University of Massachusetts Amherst

Vol. 4, No. 1 • Spring 2002

Alum Matters

A newsletter for alumni of the Department of Computer Science

Using NLP to extract information

With virtually unlimited amounts of information at our fingertips via the World Wide Web, there is an increased need for systems that can efficiently find relevant answers to our questions. This research area has risen to the forefront, partially due to the increased emphasis on developing technology to combat terrorism. University groups like the UMass Center for Intelligent Information Retrieval (CIIR) are leading the way in this language technology research. Associate Professor Ellen Riloff (Ph.D. '94), one of Professor Wendy Lehnert's students, is working on natural language processing (NLP) research at the University of Utah's School of Computing to develop systems for identifying terrorist activity as well as applications for the corporate world.

Professor Riloff has developed information extraction (IE) systems that can recognize news stories about terrorism from a variety of incoming news reports and identify the perpetrators, victims, and physical targets involved in the attack, the date and location of the incident, and the type of weapons used. The extracted information could be automatically added to a database, which would be a central repository of all terrorist information being accumulated from around the world. If the information required immediate attention, the IE system could instantly notify appropriate personnel.

"Natural language processing is essential for this task because there are no magic keywords to identify perpetrators and victims," says Riloff. "The information extraction system must look for sentences containing phrases associated with terrorism, such as 'was bombed,' and extract information from syntactic positions surrounding the phrase."

The goal of natural language processing research is to endow computers with the ability to understand language. Currently, most text processing software uses superficial techniques such as keyword search to retrieve and process textual information. At Utah's Natural Processing Laboratory, Riloff is developing the next generation of information extraction systems, which will incorporate syntactic and semantic knowledge to understand what a text actually means.

Current systems have problems dealing with ambiguity and synonymy in natural language. Ambiguity is pervasive in language because most common words have several meanings. Synonymy, in which words and phrases have the same general meaning, is a major problem for current software because only the search terms provided by the user are matched against the texts. "Most people are aware of the successful matches found by their system, but are blissfully unaware of how much relevant information their software did not find," says Riloff.

To understand the meaning of a sentence, a computer must have both a syntactic and semantic representation of the text. The main bottleneck in achieving this goal has been a lack of computer knowledge: an NLP system needs syntactic and semantic information for every word that it might encounter. In the last few years, natural language processing researchers have begun to overcome this knowledge engineering bottleneck by focusing on two things: 1) NLP applications that operate in a specific domain, such as "information extraction" tasks, and 2) the development of techniques that can automatically learn syntactic and semantic knowledge from existing text resources.

Riloff's NLP Laboratory is at the forefront of research on the information extraction problem. Along with the application to combat terrorism, IE technology also has many applications in the corporate world. Riloff and her researchers have created IE systems to extract facts about joint ventures and corporate acquisitions. These systems could be connected to an incoming news feed so that corporate executives can be notified immediately when a new joint venture or acquisition is announced.

Over the last few years, a new research area called empirical natural language processing has emerged, which is essentially a form of text mining. Empirical techniques collect samples of language and identify patterns and associations in the data. Empirical NLP techniques have been stunningly successful, producing substantially better coverage and accuracy than the previous generation of systems.

The NLP group at Utah has been actively engaged in

Riloff

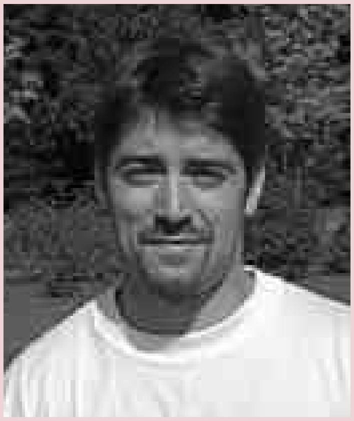
research on bootstrapping methods to acquire semantic information automatically. The idea behind a bootstrapping algorithm is to give the computer a small amount of information that it can use to automatically discover new information (i.e., pull itself up by its own bootstraps). Riloff's group has developed several bootstrapping algorithms that enable the computer to automatically acquire semantic knowledge.

"The field of natural language processing has undergone enormous change in the last ten years, and we are now able to contemplate some scenarios that seemed impossible just a short time ago," says Riloff. One area of growing importance to NLP is question answering. One of NLP's long-term goals is to develop search engines that would allow users to type in specific questions instead of keywords, not the keyword search in disguise that is currently used in some Web sites that allow users to type in questions. While intelligent question answering is probably still a long way off, researchers are beginning to make substantial progress in handling certain types of questions, such as who, where, and when questions.

"We can expect to see natural language processing working its way into more commercial products and opening up new possibilities for intelligent text processing," says Riloff. "The next generation of text analysis software will allow users to manage large text collections more effectively and more efficiently, and with greater confidence that the proverbial needle in the haystack can indeed be found."

Dr. Riloff joined the Computer Science faculty at the University of Utah in 1994. In 1997, she received a National Science Foundation CAREER award for research in building conceptual natural language processing systems for practical applications. More information on Professor Riloff and her research can be found at: **www.cs.utah.edu/~riloff.**

Rubenstein receives NSF CAREER Award



Columbia University Assistant Professor Dan Rubenstein (Ph.D. '2000), of the Electrical Engineering and Computer Science Department, received the National Science Foundation (NSF) CAREER award for his proposal "Flexible, Large-Scale Best-Effort Quality of Service in the Internet."

Rubenstein's research will investigate how to implement a service on top of existing Internet networks that attempts to meet specific needs of applications, such as delay-bounded delivery of data, or coping with limited receiving capabilities of client devices. "This service differs from the traditional approach considered over the past several years that attempts to deploy such a service within the guts of the network," says Rubenstein. "The challenge will be to find ways to scalably coordinate access to these services from the large set of network users that wish to use them."

Professor Rubenstein focuses his research on building protocols that allow large numbers of users to communicate effectively. He is very interested in devising solutions to alleviate problems associated with network flash crowds, a situation where a large number of users suddenly and unexpectedly overwhelm a particular network site with traffic. Such a problem occurred on September 11, when news web sites were overloaded with requests for network traffic.

Says Rubenstein, "One of the nice things about being at Columbia is that it gives me plenty of opportunities to come back and visit UMass. This has allowed me to continue to get advice firsthand from my advisors Jim Kurose and Don Towsley, and from Brian Levine and Micah Adler."

Alumni Connections

In January 2002, **Wei Zhao** (Ph.D. '86) received "the spirit of technology transition award" from DARPA for his work in the DARPA fault tolerant networking program. In February 2002, two of his graduate students won 2nd place in the ACM International Graduate Research Competition. He currently is an Associate Vice President for Research at Texas A&M University, College Station, TX.

The Department is trying a distance education experiment this semester. In a partnership between The MITRE Corporation and the UMass Center for Intelligent Information Retrieval (CIIR), Assistant Professor **James Allan** is coordinating a new class this semester that is being taught by members of a team of MITRE researchers. The lectures for Statistical Natural Language Processing (CMPSCI 691L) are presented by those researchers, with some course sessions presented by live video feed from MITRE. Supervision of class projects is shared between UMass and MITRE. UMass alums **Jay Ponte** (Ph.D. '98) and **Warren Greiff** (Ph.D. '99) are two of the MITRE lecturers.

The UMass Alumni Association provides a connection
between alumni, faculty, friends, and the University.
Stay connected by becoming a Member.
Membership matters.
800-456-UMASS (8627) umassalumni.com

Thank you to all who so generously gave to the Department of Computer Science. For a recent list of who you generous people are, please see the back page of this issue of *Significant Bits.*