Proceedings of the

Third Conference on Empirical Methods in Natural Language Processing

Sponsored by

The Association for Computational Linguistics
ACL/SIGDAT

Edited by

Nancy Ide
and
Atro Voutilainen

2 June 1998
Palacio de Exposiciones y Congresos
Granada, Spain

Order additional copies from:

ACL
P.O. Box 6090
Somerset, NJ 08875
USA
1-732-873-3898 (phone)
1-732-873-0014 (fax)
acl@aclweb.org

SPONSORS:

The Association for Computational Linguistics (ACL)
SIGDAT (ACL's SIG for Linguistic Data and Corpus-based Approaches to NLP)

INVITED SPEAKER:

Kevin Knight (USC Information Sciences Institute)

ORGANIZERS:

Nancy Ide (Vassar College), Chair
Atro Voutilainen (University of Helsinki), Co-chair

PROGRAM COMMITTEE:

Steven Abney	AT&T Laboratories-Research, USA
Susan Armstrong	(ISCCO, Switzerland)
Pascale Fung	Hong Kong Univ. of Science and Technology, Hong Kong
Gregory Grefenstette	Xerox Research Centre Europe, France
Eduard Hovy	USC/ISI, USA
Dan Jurafsky	University of Colorado, Boulder, USA
Kimmo Koskenniemi	University of Helsinki, Finland
Hwee Tou Ng	DSO National Laboratories, Singapore
Kemal Oflazer	Bilkent University, Turkey
Peter Schauble	ETH Zürich, Switzerland
Keh Yih Su	Tsing-Hua University, Taiwan
Dan Tufis	Romanian Academy of Sciences, Romania
Evelyne Viegas	New Mexico State University, USA

FURTHER INFORMATION:

Nancy Ide
Department of Computer Science
Vassar College
124 Raymond Avenue
Poughkeepsie, New York 12604-0520 USA
email: ide@cs.vassar.edu

FOREWORD

The Third Conference on Empirical Methods in Natural Language Processing offers a general forum for novel research in corpus-based and statistical natural language processing. This year, EMNLP is held in conjunction with the First International Language Resources and Evaluation Conference in Granada, Spain, which is concerned with existing and required resource development to support language processing work in an increasingly multi-lingual setting. Indeed, the development of natural language applications that handle multi-lingual information is the next major challenge facing the field of computational linguistics.

Given this context, this year's EMNLP conference is focused on work that describes and evaluates the strengths, weaknesses, and recent advances in corpus-based NLP as applied to multi-lingual applications. In particular, many of the papers in this volume consider questions such as the following: how well do techniques for lexical tagging, parsing, anaphora resolution, etc., handle the specific problems of multi-lingual applications? What new methods have been developed to address the deficiencies of existing algorithms for these tasks or to address problems specific to handling multi-lingual applications? What problems still lack an adequate empirical solution? Conversely, how can data-driven NLP methods be improved with the help of multi-lingual data?

It is appropriate that this is the first EMNLP conference to be held outside the U.S. We are very encouraged to see the participation of so many researchers from Europe and Asia, which will result, we hope, in greater communication and collaboration across the international NLP community.

Many people are owed thanks for their contributions to setting up this conference. In particular, Atro Voutilainen, EMNLP3 co-chair, and David Yarowksy, SIGDAT chair, provided continual and indispensable help and support throughout. The EMNLP3 Program Committee enabled us to work within a very brief time frame, by quickly turning around all the reviews for the substantial number of submissions to the conference. Finally, the LREC conference organization committee at the University of Granada, the LREC program organizers at the Istituto di Linguistica Computazionale in Pisa, and the Department of Computer Science at Vassar College provided administrative and organizational support. All of them are responsible for the success of EMNLP3.

Nancy Ide, EMNLP3 Chair
Poughkeepsie, New York
May, 1998

TABLE OF CONTENTS

CONFERENCE PROGRAM

8:45 - 9:00	Welcome
9:00 - 9:30	Breck Baldwin and Thomas S. Morton *Dynamic Coreference-Based Summarization*
9:30 - 10:00	Ruslan Mitkov, Lamia Belguith, and Malgorzata Stys *Multilingual Robust Anaphora Resolution*
10:00 - 10:30	Sotiris Boutsis and Stelios Piperidis *Aligning Clauses in Parallel Texts*
10:30 - 11:00	Michel Simard *Automatic Insertion of Accents in French Text*
11:00 - 11:30	Break
11:30 - 12:00	Glenn Carroll and Mats Rooth *Valence Induction with a Head-Lexicalized PCFG*
12:00 - 12:30	Adam Kilgarriff and Tony Rose *Measures for Corpus Similarity and Homogeneity*
12:00 - 12:30	Rebecca Bruce and Janyce Wiebe *Word-Sense Distinguishability and Inter-Coder Agreement*
1:00 - 2:45	Lunch
2:45 - 3:30	Invited Speaker : Kevin Knight (USC Information Systems Institute) *Statistical Translation: Where It Went*
3:30 - 4:00	Stephen D'Alessio, Keitha Murray, Robert Schiaffino, and Aaron Kershenbaum *Category Levels in Hierarchical Text Categorization*
3:30 - 4:00	Fumiyo Fukumoto and Yoshimi Suzuki *An Empirical Approach to Text Categorization based on Term Weight Learning*
4:30 - 5:00	Break
5:00 - 5:30	Shirai Kiyoaki, Inui Kentaro, Tokunaga Takenobu and Tanaka Hozumi *An Empirical Evaluation on Statistical Parsing of Japanese Sentences Using Lexical Association Statistics*
5:30 - 6:00	Fujio Masakazu and Matsumoto Yuji *Japanese Dependency Structure Analysis based on Lexicalized Statistics*
6:00 - 6:30	Adam Berger and Harry Printz *A Comparison of Criteria for Maximum Entropy / Minimum Divergence Feature Selection*