WRITTEN LANGUAGE SYSTEM EVALUATION

Beth M. Sundheim

Naval Command, Control and Ocean Surveillance Center
RDT&E Division (NRaD), Code 44208
San Diego, CA 92152-7420

PROJECT GOALS

Current efforts support the NLP community in the definition and implementation of a range of performance evaluations of written language technology. Goals include offering a variety of evaluations that will appeal to a broad range of NLP research groups, minimizing the task-specific system tailoring required of evaluation participants, and ensuring that the task designs facilitate scoring.

RECENT RESULTS

The final evaluation of the information extraction portion of phase one of the ARPA Tipster Text program was conducted in July, 1993. Participants in this evaluation included not only the Tipster-supported information extraction contractors but thirteen other sites as well. This evaluation was the topic of the Fifth Message Understanding Conference (MUC-5), which was held in August, 1993, and chaired by NRaD. A proceedings will be published in Spring, 1994.

With particular respect to the research and development tasks of the Tipster contractors, the goal of the evaluation was to assess success in terms of a system's ability to work in both English and Japanese (BBN, GE/CMU, and NMSU/Brandeis) and/or in both the joint ventures and microelectronics domains (BBN, GE/CMU, NMSU/Brandeis, and UMass/Hughes). The evaluations measured the completeness and accuracy of systems on information extraction tasks and used an examination of the role of missing, spurious and otherwise erroneous output as a means of diagnosing the state of the art.

Viewed as a set of performance benchmarks for information extraction technology, the MUC-5 evaluation results on the English joint ventures (EJV) task are at least as good as the MUC-4 level of performance. This comparison takes into account some measurable differences in difficulty between the EJV task and the MUC-3/MUC-4 terrorism task.

However, even a superficial comparison of task difficulty is hard to make because of the change from the flat-format design of the earlier MUC templates to the object-oriented design of the MUC-5 templates. Comparison is also made difficult by the many changes that were made to the alignment and scoring processes and to the performance metrics. Therefore, it is more useful to view the results of MUC-5 on its own terms rather than in comparison to previous MUC evaluations.

Viewed on its own terms, MUC-5 yielded very impressive results for some systems on some tasks. Error per response fill scores as low as 34 (GE/CMU optional test run using the CMU TEXTRACT system) and 39 (GE/CMU Shogun system) were obtained on the Japanese joint ventures (JJV) core-template test. The only other error per response fill scores in the 30-40 range were achieved by *humans*, who were tested on the English microelectronics (EME) task; however, machine performance on that EME test was only half as good as human performance. Thus, while the JJV core-template test results show that machine performance on a constrained test can be quite high, the EME results show that a similar level of machine performance on a more extensive task could not be achieved, at least not in the relatively short development period allowed for ME.

Not only do results such as those cited for the JJV core-template test show how well some approaches to information extraction work for some tasks, they also show how manageable languages other than English can be. A cross-language comparison of results showed fairly consistent advantage in favor of Japanese over English. Comparison of results across domains does not show an advantage in favor of one domain over the other, and it is quite likely that differences in the nature of the texts, the nature and evolution of the extraction tasks, and the amount of time allowed for development all had an impact on the results.

The quantity and variety of material on which systems were trained and tested presented challenges far beyond those posed by earlier MUC evaluations. The scope of the evaluations was broad enough to cause most MUC-5 sites to skip parts of the extraction task, especially types of information that appear relatively rarely in the corpus. Since no type of information is weighted in the scoring more heavily than any other, the biases that exist in the evaluation reflect the distribution of relevant information in the text corpus and result in a natural emphasis on handling the most frequently-occurring slot-filling tasks. These tasks turn out to be the ones that are less idiosyncratic and therefore more important to the development of generally useful technology.

PLANS FOR THE COMING YEAR

- Collaborate with other representatives of the written language NLP community to define and implement new performance evaluations.
- Coordinate a dry run of the evaluations.
- Issue call for participation in the formal evaluation and conference.