SESSION 3: SPOKEN LANGUAGE SYSTEMS III

John Makhoul, Chair

BBN Systems and Technologies, 10 Moulton St., Cambridge, MA 02138

BACKGROUND

Two years ago, the DARPA Spoken Language Systems (SLS) Coordinating Committee made a decision to develop the ATIS (Airline Travel Information System) database for use as a common domain in which spoken language systems will be developed and evaluated. Since then, there has been significant work done to develop

- a consistent and rich ATIS database,
- data collection methodologies and scenarios, and
- methods for use in the common evaluation of spontaneous speech recognition and understanding of text and speech input.

Previously, there had been two sets of evaluations, in June 90 and February 91, of initial versions of the ATIS database. In both cases, the available data to be used for training was minimal and most of the speech training data was read, with only a small amount of spontaneous training data.

Since February 91, the ATIS database has been updated and, in an effort to quickly collect a larger amount of training and test data, a concerted effort has taken place in collecting data at five different sites (AT&T, BBN, CMU, MIT, and SRI). (See [1] for details.) About 10,000 spontaneous utterances were collected, of which about half were annotated (text transcriptions, reference answers, etc.) by December 20. Thus, for the first time since the decision to adopt ATIS as the common task for evaluation, the different sites had available to them sufficient amounts of training data that is similar in nature to the data to be used in testing the systems, albeit the different sites did not have much time to work on the new data before the evaluation was performed.

Also, in the last two years, there have been changes in the evaluation methodologies. For the evaluation of spontaneous speech, the methodology has not changed much. The error rate is still computed as the sum of substitutions, deletions, and insertions, given a transcription of the speech. (Word fragments and nonspeech events are not included in the evaluation.) Since the percentage of new words in the test data has been quite minimal, no special consideration for new words is made. For evaluating natural language understanding from text and spoken language understanding from speech, the answer to a query is compared against a reference answer. The understanding error rate is then computed as the sum of the percentage of queries for which a system gives 'no answer' and twice the percentage of queries for which the system gives a false answer.

THE SESSION

This session was devoted to presentations from the six sites that performed evaluations on the February 92 ATIS speech, natural language, and spoken language tests. These sites included AT&T, BBN, CMU, MIT, Paramax, and SRI.

The results show considerable performance improvements since a year ago. In speech recognition, much of the improvement in performance is attributable to the significant increase in the amount of appropriate training data, which allowed the development of better acoustic models and better language models. In natural language understanding, there has also been substantial improvement in performance, due to further system development as well as the availability of more appropriate training data.

Much of the discussion period centered on the differences in performance on data collected from the different sites. For example, the error rates on the data collected at MIT were significantly lower than the others, while the data from AT&T and SRI resulted in higher error rates. These differences may have been due to the differences in the amounts of training data collected at the different sites [1]. Also, the AT&T and SRI data appeared to possess a larger amount of spontaneous speech effects. In general, the fact that all subjects who were employed in the collection of data were unexperienced may have resulted in a higher overall error rate. There were calls to bring back some of the subjects for further testing to test the effects of subject experience on performance.

Now that a significant amount of training data is available, it will be interesting to see how much improvement in performance can be achieved by working on this data for a reasonable amount of time.

REFERENCES

[1] MADCOW, "Multi-Site Data Collection for a Spoken Language Corpus," Session 1 in this workshop.