THE TIPSTER/SHOGUN PROJECT

Paul S. Jacobs, Principal Investigator

GE Research and Development Center
1 River Rd.
Schenectady, NY 12301

PROJECT GOALS

TIPSTER/SHOGUN, part of the ARPA TIPSTER Text (Phase I) program, was led by GE Corporate Research and Development, with Carnegie Mellon University and Martin Marietta Management and Data Systems (formerly GE Aerospace). The project ended at the beginning of 1994, with TIPSTER Phase II expected to begin in March. The TIPSTER/SHOGUN system is thus the end result of a two-year research effort. The project's main goals were: (1) to develop algorithms that would advance the state of the art in coverage and accuracy in data extraction, and (2) to demonstrate high performance and adaptability across languages and domains.

The team concentrated its research on the development of a model of finite-state approximation, within which the performance of more detailed models of language could be realized in a simple, efficient framework, and on automated knowledge acquisition. The ability of programs to extract data from free text is, in general, limited by the coverage of domain and world knowledge. We chose to focus on knowledge acquisition from corpus data, thereby expanding the coverage of the system while also helping to tune each configuration.

Like other TIPSTER contractors, the TIPSTER/SHOGUN team ran its system on a series of benchmarks, ending with the MUC-5 evaluation in August, 1993. MUC-5 included tests in four configurations, comprised of two domains (joint ventures and micro-electronics) in each of two languages (English and Japanese). Although many of the research results of SHOGUN had little or no impact on the benchmarks, MUC-5 provided a comprehensive test of system performance.

RECENT RESULTS

The finite-state approximation method developed under TIPSTER was inspired by earlier work at GE and SRI, along with experiments near the mid-point of our project, which showed that tighter control, particularly in parsing, contributed very little to text interpretation while greatly inhibiting knowledge acquisition. This shift was also influenced by the demands of Japanese language processing, where our existing knowledge resources were less refined than in English.

The relationship between representation, i.e., the finite-state patterns in our system, and acquisition, i.e., the method by which new knowledge is added, is critical. In our system, we chose to emphasize the finite-state patterns in part because they help to take advantage of the most critical source of knowledge we have available—the corpus.

The corpus-based acquisition strategy used statistical methods to help identify key phrases and other lexical relations in the corpus, and to assign these lexical relations to word groups with similar interpretations. This approach worked best for task components that required large amounts of knowledge, particularly determining the product or service of each joint venture. We believe this accounts for some of the large differences in coverage between SHOGUN and other systems.

In addition to helping coverage, the corpus-based acquisition strategy greatly eased portability across languages. In most cases, we did each English component first, then used the English as a way of bootstrapping the Japanese. For example, we would take each important "pivot" word in English, try to identify the corresponding "pivot" in Japanese, then use the corpus to identify the relevant contexts in which that word occurred in Japanese. SHOGUN's accuracy in Japanese was somewhat higher, on average, than in English.

SHOGUN, on average, extracted 37% more information correctly (37% higher recall) than any other system in each of the four MUC-5 configurations. On average, SHOGUN's precision was 13% lower than the next best system. Recall advanced 37% on average between the TIPSTER 18-month evaluation and the MUC-5 test 6 months later, and was 10% higher in the TIPSTER final test than in MUC-4 (which was a much simpler task). We are particularly satisfied by the consistently improving coverage of our system across languages and domains.

PLANS FOR THE COMING YEAR

As TIPSTER Phase II begins this year, the emphasis will be on developing an architecture that incorporates some of our Phase I results into an open framework that promotes delivery as well as further technical advances. In addition, our research will continue to integrate methods from information retrieval (detection) with more detailed language processing strategies.