A multi-party attentive listening robot which stimulates involvement from side participants

Koji Inoue, Hiromi Sakamoto, Kenta Yamamoto, Divesh Lala, and Tatsuya Kawahara
Graduate School of Informatics, Kyoto University, Japan
[inoue,sakamoto,yamamoto,lala,kawahara]
@sap.ist.i.kyoto-u.ac.jp

Abstract

We demonstrate the moderating abilities of a multi-party attentive listening robot system when multiple people are speaking in turns. Our conventional one-on-one attentive listening system generates listener responses such as backchannels, repeats, elaborating questions, and assessments. In this paper, additional robot responses that stimulate a listening user (side participant) to become more involved in the dialogue are proposed. The additional responses elicit assessments and questions from the side participant, making the dialogue more empathetic and lively.

1 Introduction

One of the expected dialogue tasks for spoken dialogue systems is *attentive listening*, which is when an automated system carefully listens to the user and then generates a response. This task has been found to be useful for elderly people living alone who desire social interaction. We have so far developed an attentive listening dialogue system using an autonomous android ERICA (Inoue et al., 2020) that is capable of generating listener responses such as backchannels (e.g., "*Yeah*"), repeats of focus words, elaborating questions, and assessments (e.g., "*That is nice*").

Although the previous system was designed for one-on-one dialogue, in this demonstration, the system is extended to the multi-party scenario, which has previously been considered in other applications such as quiz games (Klotz et al., 2011), meetings (Fernández et al., 2008), and discussions (Skantze et al., 2015; Matsuyama et al., 2015). In our situation, the system attentively acts as the moderator that listens to dialogue from multiple people in turn, as shown in Figure 1. This *group attentive listening* scenario has been found to be relatively common in elderly care facilities. In this scenario, the behaviors of the main speaker and

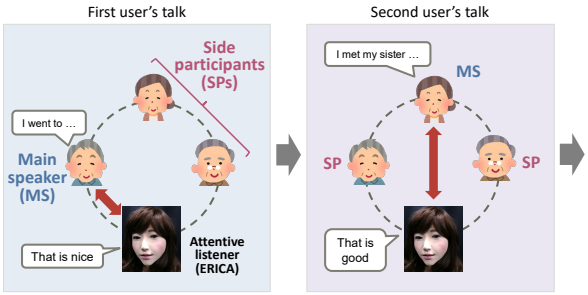

Figure 1: Scenario of multi-party attentive listening (group attentive listening)

the main listener (system) are important, and the involvement of other listeners (side participants) is also important to make the dialogue more lively. As shown in Figure 1, the side participants are people who can participate in the dialogue but are not being addressed by the current speaker (Goffman, 1981). In this scenario, the side participants can either be silent while the main speaker talks or can express their reactions towards the main speaker. In the latter case, it is expected that the main speaker will feel that he/she is listened to and understood more and also feel empathy from others. Therefore, in multi-party attentive listening, the system needs to act as a moderator to involve the side participants in the dialogue.

To promote the involvement of the side participants, this paper proposes a new type of attentive listening system utterances called *involvement-stimulating utterances*. Specifically, when the system is ready to give an assessment such as "*That is nice*" towards the current speaker, it can now also says "*That is nice, isn't it?*" aimed at one of the side participants. It is then expected that the target side participant would give an assessment and be involved in the dialogue. With more persons involved in the dialogue, the overall dialogue session is more activated and fruitful.

Another advantage of this new type of utterance is that the system can elicit human assistance when

Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 261–264

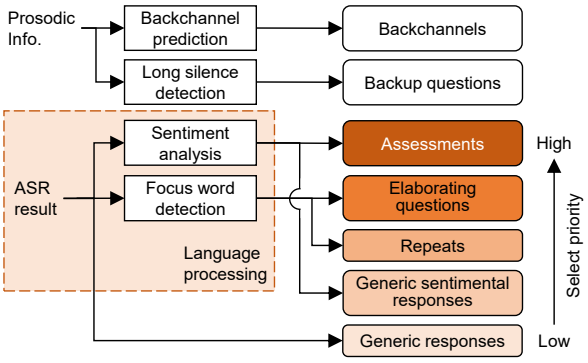

Figure 2: Diagram for the listener response generation

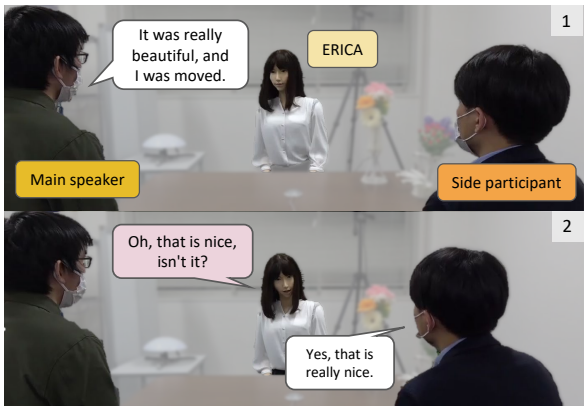

Figure 3: Proposed involvement-stimulating utterance

it is difficult for it to generate a proper assessment utterance by itself. Therefore, it can be said that this paper demonstrates cooperation between the system and users in the multi-party dialogue.

2 Multi-party attentive listening system

First, the basic attentive listening system (Inoue et al., 2020) used in this study is explained. As illustrated in Figure 2, the system generates listener responses such as backchannels, assessments, elaborating questions, repeats, and generic responses, with the speech enhancement and automatic speech recognition implemented through a 16-channel microphone array. A smooth turn-taking function is also realized through a machine-learning-based turn-taking model.

This study extends the system to a multi-party scenario in which there is more than one user and each user tells a story to the group in turn. We made a dialogue flow for the system acting as both the moderator and the main listener. The system first designates the main speaker from the participants and begins to attentively listen to this speaker. When a fixed time period has passed, the system promotes the speaker to stop talking and asks a second participant to start talking. This process is applied to all participants in turn, and after all participants end their individual talks, the dialogue finishes.

3 Eliciting assessments from side participants

In the previous one-on-one attentive listening system, the assessment responses such as "*That is nice*" had been generated on the basis of sentiment analysis (positive, negative, or neutral) using sentiment word dictionaries (Inoue et al., 2020). The assessment responses have been used to express empathy towards the speaker, which is an important role in attentive listening tasks.

In this work, during each of the user talks, the system attempts to use the proposed involvement-stimulating utterance to elicit an assessment from the side participants and then to make the dialogue more empathetic. As illustrated in Figure 3, when the proposed system stimulates involvement from the side participant, the assessment response is now "*Oh, that is nice, isn't it?*" as the involvement-stimulating utterance. As the system is designed to work with robots such as android ERICA, the eye-gaze (head) direction can be controlled to shift it to the target side participant to indicate the addressee for the involvement-stimulating utterance. If the system does not use an involvement-stimulating utterance, it only utters a conventional assessment such as "*That is nice*".

In the following, a dialogue example is given (with the original Japanese sentences), where S is the system; U1 is a user who is the main speaker; U2 is another user who is the side participant. The bolded parts identify the type of listener response and the underline marks the involvement-stimulating utterance.

U1: Last year, I went to a park in Kyoto.
(去年、京都の公園へ行きました。)
S: I see. **(generic response)**
(そうなんですね。)
U1: There is a famous cherry blossom.
(そこには、有名な桜があります。)
S: A famous cherry blossom. **(repeat)**
(有名な桜ですか。)
U1: It was nice timing to see the cherry blossom.
(ちょうど、その桜が見ごろでした。)
S: That is nice. **(normal assessment)**
(いいですね。)
U1: It was really beautiful and I was moved.
(本当に綺麗で感動しました。)

Table 1: Averages scores for the video subjective evaluation of the involvement-stimulating utterances (ISU) and the results of a t-test (7-point scale from 1 to 7)

	Evaluation item	w/ ISU	w/o ISU	p-value
(Q1)	The behavior of the robot was natural.	3.81	3.38	.007**
(Q2)	The robot was attentive to the side participant.	4.67	1.86	<.001**
(Q3)	The main speaker seemed to speak easily.	4.16	3.90	.051
(Q4)	The side participant seemed to participate easily.	4.08	1.89	<.001**
(Q5)	The whole conversation was lively.	3.92	3.08	<.001**

(** $p < .01$)

S: Oh, that is nice, isn't it?
(involvement-stimulating utterance)
(へー、いいですよね。)
U2: Yeah, that is really nice.
(うん、本当にいいですよね。)
U1: Yes, I stayed there for more than one hour.
(そうなんです、そこに 1 時間以上滞在しました。)

To realize this dialogue, the system needs to decide whether to use the involvement-stimulating utterance or a normal assessment utterance when detecting the positive sentiment. Using these two utterances properly is important because if the system uses the involvement-stimulating utterances all the time, the speaker would feel that his/her talk is being frequently interrupted and may become annoyed. Note that negative sentiment is not considered in the current system as it is thought that negative reactions should not be shared with the side participants.

3.1 Fine-grained sentiment detection

To ensure that the system properly employs the involvement-stimulating utterances, a fine-grained sentiment detector that can identify both *explicit* and *implicit* positive sentiment levels is built. The *explicit* sentiment means that there are emotional expressions such as "*moved*" in the aforementioned example sentence – "*It was really beautiful and I was moved*". The *implicit* sentiment means that there are no emotional expressions but it represents a positive emotion such as "*It was nice timing to see the cherry blossom*", which requires a higher level of inference to interpret. In this demonstration, if the system detects the explicit positive sentiment, it utters the involvement-stimulating utterances because explicit positive sentiments can be more shared with other people.

These fine-grained positive labels were manually annotated on a human-robot dialogue corpus when android ERICA was being teleoperated by a human operator and talking with a human subject in an attentive listening scenario. The dataset contained 120 5-to-8-min Japanese dialogue sessions. The sentiments in the subjects' long utterance units (Den et al., 2010) were labeled as explicitly positive, implicitly positive, or neutral. At first, to confirm the label agreements between the annotators, two annotators conducted this process in parallel over four dialogue data sessions, with the agreement score (Kappa coefficient) being measured at $\kappa = 0.788$ which indicated high agreement. Then, only one person annotated the rest of the dialogue data. The numbers of final samples for explicitly positive, implicitly positive, and neutral utterances were 390 (9.8%), 821 (20.6%), and 2,779 (69.6%), respectively.

A three-class classification model was trained using a pre-trained model BERT[1]. To evaluate the model accuracy, a 5-fold cross-validation was conducted; the results from which were a macro F-score of 66.9% and explicitly positive, implicitly positive, and neutral F-scores of 71.7%, 43.8%, and 85.1%, respectively. As expected, it was difficult to correctly detect the implicitly positive utterances because there were no emotional expressions on the surface level of utterances, therefore, it is planned to increase the amount of training data and use other sentiment label datasets as additional pre-training. In this demonstration, the BERT-based sentiment detector is used to determine the timing for the use of the involvement-stimulating utterances, corresponding to the detected sentiment label: explicit or implicit.

3.2 Subjective evaluation

A video-based subjective evaluation was conducted to confirm the effectiveness of the involvement-stimulating utterances. Using the proposed multi-party attentive listening system with android ER-

[1] https://github.com/cl-tohoku/bert-japanese

ICA, several multi-party dialogue videos were recorded with the viewpoint being as shown in Figure 3. Videos that did not use the involvement-stimulating utterances were also recorded as baseline to compare with the existing attentive listening system. We manually scripted six different scenarios and ask people from the authors' laboratory to play the role in the scenarios. Therefore, we used 12 videos (2 systems $\times$ 6 different scenarios) for this evaluation.

After the videos were recorded, other evaluators (20 university students) were asked to watch each video and then give scores based on the item listed in Table 1. It was generally felt that the robot behavior in the involvement-stimulating utterances (w/ ISU) was more natural (Q1), the robot was more attentive to the side participant (Q2), the side participants seemed to participate more easily in the dialogue (Q4), and the whole conversation was more lively (Q5). Note that no significant difference for Q3 was found, which indicated that the proposed the involvement-stimulating utterances had not interfered with the main speaker's talk. Therefore, the effectiveness of the proposed the involvement-stimulating utterances in the multi-party attentive listening scenario was confirmed.

4 Eliciting questions from side participants

Another type of involvement-stimulating utterance has been implemented using *focus words* that were originally used for repeats and elaborating questions in the attentive listening system. During the dialogue, the system detects and stores the focus words of user utterances, and when the main speaker is silent for a longer period (e.g. 5 seconds), the system requests the side participant to ask a question using the focus words.

A dialogue example is given in the following, in which S is the system, U1 is the main speaker, and U2 is the side participant. The bolded parts identify the type of listener response and the underline marks the involvement-stimulating utterance and also the focus word.

U1: Last year, I went to Kyoto.
(去年、京都へ行きました。)
S: Kyoto. **(repeat)**
(京都ですか。)
(U1 talks for a wile and then be silence)
S: It was about Kyoto.
Do you have any question?
(involvement-stimulating utterance)
(京都のお話がありましたが、何か質問はありますか。)
U2: Well, where did you go else in Kyoto?
(京都では他にどこへ行きましたか？)
U1: I also went to a famous temple.
(有名なお寺へ行きました。)

Instead of asking a question without the focus words such as "*Do you have a question?*", specifying the focus words related to the context makes it easier for the side participant to come up with a proper question. This type of involvement-stimulating utterance is also demonstrated in the multi-party attentive listening scenario.

5 Conclusions

This paper demonstrated a multi-party attentive listening system that generates involvement-stimulating utterances to better involve side participants and express listener responses, which made the dialogue livelier and more empathetic. Future research will be focused on conducting a dialogue experiment to confirm the effectiveness of the proposed system with real users.

Acknowledgments

This work was supported by JSPS KAKENHI Grant numbers (JP19H05691，JP20K19821).

References

Yasuharu Den et al. 2010. Two-level annotation of utterance-units in Japanese dialogs: An empirically emerged scheme. In *LREC*.

Raquel Fernández et al. 2008. Modelling and detecting decisions in multi-party dialogue. In *SIGdial*, pages 156–163.

Erving Goffman. 1981. *Forms of talk*. University of Pennsylvania Press.

Koji Inoue et al. 2020. An attentive listening system with android ERICA: Comparison of autonomous and WOZ interactions. In *SIGdial*, pages 118–127.

David Klotz et al. 2011. Engagement-based multi-party dialog with a humanoid robot. In *SIGdial*, pages 341–343.

Yoichi Matsuyama et al. 2015. Four-participant group conversation: A facilitation robot controlling engagement density as the fourth participant. *Computer Speech & Language*, 33(1):1–24.

Gabriel Skantze et al. 2015. Exploring turn-taking cues in multi-party human-robot discussions about objects. In *ICMI*, pages 67–74.