Technical and Design Challenges in Multimodal, Situated Human-Robot Dialogue

Allison Sauppé, Bilge Mutlu
Department of Computer Sciences
University of Wisconsin–Madison
1210 West Dayton Street
Madison, WI 53706 USA
asauppe@cs.wisc.edu, bilge@cs.wisc.edu

Abstract
As robotic products become more ubiquitous in society, fulfilling such roles as teachers and assembly-line workers, they will need to be capable of conversing with their users using spoken and nonverbal language. In this position paper, we discuss three challenges that researchers and designers face in creating interfaces for multimodal, situated human-robot dialogue: *language synthesis*, *multimodal input*, and *task modeling*. For each challenge, we present an envisioned scenario of use, discuss prior work in the area, and highlight currently open questions.

Author Keywords
Dialogue, human-robot interaction, multimodal input, task modeling

ACM Classification Keywords
H.5.2 [User Interfaces]: Natural language

Introduction
Prior work in the intersection of human-computer interaction and dialogue has traditionally focused on question-answer dialogue systems [6]. Applications of these systems, such as those in mobile computing, are primarily used as *tools* by their users to search and seek information and offload tasks such as typing to the computing platform. The vision for robotic products,

however, involves robots playing roles, such as teachers in classrooms, collaborators on assembly lines, and receptionists in offices, that require more agency and autonomy. In these roles, robots will need to communicate with their users using spoken and nonverbal language and engage in situated interactions. In this paper, we discuss some of the key technical and design challenges in achieving human-robot dialogue systems that support such interactions.

Challenges in Creating Interfaces for Human-Robot Dialogue

Dialogue interfaces that support multimodal, situated human-robot interactions will require the effective functioning of and a seamless coordination among a number of subsystems. For example, if a robot is assisting its user with assembling a set of pipes by giving assembly information, the robot will need to be capable of understanding the steps required to complete the task, how to comprehend the current task space in order to determine what steps have been completed, and how to communicate the remaining steps to the user. While there are numerous issues facing the design and development of a dialogue interface that supports such functionality, below, we highlight three technical and design challenges involved in creating systems for multimodal, situated human-robot dialogue: *language synthesis*, *multimodal input*, and *task modeling* (Figure 1).

We will contextualize these challenges in a running scenario of use that takes place in a manufacturing setting. Here, an informational robot and its user are working together to replace a faulty pipe in a manufacturing machine. This setting and partnership reflect one of the many future roles robots are envisioned to fulfill such as occupational training.

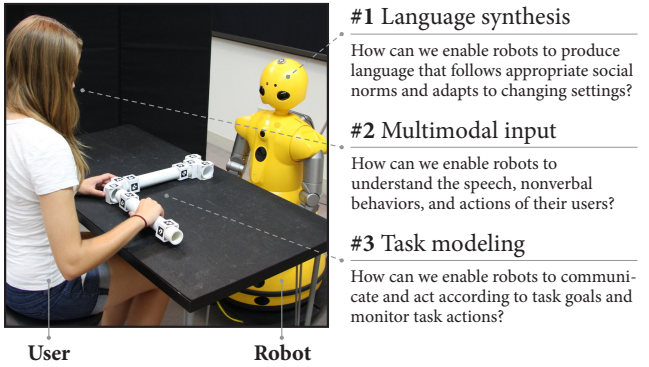

Figure 1: A summary of the three main challenges for multimodal, situated human-robot dialogue in the context of a robot helping its user repair a faulty pipe.

Language Synthesis

> The user leads the robot over to the machine that they will be working together to repair. Showing the robot the faulty pipe, the user asks the robot to find out and describe instructions on what supplies they will need in order to extract the faulty pipe and install a new pipe. After the user retrieves the necessary supplies, the robot narrates, using speech and nonverbal behaviors, the steps that its user must perform to complete the repair.

Robots will need to be capable of robust, accurate dialogue in order to effectively communicate ideas to their users. Synthesizing effective language involves multiple facets. At its lowest level, the robot will need to be capable of generating the appropriate content to communicate ideas accurately. On the other hand, it will need to follow the appropriate norms of social interaction in communicating with its user [2, 5]. Additionally, due to

the diversity of scenarios in which robots will interact with their users, they will need to be capable of dynamically adapting to changing settings and norms. In our previous work, we have provided an initial foundation for formalizing human dialogue patterns to enable robots to adapt to a variety of settings [10].

Multimodal Input

> As the user and the robot collaboratively perform the repair operations, the user picks up one of the replacement pipes and asks the robot for clarification on how the pipe should be attached to the current structure. The user poses the pipe where it would be inserted, motioning how the pipe would rotate, and asks the robot whether these actions would be the correct way of inserting the pipe.

Spoken language is often augmented with or replaced by numerous social cues, including gaze [4], gestures [3], and head movements [7]. While natural language processing is already a challenge for human-robot dialogue, an accurate understanding of user actions and language will require processing these additional nonverbal cues to better contextualize or interpret utterances. In particular, robots will need to integrate natural language processing with vision processing to detect and process the user's nonverbal cues and objects in the environment to which their user might be referring. This capability would enable robots to achieve multimodal dialogue in situated interaction settings. Prior work has taken initial steps to enable robots to understand user gestures, particularly gestures that point toward objects [1, 11].

Task Modeling

> After the robot and its user agree on how the new pipe should be inserted, the user begins to work on adding the component to the current pipe structure. The robot, after examining the structure, plans out the next several steps. Once the robot ensures that the user has correctly added the component, it informs the user of what step to take next in order to most efficiently complete the task.

Many of the roles that robots are envisioned to fulfill involve collaborating with people in physical tasks, such as working together on assembly lines or repairing machines. To effectively communicate with their users in these settings, robots will need to be capable of understanding the task at hand. Such an understanding would include comprehending task goals, understanding how the task space changes, and acknowledging how these changes relate to task goals (e.g., does this change bring us closer to finishing the task?). The ability to understand the current state of the task would allow the robot to discuss current and future actions, detect and correct mistakes during task execution, and adapt its language to reflect changes and deviations in how its user performs task steps. Previous research on human-robot collaboration has explored how robots might be enabled to recognize task progress and to adapt their utterances to reflect these changes [8, 9].

Workshop Interest

As demonstrated in this paper, human-robot interaction researchers face a number of challenges in enabling multimodal, situated human-robot dialogue. Without the capability to understand and display natural language and plan according to task models, robots will be limited in

their future roles. Our interest in this workshop stems from wishing to better understand current work in designing dialogue-based interfaces in human-computer interaction and to apply what we learn to the domain of human-robot interaction.

Conclusion

Robots are expected to serve in a wide range of roles, from teachers in labs and classrooms to working alongside humans in manufacturing settings. In these roles, robots will need to be capable of conversing with their users using spoken and nonverbal language and engage in situated interactions. In this paper, we have highlighted three challenges facing researchers and designers of interfaces for human-robot dialogue: *language synthesis*, *multimodal input*, and *task modeling*. Using our example of a human and robot working together to replace a faulty pipe in a machine, we demonstrate the need for solutions to each of these challenges, discuss prior work addressing these challenges, and highlight areas for future research. We believe that progress in these three areas will enable more natural, intuitive and effective interactions with robots.

Acknowledgments

The authors' research on human-robot dialogue is supported by the National Science Foundation award 1149970.

References

[1] Brooks, A., and Breazeal, C. Working with robots and objects: Revisiting deictic reference for achieving spatial common ground. In *Proc. of HRI* (2006), 297–304.

[2] Brown, P., and Levinson, S. C. *Politeness: Some universals in language usage*. Cambridge University Press, 1987.

[3] Clark, H. Coordinating with each other in a material world. *Discourse studies 7*, 4-5 (2005), 507–525.

[4] Gergle, D., and Clark, A. See what i'm saying?: using dyadic mobile eye tracking to study collaborative reference. In *Proc. of CSCW* (2011), 435–444.

[5] Goffman, E. The interaction order: American sociological association, 1982 presidential address. *American sociological review 48*, 1 (1983), 1–17.

[6] Gustafson, J. *Developing Multimodal Spoken Dialogue Systems: Empirical Studies of Spoken Human–Computer Interaction*. PhD thesis, KTH, 2002.

[7] Kraut, R., Fussell, S., and Siegel, J. Visual information as a conversational resource in collaborative physical tasks. *Journal of Human-Computer Interaction 18*, 1 (2003), 13–49.

[8] Mutlu, B., Terrell, A., and Huang, C.-M. Coordination mechanisms in human-robot collaboration. In *Proc. of Workshop on Collaborative Manipulation at HRI* (2013).

[9] Peltason, J., and Wrede, B. Pamini: A framework for assembling mixed-initiative human-robot interaction from generic interaction patterns. In *Proc. of SIGDIAL* (2010), 229–232.

[10] Sauppé, A., and Mutlu, B. Design patterns for exploring and prototyping human-robot interactions. In *Proc. of CHI* (2014).

[11] Sugiyama, O., Kanda, T., Imai, M., Ishiguro, H., and Hagita, N. Natural deictic communication with humanoid robots. In *Proc. of IROS* (2007), 1441–1448.