INFORMATION EXTRACTION AND EVALUATION

Lisa F. Rau
Information Technology Laboratory
GE Research and Development
Schenectady, NY 12301 USA
rau@crd.ge.com

This topic session focussed on a variety of issues in evaluation (three presentations) and extraction (one presentation). For the extraction presentation, Tsuyoshi Kitani, a visiting researcher at the Center for Machine Translation at Carnegie-Mellon University gave a presention entitled "Overview of TEXTRACT Template-Filling Solutions". This talk gave an overview of TEXTRACT, which processes articles in the Japanese joint venture and microelectronics domains. Although TEXTRACT was developed at CMU as an optional system of the GE-CMU SHOGUN system, the systems share no code beyond the MAJESTY morphological analyzer, and some of the knowledge in TEXTRACT was used to develop SHOGUN.

TEXTRACT is comprised of four major components: preprocessing, pattern matching, discourse processing, and template generating. A method of identifying company names was discussed, as the correct identification of company names is key to achieving a high performance level under the template structure of the joint venture domain. The discourse processing, which merges individual pieces of information identified by the sentence level pattern matcher, was also described.

Nancy Chinchor, SAIC, talked about "Balancing the elements of evaluation". The successful evaluation of systems requires the balancing of elements of the evaluation. She defined the elements of evaluation, the opposing forces within each element and between elements, and methods used to resolve these opposing forces. The dangers of not balancing the elements individually and altogether were pointed out. Her reflection on the evaluation offered unmathematical measures of our success and thoughts for future endeavors.

Jerry Hobbs, SRI International, in his talk "In Defense of Recall and Precision" gave some excellent arguments why the older measures of recall and precision were more appropriate for data extraction systems than the newer error rate. First he defined how information extraction sets up a correspondence between the world or text and facts within it, and a database and the items in the database. With this correspondence, recall naturally corresponds to the question "for every fact in the world/text, is there a corresponding item in the database?". Similarly, precision answers the question "for every item in the database, is there a corresponding fact in the world/text?". Recall and precision have natural correspondences to both the development cycle and the user's environment. In the development cycle, examining the corpus to determine how to modify the system increases recall. Testing the system on the corpus to put constraints on the system increases precision. In the user's environment, low recall can be fixed by increasing the redundancy of the corpus, and low precision can be improved by adding a user in the system processing loop. Moreover, the F-measure exhibits the desirable properties of being highest when both recall and precision are high. Jerry went on to claim that error rate is an appropriate evaluation measure when there is a one-to-one mapping between "key" and "response", as in speech recognition, but that with data extraction, where there are multiple possible fills for some slots, the measure is not appropriate. However there the differences between the two measures are so small that at least for MUC-5, error rate is identical to the F-measure.

Finally, Lisa Rau, GE R&D, noted the importance of defining requirements in advance, and as an integrated team including funding agencies, end users, evaluators and contractors to prevent wasted time and money in system redesign. Also, the implications of the template design on the system design were addressed. The frequency of occurrence of each slot, how easy or hard each slot is to fill and the interdependencies among slots all have an influence system design and should be addressed during the template design phase. It was noted that the differences in scores caused by changes in the algorithms used in the scoring program were dwarfed by differences in score attributable to the template design, such as default fills, the interpretation of the "correct answer" and the decision to copy objects in sentences such as "This process is similar to that used in France, Germany and Japan.". Finally, there was a discussion about the tradeoff between realistic tasks that require large amounts of non-language processing system engineering (such as TIPSTER), and simpler tasks that might take less system engineering and push research in natural language text interpretation more.