Evaluating Real-Time Responsiveness in Dialog

Nigel G. Ward

Department of Computer Science
University of Texas at El Paso
nigelward@acm.org

Abstract

This paper discusses ten difficulties in extending spoken dialog systems to exhibit real-time social skills.

1. Introduction

Over the past twelve years I have been modeling various aspects of 'real-time responsiveness', focusing on how interlocuters infer and respond to each other's needs, intentions, and feelings at the sub-second level. In dialog this is done by attention to the non-verbal cues produced unconsciously by speakers.

This research effort is one among many which seek to enable the development of spoken dialog systems capable of more 'sensitive', 'natural', 'satisfying', 'attentive', 'supportive' and 'responsive' interactions. So far we have built systems which produce back-channels (*uh-huh* etc.) at natural timing, which chose appropriate acknowlegements (*right*, *yeah*, *good*, etc.) based on the user's ephemeral emotions, which pace an explanation using turn-taking signals, and which adapt the speech rate to the user's needs [1, 2, 3, 4, 5]. One challenge in building systems with such abilities lies in discovering the cues and rules that people use, since the details of interaction at this level are below conscious attention and difficult to uncover [6].

Evaluation is another challenge; it is difficult to measure the utility of real-time responsiveness. This position paper contributes to the theme of this workshop by recounting some personal experiences and observations, presented in the form of ten observations regarding the immaturity of the field and the difficulty of working in this area.

2. Ten Laments

1. *Calibrated Evaluations are Unknown*

Even the best evaluations of advanced capabilities for spoken dialog systems generally demonstrate only a statistically significant improvement in usability or perceived quality from the addition of a new capability; which begs the question of whether this improvement is barely noticeable or something that really matters. Ideally one would like to have a standard metric of system quality, so that one could determine, for example, that innovation X is worth 1.7 times as much as innovation Y.

[0]This research was sponsored in part by DARPA and in part by NSF Grant No. 0415150.

2. *Discount Usability Evaluations are Unreliable*

In the GUI world it is possible to do rough usability evaluations by having an expert examine the properties of a system, perhaps using a checklist. Unfortunately the value of novel features for spoken dialog systems is difficult to estimate in this way, in part because there are differences between the perspective of a dialog participant and a dialog observer. For example, turn-taking timing that is normal for participants frequently appears too slow to observers. This problem, among others, plagues attempts to judge the value of an innovation by viewing system demos.

While some techniques in speech processing can be evaluated by how well the outputs match the behavior of human labelers or speakers, this is not the case for responsive behaviors, where a good match to corpus data does not always predict high satisfaction by users dealing with a live system.

A related problem is the lack of a method for estimating the overall quality of a system as a combination of the contributions of specific capabilities or directly measurable properties of dialogs. Since this is not possible even for systems built using mainstream technology for traditional applications [7], we are clearly a long way from being able to predict the value of an innovative method without actually building it into a system and testing it with uers.

3. *Real-Time Interpersonal Interaction is Below the Level of Conscious Attention*

When people are engaged in a task they are often not aware of the finer points of the interaction. It is possible to sharpen the perceptions of experimental subjects by subsequently having them listen to a recording of their interaction with the system, but this is time-consuming.

4. *Large Individual Differences Exist*

Although people with unusually good communication skills are highly valued in general, this is not universally true: not everyone enjoys interacting with highly responsive, socially sensitive dialog partners. Even if people come to expect machines to behave more like people, systems with better social skills will not be universally welcomed. A challenge for the future is the quick identification of the preferred communication style for each user.

5. *The Potential Value of Real-Time Responsiveness is Unestimated*

It is hard to make an economic case for more research on real-time responsiveness in its varios aspects, since there is no way to generate quantitative estimates of the value of such features. For example, it seems worthwhile to model the pragmatics of non-lexical utterances in dialog [8], but despite high hopes [9] there

is no easy way to estimate the value of this knowledge for dialog systems. Indeed, even the value of such abilities in human-human interaction is unknown, with rare exceptions [10].

6. *Low Quality Systems Still Sell*

In most cases the end users of spoken dialog systems have no choice in the matter, so the market pressures for better dialogs and improved usability are weak.

7. *Dialog Design is not Modular*

The final projects of my latest class, on developing spoken dialog systems in VoiceXML, showed that even to approach the commercial state of the art is a huge undertaking; one has to go a long way indeed before the lack of real-time responsivenss becomes the limiting factor in system quality. This is another reason why there is no clamor from those designing commercial systems for more advanced capabilities: they already have enough to worry about. Thus today those working on dialog systems fall into two camps: developers focusing on improving systems within the limits of today's technology [11], and researchers focusing on developing innovative capabilities, and the two camps have almost nothing to say to each other.

The situation is exacerbated by the difficulty of cleanly introducing new capabilities. For example, last year S. Kumar Mamidipally and I set out to develop a module that would choose an appropriate speaking rate for each utterance by the system. We wanted this to be autonomous, so that it could easily be plugged-in to existing systems and improve them with no need to redesign the dialog flow or other dialog features. We thought this would be possible because we expected this functionality to exemplify that dimension of social dynamics which is partly independent of the semantic and pragmatic dimensions. However this independence proved elusive: beyond the technical difficulty of integrating such functionality into existing dialog managers, the problem of choosing of an appropriate speaking rate turned out to be bound up with other dialog design choices, such as information elicitation strategy, dialog act choice, and choice of prompt wording.

8. *Formative Evaluation is Uncommon*

In the research sphere, evaluation of dialog systems is generally summative, done for the purpose of demonstrating that some proposed improvement does have value. In the industrial sphere, formative evaluation, that done for the purpose of determining what needs to be improved in an existing system, is more common. However there is also a need for formative evaluation in the research sphere, as a means for setting priorities for further research. Lacking this, research directions are often influenced more by visionary insights and pronouncements than by sober consideration of what is needed.

Detailed evaluation of the strengths and weaknesses of the state of the art is difficult, and good empirical methods are not known. We have explored the method of identifying "usability events" in human-computer dialogs and inferring the underlying problems [12], leading to some unexpected and potentially useful results, for example that swift turn-taking is more important than often thought, and that (for the billing domain) state-based dialog management is not a major limiting factor in usability. However there is a clear need for less labor-intensive and more replicable methods.

9. *The Path Forward is Unclear*

There seems to be no consensus about the mid-term future of spoken dialog systems. Ideally there should be a widely shared vision of the capabilities desired for VoiceXML 4.0 and what types of applications this will enable . . . and also for VoiceXML 5.0 and so on, for the next ten or twenty years.

10. *The Barriers to Entry are High*

Many researchers in social psychology, sociolinguistics, and conversation analysis have expertise analyzing the subtle phenomena of human communication. Few, however, are contributing to the improvement of spoken dialog systems. Ideally there would be shared, accessible testbeds and tools to make it easier for workers in neighboring fields to contribute in this area.

3. Prospects

This paper has noted some problems and issues that are holding back the development, evaluation, and adoption of advanced dialog capabilities. Fortunately none seem insurmountable.

4. References

[1] Ward, Nigel. Responsiveness in Dialog and Priorities for Language Research. *Systems and Cybernetics*, 28, pp 521–533, 1997.

[2] Iwase, Tatsuya and Nigel Ward. Pacing Spoken Directions to Suit the Listener. International Conference on Spoken Language Processing, pp 1203–1206, 1998.

[3] Ward, Nigel and Wataru Tsukahara. Prosodic Features which Cue Back-Channel Feedback in English and Japanese. *Journal of Pragmatics*, 23, 1177–1207. 2000.

[4] Ward, Nigel and Wataru Tsukahara. A Study in Responsiveness in Spoken Dialog, *International Journal of Human-Computer Studies*, 59, pp 603-630, 2003.

[5] Ward, Nigel and Satoshi Nakagawa. Automatic User-Adaptive Speaking Rate Selection, *International Journal of Speech Technology*, 7, pp 235-238. 2004.

[6] Ward, Nigel and Yaffa Al Bayyari. A Case Study in the Identification of Prosodic Cues to Turn-Taking: Back-Channeling in Arabic. in Interspeech, 2006.

[7] Möller, Sebastian. *Quality of Telephone-Based Spoken Dialogue Systems*. Springer Verlag, 2004.

[8] Ward, Nigel. Non-Lexical Conversational Sounds in American English. *Pragmatics and Cognition*, 14:1 (2006), 113-184.

[9] Ward, Nigel. The Challenge of Non-lexical Speech Sounds. International Conference on Spoken Language Processing, pp. II: 571–574, 2000.

[10] Bavelas, Janet, Linda Coates and Trudy Johnson. Listeners as Co-Narrators. *Journal of Personality and Social Psychology*, 79, pp 941–952, 2000.

[11] Heisterkamp, Paul. "Do not attempt to light with match!": Some Thoughts on Progress and Research Goals in Spoken Dialog Systems, Eurospeech 2003.

[12] Ward, Nigel, Anais G. Rivera, Karen Ward, and David G. Novick. Root Causes of Lost Time and User Stress in a Simple Dialog System. Interspeech 2005.