

Improving Asynchronous Interview Interaction with Follow-up Question Generation

Pooja Rao S B, Manish Agnihotri, Dinesh Babu Jayagopi *

International Institute of Information Technology, Bangalore (India)

Received 28 October 2020 | Accepted 23 December 2020 | Published 24 February 2021



ABSTRACT

The user experience of an asynchronous video interview system, conventionally is not reciprocal or conversational. Interview applicants expect that, like a typical face-to-face interview, they are innate and coherent. We posit that the planned adoption of limited probing through follow-up questions is an important step towards improving the interaction. We propose a follow-up question generation model (*followQG*) capable of generating relevant and diverse follow-up questions based on the previously asked questions, and their answers. We implement a 3D virtual interviewing system, *Maya*, with capability of follow-up question generation. Existing asynchronous interviewing systems are not dynamic with scripted and repetitive questions. In comparison, *Maya* responds with relevant follow-up questions, a largely unexplored feature of virtual interview systems. We take advantage of the implicit knowledge from deep pre-trained language models to generate rich and varied natural language follow-up questions. Empirical results suggest that *followQG* generates questions that humans rate as high quality, achieving 77% relevance. A comparison with strong baselines of neural network and rule-based systems show that it produces better quality questions. The corpus used for fine-tuning is made publicly available.

KEYWORDS

Asynchronous Video Interview, Follow-up Question Generation, Language Model, Question Generation, Virtual Conversational Agent.

DOI: 10.9781/ijimai.2021.02.010

I. INTRODUCTION

CONVENTIONAL hiring process is laden with challenges. The amount of time required to hire candidates, lack of interviewers, expensive labour costs, scheduling conflicts are a few examples. Traditionally, at the employer's location, candidates take tests in a calm, distraction-free environment chosen by the employers where their presence is required. It includes various costs like scheduling, infrastructure, workspace and many more. To reduce these costs and challenges, recruiters are heading to futuristic choices like social recruitment, online assessments, and video interviews [1]. Organisations are adopting innovative methods like social media, proctored assessments, asynchronous or one-way interviews.

Online interviews for hiring are conducted using computer-mediated communication like instant messaging, email or video. Online interviews can be of the types synchronous, near-synchronous and asynchronous. [2] Synchronous interviews happen in real-time with simultaneous communication exchange. Near-synchronous interviews are near-immediate, on-going post and response. In the case of asynchronous interviews, there is a time-lapse between the communicating parties. These also called one-way interviews, are usually conducted via online video interviews using internet-enabled

digital devices. The candidates can take the interview whenever and wherever it is convenient for them.

Asynchronous video interviews (AVI) have evolved as a tool to conduct first round of screening as well as interview coaching. These are gaining increased attention due to its scalability and ease of use. Many automatic talent assessment solutions like Talview¹, Hirevue², Sonru³ offer asynchronous, ubiquitous interviewing and screening. Automatic interview and coaching systems simulate the behaviour of an interviewer helping interviewees with mock interviews. The feasibility and ease of automatic assessment of the AVIs when compared to in-person interviews [3] is persuading the wide spread use of the system. They provide the advantage of taking the test at candidate's convenience and facilitate efficient screening with minimal human intervention.

Conventional AVIs adopt structured interviews which are the standardized way of interviewing job candidates. Candidates are asked same questions in the same order with limited or no prompting and follow-up, and no elaboration on questions [4]. However, with large scale implementation of these systems, it will inevitably become predictable and uninteresting for recruiters and candidates alike. The dialogue will be monotonous and far from human-human interviews. Therefore, seeking the right balance between structure and probing is the imperative next step.

* Corresponding author.

E-mail addresses: pooja.rao@iiitb.org (P. Rao S B), manish.agnihotri@ii-itb.ac.in (M. Agnihotri), jdinesh@iiitb.ac.in (D. B. Jayagopi).

¹ www.talview.com

² www.hirevue.com

³ www.sonru.com

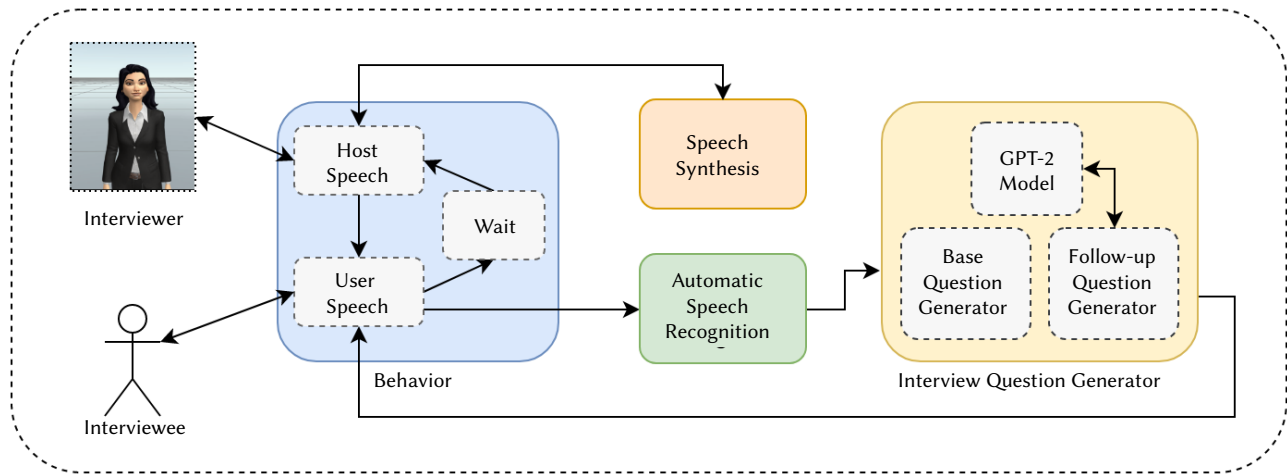


Fig. 1. Framework of Interviewing System.

An effective interviewing agent is one that has qualities similar to that of an effective human interviewer [5]. The ability to understand and respond to a conversation partner properly is one such inherent ability of human interviewer. This ability, also known as active listening, is shown to assist the progress of interviews [6]. Building a fully conversational interviewing agent with these abilities would be very challenging. A follow-up question in an interview can act as a proxy to such a human-like understanding of the answer in a crude way. It is crucial that an asynchronous system is conversational and acknowledges the candidates' response. We have enough evidence to suggest that the asynchronous interviewing systems are the need of the hour and the usage of such intelligent interviewing systems can only be scaled, if it's experience is interactive and personalized.

Levashina et al. [4] define follow-up question as the one that is intended to augment an inadequate or incomplete response provided by the applicant, or to seek additional or clarifying information. A relevant follow-up question not only improves the interaction between the interviewer and the interviewee but also makes it less predictable as the follow-up question is dynamic based on the interviewee's answer.

The methods for building a question generation model can be adapted for follow-up question generation, with one major difference. In the task of question generation, the question generated will already be answered in the input to the model. A follow-up question should not be already answered and should seek additional information not present in the input. Traditionally, the task of question generation has been tackled using hand-engineered features and crafting pattern-based templates and linguistic rules. More recently, with the advent of Deep Learning, question generation is treated as a sequence-to-sequence [7] problem with the reader sequentially parsing the input answer word by word and the generator producing the probabilities of the next word until the entire question is generated. This is a data intensive process and the availability of interview question answers with follow-ups is highly scarce. We address this challenge of data scarcity by leveraging the implicit knowledge from a pre-trained language model and adapting it with a small in-domain interview corpus.

In our work, we describe *Maya*⁴, an interviewing virtual agent that is capable of generating follow up questions. This paper is an extended version of [8] published in IntelLang workshop of ECAI 2020. Our main contributions are as follows. First, we describe our follow-up question generation model - *followQG*, and each of the individual modules. Our contributing module is the use of a large-scale transformer language model to generate relevant and diverse follow up questions. Second,

we benchmark the follow up generation output against other strong question generation/selection models, using human evaluation. Third, in order to show how well the system scales with regard to speech input vs text input, we perform experiments with manually transcribed text vs automatically transcribed spoken text, where an Automatic Speech Recognition (ASR) engine does the speech transcription. The results show the effectiveness of *Maya* even with speech input. Overall *Maya* works real-time and is able to hold an interactive interview with a candidate. Finally, we also investigate how robust is *Maya* with multiple follow up turns, and present qualitative results.

II. RELATED WORK

We organise the related work in three categories: Research addressing language model pretraining (as our solution is based on neural network based text generation models), question generation task (of which follow up question generation is a special task), and agent based interviewing systems (which is exactly the task we are addressing).

A. Language Model Pretraining

Training the Deep Learning models from scratch, starting with random layer initializations, with large datasets taking a long time to converge is the de facto standard for tackling various NLP tasks. The effective breakthrough strategy to this is the greedy layer-wise training using an unsupervised learning criterion (pre-training) followed by tuning all parameters of the network on a global supervised cost function (fine-tuning) [9], [10]. Pre-training in an unsupervised fashion on vast quantities of text has resulted in state-of-the-art development on various tasks of natural language processing [11] [12]. These pre-training objectives are mostly variants of language modelling.

ULMFIT [13] is a transfer learning method for text classification tasks. A language model is pretrained on Wikipedia data and fine-tuned for a target task with a smaller amount of labelled in-domain data. Several works follow this fine-tuning approach and produce remarkable outcomes. ELMo [14] is a bidirectional language model that uses bi-LSTM networks to predict the next and previous tokens. OpenAI's GPT [12] trains huge text data in a unidirectional language model. BERT [11] is a masked language model trained with the next sentence prediction as an additional objective. On several downstream NLP functions, like the GLUE [16] benchmark, these models have achieved state-of-the-art results. Generative tasks such as end-to-end dialogue systems [17] and automated knowledge base completion [18], use pre-training with the GPT model, obtaining significant improvements over the models trained only with in-domain data. Both the works use the

⁴ The demo of the system can be found at - <https://www.youtube.com/watch?v=gdPxdi82nV0>

TABLE I. A COMPARISON OF ASYNCHRONOUS INTERVIEW SYSTEMS. THE VERBAL INTERACTION IN MAYA DIFFERS FROM OTHER WORKS WITH A FOLLOW-UP QUESTION MECHANISM AS IT USES A QUESTION GENERATION MODEL RATHER THAN USING TEMPLATE-BASED QUESTION SELECTION METHOD

System	Agent	Nonverbal Interaction	Verbal Interaction	Follow-up Q
Rao S B et al. [19]	Text Medium	No interaction	Fixed Script of Questions	No
SPECIES [20]	Embodied Agent	Head Movement and Facial Expressions	Template based	Yes
MACH [21]	Embodied Agent	Head Nodding and Smile Sharing	Fixed Script of Questions	No
TARDIS [22]	Embodied Agent	Body Motions, Gestures and Facial Expressions	Fixed Script of Questions	No
ERICA [23]	Robotic Agent	Head Movement, Gestures and Eye Gaze	Template based	Yes
Maya (Ours)	Embodied Agent	Gestures, Facial Expressions and Follow-up Question	Dynamic Question Generation	Yes

transformer language model GPT for initialization. Our work builds on this to develop a Follow-up Question Generation model.

B. Natural Language Question Generation

The goal of the Question Generation (QG) task is to automatically generate questions based on some form of text input [24]. This task became popular ever since the First Question Generation Shared Task Evaluation Challenge [25]. Recently, neural networks have enabled end-to-end training of question generation models influenced by the sequence-to-sequence (Seq2Seq) data-driven learning methods [7]. Serban et al. generate simple natural questions from structured triples - subject, relation, object using a neural system [26]. This has been successfully extended to unstructured data. Du et al. [27] generate question to test comprehension, using the encoder-decoder model with attention on the machine comprehension dataset SQuAD [28]. Wang et al. generate questions from educational content using an RNN-based encoder-decoder model, trained on SQuAD [29].

Follow-up question generation in interviews is a relatively new task, addressed first by Su et al. [30]. Instead of using a text generation model, they generate question patterns filling it up with words from a list. They adopt a pattern-based Seq2Seq model on a small interview corpus in Chinese. To create a word class table and turn all sentences in the corpus into patterns, they use a word clustering based process. In order to select a question-worthy sentence from the answer, they use a convolutional neural tensor network [31] and generate follow-up question patterns. These patterns are filled with words from the word class table to obtain potential follow-up questions. In a subsequent work, Su et al. [32] utilize the domain knowledge from ConceptNet to fill up relevant words in the follow-up template.

In contrast, we develop a follow-up question generation model utilizing knowledge from large-scale language model and a small corpus which does not involve pattern matching and template filling.

C. Agent-based Interviewing Systems

A recent trend in Interviewing Systems is the use of Intelligent Virtual Agents. Asynchronous Video Interviews (AVIs) are more common, where questions are posed by an interface. The use of intelligent virtual agents in AVIs allows for a more interactive and immersive experience than traditional voice and text-based systems [33], [34]. A job interview is aimed to analyze the hiring feasibility of an interviewee, while a training interview gives accurate feedback about their performance. While the initial works in AVIs were restricted to the skill assessment [35], [19], improving the interview experience has gained momentum. One standard approach is the usage of virtual agents as interviewers instead of textual prompts to conduct interviews [20]. This approach makes the interview experience more interactive.

In an early work, Nunamaker et al. introduced the usage of Embodied Conversational Agents (called SPECIES) in automated interviews [20]. One of their goals was to study the difference in perceptions with varying attributes of agent, and hence their work concerned agent design aspects.

Later, two coaching-based conversational agents were proposed - MACH [21] and TARDIS [22]. Both of them focus on skill assessment and non-verbal behavior analysis to improve the feedback to interviewees significantly, but the questions are taken from a small fixed pool of questions and do not take into account the interviewee's response. In these works, the virtual agents acknowledge the interviewee's answers to questions by smiling and nodding at a randomly generated frequency. Though this might make the agent credible, it can soon become superficial. Acknowledging the interviewee's answer by understanding the context and reacting appropriately through verbal means can be a better alternative to increase the interactivity and is the aim of our work.

Apart from the use of Virtual agent, a robotic agent (called ERICA) [23] was also built for spoken dialog. ERICA had the capabilities of human-like eye gaze, head movement and gestures, and a statement-response system which is response retrieval method based on pattern and focus token matching. Although the behavior synthesis is a notable improvement, it still lacks robustness in dialogue generation.

Automated Social Skills Trainer (ASST) [36] focused on Social Skills Training as their interview objective. The embodied agent in ASST is capable of head nodding, and eye blinking, and the dialogue is based on MMDAgent, a Japanese spoken dialogue system which selects an appropriate response using pattern matching.

While a lot has been done in automatic analysis of interviewee's response [19], [37] to improve the quality of the interview, not much has been done to make the interview more verbally interactive. All the previous works have either used a fixed script of questions or used a pattern matching based question selection (see Table I). We aim to improve the question generation system to make it more personal and response-based by generating relevant and grammatically correct follow-up questions.

III. FOLLOW-UP QUESTION GENERATION - FOLLOWQG

FollowQG is an adaptation framework for generating follow-up questions using language models by training it on an in-domain corpus of question, response and follow-up triplets. These data triplets help followQG to understand the structure of the question and the relationship between the triplets, and novel questions arise from the knowledge of the language model pre-training.

A. Task

The training samples consist of $\{q, r, f\}$ in natural language, where q is the interviewer question, r is the candidate response and f is the follow-up question. The task is to generate f given q and r as inputs.

B. Transformer Language Model

In this work, we use the transformer language model architecture, Generative Pre-trained Transformer (GPT-2) introduced in Radford et al. [38]. This uses an architecture similar to the decoder part of the original transformer encoder-decoder model of Vaswani et al. [39].

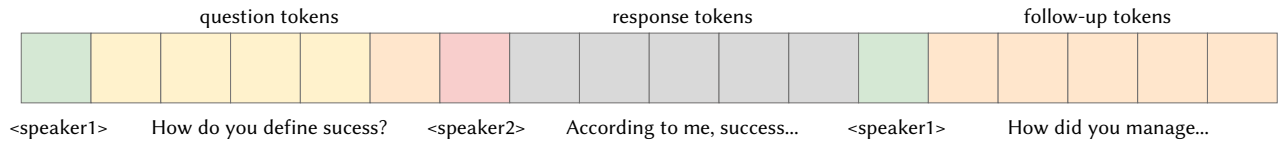


Fig. 2. Input representation for training Follow-up Question Generation model.

It uses several layers, each comprising two sub-layers. The multi-headed self-attention process is employed in the first followed by position-wise feed-forward layers to generate an output distribution over target vocabulary tokens. Our model is based on the recently published adaptation of GPT-2 from HuggingFace⁵.

We initialise followQG with 12-layer decoder-only transformers with 12 heads of self-attention comprising of 768 dimensional states. Parameters are initialised to the smallest version open-sourced by Radford et al. 2019, GPT-2 small [38]. A WebText dataset containing the text of 45 million internet links is used to pre-train the GPT-2 model.

C. Dataset

We need the training samples of $\{q, r, f\}$ triplets to train followQG. We employ the Rao S. B et al. [19] asynchronous interview dataset. This dataset consists of behavioural interviews of university students, referred to as the Asynchronous Video Interview dataset-AVI dataset and Asynchronous Written Interview dataset-AWI dataset, respectively, through asynchronous video and written media. Using AWI dataset interview excerpts, we perform a restricted crowdsourcing to obtain follow-up questions. We advise the volunteers to compose a follow-up question based on the posed interviewer question snippet and the answer of the applicant. An instruction video of the same can be found here⁶. Therefore, with more than 1000 samples, each sample containing the triplet of a question, response and a follow-up, we obtain a follow-up question dataset. Some samples from the dataset can be seen in Table II. You can find the complete dataset here⁷.

TABLE II. EXAMPLES OF HUMAN WRITTEN FOLLOW-UP QUESTIONS FROM THE DATASET

<p>Q: Are you a self-motivator?</p> <p>A: Absolutely. For me, internal motivation works far more than external motivation ever could. Yes, at first, it may seem like I want some sort of external motivation, but the very end, my heart goes into the work assigned only when my own self pushes me to do it.</p>
<p>FQ: Awesome. How would you spread motivation to others?</p>
<p>Q: What matters to you more - job satisfaction or salary?</p> <p>A: According to me, job satisfaction covers all - the quality and quantity of work, salary, company environment and others. Yes, cooperation and adjustment is needed in a company.</p>
<p>FQ: Are you ready to work in a company who offers you 5 times more than what you get now but the area may not be very interesting to you?</p>
<p>Q: Have you worked with someone unprofessional, how did you handle it?</p> <p>A: During my B-Tech final semester internship, I have experienced unprofessional behavior. I did not do anything to show my displeasure, instead, I kept behaving professionally. It didn't solve the issue, but did reduce the magnitude to some extent.</p>
<p>FQ: Do you think not showing the displeasure is the only way to tackle that situation?</p>

⁵ <https://github.com/huggingface/transformers>

⁶ https://youtu.be/KbHF7_kMaA8

⁷ <https://ms-by-research-thesis.s3.amazonaws.com/followMLdata.xlsx>

D. Fine-tuning

Using the dataset mentioned above, we fine-tune the GPT-2 language model. For training, 80% of the data is used and the rest is used for validation. Model input consists of tokens each from sequence of $\{q, r, f\}$ concatenated and embedded in order. For this sequence, a set of input embeddings is constructed. The word and position embeddings are derived from the GPT-2 model learnt during the pre-training phase. To indicate whether the token belongs to the question, answer or the follow-up, we use an additional set of embeddings, speaker embeddings. Fig. 2 illustrates how the tokens in $\{q, r, f\}$ are organised to form the speaker embeddings. These embeddings are learnt during the fine-tuning phase. The input to the model is the sum of all three forms for each token— word, position and speaker embedding.

1. Multi-task Objective

Following [17], [11], the fine-tuning is done by optimizing two loss functions – a language modelling loss, and a next-question classification loss. We use a multi-task objective where the total loss is the weighted sum of two losses. The language modelling loss is the commonly used cross-entropy loss. The last hidden state of the self-attention model is fed into a softmax layer over all the tokens in the vocabulary to obtain next token probabilities. These probabilities are then scored using the cross-entropy loss where the human written follow-up question tokens are used as labels.

With randomly sampled questions from a pool of 200 (same as those used in Section V), serving as distractors, we train a next-question classifier to recognize the correct next question. This trains the model to acquire a sense of sentence ordering. The classifier is a linear layer that applies a linear transformation to the last hidden state of self-attention model. A softmax layer obtains the classification probabilities using the computed values. We then apply a cross-entropy loss to correctly classify the follow-up question. We use $n = 2$ as the number of choices for classification making it a binary classification task. The parameters of the transformer language model and the next-question classifier layer are fine-tuned jointly to maximize the log-probability of the correct label.

2. Decoding Details

We use the top-k random sampling strategy for decoding [40]. The top-k probabilities of most likely next word is given at each point. The decoder randomly samples a word from these k candidates. Here, k is a hyperparameter determined experimentally to be $k=10$.

E. Results

We report the results of the follow-up question generation model in terms of perplexity [41]. The classification accuracy of the next-question task is also recorded. Perplexity is typically used to measure the quality of language models. It indicates how well the next word is correctly predicted by the model. Our model obtains an average validation perplexity of 20.6 and average validation accuracy of 63.1%. Considering the small size of the in-domain dataset used for fine-tuning, these values can be considered reasonable. The questions generated are novel and relevant and are not present in human written follow-up questions as we'll see in Section D.

IV. EXPERIMENTS

In this section, we demonstrate the effectiveness of followQG with quantitative and qualitative experiments. First, we quantitatively do a relative comparison of followQG with strong baselines through human evaluation. These baselines loosely mimic the different interviewing agents discussed in Section C. We then individually evaluate followQG with human annotations on relevance and grammar. We also investigate the robustness of followQG model to errors in speech recognition. Finally, we qualitatively validate the single and multiple follow-up questions on same interview question-answer pairs.

A. Baselines

We compare followQG with two strong baselines. One is a rule-based system based on similarity measure and other is the reader-generator based QG-Net model [29]. We choose these baselines as representatives of the existing asynchronous video interview systems (Table I) which either pose questions from a fixed set or template based generation.

1. Similarity-based Question Selector

This is a rule-based question selector with a rule on cosine similarity to select questions from a fixed pool of 200 behavioural questions (same as the ones used in Section V). The original interview questions and the pool of questions are all represented using GloVe vectors [42]. We calculate the cosine similarity metric between the original interview question and each of the questions from the pool. We take into account the top-10 questions with highest similarity values and randomly select one to be the follow-up question. The selected follow-ups are based only on the question and not the candidate response. This question selector loosely mimics the different rule-based selectors in the existing systems which pose a fixed question and do not take the response into account.

2. QG-Net

With a context reader and question generator, QG-net is a Seq2Seq model. The context reader is a bi-LSTM network that processes and transforms each word into a fix-sized representation in the input context. The question generator is a uni-directional LSTM which generates the question word-by-word incorporating pointer network [43] into the vocabulary of the generator. This model design enables the generator to output questions that focus on specific parts of input text. The *focus tokens* are encoded as an additional feature with each input word, using one-hot encoding to indicate if the word is a focus token. QG-Net is trained on SQuAD dataset consisting of context, query and span of answer tokens within the context. These span of answer tokens are used by QG-Net as focus tokens. For a detailed description, we refer the readers to the original paper . QG-Net effectively adapts a general purpose question generation model trained on SQuAD to generate questions from educational content, addressing the problem of insufficient training data. We therefore select this as our baseline model of the neural network. The candidate's response, in our case, is the context and the question to be generated is the follow-up question.

The interview question-answer pairs have to undergo preparatory techniques like finding focus of the answer and extractive summarization to make the input format compatible with QG-Net model. QG-Net model trained on SQuAD dataset released by Wang et al. [29] is our second baseline.

Finding Focus of the Answer QG-net uses a binary valued indicator to indicate whether a word in context is important to generate a question, regarded as *focus tokens*. To automatically find these tokens in candidate responses, we employ a simple technique similar to Hu et al., [44]. In interview question (Q) and response (A),

there are overlapping tokens seen as topics exchanged between the interviewer and candidate, that can be considered as focus tokens.

After removal of the stop words, A and Q are represented as a sequence of tokens $[a_1, \dots, a_n]$ and $[q_1, \dots, q_m]$ respectively. We consider all the tokens in A as candidates for focus tokens and all the tokens in Q as voters polling for the candidates. GloVe [42] vectors are used to represent tokens from Q and A. The i^{th} answer token a_i gets a cumulative score S_i from all the tokens in the question calculated as

$$S_i = \sum_{j=1}^m p_{ij} \cdot \text{sim}(a_i, q_j)$$

$$p_{ij} = \begin{cases} 1, & \text{sim}(a_i, q_j) > \lambda \\ 0, & \text{otherwise} \end{cases}$$

where $\text{sim}(a_i, q_j)$ is the cosine similarity between a_i and q_j . If the averaged S_i is above a certain threshold, a_i is included in the *focus*. This process is repeated for every answer token. We allow non-contiguous and multiple focus tokens which aid in the generation of distinct follow-ups.

Extractive Summarisation The input to the QG-Net model should be a sentence worthy of a follow-up representing the answer. To find this representative sentence, we employ a simple extractive summarization technique on the answer. The aim is to iteratively identify similar sentences in the answer using the focus tokens of those sentences and consider the most similar sentence as the summary sentence. We use the method described above to find the focus of each sentence. We then compare the focus of each sentence with the focus of other sentences using the cosine similarity measure.

R and S are two sentences from the candidate response with their focus tokens represented as $[fr_1, \dots, fr_p]$ and $[fs_1, \dots, fs_q]$ respectively. The cumulative score for each focus token of R is calculated as

$$W_i = \sum_{j=1}^q p_{ij} \cdot \text{sim}(fr_i, fs_j)$$

$$N = \sum_{i=1}^p W_i$$

where p_{ij} is the indicative variable same as described above. If N crosses a certain percentage of the mean length of two sentences R and S, they are considered to be similar.

Once we have the pair(s) of similar sentences, we choose the one with more information content (more number of focus tokens) as the summary sentence. If more than one pair of sentences are similar to each other, S (pre-determined) number of sentences with the highest frequency of similar sentences is considered. The summary sentence along with the focus words is fed to the trained QG-Net model to generate questions.

B. Quantitative Human Evaluation

We obtain human annotations to evaluate the quality of the generated follow-up questions and compare them with the baselines⁸. Graduate students (non-native English speakers) with a background in Computer Science and Digital Society are the human annotators involved in this research. We sample 100 unseen question-answer pairs randomly from the AWI dataset and generate one follow-up question from all three models per QA pair – Similarity-based Question Selector, QG-Net question generation and followQG. We present the QA pair to three human annotators along with the follow-up questions produced by each model. Based on their choice, they are asked to rank the questions in the order of two metrics – the relevance of follow-up question to the given QA pair of interviews

⁸ The customized web interface used for human annotation can be found here <https://poorao.github.io/followML/>

and their grammar. We define relevance as to how closely connected or appropriate the follow-up question is to the question-answer pair. We refer to this definition in the whole paper. The annotators are instructed to rank the ones with high relevance and high grammar as Rank 1, low relevance and low grammar as Rank 3 among the three choices. It is to be noted that these rankings are relative in measure.

For each follow-up question, we consider the statistical mode of the three annotators ranking. When the mode is not unique, i.e. when a different rank is selected by all three annotators (10% of the annotations), we resolve the disagreement by getting an additional set of rankings from an experienced interviewer. The statistical mode is then calculated using the extra annotations.

The findings can be seen in the Fig. 3. For each of the baseline and followQG, we calculate the frequency of the mode ranking for all three ranks. FQG model significantly outperforms (well beyond $p=0.01$ level) the other two models. With 54% of questions generated securing Rank 1, followQG is capable of high quality follow-up question generation. 34% of the questions generated by QG-Net obtain Rank 1. 50% of the questions from SQS secure Rank 2. It can be observed that the grammatically correct SQS selected questions are preferred second to the grammatically incorrect and somewhat relevant QG-Net model questions. We conclude that the FQG model, more frequently than the current baselines, produces valid and grammatically correct follow-up questions.



Fig. 3. Human ranking of preferred follow-up questions from followQG comparing with two other baseline models based on relevance and grammar. The bar indicates the frequency of rankings, indicating that the followQG model is the most preferred for highest ranking.

To improve our evaluation further, we obtain human ratings for followQG individually. Three human annotators measure the quality of the questions on a scale of 1-3, 1 being the lowest. The annotators are directed to annotate the follow-ups based on the grammar and relevance to the original interview QA pair. A score of 3 meant the follow-up question is grammatically acceptable and contextually relevant. Either a grammatically unacceptable question or not a follow-up question due to its low relevance meant a score of 1. And anything in between was considered a score 2.

The average scores from three annotators are considered for evaluation. Fig. 4 shows the statistics of the average ratings for the follow-up questions generated. 77% of the questions are scored ≥ 2 . And 27% are rated ≥ 2.5 . This shows that the followQG generates superior quality follow-up questions and are scored well by humans.

C. Robustness to Errors in Speech

Investigating the robustness of Follow-up Question Generator has two important motivations. First, the model is trained on human-written triplets of $\{q, r, f\}$ whereas it will be inferred on the candidates's response obtained from ASR transcript in the virtual interviewing system. Secondly, ASR engines are generally prone to

Average rating	Avg Rating on written QA pair	Avg Rating on manual transcripts	Avg Rating on automatic transcripts
1	2	0	4
1.3	9	11	15
1.67	12	21	18
2	23	22	22
2.3	27	21	21
2.6	20	17	20
3	7	11	3

Fig. 4. Frequency distribution of average human ratings on the quality of generated follow-up questions from followQG on a scale of 1-3 on the different types of question-answer pair inputs (hand-typed text, manually and automatically transcribed spoken text).

errors in recognition. Generation on misrecognized words may lead to an irrelevant question. Hence, analyzing how follow-up question generation varies for ASR transcripts when compared with human transcripts helps to investigate the robustness of followQG.

To this end, as they have manual transcripts of the interviews, we use the asynchronous interface-based video interview dataset from Rasipuram et al [3]. We randomly choose 103 interview QA pairs. Automatic transcripts for the same pair of 103 question answers are also obtained using the Web Speech API [45]. For each of these pairs, we generate a follow-up question. This brings us 206 triplets of questions, responses and follow-up questions, 103 for manual and automated transcripts each.

Three human annotators evaluate the quality of the question on a scale of 1-3, 1 being the lowest. Same instructions given in Section B are used. The annotators are displayed with the questions and answers from the manual transcripts and the follow-up questions generated on both manual and automatic transcripts to rate. We consider the average rating of the three annotators for evaluation.

The Fig. 4 indicates the number of average ratings for manual and automated transcript follow-up questions. 69% of the questions generated on manual transcripts and 64% of the questions generated on ASR transcripts get a score of ≥ 2 . This means that, on both manual and automated transcripts, followQG produces reasonably high quality follow-up questions.

D. Qualitative Analysis

We present some examples generated by the followQG model on randomly sampled QA pairs in Table III.

Validation Examples Results show that the questions generated on the validation samples are relevant and grammatically correct but not same as the ground truth questions. We can see few examples in the first part of Table III where the generated follow-up questions are different from the ground truth follow-up questions. This may be due to the fact that followQG model leverages the implicit knowledge from the pre-trained language model to generate previously unseen questions different from ground-truth.

Unseen Examples The second part of the table shows the follow-up questions generated on completely unseen data. The model generalizes well on the unseen data and generates grammatically correct and suitable questions.

Unseen ASR Examples The third part of the table shows the questions generated on the text from the ASR. Despite the errors in transcription, the questions are relevant and appropriate. It is to be noted that the original base question is not exposed to the errors of the ASR and can be a reliable input to followQG.

TABLE III. EXAMPLES OF THE FOLLOW-UP QUESTIONS GENERATED ON THE INTERVIEW SNIPPETS FROM THE VALIDATION DATA, UNSEEN DATA AND UNSEEN DATA FROM ASR

Validation Examples			
Question	Candidate Response	Generated Follow-up	Human Generated Follow-up
How would you be an asset to this organisation?	I can be an asset as I am patient while handling people and situations. I can participate and coordinate well with teams I get to work with. I can take in criticisms and work on to improve myself. I like to learn new things from anybody and will take necessary help whenever needed. I quite friendly with everyone.	How would you handle criticism?	Can criticism always be seen as a means for self-improvement?
Have you ever had to sell an idea to your co-workers? How did you do it?	I had to, when we were brainstorming ideas for a hacking session. I first tried to understand their concerns and requirements. Later I pitched my idea in accordance to that. I ensured that I understood and answered all their queries.	How did you handle the stress?	What were their first reactions?
What obstacles or difficulties have you ever faced in communicating your ideas?	I am not very expressive and can't put thoughts into words, so i have failed in communicating my ideas and explaining it or even getting work done.	How do you cope with a difficult situation?	And what steps have you taken to overcome this hardship?
What do people most often criticize about you?	I am not very social, an introvert and can't get along with people easily. I hvae been criticized about this.	What do you think about people who criticize you?	Have you ever tried to work on these drawbacks?
Unseen Examples			
Question	Candidate Response	Generated Follow-up	
What are your salary expectations?	As of now, I haven't thought much about it. I'm more focused on learning the requirements for this position that I am applying for.	How much do you think you will be able to meet these requirements?	
Are you a team player?	Definitely. While I will deny the fact that I can work independently with minimal supervision, I'm also one companion every leader would ever want to be in his team. Whatever the task is, I make sure it meets and exceeds what is expected of me. I also make it a point to reach out to teammates whenever needed.	Do you have any tips for team members?	
Why do you want to work for this company?	I would be proud to work for a company like yours with such a long history of leadership in the industry. I have carried out web research and believe that the company's products and its future projections are very impressive.	Why do you think this company is successful?	
Unseen Examples with ASR Transcriptions			
Question	Candidate Response	Generated Follow-up	
Do you consider yourself successful?	I do consider myself successful because I have been able to achieve the things that I wanted for myself so ja in the respect yes I too feel them successful	Do you think you can achieve the things you wanted for yourself?	
Tell me something about yourself.	I am a research assistant at the <name>... institute I work in the area of human interaction and machine learning	How did you learn to work with machine learning?	
How do you think this job will help you fulfill your career goals?	when I apply for this job applied for this job and applied for this job many things which I like to do and I will do in the stop my career goals are just to have some work satisfaction and study whatever I want to so I think this job will help me do that and	How would you handle a situation where you have no work satisfaction?	

TABLE IV. COMPARISON OF EXAMPLES GENERATED FROM THREE DIFFERENT MODELS FOR VARIOUS QUESTION-ANSWER PAIRS. SQS: SIMILARITY-BASED QUESTION SELECTOR, QG-NET: QG-NET, FQG: FOLLOW-UP QUESTION GENERATION

<p>Q: What is the biggest challenge you have faced so far? A: The biggest challenge that i have faced so far is IIT- JEE examination after completing my class 12. I prepared for it for two years. Unfortunately I didn't cracked it. SQS: Where do you see yourself 3 years from now? QG-Net: did you think was cracked ?</p>
<p>FQG: What are the best lessons you have learned from this experience?</p>
<p>Q: Short self-presentation: Introduce yourself. A: Hello, I am X, a passionate software engineer. Till my 12th standard I studied at my native place itself. My strongest plus point according to me is that I dont trust anyone easily and negative point is that I cant resist anything wrong happening to anyone around me... SQS: What has been the greatest disappointment in your life? QG-Net: what is you that you can't resist happening?</p>
<p>FQG: How do you manage your negative point?</p>
<p>Q: What has been the greatest disappointment in your life? A: I didn't get any high disappoint from my life. I was disappoint from myself when I didn't performed that well in GATE exam and not able to get admission in X, but I got admission in Y and now I feel that it was okay and not feel disappoint. SQS: What changes would you make if you came on board? QG-Net: what did you do in gate exam ?</p>
<p>FQG: How would you explain your failure in GATE exam?</p>
<p>Q: How would you be an asset to this organisation? A: I think I have a hunger to learn and solve a problem. For me, each problem is a puzzle which I need to solve and I will try till I finish it. If I don't know something, I will learn it and get the job done. My hardworking and passionate qualities can be asset to your company. SQS: How quickly do you adapt to new technology? QG-Net: what do you have a hunger to do ?</p>
<p>FQG: How do you think you can improve your skills?</p>

Comparison Table IV compares the follow-up questions generated by the followQG with the two strong baselines, similarity-based question selector and QG-Net. It is seen that question generated by followQG is a grammatically correct and relevant follow-up in all cases. We can see that the QG-Net model fails the most in terms of grammar. FollowQG, despite the grammatical errors in the reply, always adheres to grammar. Grammatical adherence is irrelevant with SQS model as it selects from a pool of grammatically correct questions.

In terms of relevance, we find that the follow-up related to the question or answer or both is often generated by the followQG model. SQS generates follow-ups that are remotely connected or unrelated to the responses. Follow-ups by QG-Net model do not always adhere to relevance. It produces questions that are irrelevant or incomplete or grammatically incorrect. We should also notice that the QG-Net model produces already answered questions (e.g. Questions 2, 4).

Variety In addition to being high quality, the generated follow-up questions from followQG are diverse and novel. Majority of the questions are not present in the training set, showing that the model is capable of generating new questions. The questions generated can be modified versions of training samples or completely different from the training samples addressing unique details in the answer. For example in Table V, the last two questions address very specific details in the responses and give an impression to the candidate that the agent has understood and responded to the answer advancing the listening skills of the agent. Table V illustrates sample follow-up questions generated on the same question but different candidate responses.

TABLE V. QUESTIONS GENERATED BY FOLLOWQG FOR THE INTERVIEW QUESTION "DESCRIBE A DECISION YOU MADE THAT WAS A FAILURE. WHAT HAPPENED AND WHY?" BUT WITH DIFFERENT CANDIDATE RESPONSES. WE CAN SEE THAT THE FOLLOW-UPS ARE NOVEL AND DIVERSE

<p>Describe a decision you made that was a failure. What happened and why?</p>
<p>A: My initial plan for post-graduation was to join MS by research in IIT Madras. But I have failed in my first attempt due to some medical issues. So I was very determined to give it a second try. Then I took admission in IIIT-B for M.tech. I planned to prepare for the interview along with my course work during first semester. But unknowingly, I got into some bad company and ignored the preparation. I realised my mistake when I finally got the interview call again. It was too late by then. I got depressed because of this and I made a lot of effort to motivate myself again and do my best in my M.tech and work on my area of interest. If I achieve the goals I've set for the next year, hopefully I'll be pursuing my Ph.D too.</p>
<p>FQ: How did you cope with this situation?</p>
<p>A: During my final year B.E. project, we had decided to do a project on detection sign language using image processing and Machine Learning. I as a team lead, decided to use Neural Networks to be used as a classification mechanism for classifying images based on some features. After proceeding with the NN we found out that it was not feasible to implement it on a wearable device. Later we changed the decision from NN to SVM and achieved our goal.</p>
<p>FQ: What was your experience when you realized that you did not have the skills to implement a machine learning algorithm</p>
<p>A: When i passed my 12 standard i chose to take admission in one private engineering college rather to follow my family decision to take coaching for iits .I thought this was my failure because i thought all engineering colleges are not much different but later i realized that iits are way better than these colleges so on that very day i decided to qualify GATE exam to do my masters from one of the best college of India.</p>
<p>FQ: Why did you decide to take admission in one private engineering college rather than another?</p>

To explore further, we investigate followQG by generating many follow-ups on the same interview question-answer pair. As we see from Table VI, followQG generates a variety of questions addressing different parts of the question and answer. Though example questions 1 and 3 focus on the "challenges in the organization", they are rephrased differently. The other example questions attend to other concepts related to "organization".

TABLE VI. QUESTIONS GENERATED BY FOLLOWQG FOR A SINGLE QUESTION-ANSWER PAIR. WE CAN SEE THE VARIETY IN FOLLOW-UPS THOUGH THE QUESTION AND ANSWER INPUT IS THE SAME

<p>Explain how would you be an asset to this organisation?</p>
<p>A: My skills are a perfect match for the company's requirements. The company would definitely benefit from me as I have domain experience. I have a good idea about the company's goals. For me it's a place I can learn and be productive. I am young and skilled.</p>
<p>FQ: Do you think you will be able to overcome the challenges of the organisation?</p>
<p>FQ: What would you learn from the company?</p>
<p>FQ: Do you think that your skills are enough to be able to handle the challenges in the organization?</p>
<p>FQ: How would you handle the pressure?</p>

Multiple Follow-ups We further test followQG by examining its capability to generate multiple follow-up questions and hold a two or three turn interview. Since the model is trained only on a single pair of QA, the input to the model is only the previous question answer pair. Table VII shows a series of follow-up questions generated by

followQG. The first question is a pre-determined interview question. This is followed by follow-up generated with the original question and candidate response. The next follow-up is generated based on the previous follow-up and answer and so on. We can note that questions generated still hold relevant to the overall theme of the interview i.e., about the teamwork.

TABLE VII. MULTIPLE FOLLOW-UP QUESTIONS GENERATED BY FOLLOWQG AS A CONVERSATION AGENT. THE INPUT TO THE MODEL IS ONLY THE PREVIOUS QUESTION-ANSWER PAIR

Give an example of how you worked in team.
A: I have handled three teams together during my second semester of M.Tech for my course projects. I used to have complete discussion with the team before landing to a conclusion. When even I or someone got stuck at something we used to help each other to solve the issue. I managed the time to work flawlessly with all the teams.
FQ1: How do you manage team problems?
A: I motivate my team mates to complete the work and also in a perfect manner. It's important to be friendly with our team mates and I do the same. I divide the work into different parts and all the team members will share the equal work and I will be open to solve the doubts and get the work done.
FQ2: How do you motivate your teammates?
A: You can only motivate and inspire your team if they know what they are working towards and see the bigger picture. I will share the larger goal with the team and fire up the energy to keep motivated.
FQ3: Do you think you have enough energy?

V. MAYA - INTERACTIVE INTERVIEWING SYSTEM

Our interactive interviewing system, *Maya*, consists of two main components – 3D Virtual Interviewer and Interview Question Generator. The overall framework of *Maya* can be seen in Fig. 1. The first component is an Amazon Sumerian [46] based 3D virtual interviewing agent which asks questions and collects the interviewee's responses. We have defined behavior for the interviewer as a state machine component which makes the scene dynamic and interactive.

A. Interviewer's Behaviour

The state machine component defined for the host (interviewer) consists of 4 states: Initialisation, Maya Response, User Response and Wait.

Maya Response, which executes the HostSpeech script defines the host response and initiates the Speech Component. After configuring the speech body and voice, it plays the audio. We use Amazon Polly text-to-speech toolkit [82] service to synthesize speech at runtime. The host response is either a greeting followed with the first question or is the follow-up question based on the interviewee's response of the previous question.

Upon successful execution of the host speech, the state changes from Maya Response to User Response. The User Response state executes the UserSpeech script, which takes the interviewee's response to the question asked and returns an appropriate follow-up question. It uses the Web Speech API [80] to get the transcript of the interviewee's response. The word error rate of this ASR engine is 45.7, calculated on 5 randomly chosen videos from the Asynchronous Interview dataset at the utterance level.

Once the transcript is collected, it is fed to our Interview Question Generator hosted on a server using an API call. The response of the API call is a follow-up question which is set as the host's next response.

B. Interview Question Generator

The Interview Question Generator component contains two modules which communicates with the 3D virtual interviewer namely, Base question selector and followQG. Base question selector selects a question randomly from 200 questions commonly asked in an HR interview. Next question is a follow-up question generated by followQG. This repeats for a fixed number of times. In our experiments, we limit the number of follow-up question to one. The next base question is selected after one follow-up question. Hence the follow-up question is based on single previous response from the candidate and not the history. We consider one follow-up question as a proxy to planned or controlled probing and try to improve the interactivity and listening skills of the asynchronous interviewing system.

VI. DISCUSSION AND LIMITATIONS

After analyzing the quantitative and qualitative results of our study, we now discuss some reflections and limitations. Traditionally, asynchronous interview media do not enable interaction. To address this interactivity attribute of the medium, we propose follow-up question generation enabling one level of probing. Since structured interviews are known to reduce different biasing factors [4], we limit the level of probing to one. This balances the structure of the interview as well as conversational flow between the system and candidate.

We use a relatively small in-domain corpus of interview question, answer and follow-up question to train our model. In spite of being small, the dataset helps the model understand the nuances of the concept of follow-up question. With the power of knowledge from the pre-trained language model, the questions are generated to be appropriate follow-ups.

The use of knowledge from the large-scale transformer language model induces external knowledge generating diverse questions adhering to grammar. The data samples help FQG to learn the question structure and the relation between the triplets, and the knowledge from the language model pre-training produces novel questions. This model can be seen as the one with improved performance when compared with two strong baseline models. The model always generates grammatically correct questions and the quality of questions is enhanced.

The task of follow-up question generation differs from question generation majorly in one dimension. In question generation, the input to the model already has the answer to the question to be generated. Whereas in follow-up question generation, the question to be generated must seek more information or related information that is not present in the input to the model. The QG-Net baseline model is originally trained to address question generation and not follow-up generation. This drawback is evident in the questions generated from QG-Net which are already answered. The similarity-based question selector relies solely on the similarity techniques and is constrained to the pre-defined number of questions in the pool. We try to address these gaps in the existing asynchronous agent-based interviewing systems with followQG.

While our evaluations are encouraging, there are certain limitations to our system. Currently, we use a dataset that is small in size to fine-tune the model. Even though the model finetuned on a small dataset is producing appropriate follow-up questions, increasing the size of this dataset might improve the quality of the questions further. All the annotators and volunteers for crowdsourcing in this study are non-native English speakers. Obtaining the data or annotations from native English speakers can add variety to the mix.

A. Future Work

Below we discuss some of the future directions to our system.

1. Considering the History of the Interview Conversation

Currently, *Maya* considers only the previous question and answer to generate a follow-up. Taking into account the whole context of the interview and remembering the information from the previous answers will be helpful in improving the overall experience with the system. An immediate next step would be to train followQG with the history of the interview conversation to generate the next follow-up question.

2. Interpreting and Understanding Deeper Meaning

Though our current system asks follow-up questions addressing specific details, it can be seen as understanding the gist of the candidate response. To generate more meaningful questions, the system must extract deeper concepts and relationships among them. Conditioning the questions on the background of the candidate (like information from resume), an external knowledge source to improve commonsense reasoning can be potential directions.

3. Interrelating Question Topics

In the current system, *Maya* asks each question independently while in a real world scenario the questions are interrelated. It is necessary to explicitly bring out these relations.

4. Building Empathetic Interviewing Agents

An important aspect of an engaging human conversation that receives relatively less focus is emotional understanding and empathy. There is a need to include emotion into the interviewing agents and respond in a way that acknowledges the feelings of the candidates. The next step towards that would be to train followQG explicitly with emotion labels and emotionally coloured words.

VII. CONCLUSION

We introduce *Maya*, a virtual agent-based interviewing system equipped with verbal interactivity from follow-up question generation. We leverage the implicit knowledge of a large scale transformer language model fine-tuned on follow-up questions dataset to generate relevant, novel and diverse questions based on the candidates' response in an interview. With availability of limited data, this approach scales as it uses external knowledge from a language model trained on a huge corpus. With human evaluation, we show that the questions generated are of good quality. We can also see that the FQG model is often robust to the errors of speech recognition. We restrict the generation of follow-up questions to one as existing research suggests the advantage of limited probing and follow-up. But the model is capable of generating multiple follow-up questions based on the previous response. These positive results point to future work in extending the approach to a variety of other types of interviewing agents not limited to behavioural domain, as well as investigating whether followQG can be trained on any other domain descriptive questions to generate follow-up questions.

ACKNOWLEDGMENT

This work was partially funded by SERB Young Scientist grant (Grant no: YSS2015001074) of Dr. Jayagopi, Karnataka government's MINRO grant and a grant from Accenture Technology Labs. We would like to thank all the participants who contributed for data collection.

REFERENCES

- [1] talview.com, "Understanding recruitment troubles and trends," 2016. [Online]. Available: <https://info.talview.com/understanding-recruitment-troubles-trends-research-2016>.
- [2] J. E. Salmons, *Qualitative Online Interviews: Strategies, Design, and Skills*. Thousand Oaks, CA, USA: Sage Publications, Inc., 2nd ed., 2014.
- [3] S. Rasipuram, S. B. P. Rao, D. B. Jayagopi, "Asynchronous video interviews vs. face-to-face interviews for communication skill measurement: A systematic study," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI '16, New York, NY, USA, 2016, pp. 370–377, ACM.
- [4] J. Levashina, C. J. Hartwell, F. P. Morgeson, M. A. Campion, "The structured employment interview: Narrative and quantitative review of the research literature," *Personnel Psychology*, vol. 67, no. 1, pp. 241–293, 2014.
- [5] J. Li, M. X. Zhou, H. Yang, G. Mark, "Confiding in and listening to virtual agents: The effect of personality," in *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, IUI '17, New York, NY, USA, 2017, p. 275–286, Association for Computing Machinery.
- [6] C. R. Rogers, R. E. Farson, University of Chicago., *Industrial Relations Center, Active listening*. Chicago: Industrial Relations Center, the University of Chicago, 1957.
- [7] I. Sutskever, O. Vinyals, Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.
- [8] S. B. P. Rao, M. Agnihotri, D. B. Jayagopi, "Automatic follow-up question generation for asynchronous interviews," in *Proceedings of the 1st Workshop on Intelligent Information Processing and Natural Language Generation*, ECAI, 2020.
- [9] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, S. Bengio, "Why does unsupervised pre-training help deep learning?," *J. Mach. Learn. Res.*, vol. 11, p. 625–660, Mar. 2010.
- [10] H. Larochelle, Y. Bengio, J. Louradour, P. Lamblin, "Exploring strategies for training deep neural networks," *Journal of Machine Learning Research*, vol. 10, no. 1, pp. 1–40, 2009.
- [11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *CoRR*, 2018.
- [12] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, "Improving language understanding by generative pre-training," 2018.
- [13] J. Howard, S. Ruder, "Universal language model fine-tuning for text classification," in *ACL*, 2018.
- [14] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. S. Zettlemoyer, "Deep contextualized word representations," *ArXiv*, 2018.
- [15] Z. Huang, W. L. Xu, K. Yu, "Bidirectional lstm-crf models for sequence tagging," *ArXiv*, 2015.
- [16] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S. R. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," in *BlackboxNLP@EMNLP*, 2018.
- [17] T. Wolf, V. Sanh, J. Chaumond, C. Delangue, "Transfer-transfo: A transfer learning approach for neural network based conversational agents," *ArXiv*, 2019.
- [18] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Çelikyil-maz, Y. Choi, "Comet: Commonsense transformers for automatic knowledge graph construction," *ArXiv*, 2019.
- [19] S. B. P. Rao, S. Rasipuram, R. Das, D. B. Jayagopi, "Automatic assessment of communication skill in non-conventional interview settings: a comparative study," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017, pp. 221–229, ACM.
- [20] J. F. Nunamaker, D. C. Derrick, A. C. Elkins, J. K. Burgoon, M. W. Patton, "Embodied conversational agent-based kiosk for automated interviewing," *Journal of Management Information Systems*, vol. 28, no. 1, pp. 17–48, 2011.
- [21] M. E. Hoque, M. Courgeon, J.-C. Martin, B. Mutlu, R. W. Picard, "Mach: My automated conversation coach," in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, 2013, pp. 697–706, ACM.
- [22] K. Anderson, E. André, T. Baur, S. Bernardini, M. Chollet, E. Chryssafidou, I. Damian, C. Ennis, A. Egges, P. Gebhard, et al., "The tardis framework: intelligent virtual agents for social coaching in job interviews," in

International Conference on Advances in Computer Entertainment Technology, 2013, pp. 476–491, Springer.

- [23] T. Kawahara, “Spoken dialogue system for a human-like conversational robot erica,” in *International Workshop Spoken Dialogue Systems*, 2018.
- [24] V. Rus, A. C. Graesser, “The question generation shared task and evaluation challenge,” in *The University of Memphis. National Science Foundation*, 2009.
- [25] V. Rus, W. Brendan, P. Piwek, M. Lintean, S. Stoyanchev, C. Moldovan, “The question generation shared task and evaluation challenge,” in *The University of Memphis. National Science Foundation*, 2009.
- [26] I. Serban, A. Garcia-Durán, Çağlar Gülçehre, S. Ahn, A. P. S. Chandar, A. C. Courville, Y. Bengio, “Generating factoid questions with recurrent neural networks: The 30m factoid question-answer corpus,” *CoRR*, 2016.
- [27] X. Du, J. Shao, C. Cardie, “Learning to ask: Neural question generation for reading comprehension,” in *ACL*, 2017.
- [28] P. Rajpurkar, J. Zhang, K. Lopyrev, P. S. Liang, “Squad: 100,000+ questions for machine comprehension of text,” in *EMNLP*, 2016.
- [29] Z. Wang, A. S. Lan, W. Nie, A. E. Waters, P. J. Grimaldi, R. G. Baraniuk, “Qg-net: a data-driven question generation model for educational content,” in *L@S*, 2018.
- [30] M.-H. Su, C.-H. Wu, K.-Y. Huang, Q.-B. Hong, H.-H. Huang, “Follow-up question generation using pattern-based seq2seq with a small corpus for interview coaching,” in *Proc. Interspeech 2018*, 2018, pp. 1006–1010.
- [31] X. Qiu, X. Huang, “Convolutional neural tensor network architecture for community-based question answering,” in *IJCAL*, 2015.
- [32] M.-H. Su, C.-H. Wu, Y. Chang, “Follow-up question generation using neural tensor network-based domain ontology population in an interview coaching system,” in *INTER-SPEECH*, 2019.
- [33] W. Swartout, R. Artstein, E. Forbell, S. Foutz, H. Lane, B. Lange, J. Morie, D. Noren, S. Rizzo, D. Traum, “Virtual humans for learning,” *AI Magazine*, vol. 34, pp. 13–30, 1 2013, doi: 10.1609/aimag.v34i4.2487.
- [34] R. López-Cózar, Z. Callejas, D. Griol, J. F. Quesada, “Review of spoken dialogue systems,” *Loquens*, vol. 1, no. 2, p. 012, 2014.
- [35] L. S. Nguyen, D. Frauendorfer, M. S. Mast, D. Gatica-Perez, “Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior,” *IEEE transactions on multimedia*.
- [36] H. Tanaka, S. Sakti, G. Neubig, T. Toda, H. Negoro, H. Iwasaka, S. Nakamura, “Automated social skills trainer,” in *Proceedings of the 20th International Conference on Intelligent User Interfaces*, 2015, pp. 17–27, ACM.
- [37] L. Hemamou, G. Felhi, V. Vandenbussche, J.-C. Martin, C. Clavel, “Hirenet: a hierarchical attention model for the automatic analysis of asynchronous video job interviews,” 2019.
- [38] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, “Language models are unsupervised multitask learners,” 2019.
- [39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, “Attention is all you need,” in *NIPS*, 2017.
- [40] A. Fan, M. Lewis, Y. Dauphin, “Hierarchical neural story generation,” in *ACL*, 2018.
- [41] Y. Bengio, R. Ducharme, P. Vincent, C. Jauvin, “A neural probabilistic language model,” *Journal of machine learning research*, vol. 3, no. Feb, pp. 1137–1155, 2003.
- [42] J. Pennington, R. Socher, C. D. Manning, “Glove: Global vectors for word representation,” in *EMNLP*, 2014.
- [43] A. See, P. J. Liu, C. D. Manning, “Get to the point: Summarization with pointer-generator networks,” in *ACL*, 2017.
- [44] W. Hu, B. Liu, J. Ma, D. Zhao, R. Yan, “Aspect-based question generation,” in *ICLR Workshop*, 2018.
- [45] G. Shires, “Web speech api: Draft community group report,” July 2019. [Online]. Available: <https://w3c.github.io/speech-api/>, [Online; posted 17-July-2019].
- [46] T. Walker, “Presenting amazon sumerian: An easy way to create vr, ar, and 3d experiences,” November 2017. [Online]. Available: <https://aws.amazon.com/blogs/aws/launch-presenting-amazon-sumerian/>, [Online; posted 27-November-2017].



Pooja Rao S B

Pooja Rao S B obtained her Master’s by Research degree from IIIT Bangalore in 2019. She is currently a doctoral student at University of Lausanne. Her current research interests lie in the intersection of Human-Computer Interaction and Machine Learning with a focus on designing and building intelligent systems capable of multimodal processing, natural language processing, understanding

and generation.



Manish Agnihotri

Manish Agnihotri obtained his Bachelor of Technology degree from MIT Manipal in 2019. He is currently a Machine Learning Engineer at Merlin AI, Zycus. His research interests lie in the areas of machine learning and affective computing with a focus on natural language processing and multimedia analysis.



Dinesh Babu Jayagopi

Dr. Dinesh Babu Jayagopi is an Associate Professor at IIIT Bangalore since Dec 2013, where he heads the Multimodal Perception Lab. His research interests are in Audio-Visual Signal Processing, Applied Machine Learning, and Social Computing. He obtained his doctorate from Ecole Polytechnic Federale Lausanne (EPFL), Switzerland, beginning of 2011. He received the Outstanding paper award in the International Conference on Multimodal Interaction (ICMI), 2012, Idiaph PhD student research award for the year 2009. Subsequently, his research papers with students has received Best paper awards and nominations. He has also received funding from several agencies including DST, CAIR(DRDO) and Accenture.