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

The Role of Dialog in Human Robot Interaction

Candace L. Sidner, Christopher Lee and Neal Lesh

TR2003-63 June 2003

Abstract

This paper reports on our research on developing the abil-ity 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 per-ceived connection during interactions they jointly undertake. The paper reports on the architecture for human-robot collaborative conversation with engagement, and the significance of the dialogue model in that architecture for decisions about engagement during the interaction.

First International Workshop on Language Understanding and Agents for Real World Interaction, Hokkaido University, July 2003.

The Role of Dialogue in Human Robot Interaction

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

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. The paper reports on the architecture for human-robot collaborative conversation with engagement, and the significance of the dialogue model in that architecture for decisions about engagement during the interaction.

1. 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 maintain the interaction 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 in 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. We focus on the latter in this research.

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 but intends to continue the interaction to its logical conclusion.

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.

While other researchers in robotics are exploring aspects of gesture (for example, [1], [2]), none of them have attempted to model human-robot interaction to the degree that involves the numerous aspects of engagement and collaborative conversation that are considered here. Robotics researchers interested in collaboration and dialogue [3] have not based their work on extensive theoretical research on collaboration and conversation, as we will detail later. Our work is also not focused on emotive interactions, in contrast to [1] among others. For 2D conversational agents, researchers (notably, [4],[5]) have explored agents that produce gestures in conversation. However, they have not tried to incorporate recognition as well as production of these gestures, nor have they focused on the

full range of these behaviors to accomplish the maintenance of engagement in conversation.

2. Architecture for human robot interaction

Our research program for investigating engagement in interaction has three main tasks: to investigate how humans convey engagement in their natural everyday collaborative activities, to explore architectures and algorithms for robots that will allow them to approximate human engagement abilities in interactions with humans, and to evaluate the resulting robots in experimental interactions with people. In this paper we focus on progress in architectures and algorithms and the role of conversation in engagement, but will sketch briefly our investigations in human-human data and our evaluation efforts.

Figure 1 illustrates the architecture we are currently using for human-robot interactions. The modules of the architecture separate linguistic decisions from sensor and motor decisions. However, information from sensor fusion can cause new tasks to be undertaken by the conversational model. These tasks concern changes in engagement that are signaled by behaviors detected by sensor fusion.

- Input to Sensor fusion comes from two (OrangeMicro iBot) cameras and a pair of microphones.
- Speech and collaborative conversation (Conversation model) rely on the Collagen™ middleware for collaborative agents [6,7] and commercially available speech recognition software (IBM ViaVoice).
- Agent decision-making software in the Conversation model that determines the overall set of gestures to be generated by the robot motors.

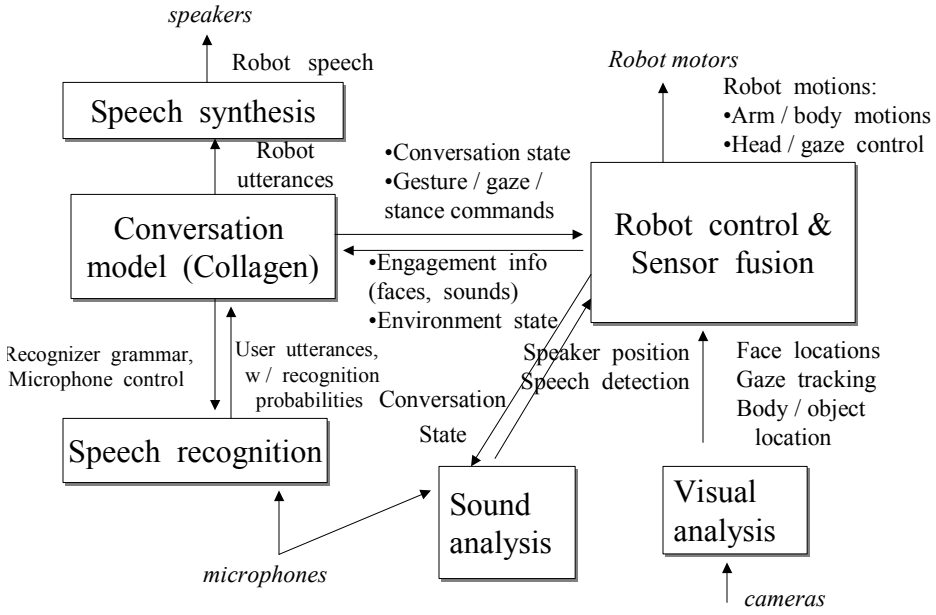

Figure 1: Architecture for robot with engagement

Sensor fusion uses the face location algorithm of [8] to find faces, notice when a face disappears, and notice the change of a face from full face to profile. It uses an object-tracking algorithm [9] to locate an object to point to and track as the object moves in the visual field. A sound location algorithm detects the source of spoken utterances, and its results together with face location permit Sensor fusion to pick out a CP from the group of people in front of the camera. Face location also provides information on direction of gaze. The result of its processing is passed to the Conversation model. The Robot control synchronizes the set of gestures from the Conversation model and controls the robot motors.

Figure 2 illustrates our robot, which takes the form of a penguin referred to as Mel.

3. The role of the conversation model

The Collagen system for collaborative dialogue is instantiated so that our robot acts as a host to a human visitor participating in a demo in a laboratory. Collagen permits the interaction to be more general and easily changed than techniques such as [3]. One such conversation taken from a conversation log is shown in Appendix 1. The conversation concerns an invention, called IGlassware (a kind of electronic cup sitting on a table), that the robot and visitor demonstrate together. Gestures that the penguin produces, which include looking at the user and sometimes at

onlookers in the room, all coordinated with turn taking, looking at the demo equipment, pointing at the equipment when it is being mentioned, and beat gestures [10] are not illustrated.

Figure 2: Mel, the penguin robot

The uses for our robot are aimed at a collaboration with a human on tasks with objects in the physical world. The Collagen model is based on extensive theory of collaboration [11] and conversation [12,13] and involves direct human-robot interaction rather than tele-operation. Our work is complementary to efforts such as [14], which was focused on sharpening the navigational skills of robots with limited human-robot interaction. Our current work extends our first effort [15] to make a robot that could simply talk about a collaborative task and point to objects on a horizontally positioned computer interface.

To accomplish natural conversation with interwoven gestures, the Collagen system has been given a set of action descriptions (called the recipe library in the Collagen system) that describe how to greet a visitor, how to perform a demo with them, and how to close an interaction. The descriptions are not scripts, but rather task models, with annotations for how to convey certain utterances. For example, the high-level task model for giving the demo consists of actions to motivate the visitor to participate in the demo, discuss the inventor of the demo object, point out each of the demo objects, and perform the actions required to use the object. The model also includes behavior (such as looking at the cup, pouring water into the cup, etc.) that the robot expects from the visitor. Recipe libraries like the one for the IGlassware demo are the means by which a developer can tailor the Collagen system to particular collaborations.

The visitor is expected to respond in English. Standard grammar techniques using JSAPI for the IBM Via Voice speech recognizer, and semantic interpretation rules provide utterance understanding. The resulting conversation is approximately 5 minutes long and has several different sub-segments depending on the visitor's actions and verbal responses to robot utterances.

To coordinate gestures, the Conversation model makes use of the agenda of next moves provided by the Collagen system. This agenda is expanded by the Collagen agent (another Collagen component), which serves to make decisions given the agenda. It uses engagement rules (discussed in the next section) to determine gestures for the robot, and to assess engagement information about the human CP from the Robot Control and Sensor Fusion module. Decisions by the agent are passed to the Robot Control module for generation of behaviors by robot motors.

The state of the conversation, which is part of the Conversation Model as implemented by Collagen, plays a significant role in determining gestures for the robot. Information in the model concerning turns, the purpose of each segment of the conversation, and information about individual utterances are needed for gesturing. Some robotic gestures must be synchronized with spoken language. For example, beak movement (the mouth of the penguin robot) must be timed closely to the start and end of speech synthesis of utterances. The robot must also produce beat gestures (with its wings) at the phrases in an utterance that represent new information. To capture this need for synchrony, the robot responds to events generated when the speech synthesis engine reaches certain embedded metatext markers in the speech text (a method inspired by [10]).

Turns in the conversation, in particular, who holds the turn and when it changes, affect gesture choices. For example, the robot must look at the CP when it passes off the turn, but during its turn, it can look freely at the CP or onlookers. However, during portions of the conversation where the robot's purpose is to discuss the cup or actions in using the cup, the robot must gaze at the cup; it may not look freely and when finished, it must return its gaze to the CP (rather than onlookers). Likewise, the conversation model provides details for when a visitor is expected to gesture in a certain fashion. Sensor fusion information contradicting such expectations will cause the conversation model to change its next choices in the conversation. Furthermore, fusion of visual face location and speech localization information (for determining the location of the human CP) must only be performed when the conversational model indicates the human has the turn. The conversation state information is therefore crucial for the gestures that are undertaken in the Robot Control module.

4. Engagement Rules and Evaluation

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, [16]) 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 conversation where the robot and visitor greeted each other and discussed a project in the laboratory. However, in hosting conversations, robots and people must discuss and interact with objects as well as each other. The principle behind the current set of gestures is to have the robot track the speaking human CP. As we have learned from careful study of the videotapes we have collected (see [17]), 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. Furthermore, when the robot is the speaking CP, it does not need to track the visitor. Rather it must balance between gazing at the human visitor and attending to the objects of the demo.

To explore interactions with such gestures, we have provided our penguin robot 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, but 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.

Evaluating a robot's interactions is a non-trivial undertaking. By observation of the robot, we have learned that some of the robot's behaviors in this interaction are unacceptable. For example, the robot often looks away for too long (at the cup and table) when explaining them, it fails to make sure it is looking at the visitor when it calls the visitor by name, and it sometimes fails to look for a long enough when it turns to look at objects.

More challenging for our work is measuring the improvement in the interaction that occurs when the robot's interactions reflect proper engagement behavior. We are at present investigating differences in visitor performance of tasks when the robot performs as described here as well as when it uses less natural gestures in the same interaction. We expect these results to inform us about the significance of gestures in human robot interaction. At the same time, we would like to improve the current rule set. The types of observations we noted previously provide some measure. However, by changing the robot's gestures automatically based on the robot's conclusions about the visitor's activities we can obtain new rules. To evaluate these rules we are exploring several different techniques, including using two simulated robots to talk to each other.

5. 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 illustrated the need for engagement models to govern the connective relations between human and robot. The paper has presented the details of an architecture for a sample hosting activity and argued that the conversation model plays a crucial role in the gestures needed to maintain engagement in interaction. The paper also discusses our investigations in developing and evaluating engagement rules.

6. Appendix 1: Sample conversation for the IGlassware demo

Agent says "Hello, I'm Mel, a robotic penguin."
User says "Hi Mel."
Agent says "What's your name?"
User says "Kathy."
Agent says "Kathy, I'd like to show you a demo. OK?"
User says "Ok."
Agent says "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."
Agent says "By the way, Kathy, do you know Paul?"
User says "Yes."
Agent says "Then you probably know how creative he is."
Agent says "I Glassware stands for instrumented glassware!"
Agent says "Kathy, right there is the I Glassware cup."
User looks at cup.
Agent says "And near it, is the table readout."
User looks at readout.
Agent says "The word re-fill on the readout means that it is empty."
User says "Ok."
Agent says "All right, now we'll see how to use the cup."
User says "Alright."
Agent says "First, you should pick up the glass from the table."
User says "Ok."
Agent says "Please pick up the glass from the table."
User picks up the glass from the table.
Agent says "Good."
Agent says "It's just a cup. It's got a pie code on it so I can track it."
Agent says "Now pour water from the pitcher into it. Then put it on the table."
User pours water into the glass from the pitcher. Then put the glass on the table.
Agent says "Good."

Agent says "Notice, the bar on the readout shows that the glass is full."
User says "Ok."
Agent says "Okay, pour the water back in the pitcher."
User says "Ok."
Agent says "I'm waiting for a reading from the table for the glass."
User pours the water from the glass back into the pitcher.
Agent says "Good."
Agent says "See, it registers needing a re-fill!"
User says "Great."
Agent says "Would you like me to explain how this works?"
User says "Yes."
Agent says "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 says "Ok."
Agent says "Do you know how this technology might be used in restaurants?"
User says "Yes."
Agent says "Okay."
Agent says "Well, Kathy, that's about all. Go see Paul Dietz, for more about I Glassware. So long!"
User says "Good bye."
Agent performs ShutDown.

7. Acknowledgements

The authors wish to acknowledge the work of Charles Rich on aspects of Collagen critical to this effort.

8. References

1. C. Breazeal, "Affective interaction between humans and robots", *Proceedings of the 2001 European Conference on Artificial Life (ECAL2001)*. Prague, Czech Republic, (2001).
2. T. Kanda, H. Ishiguro, M. Imai. T. Ono, and K. Mase, "A constructive approach for developing interactive humanoid robots. *Proceedings of IROS 2002*, IEEE Press, NY, 2002.
3. T. Fong, C. Thorpe, C. Baur. Collaboration, Dialogue and Human-Robot Interaction, *10th International Symposium of Robotics Research*, Lorne, Victoria, Australia, November, 2001.
4. J. Cassell, J. Sullivan, S. Prevost and E. Churchill, *Embodied Conversational Agents*. MIT Press, Cambridge, MA, 2000.
5. W.L. Johnson, J. W. Rickel, J. W. and J.C. Lester. "Animated Pedagogical Agents: Face-to-Face Interaction in Interactive Learning Environments," *International Journal of Artificial Intelligence in Education*, 11: 47-78, 2000.
6. C. Rich, C.L. Sidner, and N. Lesh. "COLLAGEN: Applying Collaborative Discourse Theory to Human-Computer Interaction," *AI Magazine, Special Issue on Intelligent User Interfaces*, AAAI Press, Menlo Park, CA, Vol. 22: 4: 15-25, 2001.
7. C. Rich and C.L. Sidner. "COLLAGEN: A Collaboration Manager for Software Interface Agents," *User Modeling and User-Adapted Interaction*, Vol. 8, No. 3/4, 1998, pp. 315-350, 1998.
8. Viola, P. and Jones, M. "Rapid Object Detection Using a Boosted Cascade of Simple Features," IEEE *Conference on Computer Vision and Pattern Recognition*, Hawaii, pp. 905-910, 2001.
9. Beardsley, P.A. "Piecode Detection," *Mitsubishi Electric Research Labs TR2003-11*, Cambridge, MA, February, 2003.
10. J. Cassell, H. Vilhjálmsson and T.W. Bickmore. "BEAT: the behavior expression animation toolkit," *Proceedings of SIGGRAPH 2001*. New York: ACM Press. pp. 477-486, 2001.
11. B.J. Grosz and S. Kraus. "Collaborative Plans for Complex Group Action," *Artificial Intelligence*, 86(2): 269-357, 1996.
12. B. J. Grosz and C.L. Sidner. "Attention, intentions, and the structure of discourse," *Computational Linguistics*, 12(3): 175--204, 1986.
13. K.E. Lochbaum. "A Collaborative Planning Model of Intentional Structure," *Computational Linguistics*, 24(4): 525-572, 1998
14. W. Burgard, A.B. Cremes, D. Fox, D.Haehnel, G. Lakemeyer, D. Schulz, W. Steiner, and S. Thrun. "The Interactive Museum Tour Guide Robot," *Proceedings of American Association of Artificial Intelligence Conference 1998*, 11-18, AAAI Press, Menlo Park, CA, 1998.
15. C. Sidner and M. Dzikovska. "Hosting activities: Experience with and future directions for a robot agent host," *Proceedings of the 2002 Conference on Intelligent User Interfaces*, New York: ACM Press. pp. 143-150, 2002.
16. A. Kendon. "Some functions of gaze direction in social interaction," *Acta Psychologica*, 26: 22-63, 1967.
17. C. Sidner and C. Lee. "Engagement Rules for Human-Robot Collaborative Interactions," *Proceedings of 2003 Conference on Systems, Man and Cybernetics*, 2003, forthcoming.