Investigating Effects of Multimodal Topic-continuance Recognition on Human-Robot Interviewing Interaction

Fuminori Nagasawa
s2040011@jaist.ac.jp
Japan Advanced Institute of Science and Technology
Nomi Shi, Japan

Shogo Okada
okada-s@jaist.ac.jp
Japan Advanced Institute of Science and Technology
Nomi Shi, Japan

ABSTRACT

This study's long-term goal is the development of a communication robot as a partner that can keep talking about specific things about which the user would like to talk and in which they are interested. To achieve this goal, we developed an interviewer robot that adapts topics based on the user's multimodal attitudes. The robot, utilizing the Japanese GPT-NeoX-3.6, selects questions based on the estimated topic continuance level. We regard the topic continuance level as the degree of the user's speaking willingness (willingness to continue the current topic). This paper aims to validate the multimodal topic continuance recognition model and its adaptive question selection strategy. First, we trained the model on the "Hazumi" dialog corpus, which includes user multimodal behavior in human-virtual agent interactions. Second, 10 participants were interviewed with the robot equipped with the trained model. After the interviews, we asked the participants if the topic continuance/change by the robot was appropriate and validated the estimation accuracy.

CCS CONCEPTS

• **Human-centered computing** → *Natural language interfaces*; **Human computer interaction (HCI)**.

KEYWORDS

Human robot interaction, Interview agent, Multimodal machine learning, Social signal processing, Speaker's willingness

ACM Reference Format:
Fuminori Nagasawa and Shogo Okada. 2024. Investigating Effects of Multimodal Topic-continuance Recognition on Human-Robot Interviewing Interaction. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24 Companion), March 11–14, 2024, Boulder, CO, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3610978.3640662

1 INTRODUCTION

The goal of this research is to create a dialog robot that actively encourages users to talk about their own feelings and experiences.

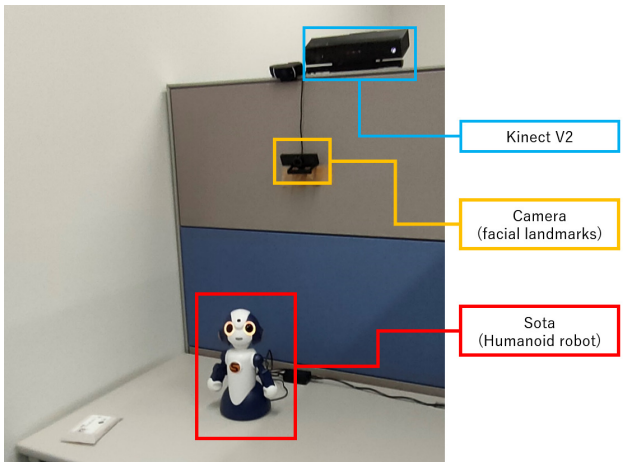

Figure 1: Interview robot system

Techniques for dialog robots to elicit detailed and personal narratives from users can be applied to motivational interviewing, life logging, and documentary production interviews. Thus, this paper presents an interview robot system with social signal sensing and adaptive interviewing strategies. The centerpiece of this system is the ability to tailor interview questions based on interviewees' willingness to talk, as inferred from social signals. These in-depth interviews differ from the standard question-and-answer format and elicit deeper information, such as personal memories and emotions, through dialog. [1]

The primary goal of an interview is to gather information through well-crafted questions [1]. Effective interviewers must interpret emotional and social cues from interviewees to foster their engagement. A key technique is to deepen the discussion, prompting spontaneous self-disclosure. Following up on topics can enhance interviewees' willingness to share, but inappropriate topics might reduce this willingness. Therefore, assessing interviewees' readiness to talk is vital for successful in-depth interviewing. Building on these principles, we have developed a robot that dynamically adjusts its interviewing strategy based on the speaker's willingness in our research. Previous work [4] in this domain defined 'willingness' as an interviewee's desire to speak, using a robot system that estimated this from acoustic and postural signals. The system used a predefined question graph to direct the conversation. Experiments showed that adapting questions based on this graph increased responses with high willingness, but the limited range of questions sometimes led to inadequate topic follow-up . Additionally, reliance

HRI '24 Companion, March 11–14, 2024, Boulder, CO, USA

ACM ISBN 979-8-4007-0323-2/24/03
https://doi.org/10.1145/3610978.3640662

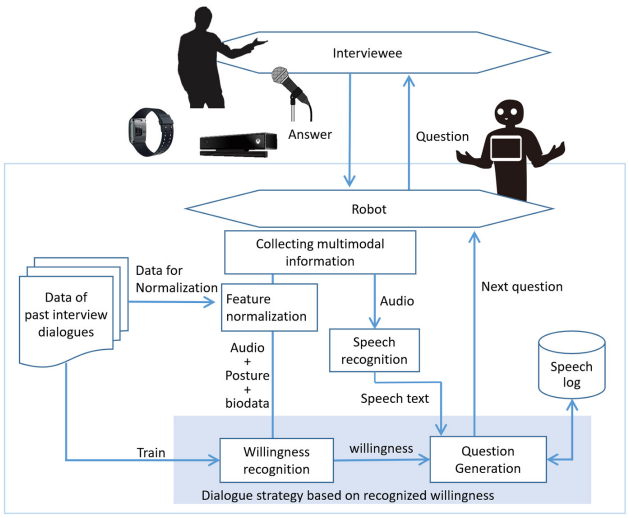

Figure 2: Overview of interview robot system

on vocal and postural features alone often resulted in poor estimation of willingness. In this study, we propose real-time generation of questions to achieve dialog without the constraints of question graphs and to improve the accuracy of willingness recognition by expanding the number of features used.

Inoue et al. [2] proposed generating deeper questions by analyzing words from interviewees' responses. Our study builds upon these findings by generating contextually relevant questions using GPT and recorded dialogs. Furthermore, Komatani et al. [3] compiled the "Hazumi" multimodal dialog corpus including human interactions with a virtual agent. While some studies, including that by Katada et al. [3], have focused on developing multimodal machine learning models to estimate topic continuance and user sentiment using the Hazumi corpus, its application in real-world robotic scenarios and the validation of the trained model through integration with conversation robots remain unexplored. Our research contributes to this field by developing a robot with a sentiment estimation model based on the "Hazumi" corpus. We collected interview data in an open environment to assess the model's real-world effectiveness.

2 INTERVIEW ROBOT SYSTEM BASED ON SSP

An overview of the proposed interview robot system equipped with a social signal (speaker's willingness level) recognition module is shown in Figure 2. The proposed interview robot aims to elicit information from the interviewee through an adaptive question selection strategy. The system consists of three modules: a humanoid conversational robot, a multimodal sensing environment, and an adaptive question generation module.

2.1 Humanoid conversational robot

The humanoid personal robot Sota was used as an interview robot to interact with the interviewees. Sota, developed by VStone, is a desktop robot with a height of 280 mm and a weight of approximately 760 g. It is equipped with voice synthesis and modules for generating hand and head movements. The content spoken by Sota is controlled by backend modules (e.g., multimodal sensing environment and adaptive question generation).

2.2 Multimodal sensing environment

During the interview dialogs, a web camera (Logicool C910, 1080p 30fps), MS Kinect V2, a wearable microphone (Shure PGA31 headset mic), and a wristband-type biometric sensor (Empatica E4) were used to collect multimodal data. The interviewees sat in front of the robot, with the web camera and Kinect sensor positioned approximately 50 cm behind and above the robot's head. The Empatica E4 was worn on the participants' left wrists.

The Kinect sensor estimated the joint coordinates of the participants, the wearable microphone captured their voices during the speech, and the E4 device measured heart rate and skin conductance during the dialog. Additionally, the web camera captured the facial landmark features of the participants. Prosodic features were extracted from the voice using OpenSMILE, and facial landmarks were extracted using dlib. These features were used to train a model for recognizing the willingness of the interviewees.

Furthermore, the voice data collected by the wearable microphone were transcribed using Whisper for speech recognition and used as input for the question generation module. The multimodal features from voice and vision and the model for recognizing willingness are detailed in Section 2.3.

In our study, we performed a binary classification task to discern interviewees' willingness, aiding in two-way question selection (topic continuation or switching). We converted the original 8-level Hazumi labels to a binary scale: scores of 5 and above were categorized as 'high' willingness, and scores below 5 were considered 'low'.

2.3 Willingness recognition model

2.3.1 Data corpus. This research utilized the Hazumi1911 dataset[3], a multimodal corpus of human-agent dialog. It encompasses 2859 exchanges from 30 participants, interacting with an agent in a Wizard-of-Oz setup. The dataset incorporates diverse data types: posture (3D joint coordinates via MS Kinect), acoustic (prosodic features), facial landmarks, and biometric signals (heart rate, skin conductance). After the experiment, participants labeled each exchange with self-assigned sentiments and topic continuation preferences. Among the annotated labels, "topic continuance" was used in this study. This determines whether the system should continue the topic or change the topic, assuming that the participant has taken the system's position. Since the purpose of the system is to determine whether a topic should be continued based on willingness, in this study, we used the "topic continuance" label (willingness to continue the current topic) as the level of willingness.

2.3.2 Multimodal Machine Learning. The willingness recognition result is used to select the next question, so the model was trained to infer the willingness level per exchange in an online manner. The input data to the model are composed of multimodal behavioral features that are observed while the user is answering the current question. The model outputs the willingness level (high/low) corresponding to the input multimodal features. To determine whether the system changes the current topic in the next question, we set

the willingness recognition problem as a binary classification task of willingness level (high or low). The binary willingness recognition model is trained with the annotated willingness label and the multimodal behavioral features observed while the user is speaking.

2.4 Adaptive question generation

The system performs adaptive question generation based on the recognized willingness level. The system performs multimodal willingness estimation, and if the recognized result of the previous utterance (the answer to the previous question) is "high willingness," the system asks questions to follow up on the topic; and if the result is "low willingness," it asks questions to change the topic.

The system uses the GPT to generate questions. In this process, instructions for question generation, appropriate example questions based on user utterances, and a record of the dialog between the system and the interviewee are entered into the GPT model as prompts. To switch the questioning mode, the system adaptively changes the appropriate question examples within the prompts. To cater to both following up on the topic and changing the topic, specific question examples are prepared for each scenario, and these examples are adaptively switched with each question generation. The GPT model used in this study is Japanese GPT-NeoX-3.6b[5], which is capable of generating Japanese text quickly in an offline environment.

3 EXPERIMENTAL SETTINGS

3.1 Evaluation of the willingness recognition model

To validate the accuracy of willingness recognition, we trained a random forest model similar to the method used in a previous study[5] and evaluated the learned models as follows. Leave-one-person-out cross-validation (LOPOCV) was used to evaluate the learned intention recognition model; in LOPOCV, the test data correspond to the sample observed in the interview session of one interviewee and the remaining sample of interviewees was used as training data.

To evaluate the change in accuracy when this system is implemented in a robot, we conducted an accuracy evaluation using an interview dialog corpus collected through dialog experiments. The model was trained using the topic continuance label assigned in the Hazumi1911 dataset as the objective variable. We evaluated the estimation accuracy of the models trained on the dialog experiment corpus collected in this study and assessed changes in estimation accuracy with the data corpus.

3.2 Effects of Adaptive Question Selection on Sentiment

To evaluate adaptive question selection in dialog and the impact of dialog breakdowns, we conducted a dialog experiment. The participants engaged in an interview dialog with the system to reflect on recent events in their daily activities. Each interviewee engaged in a 5- to 10-minute dialog with the system, and the system interacted with the subject while dynamically switching between topic continuance and topic switching.

3.2.1 Participants. Ten participants, aged 23 to 63 (average age 44), were recruited through a Japanese recruitment agency, represented a broad demographic of the Japanese population, and received a fixed payment for their participation. Prior to the experiment, the participants were informed of their right to withdraw at any time. The experiment ensured minimal physical or mental strain, and all data, including the recorded videos, were securely managed. The study and its use of the data were approved by the Research Ethics Committee of the authors' institution

3.2.2 Experimental design and procedure. To evaluate the impact of adaptive question generation on the interviewees, each participant was interviewed. They were asked beforehand to describe key words that would be used as conversation starters. After the system first made introductory remarks to the participants, the dialog began with the following question: "What has happened to (topic) recently?".

3.2.3 Measures. After the dialog experiment, the subjects were asked to answer a questionnaire and to perform a self-annotation of their own response utterances during the dialog. In the self-annotation, annotations were given by the same questions as in Hazumi1911 for each exchange in which a question by the system was paired with an answer by the subject. In addition, to evaluate the impact of dialog breakdowns, we annotated the dialog content to determine whether a breakdown had occurred. For each dialog exchange, we determined whether the robot's statements were unnatural with respect to the dialog content.

3.2.4 Analysis. The aim of this analysis was to assess the impact of a sentiment estimation model trained on the Hazumi1911 dataset and adaptive question generation by GPT on real-world conversational robots. This involved two key investigations.

First, we evaluated the accuracy of the motivation estimation model. This was done by estimating annotated topic continuance level labels in both the Hazumi1911 dataset and the experimental corpus of interview dialogs and comparing them through cross-validation. Second, we examined the effects of adaptive quality generation and dialog interruption. This involved tracking changes in estimated and annotated speech willingness, focusing on the impact of system behavior due to dialog breakdown and estimated willingness on the time-series transition of annotated willingness.

4 RESULTS

4.1 Accuracy of sentiment estimation

To examine differences in estimation accuracy by corpus, we compared the accuracy of models trained under various conditions (corpus, features). The classification accuracy of the intention estimation models is shown in Table 1.

In-corpus(P+A) is the case in which the same feature set (posture + acoustic) as in the previous study [4] is used; the Hazumi1911 dataset and In-corpus(P+A+B+F) are the cases in which the feature set (posture + acoustic + biometric signals + facial features) is used in our system. In-corpus is the accuracy of the cross-validation, and Out-corpus is the accuracy of the dataset obtained in the dialog experiment using the model trained under the conditions of In-corpus(P+A+B+F).

Table 1: Accuracy of the willingness recognition model

Corpus	accuracy
In-corpus(P+A)	0.718
In-corpus(P+A+B+F)	0.724
Out-corpus(Our interview corpus)	0.636

The results in the table show that In-corpus(P+A+B+F) has higher accuracy than In-corpus(P+A) in the In-corpus case. This result indicates that it is possible to train a model that can estimate willingness more accurately by using the features used in the Hazumi1911 dataset than by using the features used in previous studies[4]. In contrast, in the Out-corpus case, the accuracy was lower than that in the in-corpus case, suggesting that the difference between the sensor setup used to collect the Hazumi dataset and the sensor setup used in this dialog experiment was not fully absorbed by the feature normalization used to train the model. However, since we were able to estimate with an accuracy higher than the chance level of 50%, we expect that if we can estimate willingness with higher accuracy in the in-corpus, we will be able to estimate willingness with higher accuracy in the real environment as well.

4.2 Effects of Adaptive Question Generation and dialog breakdown

We investigated the impact of adaptive question generation on the interviewees' willingness. Figure 3 presents several examples of dialog. In the graph, the horizontal axis represents the time elapsed since the start of the dialog, while the vertical axis indicates the High/Low status of willingness. The estimated willingness results are marked with white circles, and the actual evaluations of willingness based on self-annotation are indicated with crosses. Points of dialog breakdown are shown with triangles. The results of the graph revealed several trends. When the estimated willingness was initially Low, the actual willingness often also transitioned to Low. Furthermore, if the estimated willingness is continuously Low, it commonly results in the actual willingness eventually becoming Low. Even if the actual willingness later became High, a persistent Low estimate tended to lead to a quick decrease in willingness. Additionally, even when initial willingness was high, it could decrease due to dialog breakdowns. Once a dialog breakdown occurred, continuous breakdowns tended to ensue.

The results of the graph revealed several trends. It was observed that when the estimated willingness was initially Low, the actual willingness often transitioned to Low as well. Furthermore, if the estimated willingness was continuously Low, it commonly resulted in the actual willingness eventually becoming Low. Even if the actual willingness later became High, a persistent Low estimate tended to lead to a quick decrease in willingness. Additionally, it was noted that even when the initial willingness was high, it could decrease due to dialogue breakdowns. Once a dialogue breakdown occurred, there was a tendency for continuous breakdowns to ensue.

5 DISCUSSION AND FUTURE WORKS

In a previous study in the same domain [4], topics were deepened or changed by adaptively selecting questions from a prelisted list.

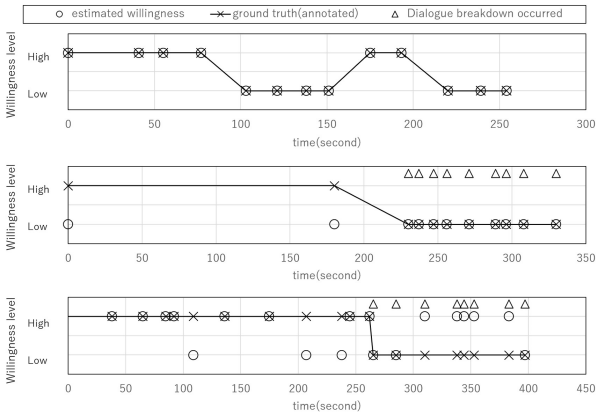

Figure 3: Timeline of willingness and Dialogue Breakdown

However, the number of questions that could be listed in advance was limited, and even when interviewees were highly motivated, topics were sometimes changed due to exhaustion of questions. In this study, question exhaustion was avoided by generating questions in real time. This may have resulted in more highly motivated interactions. In future work, we plan to evaluate whether adaptive question selection leads interviewees to give more personal responses by assessing changes in the degree of self-disclosure with question selection in a more interactive experiment with a larger number of participants. We also plan to investigate the effects of speech breakdown on self-disclosure and impressions.

6 CONCLUSION

In this initial study, which was aimed at developing an interview dialog robot that proactively encourages users to express their feelings and experiences, we explored the effects of integrating adaptive question generation with a real-world interview dialog robot system. Initially, we assessed the performance of a speech willingness estimation model trained on a precollected corpus within an interview dialog corpus involving an actual robot. We first evaluated a multimodal willingness recognition model using the Hazumi1911 dialog corpus. By conducting a dialog experiment with this model to generate an interview dialog corpus and assessing the model's accuracy within this corpus, we found that the willingness recognition model could estimate willingness with 63.6% accuracy. However, its accuracy decreased in a real-world setting. Furthermore, we examined the impact of adaptive question generation and dialog breakdowns on interviewees' willingness. Analysis of the shifts between estimated and annotated willingness during the interview dialog experiment highlighted several key patterns: a decrease in willingness was observed when the willingness recognition model was consistently misjudged as low willingness, and dialog breakdowns also led to reduced willingness.

ACKNOWLEDGMENTS

This work was also partially supported by JSPS KAKENHI (22K21304, 22H04860, 22H00536, 23H03506), JST AIP Trilateral AI Research, Japan (JPMJCR20G6) and JST Moonshot R&D program (JPMJMS2237)

REFERENCES

[1] Ben Emans. 2016. *Interviewing: Theory, techniques and training*. Routledge, London.

[2] Koji Inoue, Kohei Hara, Divesh Lala, Kenta Yamamoto, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. 2020. Job Interviewer Android with Elaborate Follow-up Question Generation. In *Proceedings of the 2020 International Conference on Multimodal Interaction* (Virtual Event, Netherlands) *(ICMI '20)*. Association for Computing Machinery, New York, NY, USA, 324–332. https://doi.org/10.1145/3382507.3418839

[3] Kazunori Komatani and Shogo Okada. 2021. Multimodal Human-Agent Dialogue Corpus with Annotations at Utterance and Dialogue Levels. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 1–8. https://doi.org/10.1109/ACII52823.2021.9597447

[4] Fuminori Nagasawa, Shogo Okada, Takuya Ishihara, and Katsumi Nitta. 2023. Adaptive Interview Strategy Based on Interviewees Speaking Willingness Recognition for Interview Robots. *IEEE Transactions on Affective Computing* (2023), 1–17. https://doi.org/10.1109/taffc.2023.3309640

[5] rinna. 2023. japanese-gpt-neox-3.6b. https://huggingface.co/rinna/japanese-gpt-neox-3.6b. Accessed: 2023-12-04.