June 15, 1993, Wendy G. Lehnert, University of Massachusetts, 413 545-3639, Lehnert@cs.umass.edu
- Using Case-Based Reasoning in Natural Language Processing, N00014-92-J-1427, 3/15/92-3/14/93

1

1. Productivity measures.

Refereed papers submitted but not yet published: 6

Refereed papers published: 9

Unrefereed reports and articles: 3

Books or parts thereof submitted but not yet published: 2

Books or parts thereof published: 1

Patents filed but not yet granted: 0

Patents granted (include software copyrights): 0

Invited presentations: 8

Contributed presentations: 2

Honors received (fellowships, technical society appointments, conference committee roles, editorships, etc.): also include descriptions of the specific honors.

Prizes or awards received (Nobel, Japan, Turing, etc.): also include descriptions of the specific prizes. University of Massachusetts Graduate School Fellowship to Claire Cardie (RA) for her continued research in Computer Science ($6,000 for 1 year).

Promotions obtained: 0

Graduate students supported >= 25% of full time: 4

Post-docs supported >= 25% of full time: 0

Minorities supported (include Blacks, Hispanics, American Indians and other native Americans such as Aleuts, Pacific slanders, etc.; do not include Asians or Asian-Americans):

2. Detailed summary of technical progress.

A variety of problems addressed in natural language processing (NLP) make this area ripe for hybrid system designs and approaches based on multiple technologies. One particularly promising crossover is the application of case-based reasoning (CBR) to NLP. There are two important facts that make CBR an obvious candidate for innovative investigations in NLP.

First, most decision processes in NLP are characterized by shades of grey and different ways of weighing preferences rather than black and white absolutes or right and wrong answers. This holds true for the lowest levels of lexical ambiguity as well as the highest levels of inference and reasoning. The affinity in language for relative preferences rather than hard and fast absolutes is fully consistent with CBR capabilities that can produce multiple solutions and then assess each one in accordance with multiple dimensions for evaluation.

Second, many aspects of NLP can be studied with the aid of large online text corpora. Full text databases are now being constructed using optical scanners and commercial OCR technologies at cost levels that make it possible for individual labs to customize their own online corpora. Access to large amounts of "real" language make it possible for NLP researchers to break out of the traditional linguistic paradigm of contrived examples in order to study language processing from a more seriously empirical perspective with replicable experiments and corpus-driven processing techniques.

At the University of Massachusetts, we have demonstrated the utility of of NLP research: lexical knowledge acquisition and content-based document classification. We will describe each of these research efforts in turn.

Lexical Knowledge Acquisition

The ability or inability of a natural language processing (NLP) system to handle gaps in lexicon coverage ultimately affects the system's performance on novel texts. Suppose, for example, that a natural language system processes a text with the goal of summarizing it or extracting relevant information, but unexpectedly encounters an unknown word. Rather than stop and wait for a knowledge engineer to enter the missing lexical information, or skip the offending word altogether, a robust sentence analyzer should infer the necessary syntactic and semantic knowledge for the unknown word and then continue processing the text. Consider the following sentence taken from the TIPSTER JV corpus for which an NLP system finds no entry in its lexicon for "Malaysia:"

> Sanyo Electric Co. and Ford Motor Co. have agreed to set up a joint
> venture by the end of this year to produce car audio parts in
> **Malaysia**, they said Thursday.

Before the NLP system can continue beyond "Malaysia," it may need to know a specific set of features for the unknown word including its

- part of speech (e.g., noun),
- general semantic class (e.g., location),
- specific semantic class (e.g., country),
- associated domain-specific concepts (e.g., "Malaysia" may activate a company-location concept in this context),

In our experiments we have demonstrated that a case-based memory can be used to infer these four features for any word in a corpus without reliance on handcrafted lexical acquisition heuristics or lexical disambiguation heuristics. We first create a case base of word definition cases each of which encodes the definition of a word as well as the context in which it occurred. A human supervisor is needed to provide values for the four features listed above that represent the definition of each word in the case base, but all other aspects of case base construction are fully automated. In particular, a semantically-oriented sentence analyzer provides a representation of the context in which the word occurred. Once training is completed, the parser can infer the definiton of unknown words by mapping the context in which the word occurs into the same representation format used by the case base and then retrieving the most similar cases from case memory. The retrieved cases vote on the part of speech, semantic features, and domain specific concepts associated with the unknown word in the current context.

We evaluated our knowledge acquisition technique in experiments that explore two different, but related applications. In the first application, we assume the existence of a nearly complete dictionary and use the case base to infer the features of occasional unknown words. In the second, more ambitious application, we assume only a small dictionary of 129 function words (e.g., determiners, prepositions, auxiliaries) and use the case base to determine the definition of {\em all} other words (i.e., all open class words) encountered by the parser. The results are shown in the table below and indicate that the case-based method performs significantly better than a system that randomly guesses a legal value for each feature based on the distribution of values across the training set and a system that chooses the most frequent value as a default.

Missing Feature	Experiment 1	Experiment 2	Random Selection	Default
part of speech	93.0%	91.0%	34.3%	81.5%
general semantic class	78.0%	65.3%	17.0%	25.6%
specific semantic class	80.4%	74.0%	37.3%	58.1%
domain-specific concept	95.1%	94.3%	84.2%	91.7%

Our technique is corpus-driven, but requires a much smaller training corpus that other corpus-based NLP efforts. Statistically-based NLP techniques typically require a corpus of at least 1,000,000 words before good results can be obtained. We can obtain our highest performance levels on the basis of less than 200 training sentences. This dramatic difference is due to the fact that we are abstracting high-level structures through the use of a semantic sentence analyzer in order to organize our case base, whereas most corpus-driven NLP work relies on a stochastic database of lexical co-occurrences. Higher levels of abstraction require a lot less data to identify reliably predictive features.

Content-Based Document Classification

Although in-depth natural language processing should be expected to produce good results in text classification, NLP techniques are typically hampered by a knowledge engineering bottleneck that makes it difficult to port these systems from one domain to another. In an effort to address this knowledge engineering bottleneck, we are currently investigating statistical and case-based text profiles that operate in conjunction with minimal NLP capabilities in order to produce practical text classification algorithms.

In our earlier work, we established strong recall and precision results for a simple text classification task where we tried to distinguish texts describing drug-related terrorist activities from texts describing other acts of violence associated with the drug cartels and military clashes in Latin American. These discriminations are often subtle and can confuse human encoders, so we were looking at a discrimination task that would not be well-served by simple keyword analysis techniques. From these experiments, we discovered that stochastic relevancy signatures using selective concept extraction are more effective than keyword techniques for sophisticated text classification tasks. Moreover, this approach could be ported to a new domain on the basis of a training corpus where texts are tagged as relevant or irrelevant, thereby obviating the need for an explicit domain definition (which can be very difficult to obtain when domains are complicated and characterized by "grey" areas).

Using stochastic signatures alone, we can take advantage of the fact that a phrase like "found dead" is strongly predicts a relevant text, while "dead" by itself does not. The first signature only appears when foul play is afoot, but descriptions of dead people occur in many contexts including natural disasters and accidents. Similarly, an instance of "no casualties" is a good indicator of a relevant text, but "casualties" is not. This is because "no casualties" is only used to describe no civilian casualties (in a context where civilians might have been harmed by some act of violence), whereas "casualties" is often used to describe military casualties (and military personnel do not qualify as victims of terrorist activities under the guidelines we were using for a domain definition). It would be very difficult to discover these signatures without some stochastic analysis of a training corpus.

However, stochastic techniques alone cannot help us distinguish situations like the following:

> "More than 100 people have died in Peru since 1980, when the Maoist Shining Path organization began its attacks and its wave of political violence."
>
> "More than 100 people have died in Peru during 2 attacks yesterday."

According to our domain guidelines, the first sentence does not describe a terrorist event (no specific event was described), but the second sentence does (the small number of attacks and the relatively tight time interval provide sufficient specificity for this to qualify).

Using case-based reasoning techniques, we can classify texts on the basis of natural language contexts instead of isolated keywords and phrases. We have developed a case-based text classification algorithm that represents a document as a set of cases, one case for each sentence. The case base is acquired automatically by applying our NLP system to a training corpus. To classify a new document, the document is converted into a set of cases and the statistical properties of the case base determine whether similar cases are highly correlated with a domain classification. We have conducted experiments with two blind test sets to compare the effectiveness of the case-based algorithm with our previous relevancy signatures algorithm. The case-based algorithm consistently performs at least as well as the relevancy signatures algorithm, and often much better. On the first test set, relevancy signatures correctly identified 30% of the relevant texts with 100% precision whereas the case-based algorithm correctly identified 61% of the relevant texts with 100% precision. The case-based approach allowed us to correctly classify documents that were inaccessible to the signature-based algorithm, without sacrificing precision. On the second test set, relevancy signatures correctly identified 24% of the relevant texts with 93% precision and the case-based algorithm correctly identified 44% with 100% precision. These results suggest that case-based techniques can be used successfully to support content-based document classification with high precision.

As storage capacities grow and memory becomes cheaper, we believe that text classification will become a central problem for an increasing number of computer applications and users. And as more documents become available on-line, we expect that high-precision text classification will become a problem of paramount importance. Although our investigation in this area has been preliminary, we are very encouraged by the success of our techniques.

3. Lists of publications, presentations and reports.
Invited Talks by Wendy Lehnert:

Invited Talk: "Information Extraction from Text," Department of Computer Science and Electrical Engineering, University of Michigan Ann Arbor, MI, Feb. 11, 1993.

Keynote Address: "Portability and Scalability for Information Extraction Systems," Ninth IEEE Conference on Artificial Intelligence for Applications. Orlando, FL. 1993, March 3, 1993.

Panel participant: "Hybrid Approaches to the Information Access Problem," First National Conference on Information and Knowledge Management. Baltimore, MD, November 8-11, 1992.

Invited Talk: "Automating the Construction of a HyperText System for Scientific Papers," AAAI-92 Workshop on Communicating Scientific and Technical Knowledge. San Jose, CA. July 15, 1992.

Invited Talk: "University of Massachusetts System Description," Fourth Message Understanding Conference. MacLean, VA. June 18, 1992.

Invited Talk: "University of Massachusetts Site Report" Fourth Message Understanding Conference. MacLean, VA. June 16, 1992.

Invited Talk: "Corpus-Driven Language Processing Using Selective Concept Extraction," Media Laboratory of the Massachusetts Institute of Technology. Cambridge, MA. April 29, 1992.

Keynote Address: "AI in the 90's: Scaling Up and Shaking Down," Artificial Intelligence Applications (sponsored by The American Defense Preparedness Association) Williamsburg, VA. March 31, 1992.

Journal Articles

Lehnert, W., Cardie, C., Fisher, D., McCarthy, J., Riloff, E. and Soderland, S. "Evaluating an Information Extraction System," Accepted pending revisions to the *Journal of Integrated Computer-Aided Engineering.*

Cowie, J. and Lehnert, W. "Information Extraction," Submitted to a special issue of the *CACM*. 1993.

Riloff, E. and Lehnert, W.G. "Information Extraction as a Basis for High-Precision Text Classification," Submitted to *ACM Transactions on Information Systems* (Special Issue on Text Categorization). 1993.

Conference and Workshop Papers

Cardie, C. (1993). "A Case-Based Approach to Knowledge Acquisition for Domain-Specific Sentence Analysis". To appear in the *Proceedings of the Eleventh National Conference on Artificial Intelligence* (AAAI-93). Washington DC.

Cardie, C. (1993). "Using Decision Trees to Improve Case-Based Learning". To appear in the *Proceedings of the Tenth International Conference of Machine Learning*". Amherst, MA.

Riloff, E. "Automatically Constructing a Dictionary for Information Extraction Tasks". To appear in *Proceedings of the Eleventh Annual Conference on Artificial Intelligence. 1993.* Washington DC.

Riloff, E. (1993). "Using Cases to Represent Context for Text Classification," in *Working Notes of* *the AAAI Spring Symposium on Case-Based Reasoning and Information Retrieval.* Palo Alto, CA.

Riloff, E. and Lehnert, W. (1993) "Automated Dictionary Construction for Information Extraction from Text," in *Proceedings of the Ninth IEEE Conference on Artificial Intelligence for Applications*, pp. 93-99. IEEE Computer Society Press.

Lehnert, W. G. (1992). "Automating the Construction of a Hypertext System for Scientific Literature," in *Working Notes of the AAAI Workshop on Communicating Scientific and Technical Knowledge.* San Jose CA.

Riloff, E. and Lehnert, W.G. (1993). "Classifying Texts Using Relevancy Signatures," in *Proceedings, Fifth DARPA Speech and Natural Language Workshop.* pp. 224-229.

Lehnert, W., Cardie, C., Fisher, D., McCarthy, J., Riloff, E. and Soderland, S. (1992). "University of Massachusetts: Description of the CIRCUS System as Used for MUC-4", in *Proceedings of the Fourth Message Understanding Conference.* pp. 282-288.

Lehnert, W., Cardie, C., Fisher, D., McCarthy, J., Riloff, E. and Soderland, S. (1992). "University of Massachusetts: MUC-4 Test Results and Analysis," in *Proceedings of the Fourth Message Understanding Conference.* pp. 151-158.

Cardie, C. (1992). "Learning to Disambiguate Relative Pronouns," in *Proceedings, Tenth National Conference on Artificial Intelligence*, San Jose, CA. July 12-16, 1992. pp. 38-43.

Cardie, C. (1992). "Corpus-Based Acquisition of Relative Pronoun Disambiguation Heuristics," in *Proceedings of the Thirtieth Annual Meeting of the Association for Computational Linguistics.* June 28-July 2, 1992. Newark, Delaware. pp. 216-223.

Cardie, C. (1992). "Using Cognitive Biases to Guide Feature Set Selection," in *Working Notes of the AAAI-92 Workshop on Constraining Learning with Prior Knowledge.* July 13, 1992. San Jose, CA. pp. 11-18.

Cardie, C. (1992). "Using Cognitive Biases to Guide Feature Set Selection," in *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society.* Indiana University, Bloomington, IN. pp. 743-48.

Fisher, D., and Riloff, E. (1992). "Applying Statistical Methods to Small Corpora: Benefiting from a Limited Domain," in *Working Notes of the AAAI Fall SYmposium on Statistical Natural Language Processing.* Cambridge, MA.

Book Chapters

Lehnert, W.G. "Cognition, Computers and Car Bombs: How Yale Prepared Me for the 90's". To appear in *Belief, Reasoning, and Decision Making: Psycho-logic in Honor of Bob Abelson* (eds: Schank & Langer), Lawrence Erlbaum Associates. (in press)

Wermter, S. and Lehnert, W.G. "A Parallel Model for Compositional Similarity of Natural Language Concepts," in *Parallel Natural Language Processing* (eds: U. Hahn and G. Adriaens). (in press).

Wermter, S., and Lehnert, W.G. "A Hybrid Symbolic/Connectionist Model for Noun Phrase Understanding," in Connectionist Natural Language Processing. Kluwar Academic Publishers, Norwell, MA. pp. 101-118.

4. Transitions and DoD interactions.

TIPSTER Phase II Planning Meeting, Schenectady, NY. April 12-14, 1993.

TIPSTER 18 Month Meeting "UMass/Hughes site report and test results" Williamsburg, VA, Feb 22-24, 1993.

Steve Dennis and Rita McCardell Doerr (NSA site visit) Feb. 12, 1993.

"Natural Language Processing at the University of Massachusetts" at the 12-month TIPSTER evaluation meeting in San Diego, California. September 17-21, 1992.

ISAT Study Group Annual Review Meeting in Woods Hole, MA Aug 11-16, 1992. At this meeting I contributed to the final reports associated with two study groups: "Making Computers Easier to Use" (chaired by Michael Dertouzos of M.I.T.) and "Multi-Modal and Language -Based Systems" (chaired by Victor Zue of M.I.T.)

ISAT Study Group meeting on "Multi-Modal Language-Based Systems" in Dedham, MA. July 9-10, 1992.

"University of Massachusetts System Description" at the Fourth Message Understanding Conference. MacLean, VA. June 18, 1992.

"University of Massachusetts Site Report" at the Fourth Message Understanding Conference. MacLean, VA. June 16, 1992.

ISAT Study Group meeting on "Making Computers Easier to Use" in Brewster, MA. May 31-June 2, 1992.

Intelligent Systems and Technology Interim Meeting (ISAT DARPA Study Group) in Washington, DC. May 7, 1992.

Hosted a visit by Robert Powell (ONR), April 21, 1992.

Keynote Address: "AI in the 90's: Scaling Up and Shaking Down" at Artificial Intelligence Applications (sponsored by The American Defense Preparedness Association) Williamsburg, VA. March 31, 1992.

5. Software and hardware prototypes.

Software systems have been implemented to conduct experiments but no general prototypes have been constructed, and no commercialization has occurred.