Ten Challenges in Highly-Interactive Dialog Systems

Nigel G. Ward* and David DeVault+
*Department of Computer Science, University of Texas at El Paso
+Institute for Creative Technologies, University of Southern California

Abstract

Systems capable of highly-interactive dialog have recently been developed in several domains. This paper considers how to build on these successes to make systems more robust, easier to develop, more adaptable, and more scientifically significant.

Introduction

Beginning about twenty years ago, dialog system researchers started building systems that go beyond rigid turn-taking and robotic interactions to exhibit human-like sensitivity to the user's behavior and state, swift and natural timing, and appropriately tailored behaviors (Ward and Tsukahara 1999; Gratch et al. 2007; Ward and Escalante-Ruiz 2009; DeVault, Sagae, and Traum 2009; Bohus and Horvitz 2011; Forbes-Riley and Litman 2011; Acosta and Ward 2011; Raux and Eskenazi 2012; Andrist, Mutlu, and Gleicher 2013; Meena, Skantze, and Gustafson 2014; Skantze, Hjalmarsson, and Oertel 2014; Ghigi et al. 2014). These systems can give users the impression of naturalness, engender rapport and improve task-related outcomes.

There are however ample opportunities for improvement. Beyond component-level performance and architectural challenges (e.g. (Buss and Schlangen 2010; Baumann 2013; DeVault and Traum 2013)), there are cross-cutting and methodological challenges in advancing this research. This position paper identifies some important issues in the construction of highly interactive systems and some possible strategies, based on our analysis of recent advances and on our own experiences in this area.

Towards More Power and Robustness

Challenge 1. Improving on Human Performance

Much research on highly-interactive dialog systems is motivated by a desire to achieve more natural and human-like system behavior. This is an important goal, but as a field we needn't stop at human-like interactive skills. Today it is easy to imagine dialog systems that are more knowledgeable, better reasoners, and use better diction than any human. Such systems are indeed a staple of science fiction movies. We can similarly imagine advanced systems that are "interactionally superior": better than any human in these respects also.

This possibility can be appreciated by listening to recordings of people in live conversation (especially yourself) and noting how inefficient and sometimes ineffective they are. While some disfluencies and awkwardnesses can be functional, to convey nuances or adjust the pace for the sake of the interlocutor, many are just regrettable. This is obvious in retrospect, when one can replay the recording to glean every detail of the interlocutors' behavior and can take the time to think about what should have been said and how, at each point in time. Future dialog systems, not subject to human cognitive limitations, might be able to do this in real time: to sense better, consider more factors, plan dialog steps further ahead, and so on, to attain superhuman charm and produce superhuman dialog.

While this is a very long-term goal, it raises a question relevant today: if corpus mining is our main resource for designing interactive behaviors, how can one produce systems that do better than the human speakers in the corpus? This issue has been, so far, only tentatively addressed. One approach is to look for consensus, to avoid modeling the "noise" inevitably seen in the behavior of any single individual (Huang, Morency, and Gratch 2010). A second approach is to selectively identify the very best interaction patterns in a corpus, and to base interaction models on these examples. A reinforcement learning framework could also prove useful, although making this work at the timescale of interactive behavior will clearly be difficult (Kim et al. 2014).

Challenge 2. Modeling Variation

Every dialog system today is carefully designed to work well with some target population of users. Recently there has also been significant work on certain topics in adaptation: how to adjust the fine details of some behavior to better suit individual users. But there is a larger question, that of how to design a family of systems, with the same basic functionality but with different personalities or behavior styles, that can be used for users of different types or preferences. We see the need for more research in interaction styles (Grothendieck, Gorin, and Borges 2011; Ranganath, Jurafsky, and McFarland 2013), so that system behavior can be better parameterized and adjusted at a high level.

Challenge 3. Using Multifunctional Behavior Specifications

Today most interactive systems require tightly constrained user behavior. While the constraints are often not explicit, a great deal of attention is paid to defining a genre that sets up expectations that lead the user to perform only a very limited set of behaviors: to follow the intended track. Under such constraints a system can behave strictly according to a role, which simplifies design and reduces the likelihood of failures due to unplanned-for inputs. However, designing around narrow tracks of interaction can lead system builders to adopt relatively impoverished models of the interactive behaviors for their systems.

Consider for example backchanneling behavior. This is a prototypical interactive behavior, probably the best studied one, and one already used successfully in several systems which backchannel at appropriate times in response to user speech (Ward and Tsukahara 1999; Fujie, Fukushima, and Kobayashi 2005; Schröder et al. 2012). Backchanneling demonstrations today work if the user has been guided to perform a specific type of dialog, such as retelling a story, answering personal-history questions, solving a puzzle, or engaging in smalltalk. Within one such activity type, good backchanneling for a specific function such as expressing agreement is possible. However, any single backchanneling policy does not generalize well to other activity types, and fails to incorporate many functions of backchannels. This can be seen from the result of a bottom-up study of where back-channels occur in unconstrained human-human dialog (Ward, Novick, and Vega 2012). This revealed 12 different activity types whose local presence strongly affects the probability of a back-channel, including rambling, expressing sympathy, negotiating the delivery of an upcoming important piece of information, deploring something, and being unsure, in addition to the obvious factors of being in a listener role and expressing empathy or agreement. Moreover, all 12 activity types had independent surface manifestations.

In general, there is a lot more going on in human interaction than we are modeling today. These aspects are not always hard to detect, and there have been good demonstrations of how to recognize user uncertainty (Forbes-Riley and Litman 2011) and various emotion-related user states (Schuller et al. 2013). Productively using such information in dialog systems, however, remains a challenge, and no work has yet demonstrated in practice how to track and deal with more than one such function at a time.

Challenge 4. Synthesizing Multifunctional Behaviors

Synthesizing multifunctional behaviors in current dialog system architectures is challenging. Today most behavior-related decisions are made in isolation. For example, a system might decide whether to produce a backchannel, and if so which word, and then for that word which prosodic form to use. While such individual decisions simplify design, making decisions jointly, optimized together, could help produce better outputs that serve multiple functions.

This need is clearest perhaps for speech synthesis. In human-human dialog, prosody is generally multifunctional. For example, in a single utterance like *"okay, well,"* prosody may serve to accept that the interlocutor's proposal is a valid one, flag that it was unexpected, indicate that there are countervailing factors that he's probably not aware of, convey that the speaker needs a second to marshal his thoughts, and project that he momentarily will propose an alternative.

Current speech synthesis techniques support concatenation but not the overlay of multiple prosodic patterns to express multiple functions simultaneously. Systems whose expressive needs are limited can make do with such synthesis (or, more commonly, rely on prerecorded speech), but not systems that have to satisfy combinations of expressive goals that are not statically-predetermined. Superpositional models of prosody do exist (van Santen, Mishra, and Klabbers 2004; Tilsen 2013; Ward 2014), but so far these are only descriptive. More generally, there is a need for the synthesis of multifunctional behaviors, including speech and animation, good enough for use directly without touch-up.

Challenge 5. Integrating Learned and Designed Behaviors

Today even the most interactive systems have a fixed skeleton specifying the overall dialog flow. Within this a few decision points may be left underspecified, for subsequent filling in with more data-driven decision rules. While ultimately dialog system behaviors might be entirely learned from data, for the foreseeable future interactive systems will include both learned and designed behaviors. Such hybrid systems involve not only technical challenges but also marketing challenges, since customers for dialog systems may not trust those that rely too much on machine learning for dialog flow (Paek and Pieraccini 2008).

We see an opportunity to explore new ways of integrating learned and designed behaviors, and in particular to develop architectures which give a larger role to behaviors learned from corpora. Perhaps the knowledge from corpora need not be distilled and shoehorned into discrete decision points. Instead behaviors in the corpus can perhaps be replayed and active at runtime, perhaps slightly abstracted and adjusted. Then the more task-oriented aspects of the system might take the form of independent constraints that, while maintaining local coherence, move the system, choice by choice, towards accomplishment of the overall dialog goals. Thus we might combine two styles of modeling, one metaphorically the kinematics, modeling motion as it follows observed patterns, and one metaphorically the dynamics, modeling motion as directed by external forces (Lee et al. 2014).

Challenge 6. Continuous State Tracking

Today's dialog systems have a lot of inertia in interaction. After making a decision (which usually happens infrequently, such as once per user turn-end), they stick with it, usually until they have delivered the full utterance and heard the user's response. Innovations in incremental processing can overcome this limitation, but in practice these are used just to add a few more decision points, for example when the user barges in, when a user's key words are recognized, or when some prosodic, gestural, or external event is detected.

Human-human interaction is different. People continuously track the current state of the dialog, not only when the other is speaking, but when speaking themselves. This involves not only fine attention to the interlocutor's gaze, gesture, and backchannels, but also self-monitoring: speakers monitor what words they say, what they sounded like after they said them, and what things are in the output buffer. For this they use their own recognition/understanding circuits to simultaneously emulate the listener's uptake and to compare their own actual performance with their intended message.

While implementing continuous state tracking won't be easy, the potential value is significant. Among other things, systems will be relieved of the pressure to make perfect decisions. If a system can track appropriateness and make mid-course corrections, then the risk associated with any individual decision is less, and initial choices of how to start turns can be more approximate.

Towards Reduced Development Costs

Challenge 7. Compositional Behavior Specification

As noted above, today most dialog systems have a fairly fixed dialog-flow skeleton, with some interactiveness around specific decision points. Such nodules of interactivity are limited in applicability to the exact context in which they appear. We would instead like to build interactive systems from more general conversational skills, as reusable components. For example, imagining that we have developed a general policy for choosing the next interview question, a general policy for showing empathy, and a general policy for supportive turn taking, we could imagine that these could be composed to produce a system capable of effective, natural, and warm first-encounter dialogs. That is, dialog systems might be built from communications skills that are decoupled from the overall dialog policy. Ultimately we would like to be able to compose behaviors learned from different corpora, to increase reuse and reduce development costs.

Challenge 8. More Unsupervised Learning

Today to develop a highly-interactive system, even one exhibiting only one or two forms of responsiveness, requires a major engineering effort. Machine learning can of course decrease the need for analysis by hand, but it brings its own costs and limitations, usually including the need to label or preprocess the data, to split it into turns or otherwise identify the decision points to which machine learning will be applied. We see the need for fully automatic discovery methods, completely unsupervised. One approach is to automatically detect recurring behavior patterns (Ward 2014). We think at run time these might be relatively easy to track and to superimpose, addressing Challenges 4, 6 and 7.

Towards Deeper Understanding

Challenge 9. Making Evaluation more Informative

Today, evaluating highly-interactive systems usually involves user studies with a final questionnaire. This is costly and not as informative as we would like (Ward 2012). In particular, it is difficult to relate user perceptions of system style — such as attentive, polite, considerate, supportive — to the details of the actual behaviors and the design choices underlying them — such as whether a certain state has a timeout of 1.2 or 1.8 seconds. We think this could be addressed in part by elaborating causal models of the relations between system properties and user perceptions (Möller and Ward 2008; Möller, Engelbrecht, and Schleicher 2008) to cover the more interactive aspects of dialog.

We also see a need to better map out the connections between interactive performance and overall system performance in highly responsive systems. For example, the virtual interviewer in the SimSensei Kiosk system (DeVault et al. 2014) is deliberately slow to take the floor after user speech ends, in support of the design goal of encouraging users to talk as much as possible. If this system's turn-taking were made lower latency and more natural, it could work against system design goals. A deeper understanding of the advantages and potential disadvantages of highly-interactive dialog behaviors across domains is needed.

Challenge 10. Engaging Social Scientists

The behaviors in today's dialog systems are seldom based on the findings of social scientists, and conversely, the results of dialog systems research are rarely noticed by them.

One reason is that the most interactive aspects of dialog systems are often not fully understandable: they may work, but it is hard to know why. There is a need for more comprehensible models. Ways to achieve this might include deeper analysis of what a learned model really has learned, more use of modeling techniques which are intrinsically more understandable, and more use of declarative representations of behaviors rather than decision algorithms. The latter may also lead to knowledge representations shareable across synthesis and recognition, addressing Challenges 3 and 4.

Regarding the other problem, the lack of social-science research contributing descriptions of interactive behaviors specific enough to use for dialog systems, we feel that computationalists could create more tools to support interaction analysis and modeling by non-technical people.

General Discussion

A mainstay of highly-interactive dialog research has been the production of "one-hit wonder" systems, and there remain many interesting issues that can explored with "just-get-it-working" methods. However, the field is maturing rapidly. As we strive to build systems that are more robust, more capable, and more understandable, the challenges and approaches discussed here will become more relevant.

Acknowledgments

This work was supported by the NSF (IIS-1449093 and IIS-1219253) and by the U.S. Army Research, Development, and Engineering Command (RDECOM). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views, position, or policy of the National Science Foundation or the United States Government, and no official endorsement should be inferred.

References

Acosta, J. C., and Ward, N. G. 2011. Achieving rapport with turn-by-turn, user-responsive emotional coloring. *Speech Communication* 53:1137–1148.

Andrist, S.; Mutlu, B.; and Gleicher, M. 2013. Conversational gaze aversion for virtual agents. In *Proceedings of Intelligent Virtual Agents (IVA)*.

Baumann, T. 2013. *Incremental Spoken Dialogue Processing: Architecture and Lower-level Components*. Ph.D. Dissertation, Universität Bielefeld, Germany.

Bohus, D., and Horvitz, E. 2011. Multiparty turn taking in situated dialog: Study, lessons, and directions. In *SIGdial*.

Buss, O., and Schlangen, D. 2010. Modelling sub-utterance phenomena in spoken dialogue systems. In *Proceedings of SemDial 2010 (Pozdial)*.

DeVault, D., and Traum, D. 2013. A method for the approximation of incremental understanding of explicit utterance meaning using predictive models in finite domains. In *NAACL-HLT*.

DeVault, D.; Artstein, R.; Benn, G.; Dey, T.; Fast, E.; Gainer, A.; Georgila, K.; Gratch, J.; Hartholt, A.; Lhommet, M.; Lucas, G.; Marsella, S.; Morbini, F.; Nazarian, A.; Scherer, S.; Stratou, G.; Suri, A.; Traum, D.; Wood, R.; Xu, Y.; Rizzo, A.; and Morency, L.-P. 2014. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of AAMAS*.

DeVault, D.; Sagae, K.; and Traum, D. 2009. Can I finish? Learning when to respond to incremental interpretation results in interactive dialogue. In *10th SigDial*.

Forbes-Riley, K., and Litman, D. 2011. Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. *Speech Communication* 53:1115–1136.

Fujie, S.; Fukushima, K.; and Kobayashi, T. 2005. Back-channel feedback generation using linguistic and nonlinguistic information and its application to spoken dialogue system. In *Interspeech*, 889–892.

Ghigi, F.; Eskenazi, M.; Torres, M. I.; and Lee, S. 2014. Incremental dialog processing in a task-oriented dialog. In *Interspeech*, 308–312.

Gratch, J.; Wang, N.; Okhmatovskaia, A.; Lamothe, F.; Morales, M.; van der Werf, R.; and Morency, L.-P. 2007. Can Virtual Humans Be More Engaging Than Real Ones? *Lecture Notes in Computer Science* 4552:286–297.

Grothendieck, J.; Gorin, A. L.; and Borges, N. M. 2011. Social correlates of turn-taking style. *Computer Speech and Language* 25:789–801.

Huang, L.; Morency, L.-P.; and Gratch, J. 2010. Parasocial consensus sampling: Combining multiple perspectives to learn virtual human behavior. In *9th Int'l Conf. on Autonomous Agents and Multi-Agent Systems*.

Kim, D.; Breslin, C.; Tsiakoulis, P.; Gasic, M.; Henderson, M.; and Young, S. 2014. Inverse reinforcement learning for micro-turn management. In *Interspeech*.

Lee, Y.; Wampler, K.; Bernstein, G.; Popović, J.; and Popović, Z. 2014. Motion fields for interactive character locomotion. *Communications of the ACM* 57(6):101–108.

Meena, R.; Skantze, G.; and Gustafson, J. 2014. Data-driven models for timing feedback responses in a Map Task dialogue system. *Computer Speech & Language* 28:903–922.

Möller, S., and Ward, N. 2008. A framework for model-based evaluation of spoken dialog systems. In *Sigdial*.

Möller, S.; Engelbrecht, K.-P.; and Schleicher, R. 2008. Predicting the quality and usability of spoken dialog services. *Speech Communication* 50:730–744.

Paek, T., and Pieraccini, R. 2008. Automating spoken dialogue management design using machine learning: An industry perspective. *Speech Communication* 50:716–729.

Ranganath, R.; Jurafsky, D.; and McFarland, D. 2013. Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates. *Computer Speech and Language* 27:89–115.

Raux, A., and Eskenazi, M. 2012. Optimizing the turn-taking behavior of task-oriented spoken dialog systems. *ACM Transactions on Speech and Language Processing (TSLP)* 9:1.

Schröder, M.; Bevacqua, E.; Cowie, R.; Eyben, F.; Gunes, H.; Heylen, D.; ter Maat, M.; Gary; Pammi, S.; Pantic, M.; Pelachaud, C.; Schuller, B.; de Sevin, E.; Valstar, M.; and Wollmer, M. 2012. Building autonomous sensitive artificial listeners. *IEEE Trans. Affective Computing* 3:165–183.

Schuller, B.; Steidl, S.; Batliner, A.; Burkhardt, F.; Devillers, L.; Müller, C.; and Narayanan, S. 2013. Paralinguistics in speech and language: state-of-the-art and the challenge. *Computer Speech & Language* 27:4–39.

Skantze, G.; Hjalmarsson, A.; and Oertel, C. 2014. Turn-taking, feedback and joint attention in situated human-robot interaction. *Speech Communication*.

Tilsen, S. 2013. A dynamical model of hierarchical selection and coordination in speech planning. *PLOS ONE* 8(4).

van Santen, J. P.; Mishra, T.; and Klabbers, E. 2004. Estimating phrase curves in the general superpositional intonation model. In *Fifth ISCA Workshop on Speech Synthesis*, 61–66.

Ward, N. G., and Escalante-Ruiz, R. 2009. Using subtle prosodic variation to acknowledge the user's current state. In *Interspeech*, 2431–2434.

Ward, N., and Tsukahara, W. 1999. A responsive dialog system. In Wilks, Y., ed., *Machine Conversations*. Kluwer. 169–174.

Ward, N. G.; Novick, D. G.; and Vega, A. 2012. Where in dialog space does uh-huh occur? In *Interdisciplinary Workshop on Feedback Behaviors in Dialog, at Interspeech 2012*.

Ward, N. G. 2012. Directions for research on spoken dialog systems, broadly defined. In *Workshop on Future Directions and Needs in the Spoken Dialog Community, at NAACL-HLT 2012*.

Ward, N. G. 2014. Automatic discovery of simply-composable prosodic elements. In *Speech Prosody*, 915–919.