Proceedings of the

Conference on Empirical Methods in Natural Language Processing

Sponsored by:

The Association for Computational Linguistics
Department of Computer and Information Science, University of Pennsylvania
Institute for Research in Cognitive Science, University of Pennsylvania
Lexis-Nexis, a Division of Reed Elsevier, Inc.
SIGDAT, a Special Interest Group of the ACL

Edited by

Eric Brill
and Kenneth Church

17-18 May 1996
University of Pennsylvania
Philadelphia, Pa. USA
Proceedings of the

Conference on Empirical Methods in Natural Language Processing

Sponsored by:

The Association for Computational Linguistics
Department of Computer and Information Science, University of Pennsylvania
Institute for Research in Cognitive Science, University of Pennsylvania
Lexis-Nexis, a Division of Reed Elsevier, Inc.
SIGDAT, a Special Interest Group of the ACL

Edited by

Eric Brill
and Kenneth Church

17-18 May 1996
University of Pennsylvania
Philadelphia, Pa. USA
Sponsors:

Association for Computational Linguistics
Department of Computer and Information Science, University of Pennsylvania
Institute for Research in Cognitive Science, University of Pennsylvania
Lexis-Nexis, a Division of Reed Elsevier, Inc.
SIGDAT, a Special Interest Group of the ACL

Organizers:

Eric Brill (Johns Hopkins University)
Kenneth Church (AT&T Research)

Program Committee:

Claire Cardie (Cornell University)
Eugene Charniak (Brown University)
Yuqing Gao (IBM)
Marti Hearst (Xerox PARC)
Gunnel Kallgren (Stockholm University)
Raymond J. Mooney (University of Texas, Austin)
Yoshinori Sagisaka (ATR)
Geoffrey Sampson (Sussex University)
David Yarowsky (Johns Hopkins University)
Joe Zhou (Lexis-Nexis)

For further information, contact:

Eric Brill
brill@cs.jhu.edu
Department of Computer Science
Johns Hopkins University
Baltimore, Md. 21218-2694
TABLE OF CONTENTS

A Geometric Approach to Mapping Bitext Correspondence
I. Dan Melamed ................................................................. 1

Parsing Chinese With an Almost-Context-Free Grammar
Xuanyin Xia and Dekai Wu .................................................. 13

Unsupervised Learning of Syntactic Knowledge: Methods and Measures
R. Basili, A. Marziali, M.T. Pazienza, P. Velardi ......................... 23

Modeling Conversational Speech for Speech Recognition
Marie Meteer and Rukmini Iyer ............................................ 33

Automatic Extraction of New Words from Japanese Texts using Generalized Forward-Backward Search
Masaaki Nagata ................................................................. 48

Better Language Models with Model Merging
Thorsten Brants ................................................................. 60

Combining Hand-crafted Rules and Unsupervised Learning in Constraint-based Morphological Disambiguation
Kemal Oflazer and Gokhan Tur ............................................. 69

Comparative Experiments on Disambiguating Word Senses: An Illustration of the Role of Bias in Machine Learning
Raymond J. Mooney ......................................................... 82

Apportioning Development Effort in a Probabilistic LR Parsing System Through Evaluation
John Carroll and Ted Briscoe .............................................. 92

The Measure of a Model
Rebecca Bruce, Janyce Wiebe, Ted Pedersen ......................... 101

Automating Feature Set Selection for Case-Based Learning of Linguistic Knowledge
Claire Cardie ................................................................. 113

Figures of Merit for Best-First Probabilistic Chart Parsing
Sharon A. Caraballo and Eugene Charniak ............................ 127

A Maximum Entropy Model for Part-Of-Speech Tagging
Adwait Ratnaparkhi ......................................................... 133

Efficient Algorithms for Parsing the DOP Model
Joshua Goodman ............................................................ 143
## AUTHOR INDEX

<table>
<thead>
<tr>
<th>Author</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>R. Basili</td>
<td>23</td>
</tr>
<tr>
<td>Thorsten Brants</td>
<td>60</td>
</tr>
<tr>
<td>Ted Briscoe</td>
<td>92</td>
</tr>
<tr>
<td>Rebecca Bruce</td>
<td>101</td>
</tr>
<tr>
<td>Sharon A. Caraballo</td>
<td>127</td>
</tr>
<tr>
<td>Claire Cardie</td>
<td>113</td>
</tr>
<tr>
<td>John Carroll</td>
<td>92</td>
</tr>
<tr>
<td>Eugene Charniak</td>
<td>127</td>
</tr>
<tr>
<td>Joshua Goodman</td>
<td>143</td>
</tr>
<tr>
<td>Rukmini Iyer</td>
<td>33</td>
</tr>
<tr>
<td>A. Marziali</td>
<td>23</td>
</tr>
<tr>
<td>I. Dan Melamed</td>
<td>1</td>
</tr>
<tr>
<td>Marie Meteer</td>
<td>33</td>
</tr>
<tr>
<td>Raymond J. Mooney</td>
<td>82</td>
</tr>
<tr>
<td>Masaaki Nagata</td>
<td>48</td>
</tr>
<tr>
<td>Kemal Oflazer</td>
<td>69</td>
</tr>
<tr>
<td>M.T. Pazienza</td>
<td>23</td>
</tr>
<tr>
<td>Ted Pedersen</td>
<td>101</td>
</tr>
<tr>
<td>Adwait Ratnaparkhi</td>
<td>133</td>
</tr>
<tr>
<td>Gokhan Tur</td>
<td>69</td>
</tr>
<tr>
<td>P. Velardi</td>
<td>23</td>
</tr>
<tr>
<td>Janyce Wiebe</td>
<td>101</td>
</tr>
<tr>
<td>Dekai Wu</td>
<td>13</td>
</tr>
<tr>
<td>Xuanyin Xia</td>
<td>13</td>
</tr>
</tbody>
</table>