SESSION I: EVALUATING SPOKEN LANGUAGE SYSTEMS

James F. Allen
Department of Computer Science
University of Rochester
Rochester, NY 14627

This session concerns the evaluation of spoken language systems. To understand the issues, it will help to briefly review the history of evaluation in the Spoken Language Systems program (see Figure 1).

The existing methods for evaluation evolved from the techniques used for the speech recognition systems. In tasks such as resource management, there is a closed vocabulary and the data is read speech. The main evaluation criteria is recognition accuracy, i.e., how many words in the test set are recognized correctly. To perform this evaluation, the researchers need a large database of read speech. Such data is relatively inexpensive to obtain in sufficient quantities.

With the ATIS domain, the task is generalized to question answering. Systems are dealing with an open vocabulary and spontaneous speech and the primary criteria for evaluating systems is the correctness of the answer given for each query. To perform this evaluation, the researchers need a large database of spontaneous questions annotated with the appropriate answer. Not only is the data collection process more complex as one needs spontaneous speech, but an order of magnitude more data is needed for training and evaluation (since there are many words in every utterance).

The ATIS task has evolved to the stage where a new concern is handling dialog. For this, the systems must deal with spontaneous natural speech in context. The evaluation criteria for such systems is not yet clear, and two of the papers in this session put forth some initial experiments with possible evaluation techniques. Whatever the evaluation technique, however, it is clear that the researchers now need an even larger database of dialogs both for training and evaluation.

At the present time, the spoken language program is in transition to the last stage described above. Systems are starting to try to deal with dialogs, but the existing evaluation techniques are only appropriate for question-answering tasks. The papers in this session offer an interesting perspective of the issues involved in making this transition.

In particular, the papers discussed three crucial issues:

- Where do we get all the data that is needed?
- How are we currently doing (at question-answering)?
- What are appropriate evaluation metrics for dialog systems?

The MADCOW paper describes the data collection effort in the last year. At each stage of development - from speech recognition, to question-answering, to dialog systems - there is an order of magnitude increase in the amount of data needed for training and evaluation. While it was possible in the early stages to have a single data collection and analysis site, it was clear that not enough dialogs could be collected rapidly enough under the old scheme. The MADCOW effort involves collecting data at all the different SLS sites, and co-ordinating the annotation of the data and its distribution.

The second paper gives the results of the latest ATIS benchmarks. Most of the results are straightforward to interpret and need no further comment here. But it is important to not confuse the full-session evaluation performed this time with a dialog evaluation. Since this is an issue that is easy to misinterpret, and since it lays the groundwork for the remaining papers in the session, I will discuss this further here.

Full session evaluation consists of testing systems on entire dialogs as they occurred in data collection. Each utterance is annotated with the correct answer. But there is no precise notion of a "correct" answer, because often many answers are possible and equally correct. For example, one answer might give more information than another because it is relevant. Often, the answer that contains the minimal amount of information requested would in fact be quite unhelpful. For example, consider a system that answered the question "What are the fares for flights from Boston to San Francisco?" by simply listing the fares without identifying what flights had what fares. The answer might be "correct" but uncooperative. On the other hand, we would not want to allow arbitrary extra information, as then the optimal scoring strategy would be for systems to list all information about anything that is mentioned in a query, or even list the entire database, not a helpful response.

As a start towards handling this problem, each query is annotated with a minimum an a maximum answer. The minimum answer contains just that data that is explicitly asked for in the query. Any system answer that does not contain all of the minimum answer would be incorrect. The maximum answer, on the other hand, includes all information that could be relevant to the query. Any system answer that includes more information than the maximum answer is incorrect.

Domain	Evaluation Measure	Data Needs
Closed vocab, read speech	Recognition Accuracy	Many words
Open vocab, spontaneous	Answer Accuracy	Many sentences with answers
Open vocab, spontaneous dialogue	? ? ?	Many dialogs

Figure 1: A mini-history of the SLS evaluation

Since the dialogs are transcripts of actual human performance, they do occasionally contain utterances that are simply not comprehensible, or are off topic. It would not be reasonable for the systems to be able to handle such utterances. To account for this, and to obtain information on how systems handle context dependency, all utterances in the dialogs are classified into one of three classes:

- Class A - queries that are answerable independent of context;
- Class D - queries that require context set by previous utterances;
- Class X - unanswerable queries.

While the type of the utterance is provided in the training data, it was not revealed on the test data. Thus, while evaluation results are tabulated using this classification, the systems did not have this information available when they were tested.

While the above procedure might seem to test dialog handling capabilities at first glance, this is misleading. It is important to remember that a transcript is just one possible dialogue between the individuals involved. If we put the same people back in the same situation with the same task, they would almost surely have a different dialog. This is because at any stage of the dialog, there are always many possible questions that could be asked and many possible answers to each question. Even if the system is restricted to only answering questions and not taking any initiative on its own, we saw that there are many possible reasonable answers. But different answers, while all reasonable, might lead to different continuations in a dialog. As a result, a transcript-based evaluation of dialog could at best test a systems ability to track an existing dialog, rather than partake in a dialog fully.

To conclude, dialog evaluation cannot be reduced to individual answer evaluation. Furthermore, there does not seem to be a plausible way to generalize the evaluation techniques based on transcripts. Think about how many dialogs would have to be collected to characterize the range of acceptable dialogs for even a simple single task! One would need a separate dialog for every possible variation that could occur in any question or answer.

Rather, we need new reliable, objective measures for dialog evaluation. To be objective means that the results are reproducible. So while some proposals discussed in this session use subjective evaluations of judges to score a dialog, if these judgements can be obtained reliably from different judges, then the measure is reproducible and thus as objective as any other measure.