SHOGUN - MULTILINGUAL DATA EXTRACTION FOR TIPSTER

P. Jacobs, Principal Investigator

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

PROJECT GOALS

The TIPSTER/SHOGUN project aims at substantive improvements in coverage and accuracy for automatic data extraction through innovative strategies in knowledge acquisition, run-time integration, and control. One of four teams in the data extraction component of the TIPSTER program, TIPSTER/SHOGUN includes GE Corporate Research and Development, Carnegie Mellon University - Center for Machine Translation, and GE Management and Data Systems.

Data extraction systems interpret the key content of natural language text, producing a structured representation of items that range from high-level business relationships to detailed knowledge coding of technologies and industry classifications. This task applies to both Japanese and English in each of two domains—joint ventures and micro-electronics. As such, TIPSTER is considerably more detailed and comprehensive than previous text interpretation experiments, including prior MUC (Message Understanding Conference) evaluations. The goals for SHOGUN are the following:

- Accuracy significantly ahead of MUC-4, with levels near those of trained human analysts at about 100 times human speed using conventional hardware and software.
- Automated knowledge acquisition and extensibility tools that support customization times of a few weeks for new applications.
- Multi-lingual performance, with comparable levels in both languages and the highest possible overlap between languages.

The project is now within a few months of completion, and is on target toward all of these goals.

RECENT RESULTS

During the early stages of the project, the team reached very good initial levels of performance on MUC-4 by successfully integrating methods used at GE and CMU, marking the first time that parsing systems of this level of coverage have been effectively combined. This provided an important testbed and also allowed for multilingual development—In recent months, Japanese performance has remained close to English performance. As the system coverage and accuracy have continued to improve, the most important recent thrust has been the incorporation of most of the knowledge and control strategies of the system into a finite-state driven analyzer, effectively replacing the traditional parsing layer with a detailed knowledge base of finite-state rules compiled from syntactic and lexical resources. While this seems close to work done in the speech community, it is an unusual approach for text, where the high perplexity and long sentence length have seemed to favor semantics-driven and high-level syntactic models.

The finite-state model allows different knowledge sources, particularly corpus-based knowledge, to have more of an impact on interpretation. Data extraction is a knowledge-intensive task, and it has been much simpler to augment the finite-state rules with corpus data than it was for the more abstract rules.

While the performance on all tasks still lags behind human analysts, closing this gap may not be as hard as we first expected. Much of the difference comes from portions of the work that are still incomplete. In addition, the ability to use automatically-acquired corpus data gives the programs a distinct advantage on certain portions of the task.

PLANS FOR THE COMING YEAR

As the project nears completion, the team is approaching the goal of near-human accuracy mostly by finishing certain key details, such as better reference resolution and word sense discrimination. At the same time, we are close to some significant advances in corpus-based training methods that will not only isolate the context required to discriminate nuances of meaning but also significantly reduce development time by acquiring domain knowledge from the corpus. This may the key to future applications of TIPSTER technology.