

This volume contains the papers presented during this joint event. More information is available on-line from <http://www.lri.fr/~cn/LLL-2000/> and <http://lcg-www.uia.ac.be/conll2000/>.

We would like to thank all the authors for submitting their papers and thus making these proceedings possible. We address special thanks to the members of the program committees for their great work which contributed to the high quality of these proceedings. We wish to extend our gratitude to the invited speakers for presenting us with their views on innovative results in Natural Language Processing and Machine Learning.

We are also grateful to the Local Chair Arlindo Oliveira, the members of the Organizing Committee, Ana Fred and Ana T. Freitas, and all other individuals who helped in the organization of this event.

Finally, we would like to thank the sponsors of LLL-2000 and CoNLL-2000 for their generous financial and moral support: the Network of Excellence in Inductive Logic Programming (ILPNet2), the Network of Excellence in Machine Learning (MLNet3), the Computational Linguistics in Flanders research community (CLIF), and SIGNLL (ACL's SIG on Natural Language Learning).

Claire Cardie
Walter Daelemans
Claire Nédellec
Erik Tjong Kim Sang

SPONSORS:

CLIF (Computational Linguistics in Flanders)
ILPNet2 (Network of Excellence in Inductive Logic Programming)
MLNet3 (Network of Excellence in Machine Learning)
SIGNLL (ACL's SIG for Natural Language Learning)

INVITED SPEAKERS:

Jörg-Uwe Kietz
Dan Roth

ORGANIZERS:

Claire Cardie (CoNLL)
Walter Daelemans (CoNLL)
Claire Nédellec (LLL)
Erik Tjong Kim Sang (CoNLL)

LOCAL ARRANGEMENTS CHAIR:

Arlindo Oliveira

CoNLL PROGRAM COMMITTEE:

Thorsten Brants	(Universität des Saarlandes)
James Cussens	(University of York)
Raymond Mooney	(University of Texas at Austin)
John Nerbonne	(University of Groningen)
Miles Osborne	(University of Edinburgh)
David Powers	(Flinders University)
Ronan Reilly	(University College Dublin)
Antal van den Bosch	(Tilburg University)

LLL PROGRAM COMMITTEE:

Pieter Adriaans	(Syllogic and University of Amsterdam, the Netherlands)
Roberto Basili	(University of Roma, Italy)
Gilles Bisson	(INRIA, Grenoble, France)
Henrik Boström	(University of Stockholm, Sweden)
Gosse Bouma	(University of Groningen, the Netherlands)
James Cussens	(University of York, United Kingdom)
Tomaz Erjavec	(Institute Jozef Stefan, Slovenia)
Daniel Kayser	(LIPN, Universit Paris-Nord, France)
Suresh Manandhar	(University of York, United Kingdom)
Guenter Neumann	(DFKI, Saarbrcken, Germany)
Steve Pulman	(University of Cambridge, United Kingdom)
Christer Samuelsson	(Xerox Research Center Europe, Grenoble, France)
Stefan Wrobel	(University of Magdeburg, Germany)

FURTHER INFORMATION:

CoNLL and SIGNLL

Walter Daelemans
CNTS Language Technology Group
University of Antwerp (UIA)
Universiteitsplein 1 (building A)
B-2610 Antwerpen, Belgium
e-mail: daelem@uia.ua.ac.be

LLL

Claire Nédellec
Laboratoire de Recherche en informatique (LRI)
UMR 8623 CNRS
Bat 490, Université Paris-Sud
F-91405 Orsay cedex, France
e-mail: cn@lri.fr

Table of Contents

CoNLL-2000 Invited Paper

<i>Learning in Natural Language: Theory and Algorithmic Approaches</i> Dan Roth	1
--	---

CoNLL-2000 Papers

<i>Corpus-Based Grammar Specialization</i> Nicola Cancedda and Christer Samuelsson	7
<i>Pronunciation by Analogy in Normal and Impaired Readers</i> R.I. Damper and Y. Marchand	13
<i>The Role of Algorithm Bias vs Information Source in Learning Algorithms for Morphosyntactic Disambiguation</i> Guy De Pauw and Walter Daelemans	19
<i>Increasing our Ignorance of Language: Identifying Language Structure in an Unknown ‘Signal’</i> John Elliott, Eric Atwell and Bill Whyte	25
<i>A Comparison between Supervised Learning Algorithms for Word Sense Disambiguation</i> Gerard Escudero, Lluís Màrquez and German Rigau	31
<i>Incorporating Position Information into a Maximum Entropy/Minimum Divergence Translation Model</i> George Foster	37
<i>Memory-Based Learning for Article Generation</i> Guido Minnen, Francis Bond and Ann Copestake	43
<i>Overfitting Avoidance for Stochastic Modeling of Attribute-Value Grammars</i> Tony Mullen and Miles Osborne	49
<i>Learning Distributed Linguistic Classes</i> Stephan Raaijmakers	55
<i>Modeling the Effect of Cross-Language Ambiguity on Human Syntax Acquisition</i> William Gregory Sakas	61
<i>Knowledge-Free Induction of Morphology Using Latent Semantic Analysis</i> Patrick Schone and Daniel Jurafsky	67
<i>Using Induced Rules as Complex Features in Memory-Based Language Learning</i> Antal van den Bosch	73

CoNLL-2000 Short Papers

<i>Using Perfect Sampling in Parameter Estimation of a Whole Sentence Maximum Entropy Language Model</i> F. Amaya and J.M. Benedí	79
<i>Experiments on Unsupervised Learning for Extracting Relevant Fragments from Spoken Dialog Corpus</i> Konstantin Biatov	83
<i>Generating Synthetic Speech Prosody with Lazy Learning in Tree Structures</i> Laurent Blin and Laurent Miclet	87
<i>Inducing Syntactic Categories by Context Distribution Clustering</i> Alexander Clark	91
<i>ALLiS: a Symbolic Learning System for Natural Language Learning</i> Hervé Déjean	95
<i>Combining Text and Heuristics for Cost-Sensitive Spam Filtering</i> José M. Gómez Hidalgo and Enrique Puertas Sanz	99
<i>Genetic Algorithms for Feature Relevance Assignment in Memory-Based Language Processing</i> Anne Kool, Walter Daelemans and Jakub Zavrel	103
<i>Shallow Parsing by Inferencing with Classifiers</i> Vasin Punyakanok and Dan Roth	107
<i>Minimal Commitment and Full Lexical Disambiguation: Balancing Rules and Hidden Markov Models</i> Patrick Ruch, Robert Baud, Pierrette Bouillon and Gilbert Robert	111
<i>Learning IE Rules for a Set of Related Concepts</i> J. Turmo and H. Rodríguez	115
<i>A default First Order Family Weight Determination Procedure for WPDV Models</i> Hans van Halteren	119
<i>A Comparison of PCFG Models</i> Jose Luis Verdú-Mas, Jorge Calera-Rubio and Rafael C. Carrasco	123

CoNLL-2000 Shared Task Papers

<i>Introduction to the CoNLL-2000 Shared Task: Chunking</i> Erik F. Tjong Kim Sang and Sabine Buchholz	127
<i>Learning Syntactic Structures with XML</i> Hervé Déjean	133
<i>A Context Sensitive Maximum Likelihood Approach to Chunking</i> Christer Johansson	136
<i>Chunking with Maximum Entropy Models</i> Rob Koeling	139
<i>Use of Support Vector Learning for Chunk Identification</i> Taku Kudoh and Yuji Matsumoto	142
<i>Shallow Parsing as Part-of-Speech Tagging</i> Miles Osborne	145
<i>Improving Chunking by Means of Lexical-Contextual Information in Statistical Language Models</i> Ferran Pla, Antonio Molina and Natividad Prieto	148
<i>Text Chunking by System Combination</i> Erik F. Tjong Kim Sang	151
<i>Chunking with WPDV Models</i> Hans van Halteren	154
<i>Single-Classifer Memory-Based Phrase Chunking</i> Jorn Veenstra and Antal van den Bosch	157
<i>Phrase Parsing with Rule Sequence Processors: an Application to the Shared CoNLL Task</i> Marc Vilain and David Day	160
<i>Hybrid Text Chunking</i> GuoDong Zhou, Jian Su and TongGuan Tey	163

LLL-2000 Invited Paper

<i>Extracting a Domain-Specific Ontology from a Corporate Intranet</i> Jörg-Uwe Kietz, Raphael Volz and Alexander Maedche	167
--	-----

LLL-2000 Papers

<i>Learning from a Substructural Perspective</i> Pieter Adriaans and Erik de Haas	176
<i>Incorporating Linguistics Constraints into Inductive Logic Programming</i> James Cussens and Stephen Pulman	184
<i>Learning from Parsed Sentences with INTHELEX</i> F. Esposito, S. Ferilli, N. Fanizzi and G. Semeraro	194
<i>Inductive Logic Programming for Corpus-Based Acquisition of Semantic Lexicons</i> Pascale Sébillot, Pierrette Bouillon and Cécile Fabre	199
<i>The Acquisition of Word Order by a Computational Learning System</i> Aline Villavicencio	209
<i>Recognition and Tagging of Compound Verb Groups in Czech</i> Eva Žáčková, Luboš Popelínský and Miloš Nepil	219

**Fourth Conference on
Computational Natural Language Learning
(CoNLL-2000)**

Preface

CoNLL-2000 is the fourth in a series of meetings organized by SIGNLL, the ACL's SIG on Natural Language Learning. Previous meetings were organized in Madrid, Sydney, and Bergen, co-located with different, but always computational linguistics-oriented, events. We are pleased that this time we could combine efforts with the grammar induction and inductive logic programming for language processing communities.

It is the explicit wish of the SIGNLL board to have the CoNLL meeting address all aspects of computational natural language learning, including issues that are not regularly discussed at computational linguistics meetings, such as computational models of human language acquisition, computational models of the origins and evolution of language, biologically-inspired learning methods, etc.

We are thrilled by the quality and quantity of the submissions, which allowed us to set up an intense but rewarding program with one invited talk, 12 long talks, and joint paper sessions with LLL-2000 and ICGI-2000. On top of that, we introduced two innovations: there are 12 *bullet presentations*, short talks accompanied by a poster presentation, and a *shared task session* in which 11 authors report on how their machine learning method performed on our shared task — the identification of syntactic constituents in text (chunking). In this part of the proceedings, you will find 37 papers providing a useful record of all presentations.

You can find out more about SIGNLL and its activities at <http://www.aclweb.org/signll/>.

Claire Cardie
Walter Daelemans
Erik Tjong Kim Sang

Ithaca and Antwerp, 2000

Second Learning Language in Logic Workshop
(LLL-2000)

Preface

LLL-2000 is the follow-up of the first LLL workshop held in 1999 in Bled (Slovenia), and co-located with the International Conference on Machine Learning and the International Conference on Logic Programming. This year LLL was integrated with the Fourth Conference on Language Learning (CoNLL) and the Fifth International Colloquium on Grammatical Inference (ICGI) with which LLL shares strong common scientific interests in language learning. The registration to ICGI, CoNLL and LLL was a joint registration so that registrants could freely move between the three events.

As in the first edition, LLL has attracted pluridisciplinary submissions from the three research fields — Natural Language Processing (NLP), Machine Learning and Computational Logic, demonstrating the growing interest in NLP methods based on ILP or non-classic logics, and hybrid methods. Relational learning more and more appears as complementary to data analysis in many NLP domains. Relational learning and logic-based learning prove here again their capacity to learn complex structured linguistic resources and knowledge such as ontology and grammar from corpora and explicit background knowledge.

The scientific program of LLL-2000 consisted of one invited talk by Jörg-Uwe Kietz on the acquisition of ontology and seven paper presentations. Six of them are reported here and the paper by Christophe Costa Florencio, accepted for presentation by both LLL and ICGI, has been published in the ICGI proceedings. The joint sessions with ICGI and CoNLL included one invited talk by Dan Roth and paper and poster presentations.

Claire Nédellec

Orsay, 2000

Author Index

Adriaans, Pieter	176	Matsumoto, Yuji	142
Amaya, F.	79	Miclet, Laurent	87
Atwell, Eric	25	Minnen, Guido	43
Baud, Robert	111	Molina, Antonio	148
Benedí, J.M.	79	Mullen, Tony	49
Biatov, Konstantin	83	Nepil, Miloš	219
Blin, Laurent	87	Osborne, Miles	49, 145
Bond, Francis	43	Pla, Ferran	148
Bouillon, Pierrette	111, 199	Popelínský, Luboš	219
Buchholz, Sabine	127	Prieto, Natividad	148
Calera-Rubio, Jorge	123	Puertas Sanz, Enrique	99
Cancedda, Nicola	7	Pulman, Stephen	184
Carrasco, Rafael C.	123	Punyakanok, Vasin	107
Clark, Alexander	91	Raaijmakers, Stephan	55
Copestake, Ann	43	Rigau, German	31
Cussens, James	184	Robert, Gilbert	111
Daelemans, Walter	19, 103	Rodríguez, H.	115
Damper, R.I.	13	Roth, Dan	1, 107
Day, David	160	Ruch, Patrick	111
De Haas, Erik	176	Sakas, William Gregory	61
De Pauw, Guy	19	Samuelsson, Christer	7
Déjean, Hervé	95, 133	Schone, Patrick	67
Elliott, John	25	Sébillot, Pascale	199
Escudero, Gerard	31	Semeraro, G.	194
Esposito, F.	194	Su, Jian	163
Fabre, Cécile	199	Tey, TongGuan	163
Fanizzi, N.	194	Tjong Kim Sang, Erik F.	127, 151
Ferilli, S.	194	Turmo, J.	115
Foster, George	37	Van Halteren, Hans	119, 154
Gómez Hidalgo, José M.	99	Van den Bosch, Antal	73, 157
Johansson, Christer	136	Veenstra, Jorn	157
Jurafsky, Daniel	67	Verdú-Mas, Jose Luis	123
Kietz, Jörg-Uwe	167	Vilain, Marc	160
Koeling, Rob	139	Villavicencio, Aline	209
Kool, Anne	103	Volz, Raphael	167
Kudoh, Taku	142	Whyte, Bill	25
Maedche, Alexander	167	Žáčková, Eva	219
Marchand, Y.	13	Zavrel, Jakub	103
Màrquez, Lluís	31	Zhou, GuoDong	163