33rd Annual Meeting of the Association for Computational Linguistics

Proceedings of the Conference

26-30 June 1995
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA

Published by the Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Morgan Kaufmann Publishers
340 Pine Street, 6th Floor
San Francisco, CA 94104
1-800-745-7323

PREFACE

This volume contains the papers prepared for the 33rd Annual Meeting of the Association for Computational Linguistics, held from 26–30 June at Massachusetts Institute of Technology in Cambridge, Massachusetts. In a separate section, it also contains the papers that were selected by the student program committee for presentation at the student poster session.

A large number of people in many places have contributed to the success of this year's meeting. First of all I would like to thank our keynote speakers Bonnie Webber and Victor Zue for accepting our invitation to present their exciting research at the meeting. As tutorial chair Yves Shabes organized an attractive and balanced pre-conference program. Special thanks go to the instructors he enlisted: Beth Levin (Northwestern University), Emmanuel Roche (Mitsubishi Electric Research Laboratories), Kenneth Ward Church (AT&T Bell Laboratories), and Lynette Hirschman and Marc Vilain (The Mitre Corporation). Robert Berwick served as the local arrangements chair. Together with his students and staff he put much thought and a tremendous amount of work into the organization of registration, rooms, exhibits, demonstrations, banquet, as well as the general logistics of the conference.

Kathy McKeown, the new Secretary/Treasurer of the ACL, assumed a very active role in the preparation of the conference. In an observant and insistent way she checked on deadlines and formats and made sure that the communication among the different organizers never stalled.

I particularly want to express my gratitude to Joanne Capstick for assisting me in a calm and efficient way during all phases of the preparation of the program and of these proceedings. She did the actual work. Thorsten Brants and other members of our local staff helped with printing and formatting.

The traditional reviewing process of the ACL puts a heavy burden on the members of the program committee. This year each member had to read and evaluate up to 20 papers. They performed this work in a very careful and efficient manner. We followed the tradition of 'blind' reviewing which, although not watertight, had proved very effective in previous years. The final selection was made in an intensive one-day meeting in New York City. I want the members of the program committee to know that their enormous work is greatly appreciated by all of us. It was a great pleasure to work with these delightful colleagues:

> Hans Uszkoreit (Universität des Saarlandes / DFKI, chair), Eric Brill (Johns Hopkins University), Phil Cohen (Oregon Graduate Institute), Ido Dagan (Bar Ilan University), Mary Dalrymple (Xerox PARC), Dan Flickinger (CSLI / Stanford University), Ralph Grishman (New York University), Eva Hajicova (Charles University), Bob Kasper (Ohio State University), Yuji Matsumoto (Nara Institute of Science and Technology), John Maxwell (Xerox PARC), Marie Meteer (BBN Systems and Technologies), Rebecca Passonneau (Columbia University), Fred Popowich (Simon Fraser University), Louisa Sadler (University of Essex), Stephanie Seneff (MIT), Paul Smolensky (Johns Hopkins University), Mark Steedman (University of Pennsylvania), Marilyn Walker (Mitsubishi Electrical Research Laboratories).

The program committee is very grateful to Ralph Grishman and his group at NYU for hosting the program committee meeting and for providing excellent facilities and assistance.

The program committee for the student session, co-chaired by Vasileios Hatzivassiloglou and Johanna Moore, invested much time and energy to make this important part of the conference a success. Before the first call for papers for ACL 95 was issued, the executive committee had engaged in a lengthy and very constructive discussion of the appropriate format for a session at which student members of ACL could present work in progress. This discussion also involved program and student program chairs from this as well as previous conferences. It was decided that a poster session would be the best forum for presenting high-quality research not yet at the state to qualify for the main sessions. We feel that the excellent selection of papers made by the student program committee justifies this decision and we hope that the students will continue to participate actively in our annual meetings.

Hans Uszkoreit, Program Chair
Saarbrücken, Germany
May, 1995

CONFERENCE PROGRAM

MONDAY, 26 JUNE

9:00–12:30 TUTORIAL SESSIONS

Lexical Semantics
Beth Levin

Finite-State Tools for Language Processing
Emmanuel Roche

2:00– 5:30 TUTORIAL SESSIONS

Ngrams
Kenneth Ward Church

Extracting Information from the MUC
Lynette Hirschman and Marc Vilain

7:00–10:00 RECEPTION: CLAMBAKE, outside Johnson Athletic Center, Building W34

TUESDAY, 27 JUNE

8:45– 9:00 OPENING REMARKS AND ANNOUNCEMENTS

9:00– 9:25 *Learning Phonological Rule Probabilities from Speech Corpora with Exploratory Computational Phonology*
Gary Tajchman, Daniel Jurafsky, and Eric Fosler

9:25– 9:50 *Automatic Induction of Finite State Transducers for Simple Phonological Rules*
Dan Gildea and Dan Jurafsky

9:50–10:15 *The Replace Operator*
Lauri Karttunen

10:15–10:40 *A Morphographemic Model for Error Correction in Nonconcatenative Strings*
Tanya Bowden and George Anton Kiraz

11:10–12:15 **INVITED TALK**
Spoken Language Interfaces to Computers: Achievements and Challenges
Victor Zue

1:45– 2:10 *Discourse Processing of Dialogues with Multiple Threads*
Carolyn Penstein Rosé, Barbara Di Eugenio, Lori S. Levin, and Carol Van Ess-Dykema

2:10– 2:35 *Robust Parsing Based on Discourse Information*
Tetsuya Nasukawa

2:35– 3:00 *Corpus Statistics Meet the Noun Compound: Some Empirical Results*
Mark Lauer

3:30– 3:55 *DATR Theories and DATR Models*
Bill Keller

3:55– 4:20 *User-Defined Nonmonotonicity in Unification-Based Formalisms*
Lena Strömbäck

4:20– 4:45 *Features and Agreement*
Samuel Bayer and Mark Johnson

4:45– 5:10 *Encoding Lexicalized Tree Adjoining Grammars with a Nonmonotonic Inheritance Hierarchy*
Roger Evans, Gerald Gazdar, and David Weir

5:10– 7:40 STUDENT POSTER SESSION AND RECEPTION: Lobby, Building 13

WEDNESDAY, 28 JUNE

9:00– 9:25 *Compiling HPSG Type Constraints into Definite Clause Programs*
Thilo W. Goetz and Walt Detmar Meurers

9:25– 9:50 *Compilation of HPSG to TAG*
Robert Kasper, Bernd Kiefer, Klaus Netter, and K. Vijay-Shanker

9:50–10:15 *Memoization of Coroutined Constraints*
Mark Johnson and Jochen Dörre

10:45–11:10 *Combining Multiple Knowledge Sources for Discourse Segmentation*
Diane J. Litman and Rebecca J. Passonneau

11:10–11:35 *Utilizing Statistical Dialogue Act Processing in Verbmobil*
Norbert Reithinger and Elisabeth Maier

11:35–12:00 *Evaluating Automated and Manual Acquisition of Anaphora Resolution Strategies*
Chinatsu Aone and Scott William Bennett

1:30– 2:35 **INVITED TALK**
Exploring Natural Language through Agent Animation
Bonnie Webber

2:35– 3:00 *Investigating Cue Selection and Placement in Tutorial Discourse*
Megan Moser and Johanna D. Moore

3:00– 3:25 *Response Generation in Collaborative Negotiation*
Jennifer Chu-Carroll and Sandra Carberry

3:25– 3:50 *A Uniform Treatment of Pragmatic Inferences in Simple and Complex Utterances and Sequences of Utterances*
Daniel Marcu and Graeme Hirst

4:20– 4:45 *D-Tree Grammars*
Owen Rambow, K. Vijay-Shanker, and David Weir

4:45– 5:10 *The Intersection of Finite State Automata and Definite Clause Grammars*
Gertjan van Noord

5:10– 5:35 *TAL Recognition in $O(M(n^2))$ Time*
Sanguthevar Rajasekaran and Shibu Yooseph

5:35– 6:00 *Extraposition via Complex Domain Formation*
Andreas Kathol and Carl Pollard

7:00–10:00 RECEPTION and BANQUET: The Boston Computer Museum
Presidential Address: Douglas Appelt

THURSDAY, 29 JUNE

9:00– 9:25 *Statistical Sense Disambiguation With Relatively Small Corpora Using Dictionary Definitions*
Alpha K. Luk

9:25– 9:50 *Unsupervised Word Sense Disambiguation Rivaling Supervised Methods*
David Yarowsky

9:50–10:15 *A Quantitative Evaluation of Linguistic Tests for the Automatic Prediction of Semantic Markedness*
Vasileios Hatzivassiloglou and Kathleen McKeown

10:45–11:10 *Quantifier Scope and Consistency*
Jong C. Park

11:10–11:35 *Using Higher-Order Logic Programming for Semantic Interpretation of Coordinate Constructs*
Seth Kulick

11:35–12:30 ANNUAL BUSINESS MEETING, ELECTIONS
Nominations for ACL offices for 1996:
President: Oliviero Stock, IRST
Vice President: Mitch Marcus, University of Pennsylvania
Executive Committee (1996-98): Paul Jacobs, SRA
Nominating Committee (1996-98): Douglas Appelt, SRI

2:00– 2:25 *New Techniques for Context Modeling*
Eric Sven Ristad and Robert G. Thomas III

2:25– 2:50 *Bayesian Grammar Induction for Language Modeling*
Stanley F. Chen

2:50– 3:15 *A Pattern Matching Method for Finding Noun and Proper Noun Translations from Noisy Parallel Corpora*
Pascale Fung

3:15– 3:40 *An Algorithm for Simultaneously Bracketing Parallel Texts by Aligning Words*
Dekai Wu

4:10– 4:35 *Two-Level, Many-Paths Generation*
Kevin Knight and Vasileios Hatzivassiloglou

4:35– 5:00 *An Efficient Generation Algorithm for Lexicalist MT*
Victor Poznański, John L. Beaven and Pete Whitelock

5:00– 5:25 *Some Novel Applications of Explanation-Based Learning to Parsing Lexicalized Tree-Adjoining Grammars*
B. Srinivas and Aravind K. Joshi

5:25– 5:50 *Statistical Decision-Tree Models for Parsing*
David M. Magerman

TUTORIALS

Lexical Semantics

Beth Levin, Northwestern University

There has been a growing interest in lexical semantics over the last ten years. This interest can be traced to the increased importance of the lexicon in many linguistic frameworks; this in turn has led to the serious investigation of the relationship between syntax and word meaning. During the same period the acquisition and representation of lexical information have become central foci of work in computational linguistics. This tutorial will introduce computational linguists to lexical semantics and review recent developments in this area. I will begin by setting out the notion of semantically-coherent verb class, which has proved valuable to the lexical classification of verbs, and by discussing its implications for the representation of lexical knowledge and the lexical semantics-syntax mapping. Organizing principles of the verb lexicon will be covered, as will certain complications presented by polysemy. Finally, I will sketch some similarities and divergences in lexical organization between languages.

Finite-State Tools for Language Processing

Emmanuel Roche, Mitsubishi Electric Research Laboratories

Finite-state tools have been part of Natural Language Processing (NLP) since the beginning. However, whereas many techniques are now considered standard, very few of the most recent mathematical results are known to non-mathematicians or have been incorporated into concrete applications. Moreover, mathematical presentations are often concerned with proofs and not necessarily constructive algorithms; and so it is only recently that many practical algorithmic issues have started to be addressed. As will be seen in this tutorial, contributions of finite-state processing to NLP range from strong efficiency improvements of old techniques to radically new solutions. The tutorial will be self-contained and should be accessible to those with basic computer science and linguistic knowledge. It will first briefly give a list of pointers to the classical results and it will then present recent results, such as decomposition of finite-state transducers and subminimal compression. Each technique will be systematically illustrated through applications to morphology, phonology, and local and global syntax.

Ngrams

Kenneth Ward Church, AT&T Bell Laboratories

Text is more available than ever before: dictionaries, corpora, email, faxes. Many laboratories have tens of millions of words, and some even have billions. What can we do with it all? It is better to do something than nothing at all. We will show some very simple Unix(TM) programs for counting words and ngrams, and generating concordances. Word and ngram counts have been used in a wide variety of applications: part of speech tagging, speech recognition, spelling correction, text compression, word-sense disambiguation, information retrieval, and author identification. In the past, we have tended to focus our attention on bigrams and trigrams. Suffix arrays make it practical to search for 5-grams, 50-grams and even 5000-grams. Long ngrams are far more common than you might have thought. In Genesis, for example, there are hundreds of repeated sequences of 50 characters or more such as: "Be fruitful, and multiply, and replenish the earth." We have to be careful when using long ngrams in prediction tasks; poor estimates of context can be worse than none. How do we estimate the probability of ngrams we haven't seen? How do we combine ngrams of different lengths? Probabilities depend on a variety of hidden variables: topic, author, genre, etc. How do we compensate for the fact that text is more than just a bag of words?

Extracting Information from the MUC

Lynette Hirschman and Marc Vilain, The Mitre Corporation

This tutorial will review the rapid progress in information extraction systems (also called message understanding systems) over the past decade. These systems have grown from fragile toy systems to robust systems achieving recall and precision of around 60% on complex extraction tasks. This progress has been closely linked to the introduction of formal evaluation methods used in the Message Understanding Conferences (MUCs). In the first half of the tutorial, we will trace the evolution from the early MUC conferences to the present, tracking the interplay of evaluation and technological progress, focusing on advances such as robust parsing, special purpose processors (e.g., phrase and name identification), and faster, simpler systems. In the second half of this tutorial, we will focus on portability, highlighting the importance of shared resources (annotated corpora, lexicons, etc.) and corpus-based techniques, including machine learning and statistical approaches. We will consider ways to bootstrap into new domains, to combine rules derived from experts with corpus-based rules, and to reduce the demands of development of training corpora. We will illustrate this in part with recent results at MITRE in transformation-based learning.

TABLE OF CONTENTS

STUDENT SESSION TABLE OF CONTENTS

PREFACE TO THE STUDENT SESSION PAPERS

These proceedings include the extended abstracts for the poster presentations at the Student Session of the *33rd Annual Meeting of the Association for Computational Linguistics.* The goal of the Student Session is to provide a forum for student members to present work in progress, rather than completed work, and to receive feedback from other members of the computational linguistics community, particularly senior researchers. The response to the ACL Student Sessions held during the previous years was very positive. The student authors consistently report that they find the Student Sessions valuable, and answers to questionnaires filled out by ACL members (most recently in 1994) indicate that the audiences find the sessions interesting and of high quality.

This year, a new format has been adopted on a trial basis. Instead of brief ten-minute presentations in parallel sessions, student authors will have the opportunity to present their work in a special poster session, with a higher potential for one-on-one discussions of the details and future directions of their research. Comments on the success of this experiment are actively sought from both the student authors and the ACL membership at large, to determine the optimal format that the sessions should have in the future.

Forty-eight papers were submitted to the ACL Student Session in 1995, thus equalling the previous highest number of submissions, from 1992. Out of these, we were able to accept nineteen papers, an all-time high. The increased number of accepted papers is partly due to the higher flexibility offered by the poster format. Nevertheless, we still had to leave out a number of interesting papers. We thank all the authors for their submissions, and hope that the reviews encourage them in their research, provide constructive criticism, and introduce them to the process of disseminating their work more broadly.

We are grateful to the reviewers for providing helpful, detailed reviews of the submissions under a very tight time schedule. We thank the student members of the ACL 1995 Student Session Program Committee, who helped plan the Student Session and served as reviewers: Jennifer Chu-Carroll, *University of Delaware*; Christy Doran, *University of Pennsylvania*; Philip Edmonds, *University of Toronto*; Mark Lauer, *Macquarie University and Microsoft Institute*; Lillian Lee, *Harvard University*; Carl de Marcken, *Massachusetts Institute of Technology*; Gerrit Rentier, *Tilburg University*; Hinrich Schütze, *Stanford University*; Wojciech Skut, *Universität des Saarlandes*; Mettina Veenstra, *University of Groningen*; Karen Ward, *Oregon Graduate Institute*; and R. Michael Young, *University of Pittsburgh.* Carl de Marcken, in addition to his duties as a Program Committee member, served as liaison to the Local Arrangements Committee; we thank Robert Berwick (*Massachusetts Institute of Technology*) for his help with these issues. We also thank the non-student members of the Student Session Program Committee, who provided reviews and their invaluable experience as senior researchers: Susan Armstrong, *ISSCO*; Nicholas Asher, *University of Texas at Austin*; Bob Carpenter, *Carnegie Mellon University*; Alison Cawsey, *Glascow University*; Michael Elhadad, *Ben Gurion University*; Marti Hearst, *Xerox Parc*; Kevin Knight, *University of Southern California and Information Sciences Institute*; K. Vijay-Shanker, *University of Delaware*; Mark Steedman, *University of Pennsylvania*; and Dekai Wu, *National Technical University of Hong Kong.* In addition, we thank Chung Hee Hwang (*University of Rochester*), Judith Klavans (*Columbia University*), and Gertjan van Noord (*Alfa-Informatica and BCN Groningen*), who served as specialist reviewers. Finally, we are grateful to last year's Student Session chairs, Beryl Hoffman (*University of Pennsylvania*) and Rebecca Passonneau (*Columbia University*), as well as to Philip Resnik (*Sun Microsystems*), for their advice and guidance.

Vasileios Hatzivassiloglou, *Columbia University*, and Johanna Moore, *University of Pittsburgh*
Student Session Co-Chairs

Peter Heeman, *University of Rochester*
Student Session Vice-Chair

AUTHOR INDEX

STUDENT AUTHOR INDEX