Towards Situated Collaboration

Dan Bohus, Ece Kamar, Eric Horvitz
Microsoft Research
One Microsoft Way
Redmond, WA, 98052, USA
{dbohus,eckamar,horvitz@microsoft.com}

Abstract

We outline a set of key challenges for dialog management in physically situated interactive systems, and propose a core shift in perspective that places spoken dialog in the context of the larger collaborative challenge of managing parallel, coordinated actions in the open world.

Multiple models for dialog management have been proposed, studied, and evaluated in the research community (*i.a.* Allen et al, 2001; Bohus and Rudnicky, 2009; Rich and Sidner, 1998; Traum and Larsson, 2003; Williams and Young, 2007). In the process, a diverse set of problems have come to light and have been pursued. These include the challenges of modeling initiative in interaction, contextual interpretation and processing, the management of uncertainty, grounding, error handling and recovery, turn-taking and, more recently, incremental processing in dialog systems. Analyses of existing approaches (Allen et. al, 2001; Churcher et. al, 1997; McTear 2002; Paek and Pieraccini, 2008) reveal a constellation of benefits but also shortcomings along multiple dimensions, where no single technique provides the benefits of all.

While taking incremental, focused steps is important for making progress within a mature discipline, we believe that the current scope and conceptual borders of work in spoken dialog constrains thinking about possibilities and gets in the way of achieving breakthrough advances. Research to date on dialog management has focused almost exclusively on dyadic settings, where *a single user* interacts with a system over a relatively narrow, *speech-only* channel. Characteristics of this dominant and shared worldview on dialog research have driven modeling and architectural choices, and often done so in an implicit, hidden manner. For instance, dialog is often viewed as a collection of dialog moves that are timed in a relatively well-structured, sequential fashion. As a consequence, dialog management models typically operate on a "per-turn" basis: inputs are assumed to arrive sequentially and are processed one at a time; for each received input, discourse understanding is performed, and a corresponding response is generated.

In reality, interactions among actors situated in the open, physical world depart deeply from common assumptions made in spoken dialog research and bring into focus an array of important, new challenges (Horvitz, 2007; Bohus and Horvitz, 2010; Bohus, Horvitz, Kanda et al., eds., 2010). We describe some of the challenges with respect to dialog management, and re-frame this problem as an instance of the larger collaborative challenge of managing parallel, coordinated actions amidst a dynamically changing physical world.

As an example, consider a robot that has been given the responsibility of greeting, interacting, and escorting visitors in a building. In this setting, reasoning about the actors, objects and events and relationships in the scene can play a critical role in understanding and organizing the interactions. The surrounding environment provides *rich, continuously streaming situational context* that is relevant for determining the best way an agent might contribute to interactions. Because the situational context can evolve asynchronously with respect to turns in the conversation, systems that operate in the open world must be able to *plan continuously*,

in stream, rather than on a "per-turn" basis. Interaction and collaboration in these settings is best viewed as a flow of *coordinated, parallel actions*. The sequential structure of turns in dyadic interactions is but one example of such coordination, focused solely on linguistic actions. However, to successfully interact and collaborate with multiple participants in physically situated settings, an agent must be able to recognize, plan, and produce both linguistic and non-linguistic actions, and reason about potentially complex patterns of coordination between actions, *in-stream*—as they are being produced by the participants in the collaboration.

We argue that attaining the dream of fluid, seamless spoken language interaction with machines requires a fundamental shift in how we view dialog management. First, we need to move from *per-turn* to continual *in-stream* planning. Second, we need to move from reasoning about *sequential* actions to reasoning about *parallel and coordinated* actions and their influence on states in the world. And third, we need models that can *track and leverage the streaming situational context*, from noisy observations, to make decisions about how to best contribute to collaborations.

Spoken dialog is an important channel for expressing coordinative information. However, we need to recognize and begin to tackle head on the larger challenge of *situated collaborative activity management*. We understand that taking this perspective introduces new complexities—and that some of our colleagues will view diving into the larger problems in advance of solving simpler ones as being unwise. However, we believe that we must embrace the larger goals to make significant progress on the struggles with the simpler ones, and that the investment in solving challenges with physically situated collaboration will have eventual payoffs in enabling progress in spoken dialog.

Making progress on the broader challenge requires technical innovations, tools, and data. Consider for instance one sub-problem of belief tracking in these systems: continuously updating beliefs over the state of the collaborative activity and the situational context requires the development of new types of models that can combine streaming evidence about context collected through sensors, with discrete evidence about the actions performed or the turns spoken collected through speech, gesture or other action-recognition components. In addition, progress hinges on identifying a set of relevant problem domains, and coordinating efforts in the community to collect data, and comparatively evaluate proposed approaches. New tools geared towards analysis, visualization and debugging with streaming multimodal data are also required.

We propose a core shift of perspective and associated research agenda for moving from *dialog management* to *situated collaborative activity management*. We invite discussion on these ideas.

References

Allen, J.F., Byron, D.K., Dzikovska, M., Ferguson, G., Galescu, L., and Stent, A. 2001. Towards Conversational Human-Computer Interaction, *AI Magazine*, **22**(3)

Bohus, D., and Rudnicky, A. 2009. The Ravenclaw dialog management framework: Architecture and systems, in *Computer, Speech and Language*, **23**(3).

Bohus, D., and Horvitz, E. 2010. On the Challenges and Opportunities of Physically Situated Dialog, *AAAI Symposium on Dialog with Robots*, Arlington, VA.

Bohus, D., Horvitz, E., Kanda, T., Mutlu, B., Raux, A., editors, 2010. Special Issue on "Dialog with Robots", *AI Magazine* **32**(4).

Churcher, G. E., Atwell, E.S, and Souter, C. 1997 Dialogue Management Systems: a Survey and Overview, *Technical Report*, University of Leeds, Leeds, UK.

Horvitz, E., 2007. Reflections on Challenges and Promises of Mixed-Initiative Interaction, *AI Magazine* 28, pp. 19-22.

McTear, M.F. 2002. Spoken dialogue technology: enabling the conversational user interface, *ACM Computing Surveys*, **34**(1):90-169.

Paek, T., and Pierracini, R. 2008. Automating Spoken Dialogue Management design using machine learning: An industry perspective, *Speech Communication*, **50**(8-9):716-729.

Rich, C., and Sidner, C.L. 1998. Collagen: A Collaboration Manager for a Collaborative Interface Agent, *User Modelling and User Assisted Interaction*, **7**(3-4):315-350, Kluwer Academic Publishers.

Traum, D., and Larsson, S. 2003. The Information State Approach to Dialogue Management. *Current and New Directions in Discourse and Dialogue*, Text Speech and Language Technology, **22:**325-353.

Williams, J., and Young, S., 2007. Partially Observable Markov Decisions Processes for Spoken Dialog Systems, *Computer, Speech and Language*, **21**(2).

Young, S. 2006. Using POMDPs for Dialog Management, in *Proc. of SLT-2006*, Palm Beach, Aruba.