MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Engagement During Dialogues with Robots

Sidner, C.L.; Lee, C.

TR2005-016 March 2005

Abstract

This paper reports on our research on developing the ability for robots to engage with humans in a collaborative conversation. Engagement is the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake. Many of these interactions are dialogues and we focus on dialogues in which the robot is a host to the human in a physical environment. The paper reports on human-human engagement and its appliction to a robot that collaborates with a human on a demonstration of equipment.

AAAI Spring Symposium Dialogical Robots

Engagement During Dialogues with Robots

Candace L. Sidner, Christopher Lee
Mitsubishi Electric Research Laboratories
201 Broadway
Cambridge, MA 02139
{Sidner, Lee}@merl.com

Cory Kidd
MIT Media Lab
Cambridge, MA 02139
Coryk@media.mit.edu

Abstract

This paper reports on our research on developing the ability for robots to engage with humans in a collaborative conversation. Engagement is the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake. Many of these interactions are dialogues and we focus on dialogues in which the robot is a host to the human in a physical environment. The paper reports on human-human engagement and its application to a robot that collaborates with a human on a demonstration of equipment.

Introduction

One goal for interaction between people and robots centers on conversation about tasks that a person and a robot can undertake together. Not only does this goal require linguistic knowledge about the operation of conversation, and real world knowledge of how to perform tasks jointly, but the robot must also interpret and produce behaviors that convey the intention to start the interaction, maintain it or to bring it to a close. We call such behaviors engagement behaviors. Our research concerns the process by which a robot can undertake such behaviors and respond to those performed by people.

Engagement is the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake. Engagement is supported by the use of conversation (that is, spoken linguistic behavior), ability to collaborate on a task (that is, collaborative behavior), and gestural behavior that conveys connection between the participants. While it might seem that conversational utterances alone are enough to convey connectedness (as is the case on the telephone), gestural behavior in face-to-face conversation provides significant evidence of connection between the participants.

Conversational gestures generally concern gaze at/away from the conversational partner, pointing behaviors, (bodily) addressing the conversational participant and other persons/objects in the environment, and various hand signs, all with appropriate synchronization with the conversational, collaborative behavior. These gestures are culturally determined, but every culture has some set of behaviors to accomplish the engagement task. These gestures sometimes also have the dual role of providing sensory input (to the eyes and ears) as well as telling conversational participants about their interaction. Our research focuses on how gestures tell participants about their interaction, but we also must address the matter of sensory input as well.

Conversation, collaboration on activities, and gestures together provide interaction participants with ongoing updates of their attention and interest in a face-to-face interaction. Attention and interest tell each participant that the other is not only following what is happening (i.e. grounding), but intends to continue the interaction at the present time.

Not only must a robot produce engagement behaviors in collaborating with a human conversational partner (hereafter CP), but also it must interpret similar behaviors from its CP. Proper gestures by the robot and correct interpretation of human gestures dramatically affect the success of interaction. Inappropriate behaviors can cause humans and robots to misinterpret each other's intentions. For example, a robot might look away for an extended period of time from the human, a signal to the human that it wishes to disengage from the conversation and could thereby terminate the collaboration unnecessarily. Incorrect recognition of the human's behaviors can lead the robot to press on with an interaction in which the human no longer wants to participate.

Learning from Human Behavior

To determine gestures, we have developed a set of rules for engagement in the interaction. These rules are gathered from the linguistic and psycholinguistic literature (for example, (Kendon 1967)) as well as from 3.5 hours of videotape of a human host guiding a human visitor on tour of laboratory artifacts. These gestures reflect US standard cultural rules for US speakers. For other cultures, a different set of rules must be investigated.

Our initial set of gestures were quite simple, and applied to a hosting activities, that is, the collaborative activity in which an agent provides guidance in the form of information, entertainment, education or other services in the user's environment and may also request that the user undertake actions to support the fulfillment of those services. Initially, human-robot conversations consisted of the robot and visitor

greeting each other and discussing a project in the laboratory. However, in hosting conversations, robots and people must discuss and interact with objects as well as each other.

As we have learned from careful study of the videotapes we have collected (see (Sidner, Lee, & Lesh 2003)), people do not always track the speaking CP, not only because they have conflicting goals (e.g. they must attend to objects they manipulate), but also because they can use the voice channel to indicate that they are following information even when they do not track the CP. They also simply fail to track the speaking CP sometimes without the CP attempting to direct them back to tracking. Our results differ from those of Nakano et al (Nakano *et al.* 2003), perhaps because of the detailed instruction giving between the participants in Nakano's experiments.

Experience from this data has resulted in the principle of conversational tracking: participants in a collaborative conversation track the other's face during the conversation in balance with the requirement to look away to: (1) participate in actions relevant to the collaboration, or (2) multitask activities unrelated to the collaboration at hand, such as scanning the surrounding scene for interest, avoidance of damaging encounters, or personal activities.

To explore interactions with such gestures, our robot acts as a host to a human visitor participating in a demo in a laboratory. The use of the COLLAGEN$^{(TM)}$ system (Rich, Sidner, & Lesh 2001) to model conversation and collaboration permits the interaction to be more general and easily changed than techniques such as (Fong, Thorpe, & Baur 2001). One such conversation taken from a conversation log is shown in Appendix 1; it shows only a few of the human's gestures and none of the robot's. There are many alternatives paths in the conversation that cannot be provided in a short space. The conversation concerns an invention, called IGlassware (a kind of electronic cup sitting on a table) (Dietz, Leigh, & Yerazunis 2002), that the robot and visitor demonstrate together. As the reader will notice, the robot's conversation are robot controlled, in large part because when a more mixed initiative style is used, participants tend to produce many types of utterances, and speech recognition becomes to unreliable for successful conversation.

The robot is a penguin (see Figure 1) with a humanoid face (eyes facing forward and a beak that opens and closes), which we hypothesize is essential to allow human participants to assume familiarity with what the robot will at least say. We have not attempted yet to test this hypothesis as doing so would require experimenting with other non-humanoid models, which we are not equipped to do. The robot is a 7 DOF stationary robot. Details of the robot's sensory devices and the architecture it uses can be found in (Sidner *et al.* 2004a).

The penguin robot has been provided with gestural rules so that it can undertake the hosting conversations discussed previously. The robot has gestures for greeting a visitor, looking at the visitor and others during the demo, looking at the IGlass cup and table when pointing to it or discussing it, for ending the interaction, and for tracking the visitor when the visitor is speaking. The robot also interrupts its intended conversation about the demo, when the visitor does not take a turn at the expected point in the interaction. Failing to take a turn is an indication of the desire to disengage, and the robot queries the visitor about his/her desire to continue. Continuing lack of response or an answer indicating desire to end the demo will lead to a closing sequence on the robot's part.

Figure 1: Mel, the penguin robot

Evaluating Human-Robot Interactions

Evaluating a robot's interactions is a non-trivial undertaking. In separate work (Sidner *et al.* 2004b), we have begun to explore both the success of the robot's behavior as well as the matter of what measures to use in order to accomplish such evaluations. We have evaluated 37 subjects in two conditions of interaction, one in which the robot has all the gestures we have been able to program (moving), and a second (talking) condition where the only movement is that of the robot's beak (after the robot locates the participant and locks onto the location of the participant's face, which it holds for the remainder of the interaction).

One of our challenges in that work was to decide how to measure the impact of the robot's behavior on the interaction. We used a questionnaire given to participants after the demo with the robot to gather information about their liking of the robot, involvement in the demo, appropriateness of movements and predictability of robot behavior. However, we also studied the participant's behaviors from video data collected during the experiment. To further measure participant's engagement, we used interaction time, amount of mutual gaze, talk directed to the robot, overall looking back to the robot, and for two pointing behaviors, how closely in time the participant tracked the robot's pointing.

Does this robot's engagement gestural behavior have an impact on the human partner? The answer is a qualified yes. While details can be found in (Sidner *et al.* 2004b), in summary, a majority of participants in both conditions were found to turn their gaze to the robot whenever they took a turn in the conversation, an indication that the robot was real enough to be worthy of conversation. Furthermore, partici-

Figure 2: Mel demonstrates IGlassware to a visitor.

pants in the moving condition looked back at the robot significantly more whenever they were attending to the demonstration in front of them. The participants with the moving robot also responded to the robot's change of gaze to the table somewhat more than the other subjects.

Another gesture that is common in conversation is nodding, which serves at least the purpose of backchanneling and grounding (Clark 1996). In collaboration with researchers at MIT, we are using the Watson system to interpret hod nods from human participants (Lee *et al.* 2004).

Most of our experiments with human participants (41 so far) have largely only provided us with further training data for the HMMs. As we have discovered, human head nodding is distinctive in conversation for being a very small motion (as little as 3 degrees), and one that is also very idiosyncratic for different people. Our plan is to improve the recognition to the point that people's nodding will be recognized. In our first study (discussed above), we discovered that people naturally nod at the robot: 55% of the participants in the moving condition did so, while 45% in the talker condition, even though the participants had no reason to do believe the robot recognized this behavior. Our subsequent studies (where participants were told that the robot could recognize nods) show an even higher incidence of head nods as backchannels and accompanying "yes" answers to questions. We are currently using that data to explore new means of interpreting head nods in conversational contexts (Morency, Sidner, & Darrell 2005).

Related Research

While other researchers in robotics have explored aspects of gesture (for example Breazeal (Breazeal 2001) and Kanda et al (Kanda *et al.* 2002)), none of them have attempted to model human-robot interaction to the degree that involves the numerous aspects of engagement and collaborative conversation that we have set out above. Recent work by Breazeal et al (Breazeal, Hoffman, & Lockerd 2004) is exploring teaching a robot a physical task that can be performed collaboratively once learned. A robot developed at Carnegie Mellon University serves as a museum guide (Burgard *et al.* 1998) and navigates well while avoiding humans, but interacts with users via a 2D talking head with minimal engagement and conversational abilities. Most similar in spirit to work reported here is the Armar II robot (Dillman, Becher, & Steinhaus 2004). Armar II is speech enabled, has some dialogue capabilities, and has abilities to track gestures and people. However, the Armar II work is focused on teaching the robot new tasks (with programming by demonstration techniques), while our work has been focused on improving the interaction capabilities needed to hold conversations and undertake tasks. Work on 2D conversational agents originally inspired our work (Cassell *et al.* 2000; Johnson, Rickel, & Lester 2000). Recent work by Traum and Rickel (Traum & Rickel 2002) have explored issues in giving attention to agents in an interaction as part of an agent architecture in 2D virtual worlds.

Future Directions

Work has just been completed in mobilizing our previously stationary robot. In addition to continuing our observations of human-human data and better recognition of nodding, we intend to use mobilization for two problems: (1) the process of undertaking the engagement of another, in particular, to have to robot move toward a group of people or an individual, "catch their eye" and begin a conversation, (2) the use of body stance by the robot in interactions. Body stance is a means to indicate the primary focus of engagement. When body stance is required to be used to face an object other than the speaker (as for example, when washing the dishes), the individual must still turn either its body or its head back to the interlocutor at points in the interaction. Our intention is to understand how people do this and apply it to the robot's behavior.

Summary

Human-robot interaction offers new challenges in the use of interaction behaviors. This paper has discussed a type of human robot interaction aimed at hosting activities, and the need for engagement models to govern the connective relations between human and robot. The paper also discusses our investigations in developing and evaluating engagement rules and evaluations of human participants in interactions with a robot on a hosting problem.

Appendix 1

Example conversation between Mel and a human user.

MEL: Hello, I'm Mel, a robotic penguin.

USER: *Hi Mel.*

MEL: What's your name?

USER: *Kathy.*

MEL: Kathy, I'd like to show you a demo. OK?

USER: *Ok.*

MEL: It would be really nice to know what kind of progress people are making in their dining. So Paul Dietz created a new product called I Glassware for this.

MEL: By the way, Kathy, do you know Paul?

USER: *Yes.*

MEL: Then you probably know how creative he is.

MEL: I Glassware stands for instrumented glassware!

MEL: Kathy, right there is the I Glassware cup. *<User looks at cup.>*

MEL: And near it, is the table readout. *<User looks at readout.>*

MEL: The word re-fill on the readout means that it is empty.

USER: *Ok.*

MEL: All right, now we'll see how to use the cup.

USER: *Alright.*

MEL: First you should pour enough water from the pitcher into the glass to fill it up. Then make sure the glass is on the IGlassware table. *<User pours water into the glass from the pitcher. Then put the glass on the table.>*

MEL: Good.

MEL: Notice, the bar on the readout shows that the glass is full.

USER: *Ok.*

MEL: Okay, pour the water back into the pitcher.

USER: *Ok.*

MEL: I'm waiting for a reading from the table for the glass. *<User pours the water from the glass back into the pitcher>*

MEL: Good.

MEL: See, it registers needing a re-fill!

USER: *Great.*

MEL: Would you like me to explain how this works?

USER: *Yes.*

MEL: The copper in the glass transmits to the readout display by inductance with the surface of the table. The readout then displays the information coming to the table.

USER: *Ok.*

MEL: Do you know how this technology might be used in restaurants?

USER: *Yes.*

MEL: Okay.

MEL: Well, Kathy, that's about all. Go see Paul Dietz, for more about I Glassware. So long!

USER: *Good bye. <Agent performs ShutDown.>*

Acknowledgements

The authors wish to acknowledge the work of Charles Rich and Neal Lesh on aspects of Collagen and Mel, and Max Makeev for mobilizing Mel.

References

Breazeal, C.; Hoffman, G.; and Lockerd, A. 2004. Teaching and working with robots as a collaboration. In *The Third International Conference on Autonomous Agents and Multi-Agent Systems AAMAS 2004*, 1028–1035. ACM Press.

Breazeal, C. 2001. Affective interaction between humans and robots. In *Proceedings of the 2001 European Conference on Artificial Life (ECAL2001)*.

Burgard, W.; Cremes, A. B.; Fox, D.; Haehnel, D.; Lakemeyer, G.; Schulz, D.; Steiner, W.; and Thrun, S. 1998. The interactive museum tour guide robot. In *In Proceedings of AAAI-98*, 11–18. AAAI Press, Menlo Park, CA.

Cassell, J.; Sullivan, J.; Prevost, S.; and Churchill, E., eds. 2000. *Embodied Conversational Agents*. MIT Press, Cambridge, MA.

Clark, H. H. 1996. *Using Language*. Cambridge: Cambridge University Press.

Dietz, P. H.; Leigh, D. L.; and Yerazunis, W. S. 2002. Wireless liquid level sensing for restaurant applications. *IEEE Sensors* 1:715–720.

Dillman, R.; Becher, R.; and Steinhaus, P. 2004. ARMAR II – a learning and cooperative multimodal humanoid robot system. *International Journal of Humanoid Robotics* 1(1):143–155.

Fong, T.; Thorpe, C.; and Baur, C. 2001. Collaboration, dialogue and human-robot interaction. In *10th International Symposium of Robotics Research*.

Johnson, W. L.; Rickel, J. W.; and Lester, J. C. 2000. Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education* 11:47–78.

Kanda, T.; Ishiguro, H.; M., I.; Ono, T.; and Mase, K. 2002. A constructive approach for developing interactive humanoid robots. In *Proceedings of IROS 2002*. IEEE Press, New York.

Kendon, A. 1967. Some functions of gaze direction in two person interaction. *Acta Psychologica* 26:22–63.

Lee, C.; Lesh, N.; Sidner, C.; Morency, L.-P.; Kapoor, A.; and Darrell, T. 2004. Nodding in conversations with a robot. In *Proceedings of the ACM International Conference on Human Factors in Computing Systems*.

Morency, L.; Sidner, C.; and Darrell, T. 2005. Towards context based vision feedback recognition for embodied agents. In *AISB Symposium in Conversational Informatics for Supporting Social Intelligence and Interaction*.

Nakano, Y.; Reinstein, G.; Stocky, T.; and Cassell, J. 2003. Towards a model of face-to-face grounding. In *Proceedings of the 41st meeting of the Association for Computational Linguistics*, 553–561.

Rich, C.; Sidner, C. L.; and Lesh, N. 2001. COLLAGEN: Applying collaborative discourse theory to human-computer interaction. *AI Magazine* 22(4):15–25. Special Issue on Intelligent User Interfaces.

Sidner, C.; C.Lee; C.Kidd; and Lesh, N. 2004a. Explorations in engagement for humans and robots. In *IEEE-RAS/RSJ Inter-national Conference on Humanoid Robots (Humanoids 2004)*. IEEE Press.

Sidner, C. L.; Kidd, C. D.; Lee, C. H.; and Lesh, N. 2004b. Where to look: A study of human-robot engagement. In *ACM International Conference on Intelligent User Interfaces (IUI)*, 78–84. ACM.

Sidner, C. L.; Lee, C. H.; and Lesh, N. 2003. Engagement when looking: behaviors for robots when collaborating with people. In Kruiff-Korbayova, I., and C.Kosny., eds., *Diabruck: Proceedings of the 7th workshop on the Semantic and Pragmatics of Dialogue*, 123–130. University of Saarland.

Traum, D., and Rickel, J. 2002. Embodied agents for multi-party dialogue in immersive virtual worlds. In *Proceedings of the International Joint Conference on Autonomous Agents and Multi-agent Systems (AAMAS 2002)*, 766–773.