Modeling Human-Agent Interaction Using Bayesian Network Technique

Yukiko Nakano[1], Kazuyoshi Murata[1], Mika Enomoto[1], Yoshiko Arimoto[2], Yasuhiro Asa[3], and Hirohiko Sagawa[3]

[1] Tokyo University of Agri-culture and Technology
2-24-16 Nakacho, Koganei-shi, Tokyo 184-8588, Japan
{nakano,kmurata,menomoto}@cc.tuat.ac.jp
[2] Tokyo University of Technology
1404-1 Katakura, Hachioji, Tokyo 192-0981, Japan
ar@mf.teu.ac.jp
[3] Central Research Laboratory, Hitachi, Ltd.
1-280, Higashi-koigakubo Kokub-unji-shi, Tokyo 185-8601, Japan
{yasuhiro.asa.mk,hirohiko.sagawa.cu}@hitachi.com

Abstract. Task manipulation is direct evidence of understanding, and speakers adjust their utterances that are in progress by monitoring listener's task manipulation. Aiming at developing animated agents that control multimodal instruction dialogues by monitoring users' task manipulation, this paper presents a probabilistic model of fine-grained timing dependencies among multimodal communication behaviors. Our preliminary evaluation demonstrated that our model quite accurately judges whether the user understand the agent's utterances and predicts user's successful mouse manipulation, suggesting that the model is useful in estimating user's understanding and can be applied to determining the next action of an agent.

1 Introduction

In application software, help menus assist the user when s/he does not understand how to use the software. Help functions are usually used in problematic situations, so their usefulness and comprehensibility are critical for overall evaluation of the software. More importantly, if the advice provided by the help function is not helpful, that may confuse the user. Therefore, help functions that give useful advice at the right time are desirable.

To solve this problem, as a design basis of conversation-based help system, this paper proposes a human-agent multimodal interaction model using the Bayesian Network technique. This model predicts (a) whether the instructor's current utterance will be successfully understood by the learner, and (b) whether the learner will successfully manipulate the object in the near future.

Clark and Schaefer [1] defined that the process of ensuring that the listener shares an understanding of what has been said is grounding. Thus, the first issue, (a), can be said as the judgment of grounding: the judgment whether the instructor's utterance will be successfully grounded or not. If the predictions by the Bayesian network are accurate enough, they can be used as constraints in determining agent actions. For

K. Satoh et al. (Eds.): JSAI 2007, LNAI 4914, pp. 5–12, 2008.

example, if the current utterance will not be grounded, then the help agent must add more information.

2 Background

2.1 Monitoring Listener's Behaviors and Adjusting Utterances

Analyzing conversations where the speaker and the listener share a workspace, Clark and Krych [2] found that speakers dynamically adjust their utterances that are in progress according to the listener's feedback expressed in multimodal manners, such as spoken language, nonverbal behaviors (e.g. gestures and facial expressions), listener's task manipulation, and change of the task environment caused by the manipulation. In particular, monitoring a listener's task performance seems to be an effective way of organizing such multimodal conversations.

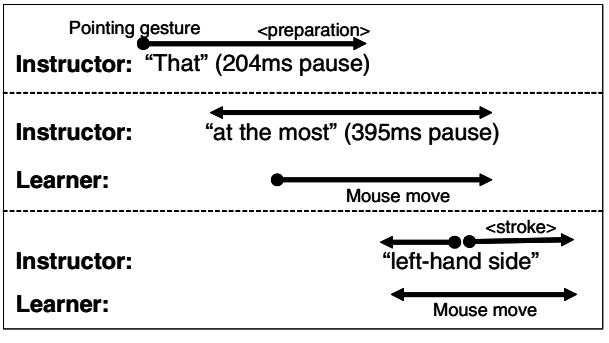


Fig. 1. Task manipulation dialogue

A software instruction dialogue in a video-mediated situation (originally in Japanese) is shown in Fig. 1. The speaker (instructor) is referring to the position of the "TV program searching button" by giving identification utterances in small pieces. Note that her gesture stroke follows the learner's mouse movements. This suggests that the speaker monitors the listener's mouse movement and adapts her verbal/nonverbal behaviors according to the listener's task manipulation. By virtue of such multimodal communication, the instructor smoothly coordinates the conversation even though there is no verbal response from the learner.

2.2 Nonverbal Information for Utterance Grounding

To accomplish such interaction between human users and animated help agents, predicting the task and conversation situation and modifying the content of instruction according to the prediction is necessary. For example, if the user does not seem to understand the agent's instruction, additional explanation is necessary. If the system predicts that the user fully understands the instruction and will conduct a proper operation in the near future, waiting for the user's operation without giving unnecessary annoying instruction would be better.

By applying a Bayesian Network technique to a dialogue-management mechanism, Paek and Horvitz [3] built a spoken-dialogue system, which takes account of the uncertainty of mutual understanding in spoken conversations. A similar technique was also applied to building user models in help systems [4]. However, there has been little study about timing dependencies among different types of behaviors in different modalities, such as speech, gestures, and mouse events, in predicting conversation status, and using such predictions as constraints in selecting the agent's next action.

Based on these discussions, this paper uses a probabilistic reasoning technique in modeling multimodal dialogues and evaluates how accurately the model can predict a user's task performance and judgment of utterance grounding [1].

3 Data Collection and Corpus

This section describes our corpus that is used for constructing a dialogue model. First, to collect dialogue data, we conducted an experiment using a Wizard-of-Oz method. In the experimental setting, an agent on an application window assists a user in operating a PC-TV application, a system for watching and recording TV programs on a PC.

3.1 Data Collection

A subject who joins the experiment as a user (hereafter, "user") and an instructor, who helps the user conducting the task and plays a role as a help agent, were in separate rooms. The equipment setting is shown in Fig. 2. The output of the PC operated by the user was displayed on a 23-inch monitor in front of the user and projected on a 120-inch big screen, in front of which the instructor was standing. The

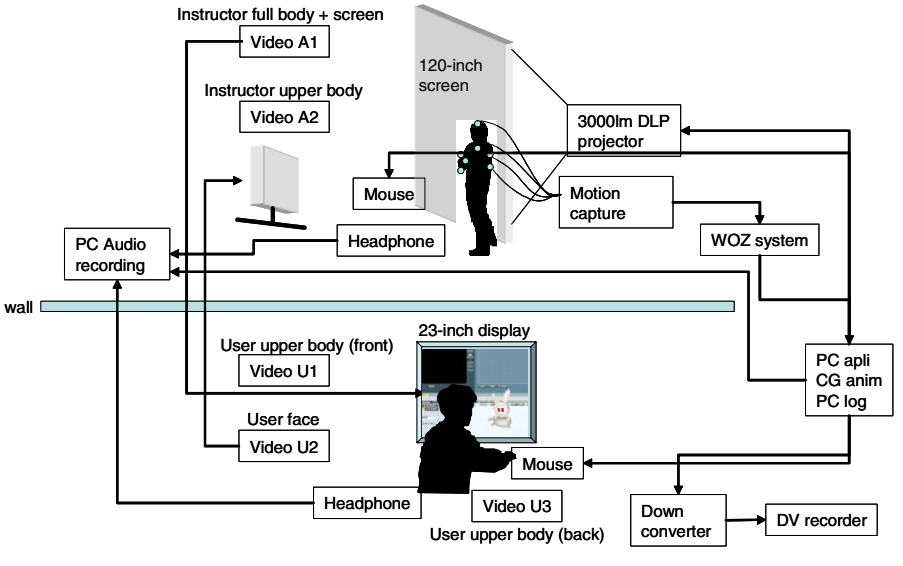
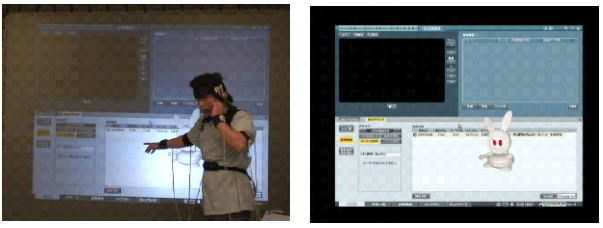


Fig. 2. Data Collection Environment



(a) Instructor (b) PC output

Fig. 3. Wizard-of-Oz agent controlled by instructor

instructor talked to the user while looking at the user's face (Video U2), which was monitored on a small display.

In addition, 10 motion sensors were attached to the instructor's body (Fig. 3 (a)) to capture the instructor's motion. The motion data was sent to the Wizard-of-Oz system and used to control a rabbit-like animated agent, which was overlaid on the user's PC-TV application. Thus, Fig. 3 (b) was displayed on the user's monitor as well as the big screen behind the instructor.

Both the user and the instructor wore headsets, and talked to each other through the headsets. The instructor's voice was changed through a voice transformation system, Herium, to make the voice sound artificial. Each participant's speech data was recorded by a headset microphone and saved in a Windows PC using a USB audio capture device. The audio was saved in the WAV audio format.

3.2 Task and Experimental Design

Each user was assigned one of two situations: recording a TV program or burning a DVD. The number of users was balanced between the situations. With the instructor's help, the user worked on two tasks for each situation.

Ten instructors and twenty users participated in the experiment. Each instructor had two sessions with two different users. Thus, we collected conversations from twenty pairs.

4 Corpus

The agent's (actually, instructor's) speech data was split by pauses longer than 200ms. We call each speech segment an inter-pausal unit (IPU), and use this as a unit of transcription. We assigned the following tags to 25 conversations using the Anvil video-annotating tool [5].

4.1 Utterance Content Tags

Focusing on the characteristics of the task, the utterance content of each IPU was categorized as follows.

- 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): identification and operation in one IPU
- State (st): referring to a state before/after an operation
- Function (fn): explaining a function of the system
- Goal (gl): utterance content for determining a purpose or a goal to be accomplished
- Acknowledgment (ack): agreement to or acknowledgement of the partner's speech

The inter-coder agreement for this coding scheme is very high, K = 0.89 (Cohen's Kappa), suggesting that the assigned tags are reliable.

4.2 Agent's Gestures and Motions

(1) Instructor's nonverbal behaviors
We annotated the shape and phase of the instructor's (Wizard-of-Oz agent's) gestures.
(1-1) Gesture Shape
- Pointing: pointing at one place on the monitor, or a hand motion that circles multiple objects.
- Trace: drawing a line to trace the words and phrases on the display.
- Other: other gestures

(1-2) Gesture Phase
- Preparation: arm movement from the beginning of a gesture to the stroke.
- Stroke: the peak of a gesture. The moment that a stroke is observed.
- Hold: holding a stroke hand shape, such as continuing to point at one position without moving a hand.
- Retract: retracting a hand from a stroke position.
- Partial retract: partially retracting an arm to go to the next stroke.
- Hesitate: a pause between a partial retract and a preparation or a pause between strokes.

(2) Agent Motions
We also annotated the positions and the gestures of the agent, which is actually controlled by the instructor.
(2-1) Agent movement: Duration of agent's position movement. If the agent does not move for longer than the time of 15 frames, that is counted as the end of the movement.
(2-2) Agent touching target as pointing (att): Duration of agent touching the target object as a stroke of a pointing gesture.

4.3 Mouse Operations

Using an automatic logging tool, we collected the following three kinds of log data as user's mouse operations.

- 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

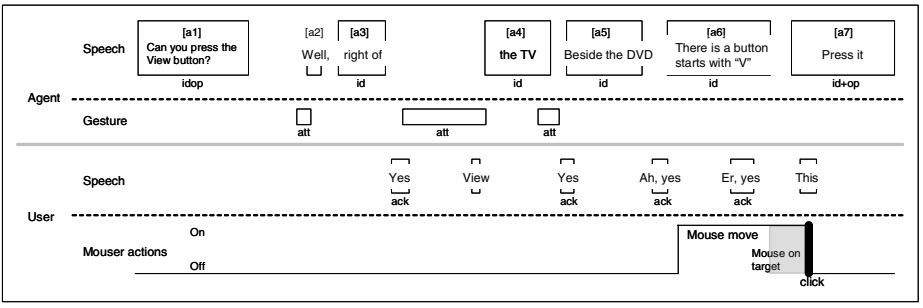


Fig. 4. Dialogue between Wizard-of-Oz agent and user

4.4 Corpus Data

An annotated corpus is shown in Fig. 4. The upper two tracks illustrate the agent's verbal and nonverbal behaviors, and the other two tracks illustrate the user's behaviors. At the first IPU, the instructor said, [a1] "Could you press the View Button?" The user did not respond to this instruction, so the instructor changed the explanation strategy: giving a sequence of identification descriptions [a2-5] by using short utterance fragments between pauses. Although the user returned acknowledgement, the user's mouse did not move at all. Thus, the instructor added another identification IPU [a6] accompanied by another pointing gesture. Immediately after that, the user's mouse cursor started moving towards the target object. After confirming that the user's mouse cursor reached the target object, the agent finally requested the user to click the object at [a7]. Note that the collected Wizard-of-Oz conversations are very similar to the human-human instruction dialogues shown in Fig. 1. While carefully monitoring the user's mouse actions, the Wizard-of-Oz agent adjusts the content of the instruction and its timing.

5 Dialogue Modeling Using Bayesian Network

In this section, a probabilistic dialogue model is constructed from the corpus data by using the 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 conversational data from the beginning of an instructor's identification utterance about a new target object to the point when the user clicks on the object. Each IPU was split at 500 ms intervals, and 1395 intervals were obtained. As shown in Fig. 5, the network consists of 9 properties concerning verbal and nonverbal behaviors for the past 1.5 seconds, current, and future interval(s).

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 information were given to the network 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

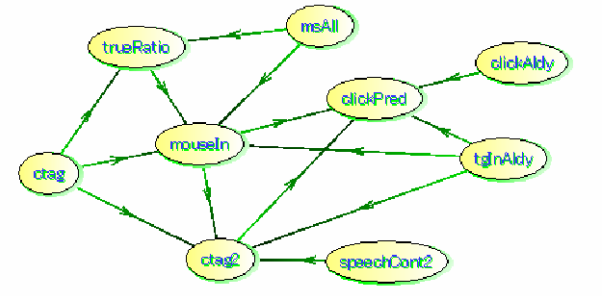


Fig. 5. Bayesian network model

the current interval is (3) content type of current IPU's, (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.

5.1 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 a 5-fold cross-validation are shown in Table 1. The prediction of "same content" is very accurate (F-measure is 0.90) because 83% of the data are "same content" cases. However, finding "content change" is not very easy because that occurs with less frequency (F-measure is 0.68). Testing the model using more balanced data would be better.

Table 1. Evaluation results

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

5.2 Predicting User's Mouse Clicks

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 five intervals (2.5 sec). If a mouse click is predicted, the agent should just wait without annoying the user with an unnecessary explanation. Randomized data is not appropriate to test mouse click prediction, so 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 the precision rate is necessary.

6 Conclusion

Aiming at building a conversational help agent, first, this paper reported our experiment and verbal and nonverbal behavior annotation. Then, we proposed a probabilistic model for predicting grounding judgment and a user's successful mouse click. Adding more data, we will conduct more precise statistical analysis to demonstrate the co-constructive process of multimodal conversations. Moreover, our next step is to implement the proposed model in a conversational agent and evaluate the effectiveness of the proposed model.

References

1. Clark, H.H., Schaefer, E.F.: Contributing to discourse. Cognitive Science 13, 259–294 (1989)
2. Clark, H.H., Krych, M.A.: Speaking while monitoring addressees for understanding. Journal of Memory and Language 50(1), 62–81 (2004)
3. Paek, T., Horvitz, E.: Uncertainty, Utility, and Misunderstanding: A Decision-Theoretic Perspective on Grounding in Conversational Systems. In: Brennan, S.E., Giboin, A., Traum, D. (eds.) Working Papers of the AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems, American Association for Artificial Intelligence, Menlo Park, California, pp. 85–92 (1999)
4. Horvitz, E., et al.: The Lumiere Project: Bayesian User Modeling for Inferring the Goals and Needs of Software Users. In: Proceedings of Fourteenth Conference on Uncertainty in Artificial Intelligence, pp. 256–265 (1998)
5. Kipp, M.: Gesture Generation by Imitation - From Human Behavior to Computer Character Animation, Boca Raton, Florida: Dissertation.com (2004)