

Proceedings of the
Fourth Workshop
on
Very Large Corpora

Sponsored by

The Association for Computational Linguistics
ACL's SIGDAT
LEXIS-NEXIS, a Division of Reed Elsevier, Inc.
AT&T

Edited by

Eva Ejerhed
and
Ido Dagan

4 August, 1996
University of Copenhagen
Copenhagen, Denmark

Proceedings of the
Fourth Workshop
on
Very Large Corpora

Sponsored by

The Association for Computational Linguistics
ACL's SIGDAT
LEXIS-NEXIS, a Division of Reed Elsevier, Inc.
AT&T

Edited by

Eva Ejerhed
and
Ido Dagan

4 August, 1996
University of Copenhagen
Copenhagen, Denmark

© 1996, The Authors

SPONSORS:

The Association for Computational Linguistics (ACL)
SIGDAT, ACL's Special Interest Group for Linguistic Data and Corpus-based
Approaches to NLP
LEXIS-NEXIS, a Division of Reed Elsevier, Inc.
AT&T

INVITED SPEAKER:

Phil Hayes

ORGANIZERS:

Eva Ejerhed, Chair
Ido Dagan, Co-chair

PROGRAM COMMITTEE:

Susan Armstrong	(ISCCO, Switzerland)
Changning Huang	(Tsinghua University, China)
Keh-Jiann Chen	(IIS Sinica, Taiwan)
Kenneth Church	(AT&T Bell Labs, USA)
Helmut Feldweg	(Universitaet Tuebingen, Germany)
Don Hindle	(AT&T Bell Labs, USA)
Fred Karlsson	(University of Helsinki, Finland)
Mark Lauer	(Macquarie University, Australia)
Ellen Riloff	(University of Utah, USA)
Kyoji Umemura	(Toyohashi University of Technology, Japan)
Mark Wasson	(LEXIS-NEXIS, USA)

FURTHER INFORMATION:

Eva Ejerhed
Dept of Linguistics
University of Umea
S 90187 Umea, Sweden
e-mail: ejerhed@ling.umu.se

Ido Dagan
Dept of Mathematics & Computer Science
Bar Ilan University
Ramat Gan 52900, Israel
e-mail: dagan@macs.biu.ac.il

WORKSHOP PROGRAM

08.00 - 09.00	Registration
09.00 - 09.05	Opening
09.05 - 09.30	Evelyne Tzoukermann and Dragomir R. Radev <i>Using Word Class for Part-of-speech Disambiguation</i>
09.30 - 09.55	Walter Daelemans, Jakob Zavrel, Peter Berck and Steven Gillis <i>MBT: A Memory-Based Part of Speech Tagger-Generator</i>
09.55 - 10.20	Akira Ushioda <i>Hierarchical Clustering of Words and Application to NLP Tasks</i>
10.20 - 10.45	Morning Break
10.45 - 11.10	Yael Karov and Shimon Edelman <i>Learning Similarity-based Word Sense Disambiguation from Sparse Data</i>
11.10 - 11.35	Atsushi Fujii, Kentaro Inui, Takenobu Tokunaga and Hozumi Tanaka <i>Selective Sampling of Effective Example Sentence Sets for Word Sense Disambiguation</i>
11.35 - 12.00	Christer Samuelsson <i>Relating Turing's Formula and Zipf's Law</i>
12.00 - 13.15	LUNCH
13.15 - 13.40	Mihoko Kitamura and Yuji Matsumoto <i>Automatic Extraction of Word Sequence Correspondences in Parallel Corpora</i>
13.40 - 14.05	Xiang Tong and David A. Evans <i>A Statistical Approach to Automatic OCR Error Correction in Context</i>
14.05 - 14.50	Tadashi Nomoto and Yuji Matsumoto <i>Exploiting Text Structure for Topic Identification</i>
14.50 - 15.20	Afternoon Break
15.20 - 15.50	INVITED TALK Phil Hayes <i>The Use of a Corpus to Help Define a Controlled Language</i>
15.50 - 16.30	PANEL DISCUSSION <i>Innovative Uses and Applications of Large Corpora</i>
16.30 - 16.40	Short Break
16.40 - 17.05	Tung-Hui Chiang and Keh-Yih Su <i>Statistical Models for Deep-Structure Disambiguation</i>
17.05 - 17.30	Rens Bod <i>Two Questions about Data-Oriented Parsing</i>
17.30 - 17.55	Hang Li <i>A Probabilistic Disambiguation Method Based on Psycholinguistic Principles</i>
17.55 - 18.00	Closing

TABLE OF CONTENTS

<i>Using Word Class for Part-of-speech Disambiguation</i> Evelyne Tzoukermann and Dragomir R. Radev	1
<i>MBT: A Memory-Based Part of Speech Tagger-Generator</i> Walter Daelemans, Jakob Zavrel, Peter Berck and Steven Gillis	14
<i>Hierarchical Clustering of Words and Application to NLP Tasks</i> Akira Ushioda	28
<i>Learning Similarity-based Word Sense Disambiguation from Sparse Data</i> Yael Karov and Shimon Edelman	42
<i>Selective Sampling of Effective Example Sentence Sets for Word Sense Disambiguation</i> Atsushi Fujii, Kentaro Inui, Takenobu Tokunaga and Hozumi Tanaka	56
<i>Relating Turing's Formula and Zipf's Law</i> Christer Samuelsson	70
<i>Automatic Extraction of Word Sequence Correspondences in Parallel Corpora</i> Mihoko Kitamura and Yuji Matsumoto	79
<i>A Statistical Approach to Automatic OCR Error Correction in Context</i> Xiang Tong and David A. Evans	88
<i>Exploiting Text Structure for Topic Identification</i> Tadashi Nomoto and Yuji Matsumoto	101
<i>Statistical Models for Deep-structure Disambiguation</i> Tung-Hui Chiang and Keh-Yih Su	113
<i>Two Questions about Data-Oriented Parsing</i> Rens Bod	125
<i>A Probabilistic Disambiguation Method Based on Psycholinguistic Principles</i> Hang Li	141
<i>A Re-estimation Method for Stochastic Language Modeling from Ambiguous Observations</i> Mikio Yamamoto	155
<i>Towards Automatic Grammar Acquisition from a Bracketed Corpus</i> Thanaruk Theeramunkong and Manabu Okumara	168

AUTHOR INDEX

Peter Berck	14
Rens Bod	125
Tung-Hui Chiang	113
Walter Daelemans	14
Shimon Edelman	42
David A. Evans	88
Atsushi Fujii	56
Steven Gillis	14
Kentaro Inui	56
Yael Karov	42
Mihoko Kitamura	79
Hang Li	141
Yuji Matsumoto	79,101
Tadashi Nomoto	101
Manabu Okumara	168
Dragomir R. Radev	1
Christer Samuelsson	70
Keh-Yih Su	113
Hozumi Tanaka	56
Thanaruk Theeramunkong	168
Takenobu Tokunaga	56
Xiang Tong	88
Evelyne Tzoukermann	1
Akira Ushioda	28
Mikio Yamamoto	155
Jakob Zavrel	14