Toward a Guide Agent who Actively Intervene Inter-user Conversation – Timing Definition and Trial of Automatic Detection using Low-level Nonverbal Features

Hung-Hsuan Huang, Shochi Otogi, Ryo Hotta and Kyoji Kawagoe
College of Information Science and Engineering, Ritsumeikan University, Shiga, Japan

Keywords: Multi-party conversation, Human-agent Interaction, Conversational Agent, Multi-modal Interface.

Abstract: As the advance of embodied conversational agent (ECA) technologies, there are more and more real-world deployed applications of ECA's. The guides in museums or exhibitions are typical examples. However, in these situations, the agent systems usually need to engage groups of visitors rather than individual ones. In such a multi-user situation, which is much more complex than single user one, specialized additional features are required. One of them is the ability for the agent to smoothly intervene user-user conversation. In order to realize this, at first, a Wizard-of-Oz (WOZ) experiment was conducted for collecting human interaction data. By analyzing the collected data corpus, four kinds of timings that potentially allow the agent to do intervention were found. The collected corpus was then annotated with these defined timings by recruited evaluators with a dedicated and intuitive tool. Finally, as the trial of the possibility of automatic detection on these timings, the use of non-verbal low level features were able to achieve a moderate accuracy.

1 INTRODUCTION

As the advance of embodied conversational agent (ECA) technology, there are more and more real-world deployed applications of ECA's (Kopp et al., 2005; Traum et al., 2012). Various kinds of kiosk information systems are used in public places, such as shopping malls, museums, and visitor centers. The typical situation in which such systems are used is that a group of people stand in front of the kiosk and operate it in order to retrieve the information while talking with another visitors in the same group. Therefore, in order to implement an ECA that can serve as an information kiosk in public places, multi-party conversation functionality for simultaneous interaction with multiple users is indispensable.

In dyadic dialogs where only two participants are involved, it can be assumed that when one participant is speaking then the other one is the addressee in most cases. In multi-party dialogs, however, the distinction among conversation participants' roles, addressee, overhearer, and speaker is necessary. Who is talking to whom about what topic should be identified and traced, the obligation to answer a question when being asked needs to be fulfilled, and whether an utterance has been grounded (Clark and Schaefer, 1989) with the addressee(s), i.e. became a shared belief between the speaker and the addressee(s), needs to be identified. Since there are potentially more interlocutors to acquire dialog turns from and transfer dialog turns to, managing the communication flow in a multi-party conversation is much more sophisticated than a dyadic one. Traum (Traum, 2003) provided the principal literature identifying some of the general issues in realizing multi-party human-agent interactions.

The conversation style of most contemporary ECA systems are either agent-initiative (e.g. the agent always asks questions and the user answers) or user-initiative (e.g. the user always ask questions and the agent answers), because these interaction designs are simpler and are easier to be implemented. However, the natural conversation occurred between human and human is mixed-initiative, i.e. both of the agent and the user may take initiative during the conversation. In this study, we are proceeding with a project that aims to build an ECA capable of mixed-initiative conversation with multiple users in a typical application, information providing for users' collaborative decision making.

Considering the case where a real estate agent is serving a recently married couple who are planning to buy a new house, the married couples consider the

Huang, H-H., Otogi, S., Hotta, R. and Kawagoe, K.
Toward a Guide Agent who Actively Intervene Inter-user Conversation – Timing Definition and Trial of Automatic Detection using Low-level Nonverbal Features.
DOI: 10.5220/0005759704590464
In *Proceedings of the 8th International Conference on Agents and Artificial Intelligence (ICAART 2016) - Volume 2*, pages 459-464
ISBN: 978-989-758-172-4

location, layout, or the price of the houses suggested by the agent, discuss with each other and collaboratively make the final decision. In this task, the decision can be made efficiently if the agent can actively intervene the discussion between the users to provide timely information while the agent is listening to the users' conversation as an overhearer. For example, the agent may suggest new candidates as "how about this apartment? It is only a five-minute walk to the train station." if it hears that the users are discussing about the transportation issue, or "how about this apartment? There is a super market which is just five-minute away from it." if it hears that the users are discussing about meal issue. This kind of interaction is nature and frequently occurs in human-human conversation. However, for a virtual agent to tackle this, additional specialized functions have to been incorporated to the agent. The agent needs to reason what information to provide from the users' conversation even their demands are not clearly described. The agent also needs to identify the timings when the user may be interested in the information being provided without making them feel disturbed. In order to realize this, we focus on nonverbal information. It has been known that nonverbal information is an important element in a conversation scene. In particular, gaze has been found to have a major impact on multiparty conversation (Subramanian et al., 2010) .

This paper describes the development of the method dealing with the timing finding issue. In order to realize the feature to find the appropriate timings for the agent, at first, a Wizard-of-Oz (WOZ) experiment was conducted for collecting a human interaction data corpus. The experiment participants were instructed to collaboratively make decisions in three tasks: travel planning, lecture registration, and part-time job hunting. They did the tasks with the information provided by a virtual agent who is controlled by one of the experimenters from remote (Section 3). Then an analysis was conducted on the collected data corpus to find the eight possible timings and the corresponding user behaviors (Section 4). A method was developed to automatically identify four of the four kinds of timings only by using nonverbal cues, gaze, body posture, and acoustic information (Section 5). Finally, the paper is concluded with the evaluation of the performance of the proposed method and the discussion on the results.

2 RELATED WORKS

Many studies on the processes of recognition and interpretation in multi-party meetings have been done in the AMI and AMIDA projects (Renals et al., 2007) and in the projects of DARPA (Waibel et al., 1998; Waibel et al., 2001). By applying speech and natural language processing technologies, a number of useful components were developed in these projects, such as speech transcription, speaker diarization, and meeting summarization.

The management of speaking turns in multi-party dialogs is very related to the goal of this work. The intervention for active information providing is a higher level task that needs to consider the situation and the context of conversation in addition. In the literature of human communication, (Kendon, 1967; Duncan, 1972; Sacks et al., 1974), it has been reported that in addition to explicit verbal utterances, people also use nonverbal signals such as their gaze, nods, and postures to regulate their conversation, e.g., to show their intention to yield or the willingness to take next speech turn. The speaker looks at the addressee to monitor her/his understanding or attitude, and, the addressee looks at the speaker in order to be able to offer positive feedbacks in return (Kendon, 1967; Argyle and Cook, 1976). When yielding his/her turns for another participant to speak, the speaker looks at the next speaker at the end of his/her utterances (Duncan, 1972). Takemae (Takemae et al., 2003) provided the evidence that the speaker's gaze indicates addressee-hood and plays a regulatory role in turn management. The Japanese spoken dialog system, DUG-1 (Nakano et al., 1999) realized rich turn-taking with rules based on linguistics features. Jonsdottir and Thorisson (Jonsdottir and Thorisson, 2009) proposed a system that learns turn-taking behavior on-line with an individual user in less 30 turns by utilizing reinforcement learning techniques. Bohus and Horvitz (Bohus and Horvitz, 2010) proposed a turn management framework in the virtual receptionist application. They showed the verbal and nonverbal cues used by the avatar can shape the dynamics of multi-party conversation. However, the application task was relatively simple, a question-answering game with trivial questions.

3 INTERACTION CORPUS COLLECTING WOZ EXPERIMENT

To collect the video corpus for an analysis of the situations when the agent can potentially intervene the users to provide timely information, a WOZ experiment on three collaborative decision making tasks was conducted. We expect that the subjects' reactions

toward the agent may differ to how they talk with a human information provider. To observe the natural interaction with humans and agents, we chose the WOZ experiment setting instead of a human-human one.

3.1 Experiment Settings

Pairs of experiment participants were instructed to interact with a life-size female character projected on a 100-inch screen. They had to retrieve information from the character in order to make a decision regarding the given tasks until the agreement between them achieved. As shown in Figure 1, the subjects stood about 1.8 m away from the screen where the character was projected. Two video cameras were used to record the whole experiment, one from the front to take the upper bodies of the participants and the other one takes the whole scene including the participants and the character from the rear. One Webcam was used for the telecommunication software Skype to connect the WOZ operator to the experiment room. The microphone array of one Microsoft Kinect sensor was to identify the voice source (left or right user) of user utterances. The other Kinect sensor was used to record body postures with depth images. The conversation experiment was conducted with the following premises:

- The participants want to make a decision based on their agreement from multiple candidates with the help of the agent who is knowledgeable about that task domain.
- The participants have a rough image of what they want, but they do not have idea about particular candidates in advance.
- The participants discuss on their own and acquire new information from the agent.
- The conversation ends when the participants made the final decision.

A total of 15 pairs of college students were recruited as the participants in the experiment, all of whom were native Japanese speakers. The students came from various departments ranging from economics, life science to engineering at average age, 19.2. 11 of the all 15 pairs were male ones and the other four were female ones. Each pair was instructed to complete three decision-making tasks described as the follows:

Travel Planning: the participants were instructed to pretend to be in a situation where they had a coupon from a travel agency that allows them to visit three of 14 sightseeing spots in Kyushu for free. The information, which includes a brief history, highlights, nearby restaurants, for each location was defined in advance. The sightseeing spots were selected from four of all seven prefectures inside Kyushu. The participants were instructed to complete their task by freely retrieving information from the travel agent, and to discuss their decisions on their own.

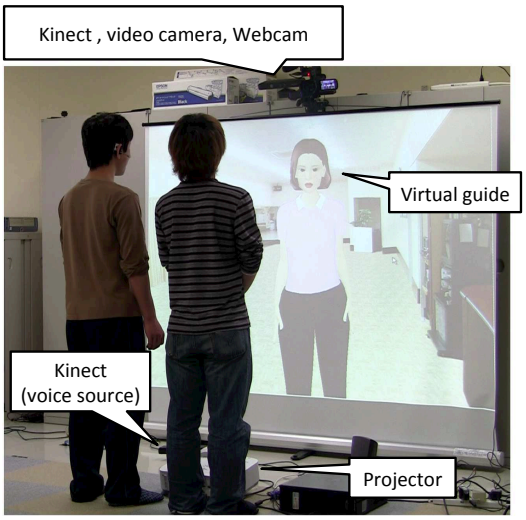

Figure 1: Setup of the WOZ experiment to collect data for the interaction corpus.

Lecture Registration: the participants were instructed to choose three out of 12 lectures to attend together in the next semester. The information about the lecturer, textbook, course difficulty, prerequisites, etc. for each lecture was defined in advance. The lectures were divided into four categories: information science, engineering, languages and communication, and social science. The subjects could freely ask the "tutor" agent for any information about the lectures or the agent itself, and then discuss this on their own in order to make the final decision.

Part-time Job Hunting: the participants were instructed to request help in choosing three out of 14 part-time jobs to work together near the university. The information about the salary, location, workload, work type, etc. for each job was defined in advance. The jobs were divided into four categories: convenient stores, book shops, restaurants, and gas stations. The subjects could freely ask the agent for any information about the part-time jobs or the agent itself, and then discuss this on their own in order to make the final decision.

These tasks were chosen because the student participants are supposed to be familiar with these issues. In order to stimulate more active discussion, the par-

ticipants were instructed to make the ranking of three final choices. All participant pairs were assigned to take all of the three tasks in three separate sessions, one task for one session. The sessions were conducted in all possible orders to cancel order effects. One student who major in computer science was recruited to operate the WOZ agent. He was chosen due to his familiarity in operating a GUI-based WOZ application, which ensured that there would be smooth interaction. The operator was asked to practice on the WOZ user interface for two hours prior to the experiment to further ensure that the agent can response as soon as possible. All the sentences that the agent could speak during the experiment were listed in a menu where relevant sentences were grouped for the WOZ operator to select from more easily. There was also a text field that allowed the operator to type arbitrary utterances, in the cases when they were needed but were not defined. The WOZ operator was instructed to try to end the interaction sessions in ten minutes, if possible.

4 TIMING EXPLORATION

The appropriate timings of when to intervene the inter-user conversation for a guide agent is a subjective issue and has no correct answers. Therefore, human evaluation was used in this work. The video clips of 12 sessions, which were randomly selected from the corpus were evaluated by 20 recruited evaluators (all male college students at average age, 20.2). Each video clip was assigned to five evaluators. The evaluators were instructed to annotate the video when they want to provide additional on-time information for the users, as if they were proactive and kind guides. Intuitive impression was considered essential for this task, a dedicated annotation tool other than widely used tools like Anvil or ELAN. The evaluators watched the video and input four possible timings with a game pad. The A, B, C, D button of the game pad were assigned to the following four possible timings respectively.

Provide-topic (P-T): switch the topic and provide new information to the users.

More-information (M-I): provide more detailed information in compensating the information provided previously.

Recall-support (R-S): remind the user about the information provided by the agent if they forgot it.

Discussion-support (D-S): sort out and conclude the discussion up to now.

In the case when the evaluator wanted to use intervene the inter-user conversation in another way, he/she can pause the video and freely type “other” types of timings.

4.1 Analysis of Annotation Data

Totally there were 436 timings labeled by the 20 evaluators upon the 12 video clips (five evaluators for one clip, averagely 36.3 timings for one clip). In order to analyze these timings, we segmented the video into the following four types of periods.

After-agent: immediately after the utterances of the agent

After-user: immediately after the utterances of the user

Silent: silent period (at least three seconds)

During-user: during the user’s utterances

The periods during the agent’s utterances were excluded. In considering what kind of support is appropriate, the annotation results of four possible timings were shown in Figure 2. The timing types did not distribute evenly in the segments, especially the type D-S (during the users’ utterances) had most instances. This is considered that the evaluators most willing to provide some kind of additional support during the users’ conversation. The other timings like, “the reply of the guide toward the question was too quick,”‘ ‘the pronunciation of the agent should be clearer” were excluded.

From these results, the timing type M-I had most instances (19.0 per group). This means that the users were not satisfied with the amount of the information provided by the agent, or the agent did not make a right answer. In these cases, timing type R-S and D-S had 4.2 and 4.8 instances averagely for each group. Type D-S was annotated around every two minutes in comparing with the average length of the video clips, 11:40. This implies that it’s important to sort the the users’ conversation in a thread if the conversation diverges.

5 AUTOMATIC CLASSIFICATION OF INTERVENING TIMINGS

5.1 Selected Features

In order to automatically classify the timing types, verbal and non-verbal features are selected and annotated if necessary. The details of the selected features are described below:

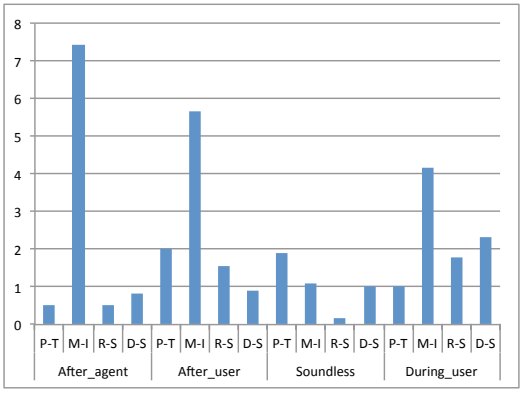

Figure 2: Summary of annotation results regarding to timing types and periods.

Speaking State: When agent itself is speaking, it can not intervene the user's conversation, and while a user is speaking, it is impolite to intervene. Among the definition of time periods regarding to speaking state in section 4.1, the three periods, After_agent, After_user, and Soundless where neither the users nor the agent is speaking can be considered as the candidates of intervening.

Speaker: In this work, in order to explore the full capability of the model which classifies the timing types, 100% correct information of the speaker of each utterance has to be prepared. In the WOZ experiment, there are three possibilities of the speaker of an individual utterance, the user standing at right hand side, the user standing at the left hand side, and the agent. During the experiment, the WOZ system recorded the logs of the utterances when the speaker is the agent. The microphone array of a Kinect sensor was used to identify the speaker via the direction of sound source. The mechanism used was so simple that the identification was interfered by noise, and the results was then manually corrected. When the realtime system is developed, this aspect should be improved with a more reliable method.

Addressee: Previous works (Huang et al., 2011; Baba et al., 2012) have shown that the addressee of a user utterance can be identified via non-verbal features at the accuracy as high as 80%. In this work, because of the requirement of correct feature values for training the classification model, the addressee of each utterance was manually labeled to the values, left user, right user, and the agent.

Face Direction: During the WOZ experiment, the face movement of each subject was logged with a WebCam and the face recognition software, FaceAPI[2]. Logged face movement information includes three dimensional positions and rotation as well as a confidence value. These raw values were then used to train a decision tree (J48 in Weka(Hall et al., 2009)) for automatically classifying face directions to front, partner, and otherwise.

5.2 Results and Discussion

Due to the fact that the evaluators in the timing annotation experiment (Section 4) might label the video clips at arbitrary positions, labels were explained as being associated with speaking state segments rather than their absolute positions. If there were multiple labels labeled by multiple evaluators at the same speaking state segments, only one was counted (i.e. multiple labels were not weighted). In addition to the four proposed intervening timing types, the class, "False" that means "not a timing" was added to the targets of automatic classification. Because there was a variety of instance numbers across timing types, the number of each timing types was balanced in the training dataset.

The classification results using random-forest algorithm are shown in Table 1. The overall accuracy was 58% (F measure: 0.53). The result was only moderate but was higher than the chance level of five-class classification (20%).

Table 1: Results of automatic classification of intervening timing types. "N" denotes the number of data instances.

Types	N	Precision	Recall	F measure
P-T	60	0.64	0.23	0.34
M-I	60	0.55	0.52	0.53
R-S	60	0.66	0.68	0.66
D-S	60	0.60	0.57	0.58
False	60	0.43	0.75	0.55
Overall	300	0.58	0.55	0.53

6 CONCLUSION AND FUTURE WORKS

This paper presents the results of the first step of an ongoing project that aims to build an information providing agent for collaborative decision making tasks, finding the timings for the agent to intervene user-user conversation to provide active support. In order to realize this, at first, a WOZ experiment was conducted for collecting human interaction data. From the collected corpus, four kinds of timings were found prob-

[2]http://www.faceapi.com/download/register.php

ably allowing the agent to do intervention in the target task. The definitions of the timings were validated with Web based questionnaires. Second, an automatic timing identifying method was developed to identify four of the timings only by using nonverbal cues including face direction and speaking state. The performance of the method is moderate (F-measure 0.53).

The current automatic estimation method is still relatively ad hoc due to the small corpus, we would like to increase the corpus size and would like to explore machine learning techniques to improve its performance. The performance should be able to be improved with additional context information. We would like to introduce the mechanism of context management and understanding in the future, this should help to estimate the four remaining timings as well. Finally, we would like to incorporate the intervention timing estimation feature into an ECA system and test it in a real-world application.

ACKNOWLEDGEMENTS

This work is partially funded by JSPS under a Grant-in-Aid for Scientific Research (B) (24300039) and (25280076).

REFERENCES

Argyle, M. and Cook, M. (1976). *Gaze and Mutual Gaze*. Cambridge University Press.

Baba, N., Huang, H.-H., and Nakano, Y. (2012). Addressee identification for human-human-agent multiparty conversations in different proxemics. In *4th Workshop on Eye Gaze in Intelligent Human Machine Interaction: Eye Gaze and Multimodality,14th International Conference on Multimodal Interaction (ICMI 2012)*.

Bohus, D. and Horvitz, E. (2010). Facilitating multiparty dialog with gaze, gesture, and speech. In *In International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*.

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

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Psychology*, 23(2):283–292.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: An update. *ACM SIGKDD Explorations*, 11(1):11–18.

Huang, H.-H., Baba, N., and Nakano, Y. (2011). Making virtual conversational agent aware of the addressee of users' utterances in multi-user conversation from nonverbal information. In *13th International Conference on Multimodal Interaction (ICMI'11)*, pages 401–408.

Jonsdottir, G. R. and Thorisson, K. (2009). Teaching computers to conduct spoken interviews: Breaking the realtime barrier with learning. In Ruttkay, Z., Kipp, M., Nijholt, A., Hogni, H., and Vilhjalmsson, editors, *9th International Conference on Intelligent Virtual Agents (IVA'09)*, volume 5773/2009 of *LNCS*, pages 446–459, Amsterdam, Netherlands. Springer Berlin.

Kendon, A. (1967). some functions of gaze direction in social interaction. *Acta Psychologica*, 26:22–63.

Kopp, S., Gesellensetter, L., Kramer, N. C., and Wachsmuth, I. (2005). A conversational agent as museum guide - design and evaluation of a real-world application. In *Proceedings of the 5th International Conference on Intelligent Virtual Agents (IVA'05)*, Kos, Greece.

Nakano, M., Dohsaka, K., Miyazaki, N., ichi Hirasawa, J., Tamoto, M., Kawamori, M., Sugiyama, A., and Kawabata, T. (1999). Handling rich turn-taking in spoken dialogue systems. In *European Conference on Speech Communication and Technology (EUROSPEECH'99)*.

Renals, S., Hain, T., and Bourlard, H. (2007). Recognition and understanding of meetings the ami and amida projects. In *IEEE Workhshop on Automatic Speech Recognition and Understanding (ASRU'07)*.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). a simplest systematics for the organization of turn-taking for conversation. *language*, 50(4):696–735.

Subramanian, R., Staiano, J., Kalimeri, K., Sebe, N., and Pianesi, F. (2010). Putting the pieces together: Multimodal analysis of social attention in meetings. In *Proceedings of the International Conference on Multimedia*, pages 659–662.

Takemae, Y., Otsuka, K., and Mukawa, N. (2003). Video cut editing rule based on participants' gaze in multiparty conversation. In *11th ACM International Conference on Multimedia*.

Traum, D. (2003). Issues in multiparty dialogues. In *Advances in Agent Communication, International Workshop on Agent Communication Languages (ACL'03)*, pages 201–211.

Traum, D., Aggarwal, P., Artstein, R., Foutz, S., Gerten, J., Katsamanis, A., Leuski, A., Noren, D., and Swartout, W. (2012). Ada and grace: Direct interaction with museum visitors. In *12th International Conference on Intelligent Virtual Agents (IVA 2012)*, pages 245–251.

Waibel, A., Bett, M., Finke, M., and Stiefelhagen, R. (1998). Meeting browser: Tracking and summarizing meetings. In *DARPA Broadcast News Transcription and Understanding Workshop*, pages 281–286.

Waibel, A., Bett, M., Metze, F., Ries, K., Schaaf, T., Schultz, T., Soltau, H., Yu, H., and Zechner, K. (2001). Advances in automatic meeting record creation and access. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing 2001 (ICASSP'01)*, Seattle, USA.