Job Interviewer Android with Elaborate Follow-up Question Generation

Koji Inoue
Kyoto University
Kyoto, Japan
inoue@sap.ist.i.kyoto-u.ac.jp

Kohei Hara
Kyoto University
Kyoto, Japan
hara@sap.ist.i.kyoto-u.ac.jp

Divesh Lala
Kyoto University
Kyoto, Japan
lala@sap.ist.i.kyoto-u.ac.jp

Kenta Yamamoto
Kyoto University
Kyoto, Japan
yamamoto@sap.ist.i.kyoto-u.ac.jp

Shizuka Nakamura
Kyoto University
Kyoto, Japan
shizuka@sap.ist.i.kyoto-u.ac.jp

Katsuya Takanashi
Kyoto University
Kyoto, Japan
takanashi@sap.ist.i.kyoto-u.ac.jp

Tatsuya Kawahara
Kyoto University
Kyoto, Japan
kawahara@i.kyoto-u.ac.jp

ABSTRACT

A job interview is a domain that takes advantage of an android robot's human-like appearance and behaviors. In this work, our goal is to implement a system in which an android plays the role of an interviewer so that users may practice for a real job interview. Our proposed system generates elaborate follow-up questions based on responses from the interviewee. We conducted an interactive experiment to compare the proposed system against a baseline system that asked only fixed-form questions. We found that this system was significantly better than the baseline system with respect to the impression of the interview and the quality of the questions, and that the presence of the android interviewer was enhanced by the follow-up questions. We also found a similar result when using a virtual agent interviewer, except that presence was not enhanced.

KEYWORDS

Job Interview System; Question Generation; Follow-up Question; Autonomous Android

ACM Reference Format:
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 (ICMI '20), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3382507.3418839

ICMI '20, October 25–29, 2020, Virtual event, Netherlands

ACM ISBN 978-1-4503-7581-8/20/10…$15.00
https://doi.org/10.1145/3382507.3418839

1 INTRODUCTION

Android robots have the potential to serve in social roles that humans currently perform. Their realistic appearance and expressions afford them a certain presence [6, 37] that may be lacking in disembodied or virtual agents. If the android also executes its role competently, then it may serve as an ideal interactive interface that can provide additional value.

In this work, we focus on one potential role for an android, that of a job interviewer. The android will ask questions to a job candidate (interviewee) and conduct the interview based on the answers to these questions. The android can conduct many interviews without fatigue, reduce bias due to factors such as gender, ethnicity, or age, and reproduce the same mannerisms towards every candidate. Additionally, an android can be used as a means to experience a job interview, a number of times to reduce the well-studied phenomenon of job interview anxiety [8, 11, 30, 33].

The use of artificial intelligence for job interviews has already been implemented for commercial use, with companies such as Hirevue[1] creating models which measure the behaviors of the candidate during the interview and provide an automatic evaluation. However, in such systems, there is no interviewer and a web camera is used for behavioral measurement. Recently, Furhat Robotics revealed their robot *Tengai*, which can conduct a structured job interview[2]. *Tengai* is not a full-bodied android but a robotic head with projected facial expressions. The motivation behind this robot is to conduct unbiased interviews so candidates can be judged fairly, so the questions are always the same.

On the other hand, we propose a robot that will be used by *candidates* to help with job interviews. An android can be physically situated in the same room and ask questions that are relevant to the answers given by the candidate. It can replicate human-like behaviors and speech which allow the user to immerse themselves in a situation. We will exploit this feature of realistic androids to simulate the experience of a real interview.

[1] https://www.hirevue.com/
[2] https://www.tengai-unbiased.com/

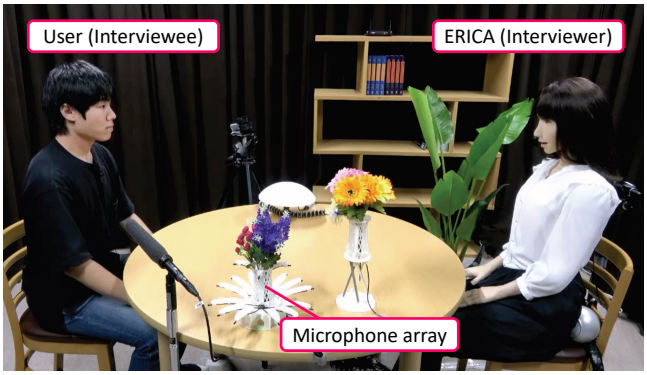

Figure 1: Job interview dialogue with android ERICA

Another issue is that these automatic systems mostly use a fixed set of questions during the interview. Since the objective is primarily to measure the behavior of the candidate, there are few natural language processing techniques used to generate follow-up questions related to the previous answer of the candidate [38, 39]. In this paper, we propose a system that can achieve this goal using the results of automatic speech recognition (ASR) and have the system generate meaningful follow-up questions for the candidate to further elaborate upon during the interview. In our system, follow-up questions are generated based on two viewpoints: quality of response and a keyword used in the response. It is expected that asking these follow-up questions makes the job interview close to a human-human job interview [29]. In this study, we implement a job interview system in a fully autonomous android robot and conduct a dialogue experiment with university students (Figure 1) to confirm the effectiveness of the follow-up questions.Our long-term goal is to implement a practice job interview system that can be used by candidates who wish to experience a job interview before having to undergo the real thing with a human interviewer.

2 RELATED WORK

Several commercial applications related to job interviews exist, such as Hirevue. These are generally targeted at companies looking to hire candidates by streamlining the interview process by measuring candidates' behaviors as they answer questions. The algorithms used for these measurements are naturally not made public, so they are difficult to compare against.

From a research perspective, there has been one large-scale project which used a virtual agent job interviewer to support training and coaching related to job interviews [3–5, 10, 12], which was followed by other related studies [2, 7, 14, 21, 31, 34, 36]. The system used in the project measured verbal and non-verbal behaviors of participants in a job interview to create and compare different types of virtual interviewers and determine user perceptions of the system for job interview training. It was reported in dialogue experiments that taking job interview training with the agents improved their interview skills more than self-learning such as reading textbooks and watching instruction videos [10, 27]. Some studies have been conducted on automatic evaluation of job interviewees by measuring their multi-modal behaviors including non-linguistic ones [31, 34]. Another work also concluded that presence in a job interview conducted in virtual reality was higher than in the real world [41]. One other recent work used an android to assess non-verbal behaviors during a job interview for users with autism [22], but the robot was tele-operated and not fully automated as in our work. In the research field of human-robot interaction, a small-sized robot NAO was used to play the role of an interviewer [1, 9].

The job interview questions used in the above studies and commercial applications are fixed before when the interviews has started since speech recognition is not used to follow up on what the subjects had said. Our system will use speech recognition as the main tool for changing the behavior of the interviewer and we will compare this against a fixed format of questions. Our proposed system generates follow-up questions based on responses of interviewees. A few studies have been made on follow-up question generation and but each module was evaluated in an offline manner [38, 39]. To our knowledge, a fully autonomous job interview system generating follow-up questions has not been made and also not been evaluated in an experiment with real users.

3 ANDROID ERICA

The android we use for this research is ERICA, who has been developed as an autonomous conversational robot [13, 17]. Her appearance is of a young Japanese woman. ERICA has a total of 46 motors in her face and body, which allows her to produce a variety of facial expressions and gestures that express her emotional state. ERICA's voice is a text-to-speech system trained on a real voice actress, which closely matches her physical appearance. She can express natural-sounding backchannel and filler utterances, which are commonly used in Japanese. Lip synchronization complements the utterances [18]. Non-verbal behaviors such as blinking, breathing, and nodding are also used by ERICA. She has been used for several research purposes, including analysis of backchannels [25], fillers [26, 32], and turn-taking [23, 24]. Currently, several social roles are considered for her, such as attentive listening [15] and as a lab guide [16, 20]. In this work, we extend her role to that of a job interviewer.

4 JOB INTERVIEW SYSTEM

The structure of the interview should not be completely fixed because we want the subjects to believe that ERICA is listening to the answers they provide and asking useful follow-up questions. This is different from other systems where the questions are largely fixed, no matter what answers are provided by the interviewee. To do this, we first define a basic structure of the interview. A diagram of the structure of the interview is shown in Figure 2. The flow of the interview is based on a topic. Each topic starts from a base question such as "What is the reason why you applied for this job?". Within each of the base questions, the system tries to generate follow-up questions depending on the responses of candidates. Two different types of follow-up question can be asked, which will be described below. Note that we made the dialogue content independent of any particular business or company, so questions from ERICA focus on the motivation and experience of interviewees. Therefore, this job interview system can be applied to interviewees of various backgrounds without modifying the contents of questions.

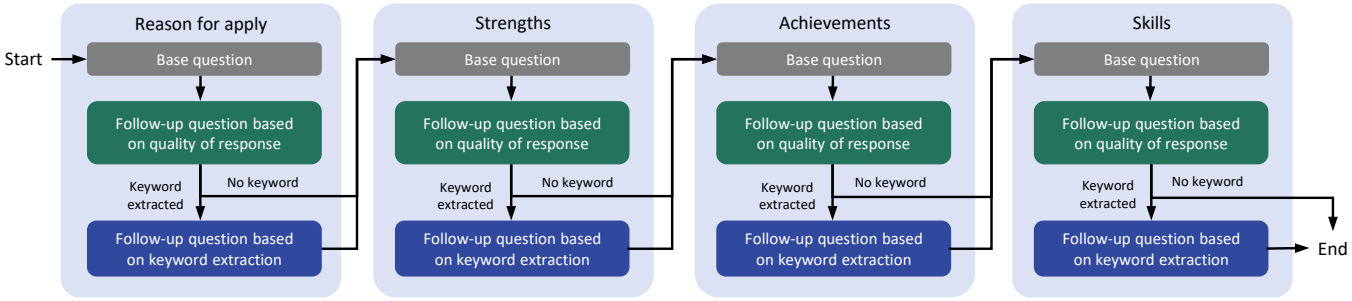

Figure 2: Dialogue flow of job interview system

Table 1: Checklist for each base question and statistics of annotation result in a human-human job interview dialogue corpus

Base question	Check item	#samples (positive/total)
(B1) What is the reason for applying?	(C1-1) Why did the candidate choose this company	35 / 63
	(C1-2) What the candidate can contribute to this company	20 / 63
	(C1-3) Suitability and strengths that can be used in the company	13 / 63
(B2) What are your strengths?	(C2-1) Which strengths can be applied in this company	8 / 31
	(C2-2) Particular examples or achievements (to confirm credibility)	15 / 31
(B3) What are your achievements?	(C3-1) Particular examples or achievements	19 / 29
(B4) What are your skills?	(C4-1) Which skills can be applied in this company	19 / 29
	(C4-2) Particular examples or achievements	23 / 29

In this study, we design two kinds of follow-up questions. The system first generates a follow-up question based on the quality of the candidate's response to the base question (section 4.1). Then, the system tries to extract a keyword from the response to the previous question in order to generate a keyword-based follow-up question (section 4.2). We hypothesize that generating these follow-up questions will make the candidates feel that the system listens and considers the candidates' responses. We will investigate the effectiveness of these follow-up questions in the later experiment.

4.1 Follow-up questions based on quality of responses

After a candidate responds to a base question, the system assesses the quality of the response to generate a follow-up question. For the assessment, we follow generic guidelines for job interviews, found in interview training manuals, and then design a checklist of the points of responses. The set of the checklist for each base question is summarized in Table 1. For example, for the base question of "reason for applying" (B1), we define 3 checklist items: (C1-1) *Why did the candidate choose this company*, (C1-2) *What the candidate can contribute to this company*, (C1-3) *Suitability and strengths that can be used in the company*. A good response to this base question will mention these items.

In order to realize an automatic assessment, we utilize a machine learning approach using dialogue data of human-human job interviews. We collected the human-human dialogue data of 14 sessions where university students played the role of candidates (interviewees) in mock job interviews. All the participants were native Japanese speakers. In advance, we gave time for the candidates to select the expected company they want to apply to and prepare some responses against prospective questions. The interviewer was ERICA who was controlled by a human operator. We gave a list of base questions to the operator and instructed them to select appropriate base questions from the list as well as sometimes giving follow-up questions. The interview lasted about 9 minutes on average.

With the collected dialogue data, we conducted a human annotation of the above checklist against all the responses uttered by the candidates. An annotator validated each response against a base question and gave a binary result on each check point. After this first annotation work, another annotator confirmed the annotation result resolved disagreements. Some annotation examples are given below. For example, the following response is annotated as mentioning *(C1-1) Why did the candidate choose this company.*

> The reason I want to work in your company is that **I sympathize with your company slogan. Your company values the personality of each person and its creativity to develop a wide range of products from home appliances to building equipment**. This is the reason why I was attracted to this company and I applied for this job.

The bold text indicates the statement marked by the annotator to make the judgement of this annotation. The following is another response annotated as mentioning *(C4-1) Which skills can be applied in this company.*

> ... I have many Chinese qualifications ... **If I was here, we would be able to communicate with clients**

Table 2: Follow-up questions on the base question of *reason for apply*. The symbols (✓and ×) represent if the item was mentioned or not mentioned in the preceding response.

Check items (C1-1)	(C1-2)	(C1-3)	Follow-up question
×	×	×	Well, I could not get any points from that response. Although there are some similar companies, why did you choose ours? (ask (C1-1))
✓	×	×	Well, I could understand why you choose our company from your answer. However, which part of our company do you think you can contribute to? (ask (C1-2))
×	✓	×	Well, from your answer, I could understand which part of our company you think you can contribute to. However, there are other companies where you can do a similar thing, so why did you choose our company? (ask (C1-1))
×	×	✓	Well, I understand your suitability for this company and strengths that can be utilized. However, there are some other companies which are similar, so why did you choose our company? (ask (C1-1))
✓	✓	×	I see. I understand why you choose this company and also which part of our company you think you can contribute to. Well, what are your own strengths that can be utilized in this company? (ask (C1-3))
✓	×	✓	I see. I understand why you choose this company and also your own strengths that can be utilized for this company. Well, which part of our company do you think you can contribute to? (ask (C1-2))
×	✓	✓	I see. I understand which part of our company you can contribute to and also your own strengths that can be utilized. Well, why did you choose this company in this industry? (ask (C1-1))
✓	✓	✓	Thank you very much. I perfectly understand why you choose this company and which part of the company you think you can contribute towards using your particular strengths. By the way, do you have any future vision after you enter this company? (ask backup question)

Table 3: Classification result of each check item

Check item	Accuracy	Precision	Recall	F1-score
(C1-1)	0.730	0.725	0.829	0.773
(C1-2)	0.524	0.372	0.842	0.513
(C1-3)	0.857	0.714	0.667	0.690
(C2-1)	0.903	0.857	0.750	0.800
(C2-2)	0.548	0.533	0.533	0.533
(C3-1)	0.724	0.824	0.737	0.778
(C4-1)	0.828	0.850	0.895	0.872
(C4-2)	0.724	0.826	0.826	0.826

who are foreigners, especially Chinese people. Then, we would be able to make a good program.
...

The statistics of the annotation result are reported in Table 1. The numbers of samples are counted in the unit of a dialogue turn. It was found that each checklist item was mentioned by around 40% to 60% of candidates. This suggests the validity and generality of the checklist used in job interviews.

We trained a binary classification model with this training data. The input feature is a bag-of-words vector of the response and the output label is the binary result of the above annotation. The training model was made for each checklist item independently. Since the amount of training data is limited, we used a simple linear regression model with a $l1$-norm regularization. The trained coefficients were also restricted to be a positive value so that we can easily confirm the effective words for the classification. We evaluated the trained model with 5-fold cross validation. The classification result is summarized in Table 3. On some checklist items, the f-score was over 70%.

Finally, based on the binary classification results, the system generates a follow-up question sentence. The question sentence reflects the classification results including both positive and negative statements. Table 2 lists a set of follow-up questions on the first base question ((B1) *reason for apply*). Note that we made the order of priority among the check items. For example, in the case of the first base question (B1), the checklist item (C1-1) has the highest priority followed by (C1-2) and (C1-3). When the system classifies that the candidate mentioned (C1-1) but not mentioned (C1-2) and (C1-3), the system generates a follow-up question asking (C1-2) due to its priority. Each question sentence was designed manually based on the definition of the checklist. Using these questions, we aim to make candidates feel that the system listens and understands the responses of the candidates and to realize more effective training of job interviews. However, if the classification fails, it would make candidates feel that the questions are redundant because the follow-up questions have already been mentioned. Therefore, it is important to correctly classify the above checklist items.

4.2 Follow-up questions based on keyword extraction

The system also generates another type of follow-up question based on the keyword extraction from the response to the previous follow-up question. To realize automatic keyword extraction, we also used the same dialogue data as the previous part. We conducted a human annotation of keywords that can be used as the basis for the next question. We obtained 367 keywords from the interviewees'

responses and trained a machine learning model. We used a neural network model that consists of one-layer bidirectional long short-term memory (BLSTM) followed by a three-layer linear transformation with an output layer. The unit sizes of BLSTM and linear transformation are 256 and 128, respectively. The input feature is a Japanese word2vec model (200 dimensions) that was trained with web-based large text data [3]. We also added the type of part of speech (12 dimensions) and idf (inverse document frequency) value calculated from Japanese Wikipedia (1 dimension). The output is a posterior probability that the corresponding input word is the keyword. If several words are regarded as keywords, we select the one that has the highest output probability. Note that if several continuous words are estimated as keywords at the same time, they are acknowledged as a compound noun (e.g. *machine learning*) and are regarded as one word. We evaluated the trained model by 4-fold cross-validation. The word-level average f1-score was 52.7% where precision was 63.1% and recall was 45.2%. For example, when an interviewee said "*I have work experience as a teacher of individual lessons*", the keyword was extracted as "*individual lessons*".

After extracting a keyword, the system fills the keyword in a pre-defined template to generate a follow-up question. For example, when an extracted keyword is *autonomous robots*, a follow-up question would be "*You mentioned autonomous robots, so could you explain them in more detail?*". Since it is a critical issue if the system extracts an incorrect keyword due to model accuracy or errors of automatic speech recognition (ASR), we made a heuristic rule that the keywords must be nouns. We also utilize the confidence score of each word, calculated by the ASR system. If we could detect a keyword, but its corresponding ASR confidence score is lower than a threshold, we do not use the keyword for the generation of follow-up questions. If any keyword is not detected, we skip this step and proceed to the next base question.

4.3 Non-linguistic features

To increase the realism of ERICA, we also implement features which are designed to make her act more human-like. These are unrelated to response generation.

Turn-taking is an important feature of not only job interviews, but all dialogues. A simple approach in a basic spoken dialogue system is to wait until the user has been silent for a set period of time before the system can take the turn. However, this requires fine tuning and is usually inflexible. For a job interview system, it is vital for the interviewer to ensure the user has finished their turn, because early interruptions will be perceived as them not listening. On the other hand, it is unnatural if the user has to wait for a long time before getting a response from the system.

We implement a machine learning turn-taking model which uses ASR as an input and supplementing this with an finite-state turn-taking machine (FSTTM) as used in previous works [23, 35] to determine how much silence from the user should elapse before the turn switches to the system. This means that utterances with a high probability of being end-of-turn are responded to quickly, while the system will wait longer if the user says utterances such as fillers or hesitations.

[3] https://github.com/hottolink/hottoSNS-w2v

To ensure that the system does not interrupt the user early, we use a heuristic rule which sets a fixed silence time threshold at 4,000 ms during the first 50 words spoken by the user during their turn. This means that at the start of the user's turn the system will not speak until 4,000 ms has elapsed. After the minimum number of words has been recognized, we switch to the machine learning model but set a minimum silence time threshold to 1,500 ms to reduce the number of interruptions by ERICA. The system will respond faster or slower according to the ASR result.

ERICA also performs non-verbal backchannels in the form of head nods, in order to express some natural listening behavior. The timing of these backchannels are not random, but determined using a machine learning model [25]. Although the original model was trained on verbal backchannels, we replaced these expressions with non-verbal nods so that the listening behavior is slightly more professional.

5 EXPERIMENT I: EFFECTIVENESS OF FOLLOW-UP QUESTION

We conducted a dialogue experiment in order to confirm the effectiveness of the follow-up questions with android ERICA.

5.1 Condition

The proposed system with follow-up question generation was compared to a baseline system that did not generate any follow-up questions, only base questions. To make a fair comparison for the length of the interview, we made an additional four base questions only for the baseline system, which resulted in 8 base questions in total. This baseline system is designed as a similar system to existing job interview systems such as Hirevue and Tengai which use fixed questions. The baseline system does not take the risk of asking inadequate or unnatural questions due to the fixed question sentences.

We used a 16-channel microphone array for automatic speech recognition so that the interviewee can speak without holding a microphone (hands-free). At first, we estimate the sound source direction based on the multi-channel speech signals, and used a Kinect v2 sensor to track the subject's position. By comparing the estimated sound source direction with the subject's position, voice activity was detected [19]. The speech signal is enhanced based on the sound source direction and fed to automatic speech recognition that was implemented by an acoustic-to-word end-to-end model [40].

We recruited 22 university students (8 females and 14 males) as subjects. Each subject talked with and evaluated both follow-up question conditions (the proposed and baseline systems) implemented in ERICA, therefore using a within-subjects design. The order of the conditions was randomized for each subject. The experiment was approved by the university's ethics committee.

Before the experiment, each subject prepared for the job interview. We asked them to choose a company (or type of industry) they were going to apply for jobs, and also to consider their answers to potential interview questions. Each subject took a job interview with one of the follow-up question conditions, then evaluated the first system using a 7-point Likert scale questionnaire (individual evaluation). Questionnaire items are listed in Table 4 and divided

Table 4: Average scores (standard deviations) and the result of paired t-test (n=22) for dialogue with ERICA (android robot). FQ represents follow-up question.

	Item	w FQ (proposed)	w/o FQ (baseline)	p-value
(Impression on job interview itself)				
Q1	I was nervous during the interview	5.3 (1.39)	4.2 (1.82)	.008 **
Q2	I took this interview seriously	6.4 (1.07)	6.3 (1.02)	.352
Q3	The interview was boring	2.3 (1.46)	3.5 (1.64)	.011 *
Q4	Thanks to the interview, I was able to notice my weak points	5.0 (1.61)	3.7 (1.86)	<.001 **
Q5	The interview was close to the real thing	4.6 (1.64)	3.2 (1.82)	<.001 **
Q6	The interview was good practice for the real thing	5.6 (1.19)	4.7 (1.66)	.005 **
Q7	Thanks to this interview, I have confidence for a real job interview	3.6 (1.61)	3.2 (1.56)	.129
Q8	The interview was real as human-human job interview dialogue	3.9 (1.59)	3.0 (1.49)	.001 **
Q9	I felt that the interviewer was listening attentively	5.0 (1.48)	3.1 (1.14)	<.001 **
(Quality of question)				
Q10	The interviewer understood my answers	4.6 (1.55)	3.0 (1.36)	.001 **
Q11	I felt the questions were suitable and well considered for me	4.7 (1.35)	3.0 (1.52)	<.001 **
Q12	Thanks to the questions, I was able to notice that my responses were insufficient and inadequate	5.0 (1.64)	3.0 (1.87)	<.001 **
Q13	I felt flustered when answering the questions	5.6 (1.67)	4.2 (1.82)	<.001 **
Q14	I felt the interviewer was able to pick out my weak points	4.3 (1.71)	2.6 (1.15)	.005 **
Q15	I think the questions were actually generated by a hidden person	3.7 (1.91)	2.7 (1.51)	.005 **
(Presence of interviewer)				
Q16	I felt the presence of the interviewer	5.2 (1.47)	4.4 (1.40)	.026 *
Q17	I consciously considered my facial expression and posture in the interview	5.1 (1.53)	5.0 (1.83)	.385
Q18	I consciously looked at the interviewer in the interview	5.1 (1.65)	5.0 (1.87)	.451
Q19	I felt I was seen by the interviewer	4.7 (1.82)	4.0 (1.82)	.007 **

(* $p < .05$, ** $p < .01$)

Table 5: The numbers of time selected by subjects in comparative evaluation and the result of the binomial test (n=22) for dialogue with ERICA (android robot). FQ represents follow-up question.

	Item	w FQ (proposed)	w/o FQ (baseline)	p-value
CQ1	Which system did offer better practice for job interviews?	19	3	.001 **
CQ2	Which system did better understand your answers?	20	2	<.001 **
CQ3	Which system did generate more appropriate questions?	14	8	.286
CQ4	Which system do you want to use again?	17	5	.017 *

(* $p < .05$, ** $p < .01$)

into three categories: *impression on job interview itself*, *quality of question*, and *presence of interview*. It is expected that the presence of the interviewer is further enhanced by the combination of the appearance of android and the follow-up questions. After the first dialogue, the same experiment and evaluation was conducted with the other condition. Finally, we asked the subject to compare and evaluate both conditions. The subject directly selected the condition that best answered the questions listed in Table 5.

5.2 Result

The result of the individual evaluation is reported in Table 4. We also conducted a paired t-test on each question, and significant differences were observed in many questions. For the first category (*Impression of job interview itself*, there were significant differences in Q1 (I was nervous during the interview), Q5 (The interview was close to the real thing), and Q8 (The interview was real as human-human job interview dialogue), which suggests that generation of follow-up questions leads to a more realistic job interview. As a result, the quality of job interview practice was enhanced, which was measured by Q4 (Thanks to the interview, I was able to notice my weak points) and Q6 (The interview was good practice for the real thing). For the second category (*Quality of questions*), significant differences were observed in all the questions, which means that the proposed system could generate effective follow-up questions without dialogue breakdown. For the third category (*Presence of interviewer*), significant differences were observed in Q16 (I felt the presence of the interviewer) and Q19 (I felt I was seen by the interviewer). Therefore, generation of follow-up questions contributed to increasing the presence of the job interview robot. There is room for improvement in the evaluation scores themselves, particularly

Table 6: The numbers of samples selected by human majority voting in comparison between the proposed follow-up question generation based on quality of responses and random choice

Topic	Proposed	Random
Reason for apply	12	3
Strengths	11	4
Achievements	10	5
Skills	9	6
Total	42	18

for questions related to the similarity of a human-human job interview (Q8 and Q15) and the gaining of self-confidence of the interviewee (Q7).

The result of the comparative evaluation among the follow-up question conditions is reported in Table 5. For all questions, most subjects preferred the job interview system with the generated follow-up questions. We also conducted a binomial test on each question and found significant difference in all except CQ3.

5.3 Comparison with random choice

Since the baseline system asked only base questions, on follow-up question based on quality of response, it can be also considered that the proposed system is compared with random choice from the set of follow-up questions such as Table 2. Therefore, we conducted another experiment where other people evaluate the follow-up questions generated in the previous dialogue experiment in an offline manner. We collected other 5 university students (2 females and 3 males) who did not attend the dialogue experiment and who have experience as job interviewees. At first, each evaluator watched a dialogue video consisting of each base question and the following subject's response. We then showed two candidates of follow-up questions: one is actually used in the experiment and one is randomly chosen from the list. Each evaluator finally selected a more appropriate one based on a criterion if the system understands the response and if the follow-up question is effective to elicit meaningful information from the interviewee. If the randomly select question is the same as the question used in the dialogue experiment, we again randomly selected it until they are different. We also randomly selected and used 60 pairs of a base question and its response where each topic has 15 pairs. Note that we evaluate only follow-up questions based on quality of responses, do not evaluate follow-up based on keyword extraction in this manner.

Table 6 reports the numbers of samples selected as more *appropriate* by majority voting among the five evaluators. In all the topics, the follow-up questions generated by the proposed system were more selected than those by random choice. We also conducted a binomial test on the total number of the majority voting (42 vs. 18) and found a significant difference among them ($p = 0.001$). This result also supports the effectiveness of the proposed follow-up question generation.

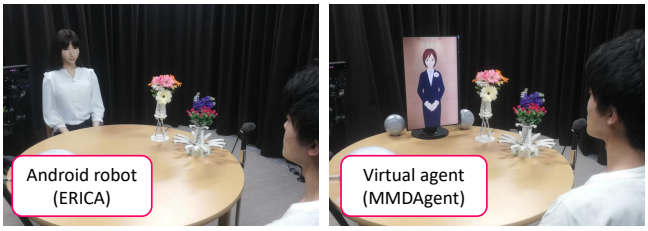

Figure 3: Difference on appearance of job interviewer (android robot vs. virtual agent)

6 EXPERIMENT II: EFFECTIVENESS IN VIRTUAL AGENT

We further investigated the effectiveness of the follow-up questions in dialogue with virtual agents as this is a more practical interface than android robots.

6.1 Condition

The virtual agent we use in this experiment is MMDAgent, which is a commonly used open-source agent toolkit [28]. The appearance of the agent compared with ERICA is shown in Figure 3. We use ERICA's text-to-speech and gesture generation system with the agent to replicate ERICA's corresponding behavior and speech as much as possible. The agent also used the same turn-taking and backchannel models as ERICA. The agent is displayed on a large screen in front of the subject. We additionally collected other 21 university students (6 females and 15 males) as subjects, and conducted the same experiment as explained in the previous section.

6.2 Result

The result of the individual evaluation is reported in Table 7. We also conducted a paired t-test on each question, and significant differences were observed in many questions, similar to the case of ERICA. However, no significant differences were observed in the third category (*Presence of interviewer*), meaning the presence of job interviewer was not affected by the follow-up questions in the virtual agent. Besides, increasing frustration by follow-up questions (Q13) was mitigated in dialogue with the virtual agent. This result can be interpreted as an advantage of making the job interview more relaxing and also make the user calm. On the other hand, this can be also interpreted as a disadvantage of making the job interview without tension and also not close to real job interviews.

The result of the comparative evaluation is reported in Table 8. Similar to the result in dialogue with ERICA, most subjects preferred the job interview with follow-up questions than those without. We also conducted a binomial test on each question and found that the differences were significant except for the third question (CQ3).

6.3 Comparison between android robot and virtual agent conditions

We also conducted a two-way mixed ANOVA for two factors: robot vs. virtual agent, and with and without follow-up questions, using the results on the individual evaluations (Table 4 and Table 7). We found cross interaction on Q5 (The interview was close to the real

Table 7: Average scores (standard deviations) and the result of paired t-test (n=21) for dialogue with MMD agent (virtual agent). FQ represents follow-up question.

	Item	w FQ (proposed)	w/o FQ (baseline)	p-value
(Impression on job interview itself)				
Q1	I was nervous during the interview	5.0 (1.46)	4.3 (1.64)	.022 *
Q2	I took this interview seriously	5.8 (1.22)	5.8 (1.33)	.500
Q3	The interview was boring	2.5 (1.22)	3.2 (1.50)	.009 **
Q4	Thanks to the interview, I was able to notice my weak points	5.3 (1.39)	4.1 (1.78)	.007 **
Q5	The interview was close to the real thing	4.1 (1.64)	3.7 (1.67)	.086 +
Q6	The interview was good practice for the real thing	5.3 (1.46)	4.2 (1.66)	.001 **
Q7	Thanks to this interview, I have confidence for a real job interview	3.7 (1.32)	3.1 (1.19)	.010 *
Q8	The interview was real as human-human job interview dialogue	3.9 (1.52)	2.9 (1.41)	.001 **
Q9	I felt that the interviewer was listening attentively	5.2 (1.50)	3.0 (1.65)	<.001 **
(Quality of question)				
Q10	The interviewer understood my answers	3.8 (1.37)	2.8 (1.53)	.009 **
Q11	I felt the questions were suitable and well considered for me	4.5 (1.40)	3.4 (1.59)	.002 **
Q12	Thanks to the questions, I was able to notice that my responses were insufficient and inadequate	5.3 (1.25)	3.2 (1.68)	<.001 **
Q13	I felt flustered when answering the questions	5.1 (1.72)	4.7 (1.32)	.112
Q14	I felt the interviewer was able to pick out my weak points	4.5 (1.33)	2.8 (1.44)	<.001 **
Q15	I think the questions were actually generated by a hidden person	3.5 (1.62)	2.0 (1.17)	.001 **
(Presence of interviewer)				
Q16	I felt the presence of the interviewer	4.0 (1.85)	3.7 (1.67)	.123
Q17	I consciously considered my facial expression and posture in the interview	4.4 (1.50)	4.5 (1.56)	.377
Q18	I consciously looked at the interviewer in the interview	4.8 (1.47)	5.0 (1.53)	.295
Q19	I felt I was seen by the interviewer	4.1 (1.83)	4.1 (1.78)	.500

(+ $p < .1$, * $p < .05$, ** $p < .01$)

Table 8: The numbers of time selected by subjects in comparative evaluation and the result of the binomial test (n=21) for dialogue with MMD agent (virtual agent). FQ represents follow-up question.

	Item	w FQ (proposed)	w/o FQ (baseline)	p-value
CQ1	Which system did offer better practice for job interviews?	18	3	.002 **
CQ2	Which system did better understand your answers?	19	2	<.001 **
CQ3	Which system did generate more appropriate questions?	15	6	.078 +
CQ4	Which system do you want to use again?	16	5	.027 *

(+ $p < .1$, * $p < .05$, ** $p < .01$)

thing) and Q13 (I felt flustered when answering the questions). We propose that these two effects were enhanced by the combination of the android robot and follow-up question generation. We also observed a main effect on the follow-up question condition on many items (Q1, Q3-16), and found another main effect on the appearance of the interviewer (android robot vs. virtual agent) on Q16 (I felt the presence of the interviewer).

7 CONCLUSION

In this paper, we proposed using an android as a job interviewer in order for people to practice in a realistic environment. At the same time, we developed a system to generate follow-up questions during the interview. The follow-up questions were made by two approaches: based on quality of responses and based on keyword extraction. We conducted the dialogue experiment in order to compare our follow-up question generation system to a fixed format baseline system. The result suggested that the follow-up question generation system significantly improved the interview. Besides, the presence of the android interviewer was enhanced by the follow-up questions. We further investigated the effectiveness of the follow-up questions in dialogue with the virtual agent. As a result, we observed the similar result in dialogue with the android robot, except that the presence of the interviewer was not enhanced by the follow-up questions. In future work, we will consider more adaptive actions of the job interviewer such as post-interview feedbacks to enhance the effectiveness of job interview practice.

ACKNOWLEDGMENTS

This work was supported by JST ERATO Grant number JPMJER1401 and JSPS KAKENHI Grant number JP19H05691 and JP20K19821.

REFERENCES

[1] Muneeb Imtiaz Ahmad, Omar Mubin, and Hiren Patel. 2018. Exploring the potential of NAO robot as an interviewer. In *International Conference on Human-Agent Interaction (HAI)*. 324–326.

[2] Mohammad R. Ali, Dev Crasta, Li Jin, Agustin Baretto, Joshua Pachter, Ronald D. Rogge, and Mohammed E. Hoque. 2015. LISSA-Live interactive social skill assistance. In *Affective Computing and Intelligent Interaction (ACII)*. 173–179.

[3] Keith Anderson, Elisabeth André, T. Baur, Sara Bernardini, M. Chollet, E. Chryssafidou, I. Damian, C. Ennis, A. Egges, P. Gebhard, H. Jones, M. Ochs, C. Pelachaud, Kaśka Porayska-Pomsta, P. Rizzo, and Nicolas Sabouret. 2013. The TARDIS framework: Intelligent virtual agents for social coaching in job interviews. In *International Conference on Advances in Computer Entertainment Technology (ACE)*. 476–491.

[4] Tobias Baur, Ionut Damian, Patrick Gebhard, Kaska Porayska-Pomsta, and Elisabeth André. 2013. A job interview simulation: Social cue-based interaction with a virtual character. In *International Conference on Social Computing (SocialCom)*. 220–227.

[5] Zoraida Callejas, Brian Ravenet, Magalie Ochs, and Catherine Pelachaud. 2014. A computational model of social attitudes for a virtual recruiter. In *International Conference On Autonomous Agents and Multi-Agent Systems (AAMAS)*. 93–100.

[6] David Cameron, Samuel Fernando, Emily Collins, Abigail Millings, Roger Moore, Amanda Sharkey, Vanessa Evers, and Tony Prescott. 2015. Presence of life-like robot expressions influences children's enjoyment of human-robot interactions in the field. In *AISB Convention*.

[7] Kirby Cofino, Vikram Ramanarayanan, Patrick Lange, David Pautler, David Suendermann-Oeft, and Keelan Evanini. 2017. A modular, multimodal open-source virtual interviewer dialog agent. In *International Conference on Multimodal Interaction (ICMI)*. 520–521.

[8] Kevin W Cook, Carol A Vance, and Paul E Spector. 2000. The relation of candidate personality with selection-interview outcomes. *Journal of Applied Social Psychology* 30, 4 (2000), 867–885.

[9] Joana Galvão Gomes da Silva, David J Kavanagh, Tony Belpaeme, Lloyd Taylor, Konna Beeson, and Jackie Andrade. 2018. Experiences of a motivational interview delivered by a robot: Qualitative study. *Journal of medical Internet Research* 20, 5 (2018), e116.

[10] Ionut Damian, Tobias Baur, Birgit Lugrin, Patrick Gebhard, Gregor Mehlmann, and Elisabeth André. 2015. Games are better than books: In-situ comparison of an interactive job interview game with conventional training. In *International Conference on Artificial Intelligence in Education (AIED)*. 84–94.

[11] Amanda R Feiler and Deborah M Powell. 2016. Behavioral expression of job interview anxiety. *Journal of Business and Psychology* 31, 1 (2016), 155–171.

[12] Patrick Gebhard, Tobias Baur, Ionut Damian, Gregor Mehlmann, Johannes Wagner, and Elisabeth André. 2014. Exploring interaction strategies for virtual characters to induce stress in simulated job interviews. In *International Conference On Autonomous Agents and Multi-Agent Systems (AAMAS)*. 661–668.

[13] Dylan F. Glas, Takashi Minaot, Carlos T. Ishi, Tatsuya Kawahara, and Hiroshi Ishiguro. 2016. ERICA: The ERATO intelligent conversational android. In *International Conference on Robot and Human Interactive Communication (ROMAN)*. 22–29.

[14] Mohammed E. Hoque, Matthieu Courgeon, Jean-Claude Martin, Bilge Mutlu, and Rosalind W. Picard. 2013. MACH: My automated conversation coach. In *International Joint Conference on Pervasive and Ubiquitous Computing (UBICOMP)*. 697–706.

[15] Koji Inoue, Divesh Lala, Kenta Yamamoto, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. 2020. An attentive listening system with android ERICA: Comparison of autonomous and WOZ interactions. In *SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*. 118–127.

[16] Koji Inoue, Divesh Lala, Kenta Yamamoto, Katsuya Takanashi, and Tatsuya Kawahara. 2019. Engagement-based adaptive behaviors for laboratory guide in human-robot dialogue. In *International Workshop on Spoken Dialog System Technology (IWSDS)*.

[17] Koji Inoue, Pierrick Milhorat, Divesh Lala, Tianyu Zhao, and Tatsuya Kawahara. 2016. Talking with ERICA, an autonomous android. In *SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*. 212–215.

[18] Carlos T. Ishi, Hiroshi Ishiguro, and Norihiro Hagita. 2012. Evaluation of formant-based lip motion generation in tele-operated humanoid robots. In *International Conference on Intelligent Robots and Systems (IROS)*. 2377–2382.

[19] Carlos T. Ishi, Chaoran Liu, Jani Even, and Norihiro Hagita. 2016. Hearing support system using environment sensor network. In *International Conference on Intelligent Robots and Systems (IROS)*. 1275–1280.

[20] Tatsuya Kawahara. 2018. Spoken dialogue system for a human-like conversational robot ERICA. In *International Workshop on Spoken Dialog System Technology (IWSDS)*.

[21] Takahiro Kobori, Mikio Nakano, and Tomoaki Nakamura. 2016. Small talk improves user impressions of interview dialogue systems. In *SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*. 370–380.

[22] Hirokazu Kumazaki, Taro Muramatsu, Yuichiro Yoshikawa, Blythe A Corbett, Yoshio Matsumoto, Haruhiro Higashida, Teruko Yuhi, Hiroshi Ishiguro, Masaru Mimura, and Mitsuru Kikuchi. 2019. Job interview training targeting nonverbal communication using an android robot for individuals with autism spectrum disorder. *Autism* 23, 6 (2019), 1586–1595.

[23] Divesh Lala, Koji Inoue, and Tatsuya Kawahara. 2018. Evaluation of real-time deep learning turn-taking models for multiple dialogue scenarios. In *International Conference on Multimodal Interaction (ICMI)*. 78–86.

[24] Divesh Lala, Koji Inoue, and Tatsuya Kawahara. 2019. Smooth turn-taking by a robot using an online continuous model to generate turn-taking cues. In *International Conference on Multimodal Interaction (ICMI)*. 226–234.

[25] Divesh Lala, Pierrick Milhorat, Koji Inoue, Masanari Ishida, Katsuya Takanashi, and Tatsuya Kawahara. 2017. Attentive listening system with backchanneling, response generation and flexible turn-taking. In *SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*. 127–136.

[26] Divesh Lala, Shizuka Nakamura, and Tatsuya Kawahara. 2019. Analysis of effect and timing of fillers in natural turn-taking. In *INTERSPEECH*. 4175–4179.

[27] Markus Langer, Cornelius J König, Patrick Gebhard, and Elisabeth André. 2016. Dear computer, teach me manners: Testing virtual employment interview training. *International Journal of Selection and Assessment* 24, 4 (2016), 312–323.

[28] Akinobu Lee, Keiichiro Oura, and Keiichi Tokuda. 2013. MMDAgent – A fully open-source toolkit for voice interaction systems. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 8382–8385.

[29] Julia Levashina, Christopher J Hartwell, Frederick P. Morgeson, and Michael A. Campion. 2014. The structured employment interview: Narrative and quantitative review of the research literature. *Personnel Psychology* 67, 1 (2014), 241–293.

[30] Julie McCarthy and Richard Goffin. 2004. Measuring job interview anxiety: Beyond weak knees and sweaty palms. *Personnel Psychology* 57, 3 (2004), 607–637.

[31] Iftekhar Naim, M Iftekhar Tanveer, Daniel Gildea, and Mohammed Ehsan Hoque. 2015. Automated prediction and analysis of job interview performance: The role of what you say and how you say it. In *International Conference on Automatic Face and Gesture Recognition (FG)*.

[32] Ryosuke Nakanishi, Koji Inoue, Katsuya Takanashi, and Tatsuya Kawahara. 2018. Generating fillers based on dialog act pairs for smooth turn-taking by humanoid robot. In *International Workshop on Spoken Dialog System Technology (IWSDS)*.

[33] Deborah M Powell, David J Stanley, and Kayla N Brown. 2018. Meta-analysis of the relation between interview anxiety and interview performance. *Canadian Journal of Behavioural Science* 50, 4 (2018), 195–207.

[34] Pooja Rao S. B, Sowmya Rasipuram, Rahul Das, and Dinesh B. Jayagopi. 2017. Automatic assessment of communication skill in non-conventional interview settings: A comparative study. In *International Conference on Multimodal Interaction (ICMI)*. 221–229.

[35] Antoine Raux and Maxine Eskenazi. 2009. A finite-state turn-taking model for spoken dialog systems. In *North American Chapter of the Association for Computational Linguistics (NAACL)*. 629–637.

[36] Matthew J. Smith, Emily J. Ginger, Katherine Wright, Michael A Wright, Julie Lounds Taylor, Laura Boteler Humm, Dale E. Olsen, Morris D. Bell, and Michael F Fleming. 2014. Virtual reality job interview training in adults with autism spectrum disorder. *Journal of Autism and Developmental Disorders* 44, 10 (2014), 2450–2463.

[37] Ilona Straub. 2016. 'It looks like a human!' The interrelation of social presence, interaction and agency ascription: a case study about the effects of an android robot on social agency ascription. *AI & society* 31, 4 (2016), 553–571.

[38] Ming-Hsiang Su, Chung-Hsien Wu, and Yi Chang. 2019. Follow-up question generation using neural tensor network-based domain ontology population in an interview coaching system. In *INTERSPEECH*. 4185–4189.

[39] Ming-Hsiang Su, Chung-Hsien Wu, Kun-Yi Huang, Qian-Bei Hong, and Huai-Hung Huang. 2018. Follow-up question generation using pattern-based seq2seq with a small corpus for interview coaching. In *INTERSPEECH*. 1006–1010.

[40] Sei Ueno, Hirofumi Inaguma, Masato Mimura, and Tatsuya Kawahara. 2018. Acoustic-to-word attention-based model complemented with character-level CTC-based model. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5804–5808.

[41] Daniela Villani, Claudia Repetto, Pietro Cipresso, and Giuseppe Riva. 2012. May I experience more presence in doing the same thing in virtual reality than in reality? An answer from a simulated job interview. *Interacting with Computers* 24, 4 (2012), 265–272.