UMass/Hughes TIPSTER Project on Extraction from Text

*Wendy Lehnert**

Department of Computer Science
University of Massachusetts
Amherst, MA 01003

*Charles Dolan***

Hughes Research Center
3011 Malibu Canyon Road
Malibu, CA 90265

PROJECT GOALS

The primary goal of our effort is the development of robust and portable language processing capabilities and information extraction applications. Our system is based on a sentence analysis technique called selective concept extraction. Having demonstrated the general viability of this technique in previous evaluations [Lehnert, et al. 1992], we are now concentrating on the practicality of our technology by creating trainable system components to replace hand-coded data and manually-engineered software.

Our general strategy is to automate the construction of domain-specific dictionaries that can be completed with minimal amounts of human assistance. Our system relies on two major tools that support automated dictionary construction: (1) OTB, a trainable part-of-speech tagger, and (2) AutoSlog, a concept node generator that operates in conjunction with the CIRCUS sentence analyzer. Concept nodes are dictionary definitions for CIRCUS that encode lexically-indexed interactions between syntactic constituents and semantic case frames. OTB and AutoSlog both require minor technical adjustments and minimal assistance from a "human in the loop" in order to create a new domain-specific dictionary, but this can generally be accomplished by a single individual in the space of one week [Riloff, 1993].

A third tool, TTS-MUC3, is responsible for the creation of a template generator that maps CIRCUS output into final template instantiations. TTS-MUC3 can be adjusted for a new domain in one day by a knowledgeable technician working with adequate domain documentation. This minimal manual engineering is required to specify objects and relationships. Once these adjustments are in place, TTS-MUC-3 uses CIRCUS and a development corpus of source texts and key templates to train classifiers for template generation. No further human intervention is required to create template generators.

RECENT RESULTS

Our emphasis has been on fast system prototyping and rapid system development cycles. In preparing for the TIPSTER 18-month evaluation, we customized a complete information extraction system for the domain of English microelectronics (EME) in the space of four weeks working from scratch without the benefit of any domain experts. This time period included the development of a new facility for keyword recognition that had not been deemed necessary for any of our previous information extraction systems. If this facility had not been added, we could have cut our EME system development time down to two weeks.

PLANS FOR THE COMING YEAR

Within the next six months we will incorporate semantic features into our system. We do not have semantic features in the current system because we would have had to acquire them through manual means, and we wanted to wait until we could acquire them through training. We now believe that we have identified a method for automated feature acquisition that should suffice for our purposes [Cardie, 1993].

We are generally satisfied with the performance of OTB, AutoSlog, and CIRCUS and we believe the addition of semantic features will significantly boost our overall performance.

REFERENCES

1. Cardie, C.T. "A Case-Based Approach to Knowledge Acquisition for Domain-Specific Sentence Analysis". To appear in *Proceedings of the Eleventh Annual Conference on Artificial Intelligence. 1993.*

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

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

*Umass: Claire Cardie, Ellen Riloff, Joseph McCarthy, Stephen Soderland, and Jon Peterson

**Hughes: Seth Goldman.