

**Proceedings of the Workshop
on
VERY LARGE CORPORA:
ACADEMIC AND INDUSTRIAL PERSPECTIVES**

**Sponsored by
Association for Computational Linguistics
American Chemical Society Chemical Abstracts
Mead Data Central, Inc.
OCLC Online Computer Library Center, Inc.**

**June 22, 1993
Ohio State University
Columbus, Ohio, USA**

Acknowledgments

Sponsors

Association for Computational Linguistics (ACL)
The Chemical Abstracts Service Division of the American Chemical Society (CAS)
Mead Data Central, Inc. (MDC)
OCLC Online Computer Library Center, Inc. (OCLC)

Invited Speakers

Robert Mercer
Mark Wasson

Program Committee

Peter Brown
Kenneth Church (chair)
Jean Godby
David Lewis
Ray Reighart
Fu-qiu Zhou

Other Reviewers

William Gale
Donald Hindle
Tim Humphrey
Karl Mitze
Richard Sproat

Workshop Program

Tuesday, June 22, 1993

Chair: Kenneth W. Church, Bell Laboratory, USA

8:30-9:30 INVITED TALK, Robert Mercer, IBM, USA

PARALLEL TEXT

9:30-9:55 Dagan, Gale and Church, "Robust Bilingual Word Alignment for Machine Aided Translation," Bell Laboratory, USA

9:55-10:20 Break

Chair: Fu-qiu Zhou, Mead Data Central, USA

10:20-11:20 INVITED TALK, Mark Wasson, Mead Data Central, USA

INFORMATION RETRIEVAL

11:20-11:45 Strzalkowski, "Robust Text Processing in Automated Information Retrieval," New York, USA

11:45-12:10 Liddy and Paik, "Document Filtering Using Semantic Information from a Machine Readable Dictionary," Syracuse University, USA

12:10-1:10 Lunch

Chair: Mitch Marcus, University of Pennsylvania, USA

PART OF SPEECH

1:10-1:35 Cloeren, "Toward a Cross-Linguistic Tagset," Nijmegen, The Netherlands

1:35-2:00 Chang and Chen, "HMM-Based Part-of-Speech Tagging for Chinese Corpora," Taiwan, Republic of China

PARSING

2:00-2:25 Voutilainen, "NPtool, a Detector of English Noun Phrases," Helsinki, Finland

2:25-2:50 Resnik and Hearst, "Structural Ambiguity and Conceptual Relations," University of Pennsylvania, USA

2:50-3:15 Break

TABLE OF CONTENT

<i>Robust Bilingual Word Alignment for Machine Aided Translation</i> Ido Dagan, Kenneth Church & Willian Gale	1
<i>Robust Text Processing in Automated Information Retrieval</i> Tomek Strzalkowski	9
<i>Document Filtering Using Semantic Information from a Machine Readable Dictionary</i> Elizabeth D. Liddy & Woojin Paik	20
<i>Toward a Cross-Linguistic Tagset</i> Jan Cloeren	30
<i>HMM-Based Part-of-Speech Tagging for Chinese Corpora</i> Chao-Huang Chang & Cheng-der Chen	40
<i>NPtool, a Detector of English Noun Phrases</i> Atro Voutilainen	48
<i>Structural Ambiguity and Conceptual Relations</i> Philip Resnik & Marti A.Hearst	58
<i>Text Recognition and Collocations and Domain Codes</i> T.G. Rose & L.J. Evett	65
<i>Extraction of V-N-Collocations from Text Corpora: A Feasibility Study for German</i> Elizabeth Breidt	74
<i>Computation of Word Associations Based on Co-occurrences of Words in Large Corpora</i> Manfred Wettler & Reinhard Rapp	84
<i>Corpus-Based Adaptation Mechanisms for Chinese Homophone Disambiguation</i> Chao-Huang Chang	94
<i>Example-Based Sense Tagging of Running Chinese Text</i> Xiang Tong, Chang-ning Huang & Cheng-ming Guo	102
<i>Experience about Compound Dictionary on Computer Network</i> Kyoji Umemura, Akihiro Umemura & Etsuko Suzuki	113

Chair: Jean Godby, OCLC Online Computer Library Center, Inc.

COLLOCATION

- 3:15-3:40 Rose and Evett, "Text Recognition and Collocations and Domain Codes,"
Nottingham, UK
- 3:40-4:05 Breidt, "Extraction of V-N-Collocations from Text Corpora: A Feasibility Study
for German," Tübingen, Germany
- 4:05-4:30 Wettler and Rapp, "Computation of Word Associations Based on Co-occurrences
of Words in Large Corpora," Paderborn, Germany
- 4:30-4:55 Break

Chair: William Gale, Bell Laboratory, USA

WORD-SENSE DISAMBIGUATION AND DICTIONARY

- 4:55-5:20 Chang, "Corpus-Based Adaptation Mechanisms for Chinese Homophone
Disambiguation," Taiwan, Republic of China
- 5:20-5:45 Tong, Huang and Guo, "Example-Based Sense Tagging of Running Chinese Text,"
Beijing, the People's Republic of China
- 5:45-6:10 Umemura, Umemura and Suzuki, "Experience about Compound Dictionary on
Computer Network," NTT, Musashino, Japan

AUTHOR INDEX

Elizabeth Breidt	74
Chao-Huang Chang	40,94
Cheng-der Chen	40
Kenneth W. Church	1
Jan Cloeren	30
Ido Dagan	1
L.J. Evett	65
Willian Gale	1
Cheng-ming Guo	102
Marti A.Hearst	58
Chang-ning Huang	102
Elizabeth D. Liddy	20
Woojin Paik	20
Reinhard Rapp	84
Philip Resnik	58
T.G. Rose	65
Tomek Strzalkowski	9
Etsuko Suzuki	113
Xiang Tong	102
Akihiro Umemura	113
Kyoji Umemura	113
Atro Voutilainen	48
Manfred Wettler	84

