Predicting Evidence of Understanding by Monitoring User's Task Manipulation in Multimodal Conversations

Yukiko I. Nakano[†]
Yoshiko Arimoto[††]
[†]Tokyo University of Agriculture and Technology
2-24-16 Nakacho, Koganei-shi, Tokyo 184-8588, Japan
{nakano, kmurata, menomoto}@cc.tuat.ac.jp

Kazuyoshi Murata[†]
Yasuhiro Asa[†††]
[††]Tokyo University of Technology
1404-1 Katakura, Hachioji, Tokyo 192-0981, Japan
ar@mf.teu.ac.jp

Mika Enomoto[†]
Hirohiko Sagawa[†††]
[†††]Central Research Laboratory, Hitachi, Ltd.
1-280, Higashi-koigakubo Kokubunji-shi, Tokyo 185-8601, Japan
{yasuhiro.asa.mk, hirohiko.sagawa.cu}@hitachi.com

Abstract

The aim of this paper is to develop animated agents that can control multimodal instruction dialogues by monitoring user's behaviors. First, this paper reports on our Wizard-of-Oz experiments, and then, using the collected corpus, proposes a probabilistic model of fine-grained timing dependencies among multimodal communication behaviors: speech, gestures, and mouse manipulations. A preliminary evaluation revealed that our model can predict a instructor's grounding judgment and a listener's successful mouse manipulation quite accurately, suggesting that the model is useful in estimating the user's understanding, and can be applied to determining the agent's next action.

1 Introduction

In face-to-face conversation, speakers adjust their utterances in progress according to the listener's feedback expressed in multimodal manners, such as speech, facial expression, and eye-gaze. In task-manipulation situations where the listener manipulates objects by following the speaker's instructions, correct task manipulation by the listener serves as more direct evidence of understanding (Brennan 2000, Clark and Krych 2004), and affects the speaker's dialogue control strategies.

Figure 1 shows an example of a software instruction dialogue in a video-mediated situation (originally in Japanese). While the learner says nothing, the instructor gives the instruction in small pieces, simultaneously modifying her gestures and utterances according to the learner's mouse movements.

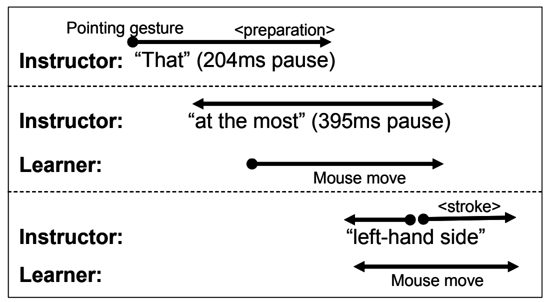


Figure 1: Example of task manipulation dialogue

To accomplish such interaction between human users and animated help agents, and to assist the user through natural conversational interaction, this paper proposes a probabilistic model that computes timing dependencies among different types of behaviors in different modalities: speech, gestures, and mouse events. The model predicts (a) whether the instructor's current utterance will be successfully understood by the learner and grounded (Clark and Schaefer 1989), and (b) whether the learner will successfully manipulate the object in the near future. These predictions can be used as constraints in determining agent actions. For example, if the current utterance will not be grounded, then the help agent must add more information.

In the following sections, first, we collect human-agent conversations by employing a Wizard-of-Oz method, and annotate verbal and nonverbal behaviors. The annotated corpus is used to build a Bayesian network model for the multimodal instruction dialogues. Finally, we will evaluate how

Proceedings of the ACL 2007 Demo and Poster Sessions, pages 121–124,
Prague, June 2007.

accurately the model can predict the events in (a) and (b) mentioned above.

2 Related work

In their psychological study, Clark and Krych (2004) showed that speakers alter their utterances midcourse while monitoring not only the listener's vocal signals, but also the listener's gestural signals as well as through other mutually visible events. Such a bilateral process functions as a joint activity to ground the presented information, and task manipulation as a mutually visible event contributes to the grounding process (Brennan 2000, Whittaker 2003). Dillenbourg, Traum, et al. (1996) also discussed cross-modality in grounding: verbally presented information is grounded by an action in the task environment.

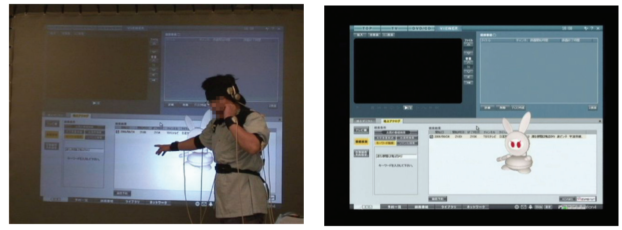
Studies on interface agents have presented computational models of multimodal interaction (Cassell, Bickmore, et al. 2000). Paek and Horvitz (1999) focused on uncertainty in speech-based interaction, and employed a Bayesian network to understand the user's speech input. For user monitoring, Nakano, Reinstein, et al. (2003) used a head tracker to build a conversational agent which can monitor the user's eye-gaze and head nods as nonverbal signals in grounding.

These previous studies provide psychological evidence about the speaker's monitoring behaviors as well as conversation modeling techniques in computational linguistics. However, little has been studied about how systems (agents) should monitor the user's task manipulation, which gives direct evidence of understanding to estimate the user's understanding, and exploits the predicted evidence as constraints in selecting the agent's next action. Based on these previous attempts, this study proposes a multimodal interaction model by focusing on task manipulation, and predicts conversation states using probabilistic reasoning.

3 Data collection

A data collection experiment was conducted using a Wizard-of-Oz agent assisting a user in learning a PCTV application, a system for watching and recording TV programs on a PC.

The output of the PC operated by the user was displayed on a 23-inch monitor in front of the user, and also projected on a 120-inch big screen, in front of which a human instructor was standing (Figure 2 (a)). Therefore, the participants shared visual events output from the PC (Figure 2 (b)) while sitting in different rooms. In addition, a rabbit-like animated agent was controlled through the instructor's motion data captured by motion sensors. The instructor's voice was changed through a voice transformation system to make it sound like a rabbit agent.



(a) Instructor (b) PC output

Figure 2: Wizard-of-Oz agent controlled by instructor

4 Corpus

We collected 20 conversations from 10 pairs, and annotated 11 conversations of 6 pairs using the Anvil video annotating tool (Kipp 2004).

Agent's verbal behaviors: The agent's (actually, instructor's) speech data was split by pauses longer than 200ms. For each inter pausal unit (IPU), utterance content type defined as follows was assigned.

- Identification (id): identification of a target object for the next operation
- Operation (op): request to execute a mouse click or a similar primitive action on the target
- Identification + operation (idop): referring to identification and operation in one IPU

In addition to these main categories, we also used: State (referring to a state before/after an operation), Function (explaining a function of the system), Goal (referring to a task goal to be accomplished), and Acknowledgment. The intercoder agreement for this coding scheme is very high K=0.89 (Cohen's Kappa), suggesting that the assigned tags are reliable.

Agent's nonverbal behaviors: As the most salient instructor's nonverbal behaviors in the collected data, we annotated agent pointing gestures:

- Agent movement: agent's position movement
- Agent touching target (att): agent's touching the target object as a stroke of a pointing gesture

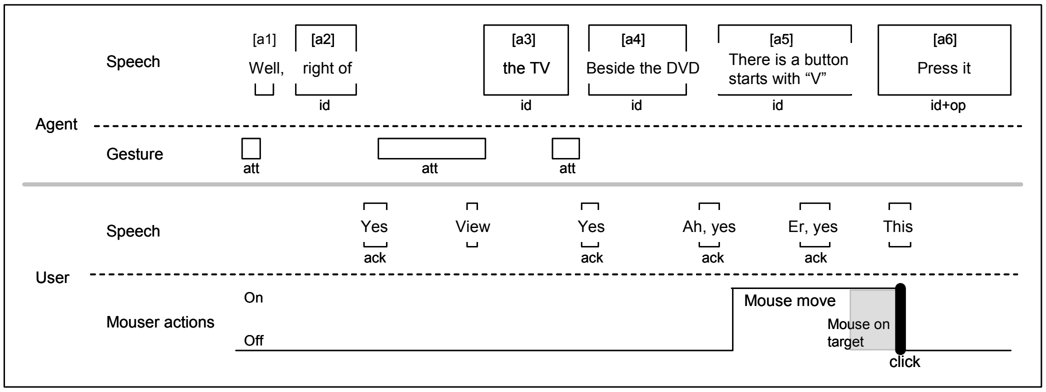


Figure 3: Example dialogue between Wizard-of-Oz agent and user

User's nonverbal behaviors: We annotated three types of mouse manipulation for the user's task manipulation as follows:

- Mouse movement: movement of the mouse cursor
- Mouse-on-target: the mouse cursor is on the target object
- Click target: click on the target object

4.1 Example of collected data

An example of an annotated corpus is shown in Figure 3. The upper two tracks illustrate the agent's verbal and nonverbal behaviors, and the other two illustrate the user's behaviors. The agent was pointing at the target (att) and giving a sequence of identification descriptions [a1-3]. Since the user's mouse did not move at all, the agent added another identification IPU [a4] accompanied by another pointing gesture. Immediately after that, the user's mouse cursor started moving towards the target object. After finishing the next IPU, the agent finally requested the user to click the object in [a6]. Note that the collected Wizard-of-Oz conversations are very similar to the human-human instruction dialogues shown in Figure 1. While carefully monitoring the user's mouse actions, the Wizard-of-Oz agent provided information in small pieces. If it was uncertain that the user was following the instruction, the agent added more explanation without continuing.

5 Probabilistic model of user-agent multimodal interaction

5.1 Building a Bayesian network model

To consider multiple factors for verbal and nonverbal behaviors in probabilistic reasoning, we employed a Bayesian network technique, which can infer the likelihood of the occurrence of a target event based on the dependencies among multiple kinds of evidence. We extracted the conversational data from the beginning of an instructor's identification utterance for a new target object to the point that the user clicks on the object. Each IPU was split at 500ms intervals, and 1395 intervals were obtained. As shown in Figure 4, the network consists of 9 properties concerning verbal and nonverbal behaviors for past, current, and future interval(s).

5.2 Predicting evidence of understanding

As a preliminary evaluation, we tested how accurately our Bayesian network model can predict an instructor's grounding judgment, and the user's mouse click. The following five kinds of evidence were given to the network to predict future states. As evidence for the previous three intervals (1.5 sec), we used (1) the percentage of time the agent touched the target (att), (2) the number of the user's mouse movements. Evidence for the current interval is (3) current IPU's content type, (4) whether the end of the current interval will be the end of the IPU (i.e. whether a pause will follow after the current interval), and (5) whether the mouse is on the target object.

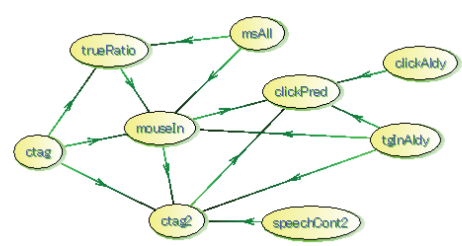


Figure 4: Bayesian network model

Table 1: Preliminary evaluation results

	Precision	Recall	F-measure
Content change	0.53	0.99	0.68
Same content	1.00	0.81	0.90

(a) Predicting grounding judgment: We tested how accurately the model can predict whether the instructor will go on to the next leg of the instruction or will give additional explanations using the same utterance content type (the current message will not be grounded).

The results of 5-fold cross-validation are shown in Table 1. Since 83% of the data are "same content" cases, prediction for "same content" is very accurate (F-measure is 0.90). However, it is not very easy to find "content change" case because of its less frequency (F-measure is 0.68). It would be better to test the model using more balanced data.

(b) Predicting user's mouse click: As a measure of the smoothness of task manipulation, the network predicted whether the user's mouse click would be successfully performed within the next 5 intervals (2.5sec). If a mouse click is predicted, the agent should just wait without annoying the user by unnecessary explanation. Since randomized data is not appropriate to test mouse click prediction, we used 299 sequences of utterances that were not used for training. Our model predicted 84% of the user's mouse clicks: 80% of them were predicted 3-5 intervals before the actual occurrence of the mouse click, and 20% were predicted 1 interval before. However, the model frequently generates wrong predictions. Improving precision rate is necessary.

6 Discussion and Future Work

We employed a Bayesian network technique to our goal of developing conversational agents that can generate fine-grained multimodal instruction dialogues, and we proposed a probabilistic model for predicting grounding judgment and user's successful mouse click. The results of preliminary evaluation suggest that separate models of each modality for each conversational participant cannot properly describe the complex process of on-going multimodal interaction, but modeling the interaction as dyadic activities with multiple tracks of modalities is a promising approach.

The advantage of employing the Bayesian network technique is that, by considering the cost of misclassification and the benefit of correct classification, the model can be easily adjusted according to the purpose of the system or the user's skill level. For example, we can make the model more cautious or incautious. Thus, our next step is to implement the proposed model into a conversational agent, and evaluate our model not only in its accuracy, but also in its effectiveness by testing the model with various utility values.

References

Brennan, S. 2000. Processes that shape conversation and their implications for computational linguistics. *In Proceedings of 38th Annual Meeting of the ACL.*

Cassell, J., Bickmore, T., Campbell, L., Vilhjalmsson, H. and Yan, H. (2000). Human Conversation as a System Framework: Designing Embodied Conversational Agents. *Embodied Conversational Agents*. J. Cassell, J. Sullivan, S. Prevost and E. Churchill. Cambridge, MA, MIT Press: 29-63.

Clark, H. H. and Schaefer, E. F. 1989. Contributing to discourse. *Cognitive Science* 13: 259-294.

Clark, H. H. and Krych, M. A. 2004. Speaking while monitoring addressees for understanding. *Journal of Memory and Language* 50(1): 62-81.

Dillenbourg, P., Traum, D. R. and Schneider, D. 1996. Grounding in Multi-modal Task Oriented Collaboration. *In Proceedings of EuroAI&Education Conference*: 415-425.

Kipp, M. 2004. Gesture Generation by Imitation - From Human Behavior to Computer Character Animation, Boca Raton, Florida: Dissertation.com.

Nakano, Y. I., Reinstein, G., Stocky, T. and Cassell, J. 2003. Towards a Model of Face-to-Face Grounding. *In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*: 553-561.

Paek, T. and Horvitz, E. (1999). Uncertainty, Utility, and Misunderstanding: A Decision-Theoretic Perspective on Grounding in Conversational Systems. *Working Papers of the AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems*. S. E. Brennan, A. Giboin and D. Traum: 85-92.

Whittaker, S. (2003). Theories and Methods in Mediated Communication. *The Handbook of Discourse Processes*. A. Graesser, MIT Press.