Estimating User's Attitude in Multimodal Conversational System for Elderly People with Dementia

Naoko Saito[1], Shogo Okada[1], Katsumi Nitta[1], Yukiko I. Nakano[2], Yuki Hayashi[3]

[1] Tokyo Institute of Technology
{saito_n@ntt., okada@, nitta@}dis.titech.ac.jp

[2] Seikei University
y.nakano@seikei.ac.jp

[3] Osaka Prefecture University
hayashi@kis.osakafu-u.ac.jp

Abstract

Toward constructing a multimodal conversation agent system which can be used to interview elderly patients with dementia, we propose a turn taking mechanism based on recognition of the subjects attitude as to whether the subject has (or relinquish) the right to speak. A key strategy in the recognition task is to extract features from pausing behavior in subject's spontaneous speech and to fuse multimodal signals (gaze, head motion, and speech). In this paper, we focus on evaluation of the recognition module used in guiding turn taking. To evaluate it, we collect multimodal data corpus from 42 dyadic conversations between subjects with dementia and the virtual agent we have developed as a prototype and annotate subject's multimodal data manually.

In experiments, we validate recognition models trained multimodal dataset by machine learning methods. Experimental results shows that pause features are effective to improve the attitude recognition accuracy and the accuracy is improved up to 88%.

Introduction

Elderly people with dementia often cause concern to their family or their care worker in a conversation because they have difficulty with their memory, speech, etc. To reduce the burden, we have developed a conversational agent. We have collected multimodal conversation data corpus including audio and video. Through video analysis, we found that pause often occurs when the subject answers the question, and the subject stops speaking after that. In this case, agent's utterance for the next question clashes with the subject's utterance. The situation prevent talking comfortably, and defeats the purpose of our agent as a conversational partner. To tackle such utterance clashing problem, we propose the framework of turn taking management based on speaker's attitude recognition using multimodal nonverbal behaviors. Objective of the attitude recognition is to estimate whether the subject is motivated to speak and want to have the right to speak. We annotate the labels of nonverbal behavior to video and extract various features from annotated labels. In experimental evaluation, we trained recognition model using machine learning techniques and evaluate whether proposed features successfully detect the correct attitude.

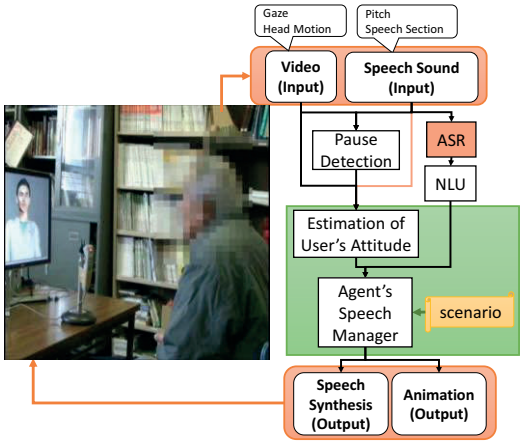

Figure 1: Overview of the Agent System

Related Work

Embodied Conversational Agent (ECA) with particular person as a target has been developed by many researchers (e.g. Vardoulakis 2012). However, very few attempts have been made at developing a conversational system for the elderly with dementia. Fuketa et al. verify an effect of topic about themselves in a conversation. To construct the system in which the patient feels able to talk comfortably, they concluded that the system needs to perform using a conversation with a topic that the patient likes in general. In past studies, we also have developed a conversational agent that can serve as a conversational partner for the elderly with dementia (Nonaka 2012).

In (Ooko 2011, Ishii 2011), the researchers developed the system estimating the subject's conversational engagement from head motion and gaze. In this study, we develop the conversational system on multimodal subject's attitude recognition, considering that the patient has a conversational feature where pauses are observed in their utterance frequently. To deal with this feature, we annotated filled pause in spontaneous utterance (Goto 1999, Jokinen 2010) to use it as features for attitude recognition.

Proposed Multimodal Architecture for ECA

The architecture of a proposed conversational agent system is shown in the Figure 1. While the process of automatic speech recognition (ASR) is running, the agent generates nodding and response as feedback, and decides the next question in the dialogue scenario (see in the next section) judging from the result of ASR. In addition, the agent's speech manager decides the timing that the agent starts to speak by detecting subject's multimodal cues including utterance, head gesture, and gaze state.

Topic Management

Topics which a subject wants to talk are different depending on the individual. For managing these topics, we introduced a scenario that has a binary tree structure for conversation. An agent's question corresponds to a node in the tree. Excepting end of tree, the agent asks a "yes-no" questions and acquire profile information of the subject. When the scenario reaches end node of the tree, the agent asks "open question", which a subject can answer in a free manner. The questions in the scenario have following topics: weather, family, health, food, and hobby. Vardoulakis et al. have reported that weather and family are important topics in conversation (Vardoulakis 2012). We also found that the elderly with dementia speaks on three topics (health, food, and hobby) for long time or frequently by analyzing the conversations with video.

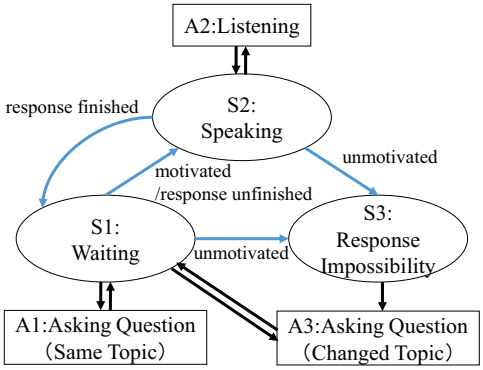

Figure 2: Dialogue Strategy Based on Speaker's Attitude Recognition

Multimodal Speaker's Attitude Recognition Framework

The subjects with dementia often pause on a word in responding and restart speaking after pausing. Since our agent decides to start speaking by detecting fixed voiceless segment, the subject's utterance often clashes with the agent's utterance and turn taking fails. To tackle this problem, we introduce a multimodal subject's attitude recognition framework.

Figure 2 shows the framework of multimodal speaker's attitude recognition. Here, we define 3 types of the attitude of the subject in conversation:

- S1 (Waiting): the subject is preparing to hold a turn and waiting for the next agent's question.
- S2 (Speaking): the subject is answering the agent's question
- S3 (Response Impossible): the subject seems to be unmotivated to speak.

The agent's action which corresponds to each attitude is defined:

- A1 (Asking Question (Same Topic))
- A2 (Listening)
- A3 (Asking Question (Changed Topic))

When the agent finishes question, the attitude of the subject transitions from S1 to S2 or S3 according to whether the subject seems to be motivated or unmotivated to speak. When the subject is motivated, the state transitions to S2. At the time, the state of agent transitions to A1 for urging and waiting the subject to speak. After the subject finishes speaking, the attitude transitions to S1 and the subject waits for the next question. In this case, the topic of the next question is equal to the topic of the previous question (A2).

When the system estimates the subject is unmotivated, the attitude is estimated to transition from S1 or S2 to S3. The agent acts A3 which the agent questions changed topic, the next question is chosen from other tree, not current tree. Then the subject returns to S1. The agent talks with the subject based on the state transition model.

Multimodal Data Setting

Data Collection

In order to train the attitude recognition model, we collected conversation data set by elderly patients and the agent. At first, we developed a prototype of conversational agent that can ask subjects questions under a fixed order. The system of agent has only speech detection module but

does not have the ASR module and Multimodal attitude recognition. In the experiment, subjects were 28 patients (16 males and 12 females), their average age is 76.4 years old. They talked with the agent for about 10 minutes in a session. As a result, we collected data on 42 conversations using a video camera and a microphone to record the conversations.

Multimodal Manual Annotation and Feature Description

To evaluate our framework, we annotated each label (Gaze, Nodding and Pause) manually by sharing with 4 annotators. Utterances by subject and agent are annotated automatically using Julius[1]. We remove misrecognition label by checking manually. Figure 3 shows that they actually annotated labels along a time axis by using ANVIL[2]. We determined the following features from label data about current time segment:

- Length : the total length of labels
- Count : the number of labels
- Timing: the number of labels existing each position

The features of "Timing" are defined by relationship between utterance label and each label (expressed in white square) which has start and end time. (Figure 4).

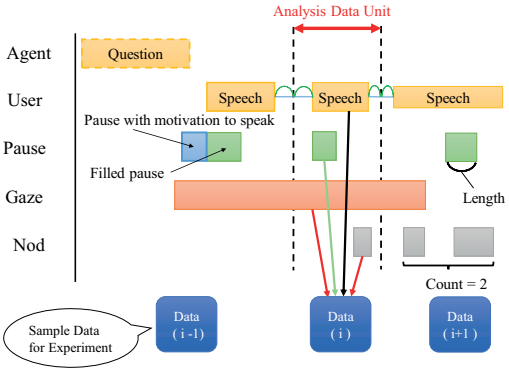

Figure 3: Annotation and Data Division

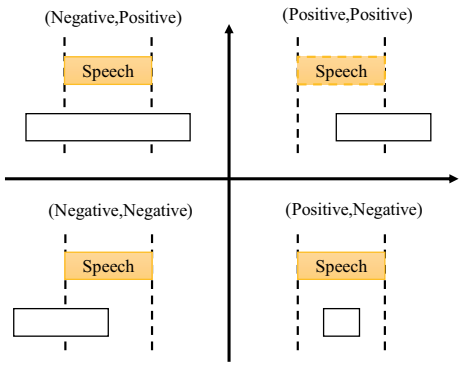

Figure 4: Concept of "Timing"

Speech Utterance of Agent (Question)

We annotated question labels because it is considered that the questions of the agent affects the subject's nonverbal behavior.

Speech Utterance of User (Answer)

We extract a Length feature (SL). User's utterance is a reference point to encode timing features.

Gaze

We annotated a state in which the subject looks at the outside of the display and use 5 features: Length(GL) and Timing(GT(P,P), GT(P,N), GT(N,P), GT(N,N))

Nod

We annotated a label during nodding including one or several times of nodding and use 5 features: Length(NL) and Timing(NT(P,P), NT(P,N), NT(N,P), NT(N,N))

Pause

We annotated two types of labels about pause. (1)The label of filled pause: This type includes intermittent utterance, restatement, etc. (2)The label of pause with motivation to speak. We annotated the labels when the subject is seemed to think about an answer to the question with nonverbal activity such as head tilting, head nodding, and changing gaze direction.

For filled pause, we defined 6 features: Length (PfL), Count (PfC) and Timing (PfT(P,P),PfT(P,N), PfT(N,P), PfT(N,N)) For pause with motivation to speak, we also defined 6 features: Length (PmL), Count (PmC) and Timing (PmT(P,P), PmT(P,N), PmT(N,P), PmT(N,N)).

Attitude of Subject

titude labels are annotated by 1 coder. Attitude label types S1, S2 and S3. These label are used as supervised data for attitude recognition model.

Experiments

ng

e objective of the experiment is to evaluate the ework of multimodal speaker's attitude recognition. It portant to estimate whether the utterance of the subject ished to prevent an utterance clashing problem. This

[1] http:

'/www.anvil-software.org/

estimation task is equivalent to classifying the subject's attitude states into S1, S2, and S3. The results are recognized based on a set rule of detecting voiceless segments.

For the evaluation, we divide annotation data by the middle point between the utterances and extract features from the annotated data. In this evaluation, we perform binary classification (S1 or S2) task using multimodal feature set. We obtained 692 sample data for S1 and S2 (614 samples of S1 and 78 samples of S2).

Statistical Analysis

In order to analyze effective feature for classifying the two attitude states, we conduct unpaired two tailed t-test (α =0.05) using the samples. The results of t-test is shown in Table 1. Some features show significant differences in Speech, Pause, and Gaze. Moreover, Length, Count and Timing of pause has significant difference and likely to be an effective feature for recognizing the attitude.

Evaluation using Machine Learning

Then, using those features, we train attitude recognition model by three kinds of machine learning techniques (DecisionTree, RandomForest, and linear SVM) by using a 10-fold Cross Validation approach. We calculate mean F-measure score which average F-measure of S1 accuracy and that of S2 accuracy rate. The result is shown in Table 2. To verify the effectiveness of features, we calculated it by using about seven kinds of feature set. As a result, highest accuracy is in the case (3) using all features. In particular, pause with motivation to speak is much effective. The cases (3), (4) and (7), which are using pause with motivation to speak, is calculated as a higher score than other cases. To improve the recognition accuracy, we need to reconsider the kinds of features for future tasks.

Table 1: T-Test Result

		df	t	P
(1)Speech	SL(*)	159	2.77	0
(2)Filled Pause	PfL(*)	89	-3.48	0
	PfC(*)	690	-4.36	0
	PfT(P,P)(*)	80	-4.33	0
	PfT(N,P)(*)	77	-2.53	0.01
(3)Pause with motivation to speak	PmL(*)	690	-8.33	0
	PmC(*)	690	-16.28	0
	PmT(P,P)(*)	83	-17.40	0
	PmT(N,N)(*)	690	-5.54	0
(4)Gaze	GL(*)	690	-2.29	0.02
	GT(N,P)(*)	87	-4.97	0

Table 2: Machine Learning Result (Mean F-Measure)

Feature set	Decision Tree	Random Forest	SVM (linear)
(1) Speech + Nod + Gaze	0.37	0.45	0.47
(2) (1) + Filled pause with utterance	0.46	0.54	0.63
(3) (2) + Pause with motivation to speak (All)	**0.84**	**0.86**	**0.87**
(4) (3) - Timing feartures	0.71	0.70	0.79
(5) Speech + Nod	0.44	0.50	0.47
(6) Speech + Gaze	0.37	0.39	0.47
(7) Speech + Filled pause	0.83	0.86	0.88

Conclusion

In this paper, we proposed the conversational agent system for elderly patients with dementia. As a dialogue strategy for the agent, we considered the framework of the agent's action management on estimating subject's attitude from multimodal nonverbal behavior. We conducted the experiment of statistical analysis and machine learning in order to evaluate the framework. While reporting the timing structure among multimodal labels, we tried to determine whether pauses are effective to improve the accuracy of estimating whether a subject finishes an utterance. These results show multimodal nonverbal features are effective to manage turn taking.

References

M. Fuketa., K.Morita, and J. I. Aoe, 2013. *Agent–based communication systems for elders using a reminiscence therapy:* International Journal of Intelligent Systems Technologies and Applications, 12(3), 254-267.

M. Goto, K. Itou, and S. Hayamizu. 1999. *A Real-time Filled Pause Detection System for Spontaneous Speech Recognition:* Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech '99), pp.227-230.

R. Ishi, Y.Shinohara, Y. I. Nakano, and T. Nishida. 2011. *Finer eye-gaze information to predict user's conversational engagement:* International Conference on Intelligent User Interfaces (IUI 2011), Workshop on Eye Gaze in Intelligent Human Machine Interaction.

K. Jokinen, and J. Allwood. 2010. *Hesitation in intercultural communication: some observations and analyses on interpreting shoulder shrugging:* In Culture and computing (pp. 55-70). Springer Berlin Heidelberg.

Y. Nonaka, Y. Sakai, K. Yasuda, Y. I. Nakano. 2012. *Towards Assessing the Communication Responsiveness of People with Dementia*: IVA 2012, pp.496-498.

R. Ooko, R. Ishii, and Y. I. Nakano. 2011. *Estimating a User's Conversational Engagement Based on Head Pose Information:* the 11th International Conference on Intelligent Virtual Agents (IVA 2011), pp. 262-268.

L. P. Vardoulakis, L. Ring, B. Barry, C. L. Sidner and T. Bickmore. 2012. *Designing Relational Agents as Long Term Social Companions for Older Adults:* Intelligent Virtual Agents conference (IVA)