Performance Measures for the Next Generation of Spoken Natural Language Dialog Systems

Ronnie W. Smith
Department of Mathematics
Computer Science Subarea
East Carolina University
Greenville, NC 27858, USA
rws@cs.ecu.edu

1 Improved Performance in Spoken Natural Language Dialog Systems

Since approximately the mid 1980's, technology has been adequate (if not ideal) for researchers to construct spoken natural language dialog systems (SNLDS) in order to test theories of natural language processing and to see what machines were capable of based on current technological limits. Over the course of time, a few systems have been constructed in sufficient detail and robustness to enable some evaluation of the systems. For the most part, these systems were greatly limited by the available speech recognition technology. Continuous speech systems required speaker dependent training and restricted vocabularies, but still had such a large number of misrecognitions that this tended to be the limiting factor in the success of the system. For example, testing in 1991 of the Circuit Fix-It Shop of (Smith, Hipp, and Biermann, 1995) required an experimenter to remain in the room in order to notify the user when misrecognitions occurred.

Fortunately, speech recognition capabilities are improving, and systems are being constructed that allow individuals to walk up and use them after a brief orientation. One example is the TRAINS system of (Allen et al., 1995) that was demonstrated at the 1995 ACL conference, where people just sat down and used the system after a brief set of instructions were given to them by the demonstrator. Another example is the current system under development at Duke University that serves as a tutor for liberal arts students learning the basics of Pascal programming. In this system, the machine itself explains how to use it. More thorough and challenging methods of evaluation are now feasible. This paper proposes some measures for evaluation based on a retrospective look at measures used in the past, analyzing their relevance in today's environment.

For the future, expect measurements of speech recognition performance and basic utterance understanding to remain important, but there should also be more emphasis on measuring robustness and measuring the utility of domain-independent knowledge about dialog. Furthermore, we should expect real-time response from evaluated systems, a sharp reduction in the amount of specialized training for using systems, and the use of longitudinal studies to see how user behavior evolves.

2 Fundamentals in Evaluation

2.1 Linguistic Coverage

A forward looking view of evaluation is offered by (Whittaker and Stenton, 1989). It is forward looking in the sense that they investigated issues in evaluation independent of building a system. Their perspective was not based on a specific SNLDS, but a general analysis of the issue of evaluation. Their main point was that evaluation needed to be placed within the context of a system's use. Consequently, they used a Wizard of Oz study in an information retrieval environment (e.g., database query) in order to identify the types of natural language inputs a typical user would use in order to gain access to needed information. Their analysis identified the following requirements for the linguistic coverage of a dialog system in the information retrieval environment: (1) operators for specifying the properties of the set of objects for which information would be requested; (2) contextual references; and (3) references to the actual source of information (e.g., the database). In general, linguistic coverage of SNLDS in the past has been limited, and to the extent that limitations will exist in the next generation of SNLDS, such limitations need to be measured and described.

2.2 Early System Performance Measures

(Biermann, Fineman, and Heidlage, 1992) report on the results of testing their VIPX system of the mid 1980's which offers users the ability to display and manipulate text on a computer terminal using spoken English supported by touches of the screen. The main dimensions which they evaluated were: (1) Learnability, (2) Correctness, (3) Timing, and (4) User Response. Learnability measures how easily subjects could learn to communicate with the machine. Correctness measures whether or not there was successful completion of the task. Timing describes the rate at which work was completed. User response measures how users felt about using the system.

These general categories of performance measures can be broken down into more precisely defined and quantifiable measures. Information on learnability and user response can be elicited via a subject survey and through comparison to alternative forms of user interface for completing the same task (e.g., discrete speech versus continuous speech, keyboard vs. speech input, and speech vs. multimodal input). In the next section we examine some of the measures relevant to correctness and timing and discuss their relevance for future evaluation of SNLDS.

3 Measures of Correctness and Timing: Past, Present, and Future

3.1 Recognition Rate

This measure has been expressed in a variety of ways. Its purpose is to describe the performance of the speech recognition component in terms of how accurately it converts the speech signal into the actual words uttered. Recognition rate and the overall success rate of the interaction are invariably highly correlated. This measure is still relevant. While recognition technology is improving, it is not perfect. In particular, telephone interactions provide a very challenging environment for speech recognition equipment. For the Dialogos system which answers inquiries about Italian Railway train schedules, (Billi, Castagneri, and Danieli, 1996) report only 68.2% word accuracy for the system in 96 dialogs. In spite of this Dialogos still understood 81.6% of all sentences, a promising result. As systems are tested in more challenging environments, the base level accuracy of the input signal remains an important benchmark in measuring system performance.

3.2 Perplexity

This measure is used to describe the amount of search that a speech recognition component must do in translating the input signal. The MINDS system of (Young et al., 1989) and the TINA system of (Seneff, 1992) represent speech systems that made use of various techniques for reducing the perplexity faced by the speech recognition component. TINA used probabilistic networks and semantic filtering to reduce perplexity. MINDS used predictions based on dialog context to reduce perplexity. While not a specific measure of a dialog system, an integrated dialog system such as MINDS can provide information that can reduce the perplexity that the speech recognition component must deal with. Consequently, comparative measures of perplexity with and without context-dependent predictions remain a valid measure for evaluating the performance of a dialog system, particularly in a complex linguistic environment where reduction of perplexity is essential for good speech recognizer performance.

3.3 Correctly Understood Utterances/Correctly Processed Queries

This measures how well a system processed utterances in isolation, but does not give the complete picture of system performance in a dialog where utterances are related through context. As the environments in which systems are tested become more challenging, the ability to handle partially understood utterances will be important. Measures for capturing the rate of success in situations where utterances are partially understood or perhaps even completely misunderstood are needed. Such measures must take the overall dialog context into account. One such measure has been proposed by (Danieli and Gerbino, 1995). They define the notion of "Implicit Recovery" (IR) as a measure of the ability of a system to filter the output of the parser and interpret it using contextual knowledge. In particular, an implicit recovery occurs when the system only partially understands an utterance, but still responds in an appropriate fashion. They also define what it means for a response to be appropriate within the context of an information retrieval situation. There is still a need for such definitions in a task assistance environment.

3.4 System Response Times

This measure was used in order to demonstrate the practical viability of systems/techniques when "the hardware gets faster." For the most part, near real-time performance was the best result obtained. However, as (Oviatt and Cohen, 1989) caution, speakers expect fast response times in a system that provides spoken interaction. If one expects to evaluate human-computer spoken language interaction, one will need a system that can give the quick responses that people normally expect in spoken interaction. It is hoped in the next generation of measuring SNLDS, system response time will no longer be a required measure, as systems will perform with real-time speed and not continually have awkward delays that break up the flow of the dialog.

3.5 Duration of the Interaction

An overall measure specifying how long it takes a user to complete the interaction, it provides a gross measure that can indicate interactional differences under different conditions, such as the level of system initiative ((Smith, Hipp, and Biermann, 1995)). Another way in which this is used is in comparing the efficiency of natural language interaction to other modes of communication that could be used for the given task. For example, (Biermann, Fineman, and Heidlage, 1992) as part of their overall evaluation of their voice and touch-driven text editor compare the time it takes to execute commands with their system to the time it takes people to complete the commands using the vi text editor. Comparing the speed at which someone can obtain information over the telephone by using a speech-based interface as opposed to the ubiquitous touch-tone interface with exhausting menu hierarchies that most businesses have (this seems to be true of businesses in the United States) might be very illuminating indeed!

3.6 Overall Interaction Success

This measures whether or not the interaction was successful (i.e., was the desired information obtained, or the required task completed?). Given the unfortunate circumstance that for the foreseeable future, some interactions will fail, this measure remains necessary. And if all interactions were successful, we might believe that the task was simply not challenging enough!

3.7 Frequency of System Failure/Error

The earliest systems were prone to frequent hardware and software failures. Robustness was measured in terms of how infrequently a system crashed. In other circumstances, system failure might be cast as "user error", because the user did not follow the allowed syntax or else spoke a word that was not in the recognizers vocabulary. As the state of the art progresses system errors are evolving into inappropriate responses rather than total system failure. It is hoped that system failure will disappear and be replaced by system robustness, that is, a measure of how well a system responds in error situations, either because of misunderstandings by itself, or because of misstatements by the human user.

4 New Issues in Evaluation

4.1 A Reduction in the Training Regimen

Due to their brittle nature and the limits of speech recognition technology, rigorous experimental evaluation of systems required extensive training by subjects before testing began. This training involved recording of voice patterns for speaker-dependent speech recognition as well as training on the restricted vocabulary and syntax that systems required. Thus, as reported in (Smith and Hipp, 1994) for the Circuit Fix-It Shop much care had to be taken to get users to speak somewhat naturally, while still remaining within the linguistic coverage capabilities of the system. In the future, we hope that such restrictions will not be necessary, or at the very least, be greatly lessened.

Speaker-independent continuous speech recognition technology is now available, so the amount of time required to enable a person to interact with an SNLDS is much less. As mentioned previously, the TRAINS system demoed at ACL 1995 did not require any particular training other than being told the task you were trying to complete, being given a brief description of the screen layout on the console you were viewing, and the encouragement to talk to the machine like you would talk to a human assistant. On the other hand, they were using the system in a "data collection" mode at that point rather than in a formal experimental evaluation of the system. Depending on the nature of the task, the amount of training required will be varied and still needs to be reported. Care must be taken in any training not to overly bias the type of linguistic behavior that users will exhibit, if claims of general capability and robustness are to be validated.

4.2 Measuring the Utility of Domain-Independent Information

SNLDS cannot succeed without a strong base of domain knowledge. Nevertheless, if our main research focus is on our theories of natural language processing, we would like to justify our theory by showing how well it performs. One way to capture this would be the development of measures that show the utility of domain-independent dialog knowledge as compared to domain-specific information which a system contains. For example, some inputs to the system will be contextually self-contained (e.g., "The red switch is in the off position" when there is only one red switch in the domain), while other inputs require the use of dialog knowledge to be understood. When reporting the percentage of utterances correctly understood, it may be illuminating to report the cause of the utterances not understood—is it because of a lack of domain knowledge, a lack of vocabulary, or a lack of ability at doing contextual interpretation? Such measures can be helpful in determining the usefulness of a theory of dialog processing as well as determining future directions for research.

5 An Idealized View of Evaluation

For the future testing of systems, I hope to see the following: systems that (1) interact with users in a complex problem-solving domain where both the user and system have knowledge about what problem is being solved; (2) do not need experimenters to act as intermediaries between system and user as the system and user will be able to collaborate via spoken natural language to a successful conclusion; and (3) allow users to be ready to use them after less than five minutes of instruction.

Due to the wide ranging motivations of funding agencies and the world-wide interest in SNLDS, it is not likely that we will find a common task for which everyone will implement their model of dialog processing and then be able to test them all on a common set of problems to see which one performs better. Consequently, when reporting evaluations, a variety of measures will be needed in order to allow ones colleagues to gain an idea of the effectiveness of the system. These measures should include (1) speech recognition accuracy; (2) the utility of domain-independent knowledge about dialog; (3) the nature and effectiveness of system error handling; and (4) comparisons of effectiveness for multiple interaction styles.

Furthermore, public access to transcripts and the production of videotapes of subjects in the actual experimental situation should also be part of the evaluation framework. In environments where one may encounter novices, experts, or individuals with intermediate expertise, the ability to interact in a variety of styles becomes essential. Longitudinal studies with subjects in such environments are the only way to gain an idea of a system's success in dealing with such a situation. Only through careful evaluation and full reporting of the results can the community of researchers as well as the general public gain an understanding of the current abilities and the future potential of SNLDS.

6 Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. IRI-9501571.

References

Allen, J.F., L.K. Schubert, G. Ferguson, P. Heeman, C.H. Hwang, T. Kato, M. Light, N. Martin, B. Miller, M. Poesio, and D.R. Traum. 1995. The TRAINS project: a case study in building a conversational planning agent. *Journal of Experimental and Theoretical Artificial Intelligence*, 7:7–48.

Biermann, Alan W., Linda Fineman, and J. Francis Heidlage. 1992. A voice- and touch-driven natural language editor and its performance. *International Journal of Man-Machine Studies*, 37:1–21.

Billi, R., G. Castagneri, and M. Danieli. 1996. Field trial evaluations of two different information inquiry systems. In *Proceedings of the Third IEEE Workshop on Interactive Voice Technologies Telecommunications Applications*.

Danieli, Morena and Elisabetta Gerbino. 1995. Metrics for evaluating dialogue strategies in a spoken language system. In *Proceedings of the 1995 AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, pages 34–39.

Oviatt, Sharon L. and Philip R. Cohen. 1989. The effects of interaction on spoken discourse. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 126–134.

Seneff, S. 1992. TINA: A natural language system for spoken language applications. *Computational Linguistics*, pages 61–86, March.

Smith, R.W. and D.R. Hipp. 1994. *Spoken Natural Language Dialog Systems: A Practical Approach*. Oxford University Press, New York.

Smith, R.W., D.R. Hipp, and A.W. Biermann. 1995. An architecture for voice dialog systems based on Prolog-style theorem-proving. *Computational Linguistics*, pages 281–320.

Whittaker, Steve and Phil Stenton. 1989. User studies and the design of natural language systems. In *Proceedings of the Fourth Conference of the European Chapter of the Association for Computational Linguistics*, pages 116–123.

Young, S.R., A.G. Hauptmann, W.H. Ward, E.T. Smith, and P. Werner. 1989. High level knowledge sources in usable speech recognition systems. *Communications of the ACM*, pages 183–194, February.