ANNOTATION OF ATIS DATA

Kate Hunicke-Smith, Project Leader
Jared Bernstein, Principal Investigator

SRI International
Menlo Park, California 94025

PROJECT GOALS

The performance of spoken language systems on utterances from the ATIS domain is evaluated by comparing system-produced responses with hand-crafted (and -verified) standard responses to the same utterances. The objective of SRI's annotation project is to provide SLS system developers with the range of correct responses to human utterances produced during experimental sessions with ATIS domain interactive systems. These correct responses are then used in system training and evaluation.

RECENT RESULTS

Since June 1991, SRI has produced classification and response files for about 9,000 utterances of training data (2900 of these since February, 1992). A dry run system evaluation and two official evaluations have been held since the project began in 1991. SRI has produced the standard responses for all of these evaluations; in all, about 2300 utterances.

These tests were performed according to the Common Answer Specification (CAS) protocol which is used in training. All systems are evaluated on a common set of data, with system responses measured against official reference answers produced at SRI in the same manner as the training data.

In addition to producing the classification and standard response files, SRI takes an active role in the adjudication of test and training data bug reports, initiates nearly all of the changes to the *Principles of Interpretation* document (a basic set of principles for interpreting the meaning of ATIS sentences agreed upon by the DARPA community), and continues to support NIST by modifying software and acting as a consultant regarding the annotation of data.

In 1992 the DARPA community developed a complementary evaluation method, referred to alternately as "end-to-end" or "logfile" evaluation, to better evaluate system-user interfaces. Using an interactive program developed by David Goodine at MIT, human evaluators from SRI and NIST used this end-to-end method in a dry run evaluation in December, 1992. For each query/response pair in 128 interactions, the human evaluators judged the correctness or appropriateness of system responses, and classified the type of user request and type of system response.

The use of human evaluators allowed more flexibility in scoring than an automatic, comparator-based method. This method allowed partial correctness judgements and the opportunity to score system responses which were not database retrievals, such as diagnostic messages and directives to the user. Because human evaluators saw the interaction from the point of view of the user, the results of the logfile evaluation method aided system developers by identifying dialogue and user interface problems which were not indicated by the usual CAS evaluation method.

PLANS FOR THE COMING YEAR

In the next year, SRI will continue to provide MADCOW annotation and other services to the DARPA community.