Proceedings of the

Third Workshop

on

Very Large Corpora

Sponsored by

The Association for Computational Linguistics
ACL's SIGDAT and SIGNLL
LEXIS-NEXIS, a Division of Reed Elsevier, Inc.

Edited by

David Yarowsky
and
Kenneth Church

30 June 1995
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA

Proceedings of the

Third Workshop

on

Very Large Corpora

Sponsored by

The Association for Computational Linguistics
ACL's SIGDAT and SIGNLL
LEXIS-NEXIS, a Division of Reed Elsevier, Inc.

Edited by

David Yarowsky
and
Kenneth Church

30 June 1995
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA

SPONSORS:

The Association for Computational Linguistics (ACL)
SIGDAT (ACL's SIG for Linguistic Data and Corpus-based Approaches to NLP)
SIGNLL (ACL's SIG for Natural Language Learning)
LEXIS-NEXIS, a Division of Reed Elsevier, Inc.

INVITED SPEAKERS:

Mark Liberman
Henry Kučera and Nelson Francis

ORGANIZERS:

David Yarowsky, Chair
Kenneth Church, Co-chair

PROGRAM COMMITTEE:

Susan Armstrong	(ISCCO, Switzerland)
Walter Daelemans	(ITK/KUB, Tilburg, Netherlands)
Marti Hearst	(Xerox PARC, USA)
Chang-Ning Huang	(Tsinghua University, China)
Pierre Isabelle	(CITI, Canada)
Yuji Matsumoto	(NAIST, Japan)
David Powers	(Flinders University, Australia)
Philip Resnik	(Sun Microsystems Laboratories, USA)
Dekai Wu	(HKUST, Hong Kong)
Joe Zhou	(LEXIS-NEXIS, USA)

ADDITIONAL REVIEWERS:

Yves Schabes	(Mitsubishi Electric Research Labs, USA)
Julian Kupiec	(Xerox PARC, USA)
Takehito Utsuro	(NAIST, Japan)

FURTHER INFORMATION:

David Yarowsky
Dept. of Computer and Info. Science
University of Pennsylvania
200 S. 33rd St.
Philadelphia, PA 19104-6389 USA
email: yarowsky@unagi.cis.upenn.edu

Kenneth Church
Room 2B-421
AT&T Bell Laboratories
600 Mountain Ave.
Murray Hill, NJ 07974 USA
e-mail: kwc@research.att.com

WORKSHOP PROGRAM

8:45 - 8:50	Welcome
8:50 - 9:35	INVITED TALK (Mark Liberman)
9:35 - 9:50	Break
9:50 - 10:15	Eric Brill *Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging*
10:15 - 10:40	Carl de Marcken *Lexical Heads, Phrase Structure and the Induction of Grammar*
10:40 - 11:05	Michael Collins and James Brooks *Prepositional Phrase Attachment through a Backed-off Model*
11:05 - 11:15	Break
11:15 - 11:40	Andrew Golding *A Bayesian Hybrid Method for Context-sensitive Spelling Correction*
11:40 - 12:05	Philip Resnik *Disambiguating Noun Groupings with Respect to Wordnet Senses*
12:05 - 1:05	LUNCH
1:05 - 1:30	Dekai Wu *Trainable Coarse Bilingual Grammars for Parallel Text Bracketing*
1:30 - 1:55	Lance Ramshaw and Mitch Marcus *Text Chunking using Transformation-Based Learning*
1:55 - 2:05	Break
2:05 - 3:00	INVITED TALK (Henry Kučera and Nelson Francis)
3:00 - 3:10	Break
3:10 - 3:35	Fernando Pereira, Yoram Singer and Naftali Tishby *Beyond Word N-Grams*
3:35 - 4:00	Jing-Shin Chang, Yi-Chung Lin and Keh-Yih Su *Automatic Construction of a Chinese Electronic Dictionary*
4:00 - 4:10	Break
4:10 - 4:35	Kenneth Church and William Gale *Inverse Document Frequency (IDF): A Measure of Deviations from Poisson*
4:35 - 5:00	Joe Zhou and Pete Dapkus *Automatic Suggestion of Significant Terms for a Predefined Topic*
5:00 - 5:25	Ellen Riloff and Jay Shoen *Automatically Acquiring Conceptual Patterns without an Annotated Corpus*

TABLE OF CONTENTS

AUTHOR INDEX