Proceedings of the

Fifth Workshop on Very Large Corpora

Sponsored by
The Association for Computational Linguistics
LEXIS-NEXIS and Reed Elsevier Technology
AT&T Laboratories - Research
National Natural Science Foundation of China
State Key Laboratory of Intelligent Technology and Systems, China
The Hong Kong University of Science and Technology
City University of Hong Kong

Edited by

Joe Zhou
and
Kenneth Church

18 August 1997
Tsinghua University, Beijing, China

20 August 1997
The Hong Kong University of Science and Technology, Hong Kong

Proceedings of the

Fifth Workshop on Very Large Corpora

Sponsored by
The Association for Computational Linguistics
LEXIS-NEXIS and Reed Elsevier Technology
AT&T Laboratories - Research
National Natural Science Foundation of China
State Key Laboratory of Intelligent Technology and Systems, China
The Hong Kong University of Science and Technology
City University of Hong Kong

Edited by

Joe Zhou
and
Kenneth Church

18 August 1997
Tsinghua University, Beijing, China

20 August 1997
The Hong Kong University of Science and Technology, Hong Kong

SPONSORS:

The Association for Computational Linguistics (ACL)
LEXIS-NEXIS and Reed Elsevier Technology
AT&T Laboratories - Research
National Natural Science Foundation of China
State Key Laboratory of Intelligent Technology and Systems, China
The Hong Kong University of Science and Technology
City University of Hong Kong

INVITED SPEAKERS:

Mitch Marcus, University of Pennsylvania, USA
John Rausch, LEXIS-NEXIS, a Division of Reed Elsevier, USA
Howard Turtle, West Group, USA

ORGANIZERS:

Joe Zhou, LEXIS-NEXIS and Reed Elsevier Technology, USA
Kenneth Church, AT&T Laboratories - Research, USA
Changning Huang, Tsinghua University, Beijing, China

PROGRAM COMMITTEE:

Susan Armstrong, ISSO, University of Geneva, Switzerland
Key-sun Choi, KAIST, Korea
Ido Dagan, Bar Ilan University, Israel
Pernilla Danielsson, University of Gothenburg, Sweden
Marti Hearst, Xerox PARK, USA
Chu-ren Huang, Academia Sinica, Taiwan
Claudia Leacock, Princeton University, USA
Sun Maosong, Tsinghua University, China
Masaaki Nagata, NTT Information and Communication Systems Labs, Japan
Daniel Pliske, LEXIS-NEXIS and Reed Elsevier Technology, USA
Benjamin Tsou, City University of Hong Kong, Hong Kong
Paul Wu, Institute of Systems Science, Singapore

FURTHER INFORMATION:

Joe Zhou
LEXIS-NEXIS and Reed Elsevier Technology
9555 Springboro Pike
Dayton, OH 45342 USA
email: joez@lexis-nexis.com

Ken Church
Room 2B-421
AT&T Laboratories - Research
Murray Hill, NJ 07974 USA
e-mail: kwc@research.att.com

WORKSHOP PROGRAM
August 18, 1997
Tsinghua University, Beijing, China

8:30 - 8:45 Welcome

8:45 - 9:10 Qiang Zhou
A Statistics-based Chinese Parser

9:10 - 9:35 Thanaruk Theeramunkong and Manabu Okumura
Grammar Acquisition Based on Clustering Analysis and its Application to Statistical Parsing

9:35 - 10:00 Seungmi Lee and Key-Sun Choi
Re-estimation and Best First Parsing Algorithms for Probabilistic Dependency Gramma

10:00 - 10:30 Break

10:30 - 10:55 Tomek Strzalkowski and Ron Brandow
A Natural Language Correction Model for Continuous Speech Recognition

10:55 - 11:20 Masaaki Nagata
A Self-Organizing Japanese Word Segmenter using Heuristic Word Identification and Re-estimation

11:20 - 12:20 INVITED TALK (Mitch Marcus)

12:20 - 2:20 LUNCH

2:20 - 2:45 Hiromi Nakaiwa
Automatic Identification of Zero Pronouns and Their Antecedent Within Aligned Sentenc Pairs

2:45 - 3:10 Xuan-jing Huang, Li-de Wu and Wen-xin Wang
Statistical Acquisition of Terminology Dictionary

3:10 - 4:10 INVITED TALK (John Rausch)

4:10 - 4:40 Break

4:40 - 5:40 Panel Discussion
Innovative Uses and Applications of Large Corpora

5:40 - 6:05 Kumiko Tanaka-Ishii and Hideya Iwasaki
Clustering Co-occurrence Graph Based on Transitivity

6:05 - 6:30 Sta Jean-David
Knowledge Acquisition: Classification of Terms in a Thesaurus from a Corpus

6:30 - 6:40 Closing

POSTER SESSION (during lunch time and breaks)

Takehito Utsuro, Takashi Miyata and Yuji Matsumoto
Maximum Entropy Model Learning of Subcategorization Preference
Scott M. Thede and Mary Harper
Analysis of Unknown Lexical Items using Morphological and Syntactic Information with the TIMIT Corpus
Jee-sun Nam and Key-sun Choi
LG-based Approach to Recognizing Proper Names in Korean
Asanee Kawtrakul, Chalatip Thumkanon
A Statistical Approach to Thai Morphological Analzer
Jun Gao and Xi-Xian Chen
Probalistic Word Classification Base on Context-Sensitive Binary Tree Method

WORKSHOP PROGRAM
August 20, 1997
Hong Kong University of Science and Technology, Hong Kong

8:30 - 8:45 Opening

8:45 - 9:10 Li Shiuan Peh and Hwee Tou Ng
Domain-Specific Semantic Class Disambiguation using WordNet

9:10 - 9:35 Jiri Stetina and Makoto Nagao
Corpus-based PP Attachment Ambiguity Resolution with a Semantic Dictionary

9:35 - 10:00 Joyce Yue Chai and Alan W. Biermann
Corpus Based Statistical Generalization Tree in Rule Optimization

10:00 - 10:30 Break

10:30 - 10:55 E. Black, S. Eubank and K. Kashioka
Probabilistic Parsing of Unrestricted English Text, with A Highly-Detailed Grammar

10:55 - 11:20 T. Rose, N. Haddock and R. Tucker
The Effects of Corpus Size and Homogeneity on Language Model Quality

11:20 - 12:20 INVITED TALK (Mitch Marcus)

12:20 - 2:00 LUNCH

2:00 - 2:25 Erika F. de Lima
Acquiring German Prepositional Subcategorization Frames from Corpora

2:25 - 2:50 Pascale Fung and Kathleen McKeown
Finding Terminology Translations from Non-Parallel Corpora

2:50 - 3:50 INVITED TALK (Howard Turtle)

3:50 - 4:20 Break

4:20 - 5:25 PANEL DISCUSSION
Innovative Uses and Applications of Large Corpora

5:25 - 5:50 Tadashi Nomoto and Yuji Matsumoto
Data Reliability and its Effects on Automatic Abstracting

5:50 - 6:15 Andrei Mikheev
Collocation Lattices and Maximum Entropy Models

6:15 - 6:25 Closing

TABLE OF CONTENTS

Invited Speech (Beijing and Hong Kong sessions)
Mitch Marcus 1

Invited Speech (Beijing session): Commercial Implementation of Text Recognition Tools for Very Large Corpora
John Rausch 2

Invited Talk (Hong Kong session): Commercial Impact of Very Large Corpora Research
Howard Turtle 3

A Statistics-based Chinese Parser
Qiang Zhou 4

Probabilistic Parsing of Unrestricted English Text, with A Highly-Detailed Grammar
E. Black, S. Eubank and H. Kashioka 16

Grammar Acquisition Based on Clustering Analysis and its Application to Statistical Parsing
Thanaruk Theeramunkong and Manabu Okumura 31

Reestimation and Best-First Parsing Algorithms for Probabilistic Dependency Grammars
Seungmi Lee and Key-Sun Choi 41

Domain-Specific Semantic Class Disambiguation Using WordNet
Li Shiuan Peh and Hwee Tou Ng 56

Corpus-Based PP Attachment Ambiguity Resolution with a Semantic Dictionary
Jiri Stetina and Makoto Nagao 66

Corpus Based Statistical Generalization Tree in Rule Optimization
Joyce Yue Chai and Alan W. Biermann 81

Clustering Co-occurrence Graph Based on Transitivity
Kumiko Tanaka-Ishii and Hideya Iwasaki 91

Knowledge Acquisition : Classification of Terms in a Thesaurus from a Corpus
Sta Jean-David 101

Data Reliability and its Effects on Automatic Abstracting
Tadashi Nomoto and Yuji Matsumoto 113

Automatic Identification of Zero Pronouns and their Antecedents within Aligned Sentence Pairs
Hiromi Nakaiwa 127

Statistical Acquisition of Terminology Dictionary
Huang Xuan-jing, Wu Li-de, and Wang Wen-xin.....142

Acquiring German Prepositional Subcategorization Frames from Corpora
Erika F. de Lima 153

TABLE OF CONTENTS

A Natural Language Correction Model for Continuous Speech Recognition
Tomek Strzalkowski and Ron Brandow 168

The Effects of Corpus Size and Homogeneity on Language Model Quality
T. Rose, N. Haddock and R. Tucker 178

Finding Terminology Translations from Non-parallel Corpora
Pascale Fung and Kathleen McKeown 192

A Self-Organizing Japanese Word Segmenter using Heuristic Word Identification and Re-estimation
Masaaki Nagata 203

Collocation Lattices and Maximum Entropy Models
Andrei Mikheev 216

Using Word Frequency Lists to Measure Corpus Homogeneity and Similarity between Corpora
Adam Kilgarriff 231

POSTER PAPERS

Maximum Entropy Model Learning of Subcategorization Preference
Takehito Utsuro, Takashi Miyata and Yuji Matsumoto 246

Analysis of Unknown Lexical Items using Morphological and Syntactic Information with the TIMIT Corpus
Scott M. Thede and Mary Harper 261

LG-based Approach to Recognizing of Proper Names in Korean Texts
Jee-sun Nam and Key-Sun Choi 273

A Statistical Approach to Thai Morphological Analyzer
Kawtrakul Asanee, Thumkanon Chalatip 289

Probalistic Word Classification Based on Context-Sensitive Binary Tree Method
Jun Gao and XiXian Chen 297

AUTHOR INDEX

Alan W. Biermann 81
E. Black 16
Ron Brandow 168
Joyce Yue Chai 81
XiXian Chen 297
Key-Sun Choi 41, 273
S. Eubank 16
Pascale Fung 192
Jun Gao 297
N. Haddock 178
Mary Harper 261
Xuan-jing Huang 142
Hideya Iwasaki 91
Sta Jean-David 101
H. Kashioka 16
Asanee Kawtrakul 289
Adam Kilgarriff 231
Seungmi Lee 41
Erika F. de Lima 153
Mitch Marcus 1
Yuji Matsumoto 113, 246
Kathleen McKeown 192
Andrei Mikheev 216
Takashi Miyata 246
Makoto Nagao 66
Masaaki Nagata 203
Hiromi Nakaiwa 127
Jee-sun Nam 273
Hwee Tou Ng 56
Tadashi Nomoto 113
Manabu Okumura 31
Li Shiuan Peh 56
John Rausch 2
T. Rose 178
Jiri Stetina 66
Tomek Strzalkowski 168
Kumiko Tanaka-Ishii 91
Scott M. Thede 261
Thanaruk Theeramunkong 31
Chalatip Thumkanon 289
R. Tucker 178
Howard Turtle 3
Takehito Utsuro 246
Wen-xin Wang 142
Li-de Wu 142
Qiang Zhou 4