

Proceedings of the
1999 Joint SIGDAT Conference
on
Empirical Methods in Natural
Language Processing
and
Very Large Corpora

Sponsored by
The Association for Computational Linguistics
SIGDAT
LEXIS-NEXIS, a Division of Reed Elsevier, Inc.
Hong Kong University of Science and Technology

Edited by
Pascale Fung
and
Joe Zhou

21-22 June 1999
University of Maryland
College Park, MD, USA

Proceedings of the
1999 Joint SIGDAT Conference
on
Empirical Methods in Natural
Language Processing
and
Very Large Corpora

Sponsored by

The Association for Computational Linguistics
SIGDAT
LEXIS-NEXIS, a Division of Reed Elsevier, Inc.
Hong Kong University of Science and Technology

Edited by

Pascale Fung
and
Joe Zhou

21-22 June 1999
University of Maryland
College Park, MD, USA

©1999, Association for Computational Linguistics

Order additional copies from:

Association for Computational Linguistics
75 Paterson Street, Suite 9
New Brunswick, NJ 08901 USA
+1-732-342-9100 phone
+1-732-342-9339 fax
acl@aclweb.org

SPONSORS:

The Association for Computational Linguistics (ACL)
 SIGDAT (ACL's SIG for Linguistic Data and Corpus-based Approaches to NLP)
 LEXIS-NEXIS, a Division of Reed Elsevier, Inc.
 Hong Kong University of Science and Technology, Human Language Technology Center

INVITED SPEAKERS:

Kenneth W. Church (AT&T Labs-Research)
 Richard Schwartz (BBN Technologies)

ORGANIZERS:

Pascale Fung, Chair
 Joe Zhou, Co-chair

PROGRAM COMMITTEE:

Jing-Shin Chang	(Behavior Design Corp.)
Ken Church	(AT&T Labs-Research)
Ido Dagan	(Bar-Ilan University)
Marti Hearst	(UC-Berkeley)
Huang, Changning	(Microsoft Research China)
Pierre Isabelle	(Xerox Research Europe)
Lillian Lee	(Cornell University)
David Lewis	(AT&T Labs-Research)
Dan Melamed	(West Group)
Mehryar Mohri	(AT&T Labs-Research)
Masaaki Nagata	(NTT)
Richard Sproat	(AT&T Labs-Research)
Andreas Stolcke	(SRI International)
Ralph Weischedel	(BBN)
Dekai Wu	(HKUST)
David Yarowsky	(Johns Hopkins University)

ADDITIONAL REVIEWERS:

Srinivas Bangalore	(AT&T Labs-Research)
Rebecca Bruce	(Univ. of North Carolina)
Michael Collins	(AT&T Labs - Research)
Gregory Grefenstette	(Xerox Research Europe)
Vasileios Hatzivassiloglou	(Columbia University)
David Hull	(Xerox Research Europe)
Peter Jackson	(West Group)
Christian Jacquemin	(LIMSI)
Liu, Xiaohu	(HKUST)
Sung Hyon Myaeng	(Chungnam National Univ.)
Shimei Pan	(Columbia University)
Ted Pederson	(Cal Poly)
Roberto Pieraccini	(AT&T Labs-Research)
Ellen Riloff	(University of Utah)
Hinrich Shütze	(Xerox PARC)
Yannis Stylianou	(AT&T Labs-Research)
Zhao, Jun	(HKUST)

FURTHER INFORMATION:

Pascale Fung
 Human Language Technology Center
 Department of Electrical and Electronic Engineering
 University of Science and Tehnology (HKUST)
 Clear Water Bay, Kowloon
 Hong Kong
 Email: pascale@ee.ust.hk

Joe Zhou
 LEXIS-NEXIS, a Division of Reed Elsevier
 9555 Springboro Pike
 Dayton, OH 45342
 USA
 Email: joez@lexis-nexis.com

CONFERENCE PROGRAM

Monday, June 21

- 8:45-9:00 Welcome
- 9:00-9:40 INVITED SPEECH
What's Happened Since the First SIGDAT Meeting?
Kenneth W. Church (AT&T Labs-Research)
- 9:40-9:50 Short Break
- 9:50-10:10 *Text-Translation Alignment: Three Languages are Better than Two*
Michel Simard
- 10:10-10:30 *Mapping Multilingual Hierarchies Using Relaxation Labeling*
J. Daudé, L. Padró and G. Rigau
- 10:30-10:50 *Improved Alignment Models for Statistical Machine Translation*
Franz Josef Och, Christoph Tillmann, and Hermann Ney
- 10:50-11:10 *Cross-Language Information Retrieval for Technical Documents*
Atsushi Fujii and Tetsuya Ishikawa
- 11:10-11:30 Break
- 11:30-11:50 *Boosting Applied to Tagging and PP Attachment*
Steven Abney, Robert E. Schapire and Yoram Singer
- 11:50-12:10 *Applying Extrasentential Context to Maximum Entropy Based Tagging with a Large Semantic and Syntactic Tagset*
Ezra Black, Andrew Finch and Ruigiang Zhang
- 12:10-12:30 *Improving POS Tagging Using Machine-Learning Techniques*
Lluís Màrquez, Horacio Rodríguez, Josep Carmona and Josep Montolio
- 12:30-14:00 LUNCH
- 14:00-14:20 *Determining the Specificity of Nouns From Text*
Sharon A. Caraballo and Eugene Charniak
- 14:20-14:40 *Retrieving Collocations From Korean Text*
Seonho Kim, Zooil Yang, Mansuk Song and Jung-Ho Ahn
- 14:40-15:00 *Noun Phrase Coreference as Clustering*
Claire Cardie and Kiri Wagstaff
- 15:00-15:20 Break
- 15:20-15:40 *Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence*
Silviu Cucerzan and David Yarowsky
- 15:40-16:00 *Unsupervised Models for Named Entity Classification*
Michael Collins and Yoram Singer
- 16:00-16:20 *Hybrid Disambiguation of Prepositional Phrase Attachment and Interpretation*
Sven Hartrumpf
- 16:20-16:40 *HMM Specialization with Selective Lexicalization*
Jin-Dong Kim, Sang-Zoo Lee and Hae-Chang Rim

Tuesday, June 22

- 9:00-9:40 **INVITED SPEECH**
Why Doesn't Natural Language Come Naturally?
Richard Schwartz (BBN Technologies)
- 9:40-9:50 Short Break
- 9:50-10:10 *POS Tags and Decision Trees for Language Modeling*
Peter A. Heeman
- 10:10-10:30 *An Information-Theoretic Empirical Analysis of Dependency-Based Feature Types for Word Prediction Models*
Dekai Wu, Zhao Jun and Sui Zhifang
- 10:30-10:50 *Word Informativeness and Automatic Pitch Accent Modeling*
Shimei Pan and Kathleen McKeown
- 10:50-11:10 *Learning Discourse Relations with Active Data Selection*
Tadashi Nomoto and Yuji Matsumoto
- 11:10-11:30 Break
- 11:30-11:50 *A Learning Approach to Shallow Parsing*
Marcia Muñoz, Vasin Punyakanok, Dan Roth and Dav Zimak
- 11:50-12:10 *Guiding a Well-Founded Parser with Corpus Statistics*
Amon Seagull and Lenhart Schubert
- 12:10-12:30 *Exploiting Diversity in Natural Language Processing: Combining Parsers*
John Henderson and Eric Brill
- 12:30-14:00 LUNCH
- 14:00-15:10 Panel Discussion
The Future of Language Technologies: Research, Development and Marketing
Ken Church (AT&T), Pierre Isabelle (Xerox Europe), Roberto Pieraccini (AT&T), John Rausch (Lexis-Nexis), Keh-Yih Su (Behavior Design Corp.), Raphael Wong (Intel)
- 15:10-15:20 Short Break
- 15:20-15:40 *Lexical Ambiguity and Information Retrieval Revisited*
Julio Gonzalo, Anselmo Peñas and Felisa Verdejo
- 15:40-16:00 *Detecting Text Similarity over Short Passages: Exploring Linguistic Feature Combinations via Machine Learning*
Vasileios Hatzivassiloglou, Judith L. Klavans and Eleazar Eskin
- 16:00-16:20 *Automated Construction of Weighted String Similarity Measures*
Jörg Tiedemann
- 16:20-16:40 *Taking the Load Off the Conference Chairs: Towards a Digital Paper Routing Assistant*
David Yarowsky and Radu Florian



TABLE OF CONTENTS

<i>What's Happened Since the First SIGDAT Meeting?</i> (INVITED TALK) Kenneth Ward Church	1
<i>Text-Translation Alignment: Three Languages are Better than Two</i> Michel Simard	2
<i>Mapping Multilingual Hierarchies Using Relaxation Labeling</i> J. Daudé, L. Padró and G. Rigau	12
<i>Improved Alignment Models for Statistical Machine Translation</i> Franz Josef Och, Christoph Tillmann, and Hermann Ney	20
<i>Cross-Language Information Retrieval for Technical Documents</i> Atsushi Fujii and Tetsuya Ishikawa	29
<i>Boosting Applied to Tagging and PP Attachment</i> Steven Abney, Robert E. Schapire and Yoram Singer	38
<i>Applying Extrasentential Context to Maximum Entropy Based Tagging With A Large Semantic And Syntactic Tagset</i> Ezra Black, Andrew Finch and Ruigiang Zhang	46
<i>Improving POS Tagging Using Machine-Learning Techniques</i> Lluís Màrquez, Horacio Rodríguez, Josep Carmona and Josep Montolio	53
<i>Determining the Specificity of Nouns From Text</i> Sharon A. Caraballo and Eugene Charniak	63
<i>Retrieving Collocations From Korean Text</i> Seonho Kim, Zooil Yang, Mansuk Song and Jung-Ho Ahn	71
<i>Noun Phrase Coreference as Clustering</i> Claire Cardie and Kiri Wagstaff	82
<i>Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence</i> Silviu Cucerzan and David Yarowsky	90
<i>Unsupervised Models for Named Entity Classification</i> Michael Collins and Yoram Singer	100
<i>Hybrid Disambiguation of Prepositional Phrase Attachment and Interpretation</i> Sven Hartrumpf	111
<i>HMM Specialization with Selective Lexicalization</i> Jin-Dong Kim, Sang-Zoo Lee and Hae-Chang Rim	121
<i>Why Doesn't Natural Language Come Naturally?</i> (INVITED TALK) Richard Schwartz	128
<i>POS Tags and Decision Trees for Language Modeling</i> Peter A. Heeman	129

<i>An Information-Theoretic Empirical Analysis of Dependency-Based Feature Types for Word Prediction Models</i> Dekai Wu, Zhao Jun and Sui Zhifang	138
<i>Word Informativeness and Automatic Pitch Accent Modeling</i> Shimei Pan and Kathleen McKeown	148
<i>Learning Discourse Relations with Active Data Selection</i> Tadashi Nomoto and Yuji Matsumoto	158
<i>A Learning Approach to Shallow Parsing</i> Marcia Muñoz, Vasin Punyakanok, Dan Roth and Dav Zimak	168
<i>Guiding a Well-Founded Parser with Corpus Statistics</i> Amon Seagull and Lenhart Schubert	179
<i>Exploiting Diversity in Natural Language Processing: Combining Parsers</i> John Henderson and Eric Brill	187
<i>Lexical Ambiguity and Information Retrieval Revisited</i> Julio Gonzalo, Anselmo Peñas and Felisa Verdejo	195
<i>Detecting Text Similarity over Short Passages: Exploring Linguistic Feature Combinations via Machine Learning</i> Vasileios Hatzivassiloglou, Judith L. Klavans and Eleazar Eskin	203
<i>Automated Construction of Weighted String Similarity Measures</i> Jörg Tiedemann	213
<i>Taking the Load Off the Conference Chairs: Towards a Digital Paper Routing Assistant</i> David Yarowsky and Radu Florian	220
<i>PP-attachment: A Committee Machine Approach</i> Martha A. Alegre, Josep M. Sopena and Agusti Lloberas	231
<i>Cascaded Grammatical Relation Assignment</i> Sabine Buchholz, Jorn Veenstra and Walter Daelemans	239
<i>Automatically Merging Lexicons that have Incompatible Part-of-Speech Categories</i> Daniel Ka-Leung Chan and Dekai Wu	247
<i>An Iterative Approach to Estimating Frequencies over a Semantic Hierarchy</i> Stephen Clark and David Weir	258
<i>Using Subcategorization to Resolve Verb Class Ambiguity</i> Maria Lapata and Chris Brew	266
<i>Improving Brill's POS Tagger for an Agglutinative Language</i> Beáta Megyesi	275
<i>Corpus-Based Learning for Noun Phrase Coreference Resolution</i> Wee Meng Soon, Hwee Tou Ng and Chung Yong Lim	285
<i>Corpus-Based Approach for Nominal Compound Analysis for Korean Based on Linguistic and Statistical Information</i> Juntae Yoon, Key-Sun Choi and Mansuk Song	292

AUTHOR INDEX

Steven Abney	38	Beáta Megyesi	275
Jung-Ho Ahn	71	Josep Montolio	53
Martha A. Alegre	231	Marcia Muñoz	168
Ezra Black	46	Hermann Ney	20
Chris Brew	266	Hwee Tou Ng	285
Eric Brill	187	Tadashi Nomoto	158
Sabine Buchholz	239	Franz Josef Och	20
Sharon A. Caraballo	63	L. Padró	12
Claire Cardie	82	Shimei Pan	148
Josep Carmona	53	Anselmo Peñas	195
Daniel Ka-Leung Chan	247	Vasin Punyakanok	168
Eugene Charniak	63	G. Rigau	12
Key-Sun Choi	292	Hae-Chang Rim	121
Kenneth Ward Church	1	Horacio Rodríguez	53
Stephen Clark	258	Dan Roth	168
Michael Collins	100	Robert E. Schapire	38
Silviu Cucerzan	90	Lenhart Schubert	179
Walter Daelemans	239	Richard Schwartz	128
J. Daudé	12	Amon Seagull	179
Eleazar Eskin	203	Michel Simard	2
Andrew Finch	46	Yoram Singer	38, 100
Radu Florian	220	Mansuk Song	71, 292
Atsushi Fujii	29	Wee Meng Soon	285
Julio Gonzalo	195	Josep M. Sopena	231
Sven Hartrumpf	111	Sui Zhifang	138
Vasileios Hatzivassiloglou	203	Jörg Tiedemann	213
Peter A. Heeman	129	Christoph Tillmann	20
John Henderson	187	Jorn Veenstra	239
Tetsuya Ishikawa	29	Felisa Verdejo	195
Jin-Dong Kim	121	Kiri Wagstaff	82
Seonho Kim	71	David Weir	258
Judith L. Klavans	203	Dekai Wu	138, 247
Maria Lapata	266	Zooil Yang	71
Sang-Zoo Lee	121	David Yarowsky	90, 220
Chung Yong Lim	285	Juntae Yoon	292
Agusti Lloberas	231	Ruigiang Zhang	46
Lluís Màrquez	53	Zhao Jun	138
Yuji Matsumoto	158	Dav Zimak	168
Kathleen McKeown	148		

