SESSION 6: SPOKEN LANGUAGE SYSTEMS

Madeleine Bates, Chair

Speech & Language Department
BBN Systems & Technologies
70 Fawcett Street
Cambridge, MA 02138

What are spoken language systems? How do they differ from the speech recognition systems that are on the verge of becoming common in everything from consumer goods to military systems?

In the way that this community uses the term, a spoken language system (SLS) is one that incorporates both speech recognition and a large amount of language understanding, generally in the context of a specific task that is being carried out by the user. A simple "voice command" system would not qualify as an SLS, since little or no language processing is needed to translate the recognized word(s) into the appropriate action(s).

A system that is capable of understanding a wide range of very natural utterances must of necessity use some form of language processing in addition to speech recognition. The goals of research in the area of SLS are to allow fluent, more "natural" communication between people and computers, particularly when they are engaged in performing non-trivial tasks, such as planning or information retrieval.

SLSs combine the power (and the problems) of recognition and understanding. But SLS is more than the sum of its two parts. Attempts to develop SLSs inevitably inspire the need to make advances in other disciplines as well, ranging from human factors of user interfaces to prosodic analysis and language generation.

Effective SLSs, even in the laboratory, must meet a demanding set of requirements imposed by users who are accustomed to having what they say understood by highly intelligent agents (other people). Some of these requirements are: allowing fluent, natural speech, including a wide range of disfluencies; real-time performance; ease of use, particularly for new users; and high performance.

The focus of effort in developing spoken language systems has shifted during the course of the ARPA program that has supported some of the work reported here. Just last year at this workshop, most of the SLS papers were concerned with data collection for the domain known as ATIS3, the evaluation methodology that was being used to evaluate SLS systems in that domain, and the language processing techniques used in those systems.

This year, the papers in this session show that although some effort has been devoted to bettering existing SLS systems, research attention is beginning to turn beyond ATIS3, to focus more on interfaces and on the development of dialogue systems.

In the area of bettering existing SLS systems, the paper by Wayne Ward and Sunil Issar of Carnegie Mellon University discusses improvements that have been made to CMU's top-performing ATIS SLS system. They describe how they made maximum use of the limited amount of training data, generalized the lexicon and parser, and improved the resolution of ambiguous parses by using context.

Many SLS systems use a list of the N best-scoring hypotheses produced by the speech recognizer (typically N is 20 or fewer) as the interface between speech and language processing. The simplest type of integration has been to have the language processor try the sentences from the N-best list one at a time, stopping at the first one that is acceptable (understandable) to the language processor.

At SRI, Rayner, Carter, Digalakis, and Price improved their ATIS SLS system by concentrating on improving the N-best the interface between the speech recognition and language understanding components. SRI's innovation was to experiment with re-ordering the N-best list using multiple knowledge sources, such as whether a complete linguistic analysis was possible or not, and discriminants based on semantic classes, grammar rules, and semantic triples that embody the linguistic analysis of the utterance. The fact that this method can be generalized to incorporate new knowledge sources, and can be automatically trained, makes it an important addition to our methods for developing SLSs.

A third system that saw improvement was the Gisting system, reported in the paper by Marie Meteer and Robin Rohlicek at BBN. This system is not an ATIS system, or even a human-computer dialogue system. It attempts to extract particular types of phrases from off-the-air recordings of the speech of air traffic controllers and pilots.

In this task, the recognition performance is almost necessarily poor (given the poor quality of the input), but it is possible nonetheless to achieve respectable precision and recall results based on extracting phrases from the noisy speech. The innovation reported on here is that the same parser is used for both speech training and the interpretation of the recognized text.

In the area of dialogue, we know that we do not know nearly enough about what kind of dialogues occur between people and machines; they are certainly different from human-human dialogues in interesting ways, but their structure is not well understood.

The PEGASUS system, described in the paper by Victor Zue and others at MIT, is being used for research focused on dialogue management. This system, which is an ATIS-like system connected to a live airline reservation system (EAASY SABRE) accessible over the telephone, features a System Manager that monitors the user's dialogue state and the state of completeness of the booking the user is trying to complete, as well as the state of the underlying application system PEGASUS has been used by several people in that laboratory during the last year to plan their actual trips and make reservations for them.

The WAXHOLM system described in the paper by Rolf Carlson of KTH is another dialogue-based system that goes beyond ATIS3. Its domain, although travel related, is travel by boat in the Stockholm archipelago. The goal of WAXHOLM to provide a set of tools for generic dialogue systems which include speech synthesis, speech recognition, language understanding, and graphical output. WAXHOLM uses a dialogue grammar that employs probabilities for topic selection, and artificial neural networks for speech recognition.. Both WAXHOLM and PEGASUS use speech as an important output modality in addition to the screen display.

The problem of disfluencies in speech has long been pointed to as one of the factors that makes understanding language that comes from speech a more difficult problem than understanding language that originates in text form. In a paper that is an interesting change from those that describe existing SLS systems, Sharon Oviatt of SRI discusses laboratory experiments aimed at improving human's dialogues with computers by preventing or minimizing disfluencies instead of simply coping with them after they occur.

Certainly, if we can design systems that result in fewer hesitations, filled pauses, restarts, mid-utterance corrections, mispronunciations, and other speech errors, the success rate of the systems will go up, and with it, user acceptance. The predictive model that Oviatt presents shows, for example, that the rate of disfluencies is a linear function of the length of the utterance. Another result that should be useful to SLS builders everywhere is that systems that allow unconstrained input have to deal with more disfluencies than systems that present a more structured interface to the user. Reducing the amount of planning the speaker must accomplish before (or while) speaking reduces the number of disfluencies produced.

Taken as a whole, the papers in this session represent a widening interest in SLSs, and a continued commitment to developing techniques that will make high-performing SLS systems broadly applicable and usable.