LANGUAGE TECHNOLOGIES 2001:
SECOND CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS

June 2 – June 7, 2001
Carnegie Mellon University
Pittsburgh, Pennsylvania USA

Invited Talks

Tom Mitchell
Carnegie Mellon University and WhizBang!

Aravind Joshi
University of Pennsylvania

Jon Kleinberg
Cornell University

SPECIAL EVENTS

2001 Conference on Empirical Methods in Natural Language Processing (EMNLP)

Full-day Industrial Exhibit Session

Student Research Workshop

PLEASE WATCH THIS WEBSITE
FOR FURTHER UPDATES:
http://www.cs.cmu.edu/~ref/naacl2001.html

NOTICE CONTENTS

REGISTRATION INFORMATION

SITE: NAACL 2001 and all co-located events will be held at the University Center of Carnegie Mellon University in Pittsburgh, Pennsylvania. Carnegie Mellon is located in the Oakland section of town, about 3 miles from downtown Pittsburgh.

REGISTRATION and MEETINGS: Registration includes one hardcopy of the Proceedings, one CD-ROM of the proceedings, as well as admission to all exhibits, demos, and the Student Research Workshop. The main conference and EMNLP will take place in the McConomy Auditorium on the first floor of the University Center. The workshops and tutorials will take place in the Rangos Ballrooms on the 2nd floor. For up-to-date program information, check the official listing of events near the registration desk and outside of each event room. Registration will be located on the , first floor of the University Center, in front of the conference auditorium. All registrations (main sessions, tutorials and workshops) will take place from 8:00am-5:00 pm.

•***Tutorials:*** Saturday, June 2 •***Workshops/ EMNLP:*** Sunday and Monday, June 3-4 •***Main Conference:*** Tuesday - Thursday, June 5-7

STUDENT RESEARCH WORKSHOP: Once again this year, student members of NAACL will be presenting their exciting work-in-progress at a Student Research Workshop. Registration for the workshop is included in your conference registration fee. This full-day workshop will take place on Monday, June 4th. Our review committee has selected six excellent student papers for presentation at the workshop based on their scholarship, originality, and technical merit. These papers cover many areas of NLP, including:

- document categorization,
- statistical parsing and tagging,
- question answering,
- corpus NLP,
- automatic ontology and lexicon construction.

In addition to audience comments, a panel of established scientists, each an expert in areas relevant to the student presentations, will be chosen to provide the students with in-depth feedback and suggestions on future directions, similar to the highly-acclaimed Doctoral Consortia at other conferences. This format debuted last year with great success and is intended to provide students with invaluable exposure to outside perspectives on their work and allow them to put their work into perspective based on feedback from the panel. Students in the audience can also benefit greatly by comparing relevant comments to their own developing work. If you would like to be considered for the scientific panel, please contact the workshop co-chairs at kczuba@cs.cmu.edu or michaud@cis.udel.edu.

PLEASE NOTE: although attendance to this workshop comes at no extra charge, pre-registration is strongly encouraged. Up-to-date information, including the titles and authors of the selected papers, is available at the URL: http://www.eecis.udel.edu/~aclstu/naacl01-student/.

EMNLP-2001: There will also be a co-located meeting of the *2001 Conference on Empirical Methods in Natural Language Processing (EMNLP).* EMNLP participants are required to register for the main NAACL meeting as well. EMNLP will take place on Sunday and Monday, June 3 and 4 in the McConomy Auditorium in the Carnegie Mellon University Center. For further information, please contact Lillian lee, email: llee@cs.cornell.edu or see the web page http://www.cs.cornell.edu/home/llee/emnlp.html.

CO-LOCATED WORKSHOP: The non-ACL affiliated Workshop on Language Modeling and Information Retrieval will be held at the same location on campus on the two days immediately preceding the NAACL conference (Thursday and Friday, May 31—June 1). Registration and all arrangements for this workshop are handled separately from NAACL. For further information, contact Jamie Callan, email: callan@cs.cmu.edu , or see the workshop web page at: http://la.lti.cs.cmu.edu/callan/Workshops/lmir01/ .

HOTEL RESERVATIONS: The NAACL has reserved a block of rooms at the Pittsburgh Holiday Inn Select – University Center Hotel. Our special conference rate is $109 per night. The hotel is located within a 15-minute walking distance from the Carnegie Mellon Campus. The hotel shuttle will also be available for commuting guests to and from the conference location on campus. For reservations, fill out the attached Hotel Registration form and fax it directly to the hotel at Fax**:** (412) 681-4749. You may alternatively call the hotel directly at Tel**:** (412) 682-6200 Toll-Free**:** (800) 864-8287. To receive the guaranteed conference rate, reservations must be made by May 18.

ON CAMPUS HOUSING: we have also reserved a block of 150 modern air-conditioned dormitory rooms on campus for accommodation during the conference. Both single occupancy and double occupancy rooms are available. To apply, please fill out the On-Campus Housing Registration Form and fax it to the CMU Conference Services Office at Fax: (412) 268-5718. Registrations will be taken on a first-come first-serve basis up to the deadline of May 18th.

EXHIBITS: A special all-day Exhibit Session, highlighting the latest exciting trends in Language Technologies in the high-tech industrial world, will be one of the special features of the conference. Exhibits will take place in the Rangos Ballroom on the second floor of the University Center on Wednesday, June 6th. To participate in this special exhibit session, apply online at the conference website, or contact Lynn Carlson, email: lmcarls@super.org , phone: +1-301-805-7477 fax: +1-301-805-7602.

EMAIL ROOM: The email room will be located on the 2nd floor and available throughout the conference during normal conference hours. It will include ten PCs with internet access and a number of Ethernet ports for connecting laptops via DHCP directly to the CMU local network.

WIRELESS NETWORK ACCESS: CMU has a campus wide wireless LAN. Conference attendees that have a WaveLAN Silver or similar 802.11 card for their laptop can have a wireless network connection activated for the duration of the conference. Those interested in this service must pre-register their machines and their MAC address (written on the back of the card) by no later than May 18. For information, please contact Alan Black, email: awb@cs.cmu.edu.

DEMONSTRATIONS: Demonstrations of NLP systems will be featured in the Connan room on the first floor, behind the conference Registration Desk. For information on demonstrations, contact Ronnie Smith, East Carolina University, email: rws@cs.ecu.edu .

CONFERENCE OPENING RECEPTION: The opening reception will be held on Monday evening, June 4th, 7:00-9:00pm, in the Rangos Ballroom on the second floor of the University Center.

BANQUET: The Conference Banquet will take place at the Carnegie Museum of Natural History on Wednesday, June 6th, 7:00-11:00pm. The Museum is located on Forbes Avenue, the main road between the CMU campus and the Holiday Inn Hotel. It is about a ten minute walk from both the hotel and the campus conference location. The famous collection of dinosaurs is anxiously awaiting our visit!

MEALS and BREAKS: Complimentary beverages and snacks will accompany the Workshops, Tutorials, Main Session, and EMNLP. For lunch and other major meals, several on and off-campus options will be available. Information about dining choices will be available at the conference, and can be found in advance of the conference on the conference web site. The selection of food in Pittsburgh is almost as diverse as the neighborhoods that make up the city. There is something to be found for all tastes. Just don't be surprised if you end up with French fries in your salad or on your sandwich.

SMOKING POLICY: Carnegie Mellon and its University Center are smoke-free environments. No smoking is permitted inside the meeting rooms or hallways. Smoking is also not permitted inside any of the on-campus dormitory rooms. The Holiday Inn hotel has both smoking and non-smoking rooms. Smoking is prohibited in restaurants, meeting facilities, and public areas. First-time visitors to the US should be alert for restrictions on smoking in most buildings and public facilities.

RECREATION: The Holiday Inn hotel has a swimming pool, sauna and exercise facilities. There is no extra charge for the use of these facilities. Modern exercise and recreation facilities are also available in the CMU University Center, literally adjacent to the conference meeting rooms. These include a swimming pool, Squash and Racket Ball courts, tennis courts, an outdoor track, a fitness center as well as a sauna and Jacuzzi. Daily access passes can be purchased at the University Center Information Desk for the price of $5 per day. Schenley Park, a large city natural park, borders the campus to the East, and has numerous walking and jogging trails.

CHILDCARE: The hotel and the conference location do not provide any organized childcare arrangements. Conference attendees interested in information about possible childcare solutions should contact Lori Levin, email: lsl@cs.cmu.edu.

GEOGRAPHIC SITUATION AND CLIMATE: Pittsburgh is located at the head of the Ohio River, which is formed by the confluence of the Allegheny River and the Monongahela River. It is also situated near the Allegheny Mountains, part of the Appalachian mountain chain separating the East Coast and the Great Plains. The city of Pittsburgh is made up of almost 90 distinct neighborhoods, each with it's own personality and charm. Early June is typically sunny with average daily highs in the low 70's Fahrenheit (low 20's Celsius), but come prepared for cooler evenings, hotter than normal days or a sudden late spring shower.

SIGHTSEEING: The conference site is located in the neighborhood of Oakland on the Carnegie Mellon University campus. The Carnegie Museums of Art and Natural History and the campus of the University of Pittsburgh are located a short walk away. Don't miss the view of downtown and Point State Park from the overlooks on Mount Washington. See the NAACL web site for suggestions and web links on things to see and do in Pittsburgh.

PARKING: Parking is available at the Holiday Inn for $14 per day. On campus parking will be available in the covered East Campus Garage which is adjacent to the University Center. Daily parking passes can be purchased on site for $5.50 per day.

MONEY AND FOREIGN CURRENCY EXCHANGE: Several commercial bank branches are located in the vicinity of the hotel and campus, and may be able to handle some foreign currency exchanges. ATMs are available in the University Center on campus and in nearby bank branches. Many ATM machines impose a surcharge ($1-3) on transactions.

ELECTRICAL POWER: It is Conference policy to use the local power source. At this conference site it is 110V 60Hz AC. Please plan ahead: power converters, extension cords and power strips will not be provided by the NAACL.

PROGRAM COMMITTEES: The **NAACL Senior Program Committee** this year included the following people: Kevin Knight, Chair (USC/ISI), Eric Brill (Microsoft Research), Ann Copestake (Cambridge University and CSLI), Marti Hearst (UC Berkeley), Aravind Joshi (University of Pennsylvania), Andrew Kehler (UC San Diego), Elliott Macklovitch (University of Montreal), Fernando Pereira (WhizBang! Labs), Owen Rambow (AT&T Research), Elizabeth Shriberg(SRI), and Ralph Weischedel (BBN). The Program Committee was aided in their decision-making by large international panels of reviewers.

STUDENT RESEARCH WORKSHOP PROGRAM COMMITTEE: Committee members are: Krzysztof Czuba (Carnegie Mellon University) and Lisa Michaud (University of Delaware), Co-Chairs, and Deborah Dahl (Unisys Corporation), Faculty Advisor. **Student Committee Members*:*** Donna Byron (University of Rochester), John Chen (University of Delaware), Anne R. Diekama (Syracuse University), Mary Ellen Foster (University of Edinburgh), Janna Hamaker (Mississippi State University), Maria Lapata (University of Edinburgh), Mary Xiaoyong Liu (Syracuse University), Kathleen Murray (University of Pennsylvania), Jill Nickerson (Harvard University), Lynellen Perry (Mississippi State University), Brian Roark (Brown University), Amanda Stent (University of Rochester), Klaus Zechner (Carnegie Mellon University). **Non-student Committee Members*:*** Lois Boggess (Mississippi State University), Michael Brent (Washington University in St. Louis), Jennifer Chu-Carroll (IBM), Carolyn Penstein Rose (University of Pittsburgh), Yael Ravin (IBM).

DIRECTIONS:

Air: Pittsburgh's airport (http://www.pitairport.com) is approximately 20 miles west of Oakland. A one-way taxi cab fare will range from $35-$40.

- Port Authority Transit (http://www.portauthority.org/) offers bus service between the airport, downtown and Carnegie Mellon. This service is called the 28X or the airport flyer. There is a small luggage rack for baggage and you will be responsible for your own baggage. There are stops near the conference hotel and Carnegie Mellon University. The bus runs 7 days a week and is scheduled to run every 45 minutes. The cost is $1.95 each way.

- Shuttle service from the airport is also available.

Automobile: The conference site is best approached from the north or south via Interstate 79, i.e. I-79, and from the east via the Pennsylvania Turnpike (I-76). From the north or south: From I-79, exit onto I-279. From I-279, exit onto I-376 East. From I-376, take the Oakland Exit 5 to Forbes Avenue, which runs past Carnegie Mellon and near the hotel. If you are driving from the airport, follow RT 60 to I-279, and proceed as above. From the east: From the Turnpike (I-76), exit onto I-376 at Monroeville. From I-376, take the Squirrel Hill Exit 8 (just after leaving the Squirrel Hill tunnel). Take the first left after the exit to the first stoplight (a five-way intersection). Turn left onto Murray Ave. Follow Murray Avenue to Forbes Avenue, and turn left onto Forbes.

Please see the conference website for additional information.

GETTING AROUND: A number of *Port Authority Transit buses* have stops near the Carnegie Mellon campus and the hotel. Fares are generally $1.60, with higher fares of $2.00 and $2.50 for destinations further from the city. Transfers are $0.25.

NAACL-2001 TECHNICAL SESSIONS PRELIMINARY PROGRAM

TUESDAY, JUNE 5

8:30-5:00 CONFERENCE REGISTRATION

8:45-9:05 OPENING REMARKS

SESSION 1: Natural Language Generation

9:05-9:30 *Instance-Based Natural Language Generation*
Sebastian Varges, Chris Mellish
9:30-9:55 *Corpus-based NP Modifier Generation*
Hua Cheng, Massimo Poesio, Renate Henschel, Chris Mellish
9:55-10:20 *A Trainable Sentence Planner*
Marilyn A. Walker, Owen C. Rambow, Monica Rogati

10:20-11:00 BREAK

SESSION 2: Information Retrieval and Machine Learning

11:00-11:25 *Why Inverse Document Frequency?*
Kishore Papineni
11:25-11:50 *Question Answering Using Maximum-Entropy Components*
Abraham Ittycheriah, Martin Franz, Wei-Jing Zhu, Adwait Ratnaparkhi
11:50-12:15 *Transformation Based Learning in the Fast Lane*
Grace Ngai, Radu Florian

12:15-2:00 LUNCH

2:00-3:00 INVITED TALK: Tom Mitchell (CMU)

SESSION 3: Natural Language Dialogue

3:05-3:30 *Identifying User Corrections Automatically in Spoken Dialogue Systems*
Julia Hirschberg, Diane Litman, Marc Swerts
3:30-3:55 *Learning Optimal Dialogue Management Rules by Using Reinforcement Learning and Inductive Logic Programming*
Renaud Lecœuche

3:55-4:30 BREAK

SESSION 4: Word Meaning

4:30-4:55 *A Corpus-based Account of Regular Polysemy: The Case of Context-Sensitive Adjectives*
Maria Lapata
4:55-5:20 *Tree-Cut and a Lexicon Based on Systematic Polysemy*
Noriko Tomuro
5:20-5:45 *A Decision Tree of Bigrams is an Accurate Predictor of Word Sense*
Ted Pedersen

WEDNESDAY, JUNE 6

8:30-5:00 CONFERENCE REGISTRATION

SESSION 5: Semantics

9:05-9:30 *An Algorithm for Aspects of Semantic Interpretation Using an Enhanced WordNet*
Fernando Gomez
9:30-9:55 *Class-Based Probability Estimation Using a Semantic Hierarchy*
Stephen Clark, David Weir

9:55-10:20 *Identifying Cognates by Phonetic and Semantic Similarity*
Grzegorz Kondrak

10:20-11:00 BREAK

SESSION 6: Speech Synthesis and Recognition

11:00-11:25 *Re-engineering Letter-to-Sound Rules*
Martin Jansche

11:25-11:50 *Edit Detection and Parsing for Transcribed Speech*
Eugene Charniak, Mark Johnson

11:50-12:15 *Generating Training Data for Medical Dictations*
Sergey Pakhomov, Michael Schonwetter, Joan Bachenko

12:15-2:00 LUNCH

2:00-3:00 NAACL BUSINESS MEETING

3:00-4:00 INVITED TALK: Jon Kleinberg (Cornell)

4:00-4:30 BREAK

SESSION 7: Machine Translation

4:30-4:55 *A Finite-State Approach to Machine Translation*
Srinivas Bangalore, Giuseppe Riccardi

4:55-5:20 *Information-Based Machine Translation*
Keiko Horiguchi

5:20-5:45 *Multipath Translation Lexicon Induction*
Gideon S. Mann, David Yarowsky

THURSDAY, JUNE 7

8:30-12:00 CONFERENCE REGISTRATION

SESSION 8: Parsing

9:05-9:30 *A Probabilistic Earley Parser as a Psycholinguistic Model*
John Hale

9:30-9:55 *Refining Tabular Parsers for TAGs*
Eric Villemonte de la Clergerie

9:55-10:20 *Applying Co-Training Methods to Statistical Parsing*
Anoop Sarkar

10:20-11:00 BREAK

SESSION 9: Language Modeling

11:00-11:25 *A Structured Language Model Based on Context-Sensitive Probabilistic Left-Corner Parsing*
Dong Hoon Van Uytsel, Dirk Van Compernolle, Filip Van Aelten

11:25-11:50 *Do CFG-Based Language Models Need Agreement Constraints?*
Manny Rayner, Genevieve Gorrell, Beth Ann Hockey, John Dowding, Johan Boye

11:50-12:15 *Naive Bayes Detection of Non-Native Utterances*
Laura Mayfield Tomokiyo, Rosie Jones

12:15-2:00 LUNCH

2:00-3:00 INVITED TALK: Aravind Joshi (University of Pennsylvania)

SESSION 10: Names and Coreference

3:05-3:30 *Unsupervised Learning of Name Structure From Coreference Data*
Eugene Charniak

3:30-3:55 *Text and Knowledge Mining for Coreference Resolution*
Sanda Harabagiu, Razvan Bunescu, Steve Maiorano

3:55-4:30 BREAK

SESSION 11: Morphology and Chunking

4:30-4:55 *Knowledge-Free Induction of Inflectional Morphologies*
Patrick Schone, Daniel Jurafsky

4:55-5:20 *Chunking with Support Vector Machines*
Taku Kudo, Yuji Matsumoto

5:20-5:45 *Inducing Multilingual POS Taggers and NP Bracketers via Robust Projection Across Aligned Corpora*
David Yarowsky, Grace Ngai

TUTORIALS
Saturday 2 June 2001, 9:00AM--5:00PM

"How May I Help You?": Automated Customer Service via Natural Spoken Dialog
9:00AM-12:30PM
Alicia Abella, Allen Gorin, Guiseppe Riccardi, Tirso Alonso, Jerry Wright, AT&T Shannon Laboratory

The next generation of voice-based user interfaces will enable easy-to-use automation of new and existing communication services. A critical issue is to move away from highly-structured menus to a more natural human-machine paradigm. In this tutorial we will cover the large vocabulary speech recognition, language modeling, spoken language understanding, dialog manager and logging functionalities of our system. We will show how finite state representation and stochastic modeling provide rich tools to model different language models: n-grams, word phrases, word classes, phrase grammars. We will also present our latest results on automatically learned head-dependency grammars and speech disfluency-based language models.

The Spoken Language Understanding (SLU) is based on salient grammar fragments, acquired automatically from a corpus of transcribed and labelled training utterances. Each grammar fragment represents a cluster of similarly-meaningful phrases, represented as a finite state machine. Matches of these to the recognizer output for a test utterance are grouped in semantically coherent ways, and the best interpretation of the utterance is found, taking account of dialog context. Based on the output of the SLU the dialog manager needs to determine whether to ask the customer a question, create a database query, transfer a call, etc. The dialog manager is flexible enough to be utilized in a wide variety of applications. The dialog manager is built from general dialog principles that are captured quantitatively using a Construct Algebra and a task representation that not only structures the task knowledge but also influences the behavior of the dialog manager and utilizes the object-oriented paradigm.

The HMIHY platform has an extensive array of instrumentation built-in to track its internal operation. The collected information is logged in files for later analysis. The analysis tool used on these log files is object oriented, modular, reusable, and extensible. The collected data includes such things as prompts selected by the dialog manager to play, audio fed to the ASR engine, etc. Each of the aforementioned components was initially integrated into a prototype in 1997 that automated over 10,000 customer requests for operator services. This year a wizard-of-oz version of the system for customer care conducted more than 25,000 dialogs. Based on this data collection, a fully autonomous system has been deployed in the AT&T network to handle customer care requests.

Empirical Methods in Natural Language Processing: What's Happened Since the First SIGDAT Meeting?
9:00AM-12:30PM
Kenneth Ward Church, AT&T Labs-Research

The first workshop on Very Large Corpora was held just before the 1993 ACL meeting in Columbus Ohio. The turnout was even greater than anyone could have predicted (or else we would have called the meeting a conference rather than a workshop). We knew that text analysis was a ''hot area," but we didn't appreciate just how hot it would turn out to be.

The 1990s were witnessing a resurgence of interest in 1950s-style empirical and statistical methods of language analysis. Empiricism was at its peak in the 1950s, dominating a broad set of fields ranging from psychology (behaviorism) to electrical engineering (information theory). At that time, it was common practice in linguistics to classify words not only on the basis of their meanings but also on the basis of their co-occurrence with other words. Firth, a leading figure in British linguistics during the 1950s, summarized the approach with the memorable line: "You shall know a word by the company it keeps." Regrettably, interest in empiricism faded in the late 1950s and early 1960s with a number of significant events including Chomsky's criticism of n-grams in "Syntactic Structures" (Chomsky, 1957) and Minsky and Papert's criticism of neural networks in "Perceptrons" (Minsky and Papert, 1969).
Perhaps the most immediate reason for this empirical renaissance is the availability of massive quantities of data: text is available like never before. Just ten years earlier, the one-million word Brown Corpus (Francis and Kucera, 1982) was considered large, but these days, everyone has access to the web. Experiments are routinely carried out on many gigabytes of text. Some researchers are even working with terabytes.

The big difference since the first SIGDAT meeting in 1993 is that large corpora are now having a big impact on ordinary users. Web search engines/portals are an obvious example. Managing gigabytes is not only the title of a popular (Moffat, Bell and Witten, 1999), but it is something that ordinary users are beginning to take for granted. Recent progress in Information Retrieval and Digital Libraries was worth a fortune (when stock prices were at their peak). Speech Recognition and Machine Translation are also changing the world. If you walk into any software store these days, you will find a shelf full of speech recognition and machine translation products. And it is getting so you can't use the telephone these days without talking to a computer.

Building Synthetic Voices
2:00PM-5:30PM
Alan W Black and Kevin A. Lenzo, Carnegie Mellon University

This tutorial will give an overview of the basic techniques available for building synthetic voices for speech synthesis systems, including an actual example of voice building.

The first part will describe the basic components of a speech synthesis system covering the state of the art techniques used within them. Specifically:

- Text Analysis: addressing issues of expansions of symbols, numbers, acronyms etc and resolving homographs

- Linguistic Analysis: "from words to how to say them", addressing issues in lexical entries, letter to sound rules and prosodic modeling, (phrasing, intonation and duration).

- Waveform Synthesis: "from phones and prosody to waveforms" describing basic techniques for making computers talk using recorded prompts, diphones, and general unit selection synthesis

The second part will describe the basic stages required in building new synthetic voices (in English or other languages):

- building a text analysis system
- building a lexicon and letter to sound rules
- build phrasing, intonation and duration models
- recording data for concatenative speech synthesis (diphones, unit selection and/or limited domain)

This tutorial is based on the techniques, documentation and tools freely distributed through CMU's FestVox project (http://festvox.org/) leading to voices that can be run on Edinburgh University's Festival Speech Synthesis System.

Open-Domain Textual Question Answering
2:00PM-5:30PM
Sanda Harabagiu and Dan Moldovan, Southern Methodist University

Question Answering (QA) is a fast growing area of research and commercial interest. The problem of QA is to find answers to open-domain questions by searching a large collection of documents. Unlike Internet search engines, QA systems provide short, relevant answers to questions. The recent explosion of information available on the World Wide Web makes question answering a compelling framework for finding information that closely matches user needs. The success of QA services, like AskJeeves serves as proof of the popularity of this technique. Due to the fact that both questions and answers are expressed in natural language, QA methodologies deal with language ambiguities and incorporate NLP techniques. Several current NLP-based technologies are able to provide the framework that approximates the complex problem of answering questions from large collections of texts. Ideal QA systems should have good dialog understanding, rich knowledge bases and quality text mining methods. They will certainly incorporate common sense reasoning methods and use good approximations of world knowledge. Until we have these more advanced tools, we can approximate QA with NLP enhancements of IR and IE techniques. The tutorial presents the recent results in QA research and system implementations.

WORKSHOPS – Sunday, 3 June 2001

Workshop on Automatic Summarization
Jade Goldstein and Chin-Yew Lin

The problem of automatic summarization poses a variety of tough challenges in both NL understanding and generation. A spate of recent papers and tutorials on this subject at conferences such as ACL, ANLP/NAACL, ACL/EACL, AAAI, ECAI, IJCAI, and SIGIR point to a growing interest in research in this field. Several commercial summarization products have also appeared. There have been several workshops in the past on this subject: Dagstuhl in 94, ACL/EACL in 97, the AAAI Spring Symposium in 98, and ANLP/NAACL in 2000. All of these were extremely successful, and the field is now enjoying a period of revival and is advancing at a much quicker pace than before. NAACL'2001 is an ideal occasion to host another workshop on this problem.

Workshop on MT Evaluation: Hands-On Evaluation
Eduard Hovy and Florence Reeder

Evaluation of language tools, particularly tools that generate language, remains an interesting and general problem. Machine Translation (MT) is a prime example. Approaches to evaluating MT are even more plentiful than approaches to MT itself; the number of evaluations and range of variants is confusing to anyone considering an evaluation. In an effort to systematize MT evaluation, the NSF-funded ISLE project has created a taxonomy of evaluation-related features and measures. Unfortunately, however, many prior evaluations do not include an adequate specification of important aspects such as evaluation process complexity, cost, variance of score, etc.

In an effort to drive MT evaluation to the next level, this workshop will focus on exercising with methods of acquiring such information for several important MT evaluation measures. The workshop thus embodies the challenge of Hands-On Evaluation, within the context of the framework being developed by the ISLE MT Evaluation effort. The workshop follows a workshop on MT Evaluation held at the AMTA Conference in Cuernavaca, Mexico, in October 2000, and a subsequent workshop being planned for April 2001 in Geneva.

Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations (Day 1)
Dan Moldovan, Sanda Harabagiu, Wim Peters, Mark Stevenson, and Yorick Wilks

This is a two-day merged workshop of the two previously announced workshops: WordNet: Extensions and NLP Applications; and Customizing Lexical Resources.

Lexical resources have become important basic tools within NLP and related fields. The range of resources available to the researcher is diverse and vast - from simple word lists to complex MRDs and thesauruses. The resources contain a whole range of different types of explicit linguistic information presented in different formats and at various levels of granularity. Also, much information is left implicit in the description, e.g. the definition of lexical entries generally contains genus, encyclopaedic and usage information.
The majority of resources used by NLP researchers were not intended for computational uses. For instance, MRDs are a by-product of the dictionary publishing industry, and WordNet was an experiment in modelling the mental lexicon. In particular, WordNet has become a valuable resource in the human language technology and artificial intelligence. Due to its vast coverage of English words, WordNet provides with general lexico-semantic information on which open-domain text processing is based. Furthermore, the development of WordNets in several other languages extends this capability to trans-lingual applications, enabling text mining across languages. For example, in Europe, WordNet has been used as the starting point for the development of a multilingual database for several European languages (the EuroWordNet project). Other resources such as the Longman

Dictionary of Contemporary English and Roget's Thesaurus have also been used for various NLP tasks.

The topic of this workshop is the exploitation of existing resources for particular computational tasks such as Word Sense Disambiguation, Generation, Information Retrieval, Information Extraction, Question Answering and Summarization.

WORKSHOPS – Monday, 3 June 2001

Workshop on Adaptation in Dialogue Systems

Cindi Thompson, Tim Paek, and Eric Horvitz

The purpose of this workshop is to bring together researchers investigating the application of learning and adaptation to dialogue systems, both speech and text based.

Methods for learning and adaptation show promise for enhancing the robustness, flexibility, and overall accuracy of dialogue systems. While researchers in many parts of computational linguistics who use these methods have begun to form communities, the burgeoning set of activities within dialogue has remained relatively disparate. We are interested in adaptation that includes learning procedures as well as decision making methods aimed at dynamically reconfiguring dialogue behavior based on the context. We would also like to explore techniques that allow a dialogue system to learn with experience or from data sets gathered from empirical studies. Researchers looking at methods to automatically improve different modules of dialogue systems, or the system as a whole, have not had many opportunities to come together to share their work. We thus welcome submissions from researchers supplementing the traditional development of dialogue systems with techniques from machine learning, statistical NLP, and decision theory.

Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations (Day 2)

Dan Moldovan, Sanda Harabagiu, Wim Peters, Mark Stevenson, and Yorick Wilks

See description above.

Student Research Workshop

Krzysztof Czuba and Lisa Michaud

These papers have been accepted for presentation at the NAACL 2001 Student Research Workshop:

Answer Fusion with On-line Ontology Development
Roxana Girju
Southern Methodist University, Dallas, Texas

Building a Bilingual Dictionary with Scarce Resources: A Genetic Algorithm Approach
Benjamin Han
Language Technology Institute, Carnegie Mellon University

APT: Arabic Part-of-speech Tagger
Shereen Khoja
Lancaster University

Automatic Detection of Noun Phrases in English and Estonian Electronic Texts
Maarika Traat
University of Tartu

A Probabilistic Model for Automatic Document Categorization
Takeshi Masuyama
Dept. of Computer Science and Media Engineering,
Yamanashi Univ., 4-3-11, takeda, kofu-shi,
Yamanashi-ken 400-8511, JAPAN

Lynx: Learning a Statistical Parser
Peter Venable
Carnegie Mellon

EMNLP 2001 - Sunday & Monday, 3 – 4 June 2001
2001 Conference on Empirical Methods in Natural Language Processing (EMNLP-2001)

Sponsored by SIGDAT and the Intelligent Information Systems Institute

We have solicited papers from academia, government, and industry on all areas of traditional interest to the SIGDAT community and aligned fields, including but not limited to:

- information extraction
- information retrieval
- language and dialog modeling
- lexical acquisition
- machine translation
- multilingual technologies
- question answering
- statistical parsing
- summarization
- tagging
- term and named entity extraction
- word sense disambiguation
- word, term, and text segmentation

Also, to encourage reflection on the current state of the art in corpus-based methods, the conference will have the following theme:

"What Works and What Doesn't: Successes and Challenges"

Successes --- We have solicited papers showing the success of empirical methods in and across application settings. Examples include improvements in information retrieval performance due to employing language modeling techniques; effective use of statistical word segmentation algorithms in machine translation systems; and increased speech recognition accuracy through the incorporation of statistical parsing.

Challenges --- It is clear that empirical and corpus-based methods have enjoyed many successes over the past years; but in looking to future accomplishments, the community needs to be aware of the limitations of various techniques and paradigms. We have welcomed papers that carefully expose and study such limitations. Examples include the identification and exploration of: classes of domains or problems in which popular techniques perform poorly; significant gaps between human and machine performance on tasks where statistical approaches have made great progress; and important practical situations where common assumptions fail to hold. We emphasize that we sought submissions that thoughtfully document fundamental limitations, rather than simply report on unsuccessful experiments. It is desired that such papers contain thorough examination, via careful experimentation, of the critical factors contributing to the "negative" result.