

New Methods in Language Processing
and
Computational Natural Language Learning

NeMLaP3/CoNLL98

Editor: David M. W. Powers

Organized by: Macquarie University
Sydney University
Flinders University

Held at Macquarie University, January 11-17, 1998

**Proceedings of the
Joint Conference on
New Methods in Language Processing
and
Computational Natural Language Learning**

NeMLaP3/CoNLL98

Editor: David M. W. Powers

**Organized by: Macquarie University
Sydney University
Flinders University**

Held at Macquarie University, January 11-17, 1998

**CoNLL is the annual conference of the ACL Special
Interest Group on Natural Language Learning (SIGNLL)**

Order additional copies from

ACL
P.O. Box 6090
Somerset, NJ, 08875 USA
+1-908-873-3898
acl@bellcore.com

ISBN: 0-7258-0634-6

PREFACE

Natural Language Processing is increasingly moving away from the idea of handcoding grammars and systems, and the NeMLaP series of conferences has encouraged the exploration of new approaches to all aspects of the field. The techniques NeMLaP portrays tend to move away from the symbolic toward the connectionist and the statistical, and often have a learning flavour about them.

The ACL Special Interest Group on Natural Language Learning last year initiated the CoNLL series which focus specifically on Learning, so it is highly appropriate that the two conferences, originally scheduled for two different cities in Australia, should be held jointly to maximize the interaction between the two groups.

The timing of the conference is a bit unusual for those used to conferences in the northern summer, and a bit difficult for those who only have short breaks. But attending conferences in our winter semester is something Australians are used to, and the European members of the organizing committee were keen on the idea of escaping their winter for a while. The distance is also a big factor, and combining a joint conference with a series of other events has enabled us to maintain both a good number of papers and a high level of refereeing - at the main conferences we have 32 papers being presented out of 40 odd submissions.

This volume is a bit unusual in one respect, in that we are including the papers from the associated workshops: the Workshop on Human Computer Conversation which is being held in association with the Loebner Prize in Sydney before the joint conference, and the Workshop on Paradigms and Grounding in Language Learning which is being held in Adelaide following the conferences. We felt that registrants and purchasers of the proceedings would appreciate having the additional papers, given that the cost of adding them was less than the cost of producing separate proceedings for the workshops.

We trust that you will enjoy your time in Australia and will find the events worthwhile. We look forward to seeing you.

David Powers
Flinders University

ORGANIZERS

David Powers, Flinders Uni, Australia
Sandra Williams, Macquarie Uni/Microsoft Research Inst, Australia
Harold Somers, UMIST, UK
Chris Manning, Sydney Uni, Australia
Kemal Oflazer, Bilkent Uni, Turkey
Robert Dale, Macquarie Uni/Microsoft Research Inst, Australia

PROGRAM COMMITTEE

Michael Brent, Johns Hopkins Uni, USA
Claire Cardie, Cornell Uni, USA
Walter Daelemans, Tilburg Uni, NL
Robert Dale, Macquarie Uni
Mark Ellison, Edinburgh Uni, UK
Dominique Estival, Melbourne Uni, Australia
Chris Manning, Sydney Uni, Australia
Kemal Oflazer, Bilkent Uni, Turkey
Raymond Mooney, University of Texas at Austin, USA
Kim Plunkett, Oxford University, UK
David Powers, Flinders Uni, Australia
Christer Samuelsson, Bell Labs, USA
Jeffrey M. Siskind, NEC Research Institute, Princeton, USA
Harold Somers, UMIST, UK
Junichi Tsujii, Tokyo Uni, Japan
Antal van den Bosch, Tilburg University, The Netherlands
Atro Voutilainen, Helsinki Uni, Finland
Peter Wallis, DSTO, Australia
Dekai Wu, HKUST, Hong Kong
David Yarowsky, Johns Hopkins Uni, USA
Ingrid Zukerman, Monash Uni, Australia

NeMLaP3/CoNLL98

Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning

January 11-17, 1998
Macquarie University
SYDNEY, AUSTRALIA

TABLE OF CONTENTS

Main Conference

<i>Abstraction is Harmful in Language Learning</i> Walter Daelemans	1
<i>Natural Language Learning by Recurrent Neural Networks: A Comparison with Probabilistic Approaches</i> Michael Towsey, Joachim Diederich, Ingo Schellhammer, Stephan Chalup and Claudia Brugman	3
<i>Learning a Lexicalised Grammar for German</i> Sandra Kübler	11
<i>A Lexically-Intensive Algorithm for Domain-Specific Knowledge Acquisition</i> René Schneider	19
<i>Look-Back and Look-Ahead in the Conversion of Hidden Markov Models into Finite State Transducers</i> André Kempe	29
<i>The Effect of Alternative Tree Representations on Tree Bank Grammars</i> Mark Johnson	39
<i>Automation of Treebank Annotation</i> Thorsten Brants and Wojciech Skut	49
<i>Implementing a Sense Tagger in a General Architecture for Text Engineering</i> Hamish Cunningham, Mark Stevenson and Yorick Wilks	59
<i>Knowledge Extraction and Recurrent Neural Networks: An Analysis of an Elman Network Trained on a Natural Language Learning Task</i> Ingo Schellhammer, Joachim Diederich, Michael Towsey and Claudia Brugman	73
<i>Finding Structure via Compression</i> Jason L. Hutchens and Michael D. Alder	79
<i>Linguistic Theory in Statistical Language Learning</i> Christer Samuelsson	83

<i>A Bayesian Approach to Automating Argumentation</i> Richard McConachy, Kevin B. Korb and Ingrid Zukerman	91
<i>Automatically Generating Hypertext in Newspaper Articles by Computing Semantic Relatedness</i> Stephen J. Green	101
<i>Choosing A Distance Metric for Automatic Word Categorization</i> Emin Erkan Korkmaz and Göktürk Üçoluk	111
<i>Sense Variation and Lexical Semantics Generative Operations</i> Patrick Saint-Dizier	121
<i>An Attempt to Use Weighted Cusums to Identify Sublanguages</i> Harold Somers	131
<i>Cross-Entropy and Linguistic Typology</i> Patrick Juola	141
<i>Applications and Explanations of Zipf's Law</i> David M.W. Powers	151
<i>Proper Name Classification in an Information Extraction Toolset</i> Peter Wallis, Edmund Yuen and Greg Chase	161
<i>Evolution and Evaluation of Document Retrieval Queries</i> Robert Steele and David Powers	163
<i>Generation of Simple Turkish Sentences with Systemic-Functional Grammar</i> Ilyas Cicekli and Turgay Korkmaz	165
<i>Extracting Phoneme Pronunciation Information from Corpora</i> Ian Thomas, Ingrid Zukerman and Bhavani Raskutti	175
<i>Modularity in Inductively-Learned Word Pronunciation Systems</i> Antal van den Bosch, Ton Weijters and Walter Daelemans	185
<i>Do Not Forget: Full Memory in Memory-Based Learning of Word Pronunciation</i> Antal van den Bosch and Walter Daelemans	195
<i>Natural Language Concept Analysis</i> V. Kamphuis and J.J. Sarbo	205
<i>The Present Use of Statistics in the Evaluation of NLP Parsers</i> Jim Entwisle and David Powers	215
<i>A Method of Incorporating Bigram Constraints into an LR Table and its Effectiveness in Natural Language Processing</i> Hiroki Imai and Hozumi Tanaka	225
<i>Selective Attention and the Acquisition of Spatial Semantics</i> James M. Hogan, Joachim Diederich and Gerard D. Finn	235
<i>Towards Language Acquisition by an Attention-Sharing Robot</i> Hideki Kozima and Akira Ito	245
<i>A Constructivist Approach to Machine Translation</i> Michael Carl	247
<i>Shallow Post Morphological Processing with KURD</i> Michael Carl and Antje Schmidt-Wigger	257
<i>Induction of a Stem Lexicon for Two-level Morphological Analysis</i> Erika F. de Lima	267

HCC Workshop

<i>Introducing MegaHal</i> Jason L. Hutchens and Michael D. Alder	271
<i>Methods and Tricks Used in an Attempt to Pass the Turing Test</i> Véronique Bastin and Denis Cordier	275
<i>The Total Turing Test and the Loebner Prize</i> David M. W. Powers	279
<i>Language Model and Sentence Structure Manipulations for Natural Language Applications Systems</i> Zenshiro Kawasaki, Keiji Takida and Masato Tajima	281
<i>Position Paper on Appropriate Audio/Visual Turing Test</i> Bradley B. Custer	287

PaGiLL Workshop

<i>Learning Feature-Value Grammars from Plain Text</i> Tony C. Smith	291
<i>Morphemes as Necessary Concept for Structures Discovery from Untagged Corpora</i> Hervé Déjean	295
<i>The Segmentation Problem in Morphology Learning</i> Christopher D. Manning	299
<i>Reconciliation of Unsupervised Clustering, Segmentation and Cohesion</i> David M. W. Powers	307
<i>Syntactico-Semantic Learning of Categorical Grammars</i> Isabelle Tellier	311

Tutorial TIMETABLE

MONDAY, 12 JANUARY (BLUE MOUNTAINS)

10.00	Christer Samuelsson <i>Introduction to Statistical Methods in Natural Language Processing</i>
13.00	Lunch
14.00	Guided Walk - Grand Canyon
17.00	Robert Dale <i>Introduction to Natural Language Generation</i>
20.00	Dinner and relaxation

TUESDAY, 13 JANUARY (BLUE MOUNTAINS)

7.00	Early morning stroll / Bird Watching
8.00	Breakfast
9.00	Walter Daelemans <i>Introduction to Memory-Based Language Technology</i>
12.00	Lunch
14.00	Guided Walk - Grand Canyon
17.00	Dominique Estival <i>Introduction to Grammatical Formalisms for Natural Language Processing</i>
20.00	Dinner and relaxation

TUESDAY, 20 JANUARY (MELBOURNE)

14.00	David Dowe <i>Introduction to Snob, MML and Mixture Modelling</i>
-------	---

HCC TIMETABLE

WEDNESDAY, 14 JANUARY

11.30		
	Jason L. Hutchens and Michael D. Alder	
	<i>Introducing MegaHaL</i>	271
	Véronique Bastin and Denis Cordier	
	<i>Methods and Tricks Used in an Attempt to Pass the Turing Test</i>	275
12.30	Lunch	
14.00		
	David M. W. Powers	
	<i>The Total Turing Test and the Loebner Prize</i>	279
	Zenshiro Kawasaki, Keiji Takida and Masato Tajima	
	<i>Language Model and Sentence Structure Manipulations for Natural Language Applications Systems</i>	281
	Discussion	
15.30	Break	
16.30	BrainStorming session	
17.30		

NeMLaP3/CoNLL98 TIMETABLE

THURSDAY, 15 JANUARY

9.00	INVITED LECTURE	
	Walter Daelemans	
	<i>Abstraction is Harmful in Language Learning</i>	1
10.30	Break	
11.00	LEARNING 1	
	Michael Towsey, Joachim Diederich, Ingo Schellhammer, Stephan Chalup and Claudia Brugman	
	<i>Natural Language Learning by Recurrent Neural Networks: A Comparison with Probabilistic Approaches</i>	3
	Sandra Kübler	
	<i>Learning a Lexicalised Grammar for German</i>	11
	René Schneider	
	<i>A Lexically-Intensive Algorithm for Domain-Specific Knowledge Acquisition</i>	19
12.30	Lunch	
14.00	TAGGING and TREEBANKS	
	André Kempe	
	<i>Look-Back and Look-Ahead in the Conversion of Hidden Markov Models into Finite State Transducers</i>	29
	Mark Johnson	
	<i>The Effect of Alternative Tree Representations on Tree Bank Grammars</i>	39
	Thorsten Brants and Wojciech Skut	
	<i>Automation of Treebank Annotation</i>	49
	Hamish Cunningham, Mark Stevenson and Yorick Wilks	
	<i>Implementing a Sense Tagger in a General Architecture for Text Engineering</i>	59
16.00	Break	
16.30	LEARNING 2	
	Ingo Schellhammer, Joachim Diederich, Michael Towsey and Claudia Brugman	
	<i>Knowledge Extraction and Recurrent Neural Networks: An Analysis of an Elman Network Trained on a Natural Language Learning Task</i>	73
	Jason L. Hutchens and Michael D. Alder	
	<i>Finding Structure via Compression</i>	79
17.30		

FRIDAY, 16 JANUARY

9.00	INVITED LECTURE	
	Christer Samuelsson	
	<i>Linguistic Theory in Statistical Language Learning</i>	83
10.00	Break	
10.30	SEMANTICS and PRAGMATICS	
	Richard McConachy, Kevin B. Korb and Ingrid Zukerman	
	<i>A Bayesian Approach to Automating Argumentation</i>	91
	Stephen J. Green	
	<i>Automatically Generating Hypertext in Newspaper Articles by Computing Semantic Relatedness</i>	101
	Emin Erkan Korkmaz and Göktürk Üçoluk	
	<i>Choosing A Distance Metric for Automatic Word Categorization</i>	111
	Patrick Saint-Dizier	
	<i>Sense Variation and Lexical Semantics Generative Operations</i>	121
12.30	Lunch	
14.00	MULTILINGUAL ANALYSES	
	Harold Somers	
	<i>An Attempt to Use Weighted Cusums to Identify Sublanguages</i>	131
	Patrick Juola	
	<i>Cross-Entropy and Linguistic Typology</i>	141
	David M.W. Powers	
	<i>Applications and Explanations of Zipf's Law</i>	151
15.30	Break	
16.00	INFORMATION RETRIEVAL, DIALOGUE and GENERATION	
	Peter Wallis, Edmund Yuen and Greg Chase	
	<i>Proper Name Classification in an Information Extraction Toolset</i>	161
	Robert Steele and David Powers	
	<i>Evolution and Evaluation of Document Retrieval Queries</i>	163
	Ilyas Cicekli and Turgay Korkmaz	
	<i>Generation of Simple Turkish Sentences with Systemic-Functional Grammar</i>	165
17.30		

SATURDAY, 17 JANUARY

9.00	PHONOLOGY and PRONUNCIATION	
	Ian Thomas, Ingrid Zukerman and Bhavani Raskutti	
	<i>Extracting Phoneme Pronunciation Information from Corpora</i>	175
	Antal van den Bosch, Ton Weijters and Walter Daelemans	
	<i>Modularity in Inductively-Learned Word Pronunciation Systems</i>	185
10.00	Break	
10.30	METHODOLOGY	
	Antal van den Bosch and Walter Daelemans	
	<i>Do Not Forget: Full Memory in Memory-Based Learning of Word Pronunciation</i>	195
	V. Kamphuis and J.J. Sarbo	
	<i>Natural Language Concept Analysis</i>	205
	Jim Entwisle and David Powers	
	<i>The Present Use of Statistics in the Evaluation of NLP Parsers</i>	215
	Hiroki Imai and Hozumi Tanaka	
	<i>A Method of Incorporating Bigram Constraints into an LR Table and its Effectiveness in Natural Language Processing</i>	225
12.30	Lunch	
13.30	Business Meeting/SIGNLL/Group Reports	
14.30	GROUNDING	
	James M. Hogan, Joachim Diederich and Gerard D. Finn	
	<i>Selective Attention and the Acquisition of Spatial Semantics</i>	235
	Hideki Kozima and Akira Ito	
	<i>Towards Language Acquisition by an Attention-Sharing Robot</i>	245
	Michael Carl	
	<i>A Constructivist Approach to Machine Translation</i>	247
16.00	Break	
16.30	MORPHOLOGY	
	Michael Carl and Antje Schmidt-Wigger	
	<i>Shallow Post Morphological Processing with KURD</i>	257
	Erika F. de Lima	
	<i>Induction of a Stem Lexicon for Two-level Morphological Analysis</i>	267
17.30		

PaGiLL TIMETABLE

WEDNESDAY, 21 JANUARY

11.30	
	Tony C. Smith
	<i>Learning Feature-Value Grammars from Plain Text</i>291
	Hervé Déjean
	<i>Morphemes as Necessary Concept for Structures Discovery from Untagged Corpora</i>295
12.30	Lunch
14.00	
	Christopher D. Manning
	<i>The Segmentation Problem in Morphology Learning</i>299
	David M. W. Powers
	<i>Reconciliation of Unsupervised Clustering, Segmentation and Cohesion</i>307
	Isabelle Tellier
	<i>Syntactico-Semantic Learning of Categorical Grammars</i>311
15.30	Break
16.30	BrainStorming session
17.30	

AUTHOR INDEX

Alder, Michael D.	79, 271	McConachy, Richard	91
Bastin, Véronique	275	Powers, David	151, 163, 215, 279, 307
Brants, Thorsten	49	Raskutti, Bhavani	175
Brugman, Claudia	3, 73	Saint-Dizier, Patrick	121
Carl, Michael	247, 257	Samuelsson, Christer	83
Chalup, Stephan	3	Sarbo, J. J.	205
Chase, Greg	161	Schellhammer, Ingo	3, 73
Cicekli, Ilyas	165	Schmidt-Wigger, Antje	257
Cordier, Denis	275	Schneider, René	19
Cunningham, Hamish	59	Skut, Wojciech	49
Custer, Bradley B.	287	Smith, Tony C.	291
Daelemans, Walter	1, 185, 195	Somers, Harold	131
Déjean, Hervé	295	Steele, Robert	163
de Lima, Erika F.	267	Stevenson, Mark	59
Diederich, Joachim	3, 73, 235	Tajima, Masato	281
Entwisle, Jim	215	Takida, Keiji	281
Finn, Gerard D.	233	Tanaka, Hozumi	225
Green, Stephen J.	101	Tellier, Isabelle	311
Hogan, James M.	235	Thomas, Ian	175
Hutchens, Jason L.	79, 271	Towsey, Michael	3, 73
Imai, Hiroki	225	Üçoluk, Göktürk	111
Ito, Akira	245	van den Bosch, Antal	185, 195
Johnson, Mark	39	Wallis, Peter	161
Juola, Patrick	141	Weijters, Ton	185
Kamphuis, V.	205	Wilks, Yorick	59
Kawasaki, Zenshiro	281	Yuen, Edmund	161
Kempe, André	29	Zukerman, Ingrid	91, 173
Kübler, Sandra	11		
Korb, Kevin B.	91		
Korkmaz, Emin Erkan	111		
Korkmaz, Turgay	165		
Kozima, Hideki	245		
Manning, Christopher	299		