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 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

In a previous project, SRI devised a set of procedures for transcribing and annotating utterances collected during interactions between human subjects and a simulated voice-input computer system that provided information on air travel, i.e. utterances in the ATIS domain. During spring of 1991, it was decided to expand the collection of these human-machine interactions so that most DARPA speech and natural language sites would be collecting this type of data. However, SRI was to remain the only site providing the 'standard' answers.

At the start of the project, a basic set of principles for interpreting the meaning of ATIS utterances was agreed upon by the DARPA community and documented in a network-accessible file known as the *Principles of Interpretation*. Initial annotation procedures used at that time at SRI were documented in a net note dated July 12, 1991.

During the earlier project, SRI had installed software that produced answer files in the format required by NIST. The essential component of the software used by the annotators was and is NLParse, a menu-driven program developed by Texas Instruments that converts English-like sentences into database queries expressed in SQL.

As standard responses were generated for use in system training, some aspects of the *Principles of Interpretation* were changed. This process has continued throughout the project. In July, SRI worked with NIST to establish a committee of representatives from each data collection site to modify the *Principles of Interpretation* document as needed. The SRI annotators have worked closely with this committee, contributing knowledge of the data corpus gained in the annotation process.

Software used in the production of the standard response files was modified and expanded upon. SRI modified NLParse itself to accommodate changes in the software environment and new testing rules which limited the size of legal answers. Also, a few high-level programs were written to drive and monitor the results of the various low-level routines that had been used in SRI's previous project. This consolidation eliminated the need for annotators to monitor the process at each stage, thus eliminating opportunities for human error.

A dry-run system evaluation was held in October, 1991, for which SRI produced the standard responses. The dry run offered an opportunity to measure the real accuracy of a sample of 'standard' responses in the annotated data. In the dry run test, about 6% of the annotations were incomplete or inappropriate in some way; some due to human error and some to software error. In an effort to improve data quality, SRI revised its human checking procedures and added new checking programs to the software involved in the production of the answer files. It had originally been hoped that human double checking could be decreased after an initial period of annotation, but based on the adjudication of the dry run data, 100% double checking has continued.

Since June, 1991, SRI has produced classification and response files for 8000 utterances of training data, and nearly 1000 utterances of test data. Currently, we annotate about 190 utterances per annotator-week.

For the first official system evaluations in February, 1992, SRI again worked with NIST to produce the standard response files.

PLANS FOR THE COMING YEAR

In the next year, SRI will work with NIST and the DARPA community to develop and implement more efficient evaluation procedures for SLS systems.