Proceedings of the

Sixth Workshop on Very Large Corpora

Sponsored by
The Association for Computational Linguistics
ACL's SIGDAT
West Group

Edited by Eugene Charniak

15th–16th August 1998
Université de Montréal
Montreal, Quebec, Canada

COLING-ACL'98

Proceedings of the

Sixth Workshop on Very Large Corpora

Sponsored by
The Association for Computational Linguistics
ACL's SIGDAT
West Group

Edited by Eugene Charniak

15th–16th August 1998
Université de Montréal
Montreal, Quebec, Canada

COLING-ACL'98

Current ACL members may order copies of this and other ACL-related proceedings from:

Association for Computational Linguistics (ACL)
75 Paterson Street, Suite 9
New Brunswick, NJ 08901 USA
Tel: +1-732-342-9100
Fax: +1-732-873-0014
rasmusse@cs.rutgers.edu

All other orders (from libraries, institutions and other individuals) should be addressed to:

Morgan Kaufmann Publishers
340 Pine Street, 6th Floor
San Francisco, CA 94104 USA
Tel: +1-650-392-2665, ext. 231
Fax: +1-650-982-2665
mkp@mkp.com

SPONSORS:

The Association for Computational Linguistics
SIGDAT, ACL's Special Interest Group for Linguistic Data and Corpus-based Approaches to NLP
West Group

INVITED SPEAKERS:

Jan Pedersen
Ellisa Newport

CONFERENCE CHAIR:

Eugene Charniak

PROGRAM COMMITTEE:

Steven Abney	Lillian Lee
Eric Brill	Christopher Manning
Ted Briscoe	Dan Melamed
Rebecca Bruce	Scott Miller
Claire Cardie	Raymond Mooney
Bob Carpenter	James Pustejovsky
Glen Carroll	Lance Ramshaw
Ken Church	Adwait Rathnaparkhi
Michael Collins	Ellen Riloff
Joshua Goodman	Hinrich Schütze
Vasilis Hatzivassiloglou	Ralph Weischedel
Mark Johnson	Janyce Wiebe
Andrew Kehler	Dekai Wu
John Lafferty	David Yarowsky

FURTHER INFORMATION:

Eugene Charniak
Department of Computer Science
115 Waterman Street
Brown University
Box 1910, Providence RI 02912
USA
e-mail: ec@cs.brown.edu

Programme

Saturday August 15, 1998

Session 1

9:00–9:25 *Bayesian Stratified Sampling to Assess Corpus Utility*
Judith Hochberg, Clint Scovel, Timothy Thomas and Sam Hall
Los Alamos National Laboratory

9:25–9:50 *Encoding Linguistic Corpora*
Nancy Ide
Vassar College

9:50–10:15 *Using a Probabilistic Translation Model for Cross-Language Information Retrieval*
Jian-Yun Nie, Pierre Isabelle, Pierre Plamondon and George Foster
Université de Montréal

10:15–10:40 *Using Suffix Arrays to Compute Term Frequency and Document Frequency for All Substrings in a Corpus*
Mikio Yamamoto and Kenneth W. Church
University of Tsukuba and AT&T Labs – Research

10:40–11:15 Break

Session 2

11:15–11:40 *Semantic Tagging using a Probabilistic Context Free Grammar*
Michael Collins and Scott Miller
University of Pennsylvania and BBN Technologies

11:40–12:05 *An Empirical Approach to Conceptual Case Frame Acquisition*
Ellen Riloff and Mark Schmelzenbach
University of Utah

12:05–12:30 *Semantic Lexicon Acquisition for Learning Natural Language Interfaces*
Cynthia A. Thompson and Raymond J. Mooney
University of Texas

12:30–2:00 Lunch

2:00–2:55 **Invited Talk:** *The Role of NLP in an Internet Search Engine*
Jan Pedersen
Director of Advanced Technology, Infoseek

Session 3

2:55–3:20 *The Effect of Topological Structure on Hierarchical Text Categorization*
Stephen D'Alessio, Keitha Murray, Robert Schiaffino and Aaron Kershenbaum
Iona College and Polytechnic University

3:20–3:45 *Refining the Automatic Identification of Conceptual Relations in Large-scale Corpora*
Alex Collier, Mike Pacey and Antoinette Renouf
University of Liverpool

3:45–4:20 Break

Session 4

4:20–4:45 *Generalized Unknown Morpheme Guessing for Hybrid POS Tagging of Korean*
Jeongwon Cha, Geunbae Lee and Jong-Hyeok Lee
Pohang University

4:45–5:10 *Language Identification With Confidence Limits*
David Elworthy
Canon Research Centre Europe

5:10–5:35 *Aligning Tagged Bitexts*
Raquel Martínez, Joseba Abaitua and Arantza Casillas
Universidad Complutense de Madrid, Universidad de Deusto, and Universidad de Alcalá de Henares

5:35–6:00 *Towards Unsupervised Extraction of Verb Paradigms from Large Corpora*
Cornelia H. Parkes, Alexander M. Malek and Mitchell P. Marcus
University of Pennsylvania

Sunday August 16, 1998

Session 5

9:00–9:25 *Can Subcategorisation Probabilities Help a Statistical Parser?*
John Carroll, Guido Minnen and Ted Briscoe
University of Sussex and University of Cambridge

9:25–9:50 *Edge-Based Best-First Chart Parsing*
Eugene Charniak, Sharon Goldwater and Mark Johnson
Brown University

9:50–10:15 *What Grammars tell us about Corpora: the Case of Reduced Relative Clauses*
Paola Merlo and Suzanne Stevenson
University of Pennsylvania and Rutgers University

10:15–10:40 *A Maximum-Entropy Partial Parser for Unrestricted Text*
Wojciech Skut and Thorsten Brants
Universität des Saarlandes

10:40–11:15 Break

Session 6

11:15–11:40 *Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition*
Andrew Borthwick, John Sterling, Eugene Agichtein and Ralph Grishman
New York University

11:40–12:05 *A Statistical Approach to Anaphora Resolution*
Niyu Ge, John Hale and Eugene Charniak
Brown University

12:05–12:30 *A Decision Tree Method for Finding and Classifying Names in Japanese Texts*
Satoshi Sekine, Ralph Grishman and Hiroyuki Shinnou
New York University and Ibaraki University

12:30–2:00 Lunch

2:00–2:55 **Invited Talk:** *Statistical Language Learning in Biological Devices*
Ellisa Newport
George Eastman Professor of Brain and Cognitive Sciences, University of Rochester

Session 7

2:55–3:20 *POS Tagging versus Classes in Language Modeling*
Peter A. Heeman
Oregon Graduate Institute

3:20–3:45 *Automatic Acquisition of Phrase Grammars for Stochastic Language Modeling*
Giuseppe Riccardi and Srinivas Bangalore
AT&T Labs – Research

3:45–4:20 Break

Session 8

4:20–4:45 *Linear Segmentation and Segment Significance*
Min-Yen Kan, Judith L. Klavans and Kathleen R. McKeown
Columbia University

4:45–5:10 *Improving Summarization through Rhetorical Parsing Tuning*
Daniel Marcu
University of Southern California

5:10–5:35 *Discourse Parsing: A Decision Tree Approach*
Tadashi Nomoto and Yuji Matsumoto
Hitachi Ltd. and Nara Institute of Science and Technology

5:35–6:00 *Mapping Collocational Properties into Machine Learning Features*
Janyce M. Wiebe, Kenneth J. McKeever, and Rebecca F. Bruce
New Mexico State University, and Southern Methodist University

Table of Contents

Author Index