NLP AND TEXT ANALYSIS AT THE UNIVERSITY OF MASSACHUSETTS

Wendy G. Lehnert

Department of Computer Science
University of Massachusetts
Amherst, MA 01003

PROJECT GOALS

Our group is investigating a variety of techniques centered around the use of text corpora to support natural language processing applications. We are interested in information extraction from text, text classification, and knowledge acquisition from text corpora. Our goal is to develop technologies that can be readily ported across domains and scaled up with a minimal amount of manual engineering. In particular, we are experimenting with various kinds of statistical profiles and case based reasoning systems in order to facilitate:

- semantically-oriented dictionary construction
- the analysis of complex sentence structures
- complex domain discriminations
- specific aspects of discourse analysis

Although it is doubtful that all manual knowledge engineering can be eliminated from the development cycle of practical NLP systems, we believe that minimal amounts of manual engineering can be highly leveraged when used in conjunction with a suitable text corpus. Given a specific text processing application and a text corpus that is representative of the target texts, we are experimenting with different aspects of system development that can be fully or partially automated.

RECENT RESULTS

Using the UMass/MUC-3 system implementation as a starting point, we have been looking at the problem of text classification as it pertains to domain relevancy. Based on a fully-automated semantic analysis of the MUC-3 texts, we have developed statistical profiles to distinguish texts that describe legitimate terrorist events from texts that are "near misses" with respect to the domain definition. Using these profiles, we can discriminate new texts with relatively high degrees of recall and precision (as high as 97% recall with 93% precision on one test run of 100 texts).

We have also been looking at case-based reasoning (CBR) techniques and evaluating the utility of CBR in conjunction with the MUC-3 text corpus. Drawing once again from the UMass/MUC-3 system, we have run additional experiments on our CBR-based consolidation module in order to better understand its capabilities. In one such experiment, we determined that the CBR module is capable of producing recall and precision scores for incident types that exceed the recorded performance levels of all the MUC-3 systems (85% recall with 91% precision). Our own UMass/MUC-3 system posted 77% recall with 81% precision on incident types. Unfortunately, comparable performance improvements have not been obtained for any other MUC-3 template slots.

In a separate CBR effort, we have designed a new CBR module that locates referents for the relative pronoun "who" in the MUC-3 texts. Operating with 75-90% hit rates, this system outperforms our original hand-coded heuristics. Interestingly, it tends to make most of its mistakes on convoluted sentences that are confusing to human readers.

PLANS FOR THE COMING YEAR

We expect to continue our ongoing investigations in each of the three areas mentioned above. We want to further refine our text classification profiles and investigate the integration of these capabilities back into a complete information extraction system (such as the UMass/MUC-4 system). We hope to experiment with variations on our UMass/MUC-3 consolidation component to see if aspects of that capability can assume a more prominent role in our overall system design. We will also continue our investigations with CBR-based discourse analysis and see if we can generalize this technique from relative pronoun resolution to other problems associated with scoping and structural ambiguities.

More generally, we hope to gain a greater understanding of selective concept extraction as a sentence analysis technique, both in terms of its portability across domains, and its inherent limitations within specific text processing applications.