Fifth Conference on Applied Natural Language Processing

Association for Computational Linguistics

Proceedings of the Conference

31 March - 3 April 1997
Washington Marriott Hotel
Washington, DC, 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

Little by little, we are learning how to do a better job in building natural language analyzers and generators. Our tool kit is slowly growing — adding, in particular, in the last few years, better tools for learning language patterns from corpora. Of course, our tools are still quite primitive; when we look back at this time in later years we will be amazed at how much we didn't understand about natural language.

But we are also learning to make better use of the tools we do have. We are coming to a better appreciation of how relatively simple tools — morphological analyzers, name recognizers, phrase parsers, to name a few — can be remarkably effective in particular applications. And from this appreciation has flowed a steadily increasing stream of natural language applications.

It is this growing stream that we come here this week to nurture and reflect on. The Conferences on Applied Natural Language Processing are intended to highlight the ways in which natural language processing can be applied to real tasks. With the help of the program committee and other colleagues, we have made a particular effort this year to broaden the range of applications which are presented. Conferences above all are about exchanging ideas, and by stretching the range of the conference we hope to expose people to problems, to techniques, and to applications they might not have seen before. We have also provided an extensive program of demonstrations, ranging from early research prototypes to more mature commercial systems; there is nothing like a live demo to crystallize the problems and accomplishments in our field.

In running an applied conference we are faced forever with the question of what is an "applied" paper. We have chosen to answer that question in an inclusive fashion, including several sessions which address basic technologies such as morphology, parsing, and sense disambiguation, which underlie many of our applications. As we build new applications, we are aware of how shortcomings in these basic technologies affect our design, so it is important to bring together people working on the technologies with those working on the applications.

A conference like this brings together the NLP community not only at the conference itself but also in the months before, soliciting and reviewing papers and planning the sessions. I want to thank the program committee and all the reviewers enlisted for this conference, who prepared reviews of exceptional care and detail; the presenters of the tutorials, technical papers, and demonstrations, who have together prepared the panoply of presentations of natural language processing which we are to savor this week; and John White, the local arrangements chair, and his secretary, Juanita Caldwell, who — together with Priscilla Rasmussen at the ACL office and Kathy McKeown, our secretary-treasurer — planned and organized the conference itself. I want to especially thank John for bearing the added burden of our greatly expanded program of demos this year. I want to acknowledge Victoria Mason and John Sterling at NYU, who dealt with the problems of preparing and printing the papers, and assembling the final proceedings, keeping everything in good order and on schedule. And most of all I want to thank all the authors who submitted papers and entrusted us with the fruit of their research.

Ralph Grishman
Program Chair
New York University, New York, NY

ANLP 97 Program Committee

Ralph Grishman (chair), New York University
Chinatsu Aone, SRA Corp.
Rusty Bobrow, BBN
Martha Evens, Illinois Institute of Technology
Lynette Hirschman, MITRE Corp.
Eduard Hovy, Univ. of Southern California/Information Sciences Institute
Yuji Matsumoto, Nara Institute of Science and Technology
Boyan Onyshkevych, U. S. Dept. of Defense
Tomek Strzalkowski, General Electric Corporate Res. and Dev.
Henry Thompson, Univ. of Edinburgh
Hans Uszkoreit, DFKI Saarbruecken
Marc Vilain, MITRE Corp.

ANLP 97 Reviewers

Steven Abney
Thomas Ahlswede
Chinatsu Aone
Doug Appelt
Scott William Bennett
Hans-Ulrich Block
Sean Boisen
Thorsten Brants
Eric Brill
Jonathan Calder
Jamie Callan
Jean-Pierre Chanod
Lois Childs
Sumali Conlon
Jim Cowie
Robert Dale
Thierry Declerck
Yasuharu Den
Barbara Di Eugenio
Luca Dini
John Dowding
Ted Dunning
Gregor Erbach
Helmut Feldweg
Sharon Flank
Reva Freedman
Pascale Fung
Andy Golding
Greg Grefenstette
Henry Hamburger
Daniel Hardt
Melissa Holland
Helmut Horacek
Pierre Isabelle
Mark Johnson
Doug Jones
Megumi Kameyama
Andrew Kehler
Margaret King
George Kiraz
Tibor Kiss
Lars Konieczny
Kimmo Koskenniemi
Sadao Kurohashi
K.L. Kwok
Mark Light
Kavi Mahesh
Judith Markowitz
Kathleen McCoy
Marie Meteer
Marc Moens
Tetsuya Nasukawa
Klaus Netter
Guenter Neumann
Mary Ellen Okurowski
David Palmer
Rebecca Passonneau
Jose Perez-Carballo
Fred Popowich
Owen Rambow
Lance Ramshaw
Ehud Reiter
Norbert Reithinger
Philip Resnik
Ellen Riloff
Mats Rooth
Salim Roukos
Gerda Ruge
Hinrich Schuetze
Satoshi Sekine
Brian Slator
Alan Smeaton
Harold Somers
Karen Sparck Jones
Beth Sundheim
Koichi Takeda
Axel Theofilidis
Craig Thiersch
Paul Thompson
Junichi Tsujii
Naohiko Uramoto
Takehito Utsuro
Keith Vander Linden
Evelyne Viegas
Marc Vilain
Theo Vosse
Marilyn Walker
Jin Wang
Hideo Watanabe
David Yarowsky

TUTORIALS

Creating and Using Automatic Linguistic Annotation Software

Eric Brill

Department of Computer Science
and
Center for Language and Speech Processing
Johns Hopkins University

In order to perform any sophisticated natural language processing task, it is necessary to first discover the underlying linguistic structure of the input. Depending on the task, this might include information such as parts of speech, word senses, phrase structure, different types of names, etc. Recently a number of approaches have been developed for automatically training programs to provide such annotations. We will survey these approaches and discuss their advantages and disadvantages. The most accurate automatically trained systems typically require large manually-annotated corpora for training, thereby making them expensive to port across domains or languages for which such corpora are not readily available. We will describe methods that allow rapid porting, including: learning without an annotated corpus, adapting an already-trained program to a new domain with minimal resources, and methods for combining human intuitions with automatic acquisition.

Building Applied Natural Language Generation Systems

Ehud Reiter
Computer Science Department
University of Aberdeen, Scotland

and

Robert Dale
Microsoft Institute
Macquarie University, Australia

Natural language generation systems produce understandable texts in English or other human languages from some underlying non-linguistic representation of information. NLG systems combine knowledge about language and the application domain to automatically produce documents, reports, explanations, help messages, and other kinds of texts.

The late 1990s is an exciting time for applied NLG. 10 years ago NLG was purely a research activity, but in 1997 there are several fielded NLG systems in everyday use, and many more systems under development. In this tutorial, we will describe some of the techniques that are being used to build practical working applications today; we will also provide pointers to leading-edge research developments in the field. The material is based around a popular architectural model of NLG that encompasses the three stages of text planning, sentence planning and linguistic realisation. We will include a case study showing how to construct an NLG system which produces textual meteorological summaries from underlying numeric data sets.

The tutorial should be useful for managers, implementors, and researchers. For managers, it will provide a broad overview of the field and what is possible today; for implementors, it will provide a realistic assessment of available techniques; and for researchers, it will highlight the issues that are important in current applied NLG projects.

Using Speech Recognition

Judith Markowitz

President, J. Markowitz Consultants

Talking is a fundamental and ubiquitous mode of communication between humans. The idea of extending speech to verbal interaction with machines has produced powerful icons, such as Arthur Clark's Hal; Kit, the futuristic car; and StarTrek computers.

Researchers and developers have been designing speech recognition systems for almost 50 years, and the fruit of their labor is a growing number of diverse speech-controlled systems, including speech-to-text dictation products, voice-activated dialing systems, and telephone messaging tools.

The presentation addresses three major questions about speech recognition:

- What is speech recognition?
- How does it work?
- What is it used for?

Answers to these questions include examination of speaker modeling, vocabulary creation, grammar, and input channels. The presentation will be accompanied by videotaped examples of existing systems and products.

Building Information Extraction Systems

Douglas E. Appelt and David Israel

Artifical Intelligence Center
SRI International

This tutorial will cover the what and the how of Information Extraction (IE) systems. First we characterize the range of tasks usually intended for IE techniques, and then describe the various approaches to implementing these techniques, discussing the advantages and disadvantages of each. Most IE systems process texts in sequential steps ("phases") ranging from lexical and morphological processing, recognition and typing of proper names, parsing of larger syntactic constituents, and resolution of anaphora and coreference. Finally, IE systems have a domain phase that recognizes events and relationships relevant to the specific IE task. We shall discuss various approaches to each of these phases in turn, and examine their suitability for different types of IE problems. We will discuss the problems and advantages of incorporating various external resources into extraction systems, including large lexicons, gazetteers, and part-of-speech taggers, and conclude with a discussion of template design principles that can have a significant impact on the difficulty of the IE task.

TABLE OF CONTENTS

AUTHOR INDEX

Program At a Glance

Monday, March 31
Tutorials

9:00-12:30	Creating and Using Automatic Linguistic Annotation Software	Building Applied Natural Language Generation Systems
2:00-5:30	Using Speech Recognition	Building Information Extraction Systems

Tuesday, April 1

9:00-9:15	introductions: Ralph Grishman, John White	
9:15-10:00	invited talk: Government Perspectives on the Future of Language Technologies Ruth A. David, Deputy Director, Science and Technology, Central Intelligence Agency	
10:00-10:30	break	
10:30-5:00	*Track A:* Spoken Language and Dialog	*Track B:* Syntax and Morphology

Wednesday, April 2

	Track A	*Track B*
9:00-11:10	Language Learning	Information Extraction
11:10-2:00	Text Checking and other Applications	
2:00-3:30		Document Management
3:30-5:00	demonstrations	

Thursday, April 3

	Track A	*Track B*
9:00-11:00	Text Generation	Multilingual Systems
11:00-1:30		Acquisition of Lexical Information from Corpora
1:30-4:00	Information Retrieval and Summarization	

Tutorial reception: Sunday, March 30th, 6:00-8:00, 2nd Floor
Conference reception: Monday, March 31st, 7:00-10:00, 2nd Floor
Conference banquet: Wednesday, April 2nd, 6:30-10:00 (buses leave the Marriott at 5:30)

Tuesday, April 1: Track A

SPOKEN LANGUAGE and DIALOG		
10:30	CommandTalk: A Spoken-Language Interface for Battlefield Simulations	Robert Moore, John Dowding, Harry Bratt, J. Mark Gawron, Y. Gorfu, and Adam Cheyer
11:00	Natural Language in Four Spatial Interfaces	Kenneth Wauchope, Stephanie Everett, Dennis Perzanowski, and Elaine Marsh
11:20	High Performance Segmentation of Spontaneous Speech Using Part of Speech and Trigger Word Information	Marsal Gavaldá, Klaus Zechner, and Gregory Aist
11:40	A Maximum Entropy Approach to Identifying Sentence Boundaries	Jeffrey Reynar and Adwait Ratnaparkhi
12:00	l u n c h	
1:30	Quickset: Multimodal Interaction for Simulation Set-up and Control	Philip Cohen, Michael Johnston, David McGee, Sharon Oviatt, Jay Pittman, Ira Smith, Liang Chen, and Josh Clow
2:00	Natural Language Dialogue Service for Appointment Scheduling Agents	Stephan Busemann, Thierry Declerck, Abdel Kader Diagne, Luca Dini, Judith Klein, and Sven Schmeier
2:30	Insights into the Dialogue Processing of Verbmobil	Jan Alexandersson , Norbert Reithinger, and Elisabeth Maier
3:00	b r e a k	
3:30	An Evaluation of Strategies for Selective Utterance Verification for Spoken Natural Language Dialog	Ronnie Smith
4:00	Name Pronunciation in German Text-to-speech Synthesis	Stefanie Jannedy and Bernd Möbius
4:30	Applying Repair Processing in Chinese Homophone Disambiguation	Yue-Shi Lee and Hsin-Hsi Chen

Tuesday, April 1: Track B

SYNTAX and MORPHOLOGY		
10:30	A Non-projective Dependency Parser	Pasi Tapanainen and Timo Järvinen
11:00	Incremental Finite-State Parsing	Salah Aït-Mokhtar and Jean-Pierre Chanod
11:30	Developing a Hybrid NP Parser	Atro Voutilainen and Lluís Padró
12:00	l u n c h	
1:30	An Annotation Scheme for Free Word Order Languages	Wojciech Skut, Brigitte Krenn, Thorsten Brants, and Hans Uszkoreit
2:00	The Domain Dependence of Parsing	Satoshi Sekine
2:30	Automatic Acquisition of Two-level Morphological Rules	Pieter Theron and Ian Cloete
3:00	b r e a k	
3:30	Probabilistic and Rule-based Tagger of an Inflective Language - A Comparison	Jan Hajic and Barbora Hladká
4:00	Cseg&Tag1.0: A Practical Word Segmenter and POS Tagger for Chinese Texts	Sun Maosong, Shen Dayang, and Huang Changning

Wednesday, April 2: Track A

COMPUTER-AIDED LANGUAGE LEARNING		
9:00	session introduction	Melissa Holland
9:10	The NLP Role in Animated Conversation for CALL	Michael Schoelles and Henry Hamburger
9:40	Reading more into Foreign Languages	John Nerbonne, Lauri Karttunen, Elena Paskaleva, Gabor Proszeky, and Tiit Roosmaa
10:10	Large-Scale Acquisition of LCS-Based Lexicons for Foreign Language Tutoring	Bonnie Dorr
10:40	break	
TEXT CHECKING & other applications		
11:10	A prototype of a Grammar Checker for Czech	Tomás Holan, Vladislav Kubon, and Martin Plátek
11:40	Techniques for Accelerating a Grammar Checker	Karel Oliva
12:00	lunch	
1:30	EasyEnglish: A Tool for Improving Document Quality	Arendse Bernth
2:00	Contextual Spelling Correction Using Latent Semantic Analysis	Michael Jones and James Martin
2:30	An Automatic Scoring System for Advanced Placement Biology Essays	Jill Burstein, Susanne Wollf, Chi Lu, and Randy Kaplan
3:00	Dutch Sublanguage Semantic Tagging combined with Mark-Up Technology	Peter Spyns, Ngô Thanh Nhàn, Eric Baert, Naomi Sager, and Georges De Moor

Wednesday, April 2: Track B

INFORMATION EXTRACTION		
9:00	session introduction	Sarah Taylor
9:10	A Statistical Profile of the Named Entity Task	David Palmer and David Day
9:30	Nymble: a High-Performance Learning Name-Finder	Daniel Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel
10:00	Disambiguation of Proper Names in Text	Nina Wacholder, Yael Ravin and Misook Choi
10:30	b r e a k	
11:00	An Information Extraction Core System for Real World German Text Processing	Günter Neumann, Rolf Backofen, Judith Baur, Markus Becker, and Christian Braun
11:30	Layout & Language: Preliminary experiments in assigning logical structure to table cells	Matthew Hurst and Shona Douglas
11:50	l u n c h	
1:30	Building a Generation Knowledge Source using Internet-Accessible Newswire	Dragomir Radev and Kathleen McKeown
DOCUMENT MANAGEMENT		
2:00	Using SGML as a Basis for Data-Intensive NLP	David McKelvie, Chris Brew, and Henry Thompson
2:30	Software Infrastructure for Natural Language Processing	Hamish Cunningham, Kevin Humphreys, Robert Gaizauskas, and Yorick Wilks
3:00	An Open Distributed Architecture for Reuse and Integration of Heterogenous NLP Components	Rémi Zajac, Mark Casper, and Nigel Sharples

Thursday, April 3: Track A

TEXT GENERATION		
9:00	Customizable Descriptions of Object-Oriented Models	Benoit Lavoie, Owen Rambow, and Ehud Reiter
9:30	CogentHelp: NLG meets SE in a tool for authoring dynamically generated on-line help	Michael White and David Caldwell
10:00	A Fast and Portable Realizer for Text Generation Systems	Benoit Lavoie and Owen Rambow
10:30	break	
11:00	Multilingual Generation and Summarization of Job Adverts: the TREE Project	Harold Somers, Bill Black, Joakim Nivre, Torbjörn Lager, Annarosa Multari, Luca Gilardoni, Jeremy Ellman, and Alex Rogers
11:30	Language Generation for Multimedia Healthcare Briefings	Kathleen R. McKeown, Shimei Pan, James Shaw, Desmond Jordan, and Barry A. Allen
12:00	lunch	
INFORMATION RETRIEVAL AND SUMMARIZATION		
1:30	session introduction	Donna Harman
1:40	Identifying Topics by Position	Chin-Yew Lin and Eduard Hovy
2:10	An Automatic Extraction of Key Paragraphs Based On Context Dependency	Fumiyo Fukumoto, Yoshimi Suzuki, and Jun'ichi Fukumoto
2:40	Building Effective Queries in Natural Language Information Retrieval	Tomek Strzalkowski, Fang Lin, Jose Perez-Carballo, and Jin Wang
3:10	break	
3:30	Construction and Visualization of Key Term Hierarchies	Joe Zhou and Troy Tanner
4:00	Fast Statistical Parsing of Noun Phrases for Document Indexing	Chengxiang Zhai

Thursday, April 3: Track B

MULTILINGUAL SYSTEMS		
9:00	An English to Turkish Machine Translation System Using Structural Mapping	Cigdem Turhan
9:30	An Interactive Translation Support Facility for Non-Professional Users	Kiyoshi Yamabana, Kazunori Muraki, Shin-ichiro Kamei, Kenji Satoh, Shinichi Doi, and Shinko Tamura
10:00	An Intelligent Multilingual Information Browsing and Retrieval System Using Information Extraction	Chinatsu Aone, Nicholas Charocopos, and James Gorlinsky
10:30	break	
ACQUISITION of LEXICAL INFORMATION from CORPORA		
11:00	Semi-automatic Acquisition of Domain-specific Translation Lexicons	Philip Resnik and I. Dan Melamed
11:30	Mixed-Initiative Development of Language Processing Systems	David Day, John Aberdeen, Lynette Hirschman, Robyn Kozierok, Patricia Robinson, and Marc Vilain
12:00	lunch	
1:30	Automatic Extraction of Subcategorization from Corpora	Ted Briscoe and John Carroll
2:00	Learning Probabilistic Subcategorization Preference by Identifying Case Dependencies and Optimal Noun Class Generalization Level	Takehito Utsuro and Yuji Matsumoto
2:30	A Workbench for Finding Structure in Texts	Andrei Mikheev and Steven Finch
3:00	break	
3:30	Automatic Selection of Class Labels from a Thesaurus for an Effective Semantic Tagging of Corpora	Alessandro Cucchiarelli and Paola Velardi
4:00	Sequential Model Selection for Word Sense Disambiguation	Ted Pedersen, Rebecca Bruce, and Janyce Wiebe