IMPROVING ASYNCHRONOUS INTERVIEW SYSTEM WITH AUTOMATIC ASSESSMENT AND FOLLOW-UP QUESTION GENERATION

Pooja Rao S. B

Master of Science by Research Thesis

November 2019

International Institute of Information Technology, Bangalore

IMPROVING ASYNCHRONOUS INTERVIEW SYSTEM WITH AUTOMATIC ASSESSMENT AND FOLLOW-UP QUESTION GENERATION

Submitted to International Institute of Information Technology,
Bangalore
in Partial Fulfillment of
the Requirements for the Award of
Master of Science by Research

by

Pooja Rao S. B
MS2015009

International Institute of Information Technology, Bangalore
November 2019

Dedicated to

Amma and Appa

Thesis Certificate

This is to certify that the thesis titled **Improving Asynchronous Interview System with Automatic Assessment and Follow-up Question Generation** submitted to the International Institute of Information Technology, Bangalore, for the award of the degree of **Master of Science by Research** is a bona fide record of the research work done by **Pooja Rao S. B**, **MS2015009**, under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Prof. Dinesh Babu Jayagopi

Bangalore,

The 5th of November, 2019.

IMPROVING ASYNCHRONOUS INTERVIEW SYSTEM WITH AUTOMATIC ASSESSMENT AND FOLLOW-UP QUESTION GENERATION

Abstract

The current generation of machine learning algorithms provides systems with the ability to automatically learn and improve from experience without using explicit instructions, relying on patterns and inference from the data. Given these capabilities, natural language processing is increasingly applied to new tasks, new domains, and new languages. The recent advent of machine learning in the field of recruitment has enabled automation of the process. Asynchronous interviews offer ubiquitous interviewing and coaching. They are becoming the standard for the first round of screening providing organizations with an adept, fair, and structured method for conducting interviews.

Effective communication is an important social skill facilitating us to interpret and connect with people and is of utmost importance in employment-based behavioural interviews. Assessment of the communication skill in these interviews is a vital task.

This thesis works towards adding automated elements to asynchronous interview systems. We make two major contributions towards the automation of asynchronous behavioural interviews. Firstly, we build a computational model to predict the communication skill of candidates in non-conventional asynchronous interview settings. We present a comparative analysis of the automatic communication skill prediction and expert perception of the candidates over different settings. Secondly, we propose two approaches for automatic Follow-up Question Generation given the interviewer question and the candidate response to add the interactivity element to asynchronous interviews.

Acknowledgements

This thesis is a result of a long journey through unknowns and over unfamiliar peaks. I am thankful for the many people I got to meet and who accompanied me on the way. First, I'd like to thank my supervisor Prof. Dinesh Babu Jayagopi. The freedom you gave me to pursue my interests and follow my own pace has made all the difference. Thank you for your advice, patience, support and guidance. I would also like to thank you and the IIIT Bangalore institute for nurturing a productive research environment.

I want to thank my husband Santhosh for being my guide, pillar of strength and motivator. I am also grateful to my parents for supporting me in my further studies. I would like to thank my brother for the initial motivation. I thank my family for their constant assistance.

I like to thank all the students of the Multimodal Perception Lab, who helped me in various ways all along. I am particularly grateful to Sowmya Rasipuram for her timely discussions and involvement. I sincerely thank Pooja Venkatesh, Chinchu Thomas, Annapurna Sharma, Rahul Das, Shruthi Nambiar. I personally thank all the students of our institute and my friends for their participation in the data collection process. I am grateful to Santhosh for designing and developing the web application for the follow-up data collection.

I would also like to thank the examiners in advance for their time.

List of Publications

(1) **S. B. P. Rao**, M. Agnihotri, and D. B. Jayagopi, "Incorporating Follow-Up Questioning in Virtual Agent-Based Employment Interviews" Submitted to Fourteenth International Conference on Tangible, Embedded, and Embodied Interaction 2020.

(2) **S. B. P. Rao**, S. Rasipuram, R. Das, and 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, ICMI. New York, USA: ACM, 2017, pp. 221–229.

(3) S. Rasipuram, R. Das, **S. B. P. Rao**, and D. B. Jayagopi, "Online peer-to-peer discussions: A platform for automatic assessment of communication skill," In Proceedings of Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, ACIIW. San Antonio, Texas: IEEE, 2017, pp. 68-73.

(4) S. Rasipuram, **S. B. P. Rao**, and D. B. Jayagopi, "Automatic prediction of fluency in interface-based interviews." In IEEE Annual India Conference, INDICON. Bangalore, India: IEEE, 2016, pp. 1-6.

(5) S. Rasipuram, **S. B. P. Rao**, and 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. New York, USA: ACM, 2016, pp. 370–377.

Contents

List of Figures

List of Tables

List of Abbreviations

AQAD . . . . . Amazon Question Answer Dataset

API . . . . . Application Programming Interface

ASR . . . . . Automatic Speech Recognition

BERT . . . . . Bidirectional Encoder Representations from Transformers

FQ . . . . . Follow-up Question]

FQG . . . . . Follow-up Question Generation

GPT . . . . . Generative Pre-Training

HR . . . . . Human Resource

LIWC . . . . . Linguistic Inquiry and Word Count

QG . . . . . Question Generation

Seq2Seq . . . . . Sequence-to-Sequence

SQuAD . . . . . Stanford Question Answering Dataset

CHAPTER 1

INTRODUCTION

1.1 Motivation

As technology continues to evolve, it is changing the ways businesses create and capture value, the way we work, interact and communicate. Machine Learning is one of the five technologies that is transforming the very foundations of global business and the organisations that drive it [1]. Not only it is empowering people to do things better and faster, it also facilitating profound changes in how organizations work.

Recruitment is one such area undergoing disruption by Artificial Intelligence. The 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 heeding to futuristic choices like social recruitment, online assessments, and video interviews [2]. Organisations are adopting innovative methods like social media, proctored assessments, asynchronous or one-way interviews. Many automatic talent assessment solutions are gaining attention like

Talview[1], Hirevue[2] , Sonru[3]. Such platforms offer asynchronous, ubiquitous interviewing and screening.

1.1.1 Asynchronous Interviews

Traditional face-to-face interviews are always synchronous. They need to occur at the same time and in the same place. On the other hand, 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. [3] 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 interviews allow parallel processing of candidates. They provide the advantage of taking the test at the candidate's convenience and facilitate efficient screening with minimal human intervention. They are very much scalable and time-saving. Also, with the help of machine learning techniques, automatic assessment of such interviews is feasible. Automatically assessed asynchronous interviews complete the process automation in the interview systems. They are useful for initial screening as well as interview coaching. They are good alternatives to initial phone interview screening, premature in-person interviews and resume scanning.

Automatic assessment of communication skill is an intriguing problem in social computing. Communication is the ability to convey information to others effectively

[1] www.talview.com
[2] www.hirevue.com
[3] www.sonru.com

and efficiently. Expertise in communication skill can help all aspects of life, from professional settings to social gatherings and everything in between. Excellence in various types of communication is a vital requirement for any job profile of a knowledge worker. The different types may be verbal, written and non-verbal. Communication skills are not limited to direct interaction with other people and the spoken word. The ability to write clearly and effectively is also a key. Asynchronous interviews equipped with automatic communication skill assessment can be particularly useful for initial screening of a knowledge worker.

It is also necessary that these asynchronous systems are interactive and lively to ensure the smooth progress of an interview. They should be conversational and acknowledge the response of the participant. Most of the existing systems pose questions which are predefined or randomly selected from a limited set. A follow-up question generation system promises to ameliorate this inadequacy.

In this thesis, we address these features and concerns of an asynchronous interview system. We analyse the administration medium of asynchronous interview system facilitating smooth interviews between the employer and prospective employees in terms of different attributes.

1.2 Research Objectives

Potosky [4] describes the administration of personnel tests and assessments as a communication exchange process, and defines the administration medium in terms of four attributes:

1. **Transparency** - the extent to which the medium facilitates clear and unobstructed communication exchange
2. **Social bandwidth** - the capacity for data transfer
3. **Interactivity** - the pace of mutual or reciprocal exchange between communicating parties
4. **Surveillance** - the extent to which an outside party can monitor or intercept messages carried by the medium for test administration.

The administration medium might affect the construct(s) being measured as well as the assessment outcomes [1]. The asynchronous interview is an administration medium which does not require the interviewee and the employer to be present at the same time. The interviewee will have a one-way conversation or a conversation with an agent. The recorded interviews are processed through different channels like audio, video, and text for further assessment.

In this thesis, the objective is to study these attributes of transparency, social bandwidth and interactivity in the administration medium of asynchronous interviews and address the inadequacies. The study of the surveillance attribute is beyond the scope of this thesis.

1. Potosky [4] states that ratings about the extent to which the medium enabled the individuals to ask or respond to questions as intended, might provide a reasonable

index of media transparency. We present a feedback questionnaire to obtain this rating from the participants of the asynchronous interviews. We also present a set of annotations of the interviews from expert HR personnel, which can be viewed as the perceived measure of unobstructed communication exchange. The ratings thus obtained is used to analyse and compare across the settings.

2. A measure of social bandwidth may be best represented as the number of kinds of social cues included in an assessment using a medium [4]. To estimate this, we introduce a predictive model using automatically extracted multimodal features from audio, visual and lexical cues. A comparative study of the predictions from the model allows for the investigation of this attribute.

3. Interactivity refers to the pace of mutual or reciprocal exchange (i.e., turn taking) between communicating parties. Asynchronous communication does not enable coordinated turn-taking by interactants [4]. We propose an automatic follow-up question generation model which can facilitate turn taking to ensure a conversational flow.

In addition to the objective of studying these attributes of the administration medium, we also examine the feasibility of automatic prediction of communication skill. We predict both oral as well as written communication of the candidates using a model trained on the Interview dataset (Section 1.3). This prediction module can be helpful in the initial screening of candidates or interview coaching.

1.3 Contributions

The main contributions of this thesis are

1. **Asynchronous Interview Dataset:** A dataset of 100 participants with video and written interviews with short essays annotated independently for communication

skill by two human expert annotators. An interface-based web application is used for the data collection process. We present this unique dataset of video and written interviews with short essays obtained from the same participants [5]

2. **Automated assessment of communication skill of participants:** We present a systematic study and automatic measurement of the communication skill of candidates in different modes of behavioural interviews. We demonstrate a comparative analysis of non-conventional methods of employment interviews namely

   - Interface-based asynchronous written interviews
   - A short essay
   - Interface-based asynchronous video interviews

   in terms of behavioural perception and automated predictions. We study the behaviour and performance changes of the participants over these modes. We hypothesise that the performance is independent of the administration medium [5].

3. **Automatic follow-up question generation:** We showcase two methods to generate follow-up questions given the interviewer question and the candidate response in an asynchronous HR interview setting. Follow-ups make the system more interactive and ensure the smooth progress of the interview. Since the dataset used for training is small, we employ techniques to utilize external datsets and knowledge for automatic follow-up question generation. This eliminates the need for a considerably significant training corpus.

   We also present a follow-up question corpus collected via crowdsourcing. The contributors manually composed follow-up questions for a given interviewer question and candidate answer pair. It is a small corpus, consisting of 1089 triples of question, answer and follow-up question.

1.4 Thesis Outline

In chapter 2, we present an extensive literature survey of the automatic assessment and coaching of candidates in interview scenarios, automatic content scoring and question generation.

In chapter 3, we perform an automated assessment of the communication skill of participants in different asynchronous settings. We analyse and compare the communication skill behaviour of the participants across the settings.

In chapter 4, we propose approaches to automatic follow-up question generation to maintain the interactiveness of the interview.

Chapter 5 concisely outlines the collaboration work.

Chapter 6 finally concludes this thesis, where we summarize our findings and provide an outlook into the future.

CHAPTER 2

LITERATURE SURVEY

This chapter describes the literature related to the different components that we use towards an automated asynchronous interview system. We propose studies, methods and models which require knowledge in various research fields like natural language processing and generation, machine learning and deep learning, multimodal analysis and social computing. We also derive some knowledge from social and organisational psychology to aid our research.

Cascio and Montealegre [1] state that, " A researcher who wishes to study the effects of technology on test performance must consider not only differences in the mode of administration (e.g., paper-and-pencil versus computer) but also the mode of test delivery (e.g., face-to-face interviews versus remote, videoconference interviews). Either of these might affect the construct(s) being measured as well as assessment outcomes; other contextual, strategic, or environmental effects may, as well. " In the asynchronous medium of administration, we examine the different modes of test delivery, namely oral and written.

Our variable of interest is communication skill, as this is vital for success in any employment interview [6]. Our work not only focuses on oral communication skill but also compares it with a non-conventional type of interview measuring the written communication of a candidate. We try to find out if this can be a feasible alternative to the

interface based video interview in case of infrastructure issues like network bandwidth.

The design of an automated interview scoring system for communication skills brings together several fields, including research into (computers in education) automatic video assessment, automated content scoring including automated short answer grading and automated essay scoring, affective computing and question generation. Section 2.1 describes the works in automated short answer scoing, and Section 2.2 talks about automated essay scoring. Section 2.3 briefly discusses the literature in automated video assessment. Section 2.4 explains the work in question generation.

2.1 Automated Short Answer Scoring

Automatic evaluation of written communication in interview scenarios is a less explored area. It can be aligned with the field of automatic content scoring. Research in computationally scoring natural language responses has a history dating back to the work of Page [7]. Burrows et al. give a comprehensive review of the work in the area of automatic short answer grading [8]. *C-rater* is one of the ETS's early systems, which scores constructed responses against targeted questions. Constructed responses are scored by their content rather than the quality of writing [9]. Many state-of-the-art systems and even few systems lately expect manually worked patterns to score the answers [10] [11]. Ramachandran et al. suggest an approach to extract patterns from the reference answers automatically [12]. Others involve text-to-text similarity measures between the students' answers and the standard/reference answers [13] [14]. Higgins et al. go beyond the question specific manual features and explore the usage of syntactically informed features [15]. Riordan et al. [16] investigate neural network models for short answer scoring on three different datasets.

In short answer scoring tasks, questions are domain-specific and have appropriate distinct answers. In many cases, there exists a gold standard or a reference answer

against which the student answers are compared. In our task of assessing the written communication skill of candidates, the answers to behavioural questions are unseen and new for each participant and hence cannot be trained on the manual patterns. Also, due to the high subjectivity of the domain, there can be no standard/reference answers.

2.2 Automated Essay Scoring

Automated essay scoring is an important text evaluation area and a well-researched field. Project Essay Grader [17] was one of the pioneers of automated text evaluation, solely based on the quality of writing. Many more successful models like the E-rater [18], Intelligent Essay Assessor [19], Bayesian Essay Test Scoring System [20] and were developed. E-rater, built by ETS, uses regression models coupled with natural language processing techniques of feature extraction and achieved efficiency in the range of 87-94% against human-based scores that were taken to be the ground truth.

Recently, there are works which take the neural network approach to essay grading. Aikaniotis et al. [21] explore an augmented version of the C&W embeddings [22] called score-specific word embeddings with different LSTM architectures to score the essays. Taghipour and Ng [23] take the recurrent neural network approach with an ensemble of CNN and unidirectional LSTM.

2.3 Automated Video Assessment

Automated video assessment has spurred various research interests lately. Studies address many social variables. Batrinca et al. present the automatic prediction of personality extracting acoustic features and manually extracted visual features [24]. Biel et al. predict the Big Five personality traits from facial expressions of emotions in online conversational videos [25]. Nguyen et al. present a computational framework for

the automatic prediction of hirability in real employment interviews by extracting non-verbal cues from both the interviewee and the interviewer [26]. Rasipuram et al. predict the communication skill in employment interviews using the audio-visual cues [27]. Another study [28] uses the MIT Interview dataset to predict overall job performance (overall rating and recommended hiring) and 14 different social traits like excitement, friendliness, and engagement with promising results. Rasipuram and others propose a dual dataset of asynchronous and face-to-face interviews and analyse the differences between them, predicting the communication skill [29]. There are also many tools to help with training or improving social or presentation skills. MACH [30], Rhema [31], ROC Speak [32], Automanner [33] and Automated Social Skills Trainer [34] to mention a few.

Our work is an amalgamation of all the three systems described - automated short answer scoring, automated essay scoring and automated video assessment; and the aim is to automatically predict the different types of communication skill and also to compare them across the three settings - written interview, short essay and video interview.

2.4 Question Generation

Question Generation (QG) has been defined as the task of automatically generating questions from some form of text input and has attracted attention since the First Question Generation Shared Task Evaluation Challenge [35]. More recently, neural networks have enabled the end-to-end training of question generation systems influenced by the sequence-to-sequence (Seq2Seq) data-driven learning methods [36]. Serban et al., [37] train a neural system to generate simple natural language questions from the structured triples of subject, relation, object. There have been studies which have successfully tried to expand this to unstructured text [38] [39]. Du et al., [38] use encoder-decoder model with attention to generate questions on the machine compre-

hension dataset SQuAD [40]. Yuan et al., [41] also apply reinforcement learning techniques to natural language generation, specifically question generation. QG-net [42] is an RNN-based encoder-decoder model designed to generate questions from the educational content. SQuAD machine comprehension dataset is used to train the model.

Considering that automatic follow-up question generation in an interview setting is a new task, there has been one study which investigates this [43]. Su et al., adopt a pattern-based Seq2Seq model on a small interview corpus collected in Chinese. They use a word clustering method to build a word class table and transform all the sentences in the database to patterns. CNTN-based [44] sentence selection model is implemented to select a sentence from the answer to generate follow-up question patterns. These generated patterns are filled with words from the word class table to obtain potential candidates for follow-ups. A statistical language model is used to rank these candidates and choose a question.

In contrast, we develop a follow-up question generation model utilizing knowledge from external data sources. The first approach utilizes external general-purpose dataset to train a Seq2Seq question generation model. This model adapts successfully to follow-up question generation without further fine-tuning. The second approach utilizes knowledge from large-scale language model and the small follow-up question corpus to generate follow-up questions. The questions generated in natural language avoids the restraint of a pattern-based model and a small vocabulary without the need for pattern matching and template filling.

Pre-training on enormous amounts of text data in an unsupervised fashion has led to state-of-the-art advancements on a variety of natural language processing tasks [45] [46]. Currently, these pre-training steps are all the variants of language modelling objectives. Howard and Ruder [47] train a language model on huge amounts of Wikipedia data and fine-tune this model on a specific target task with a smaller amount of labelled

in-domain data. Several other works follow this approach of fine-tuning and achieve impressive results. ELMo [48] is a bidirectional language model predicting the next and the previous tokens using bi-directional Long Short-Term Memory Networks [49]. OpenAI's GPT [46] train a unidirectional language model on massive amounts of text data. BERT [45] is a masked language model trained with an additional objective of next sentence prediction. All of these models have attained state-of-the-art results on many downstream NLP tasks including the GLUE benchmark [50].

Pre-training with GPT model has also been used in generative tasks such as end-to-end dialog systems [51] and automatic knowledge base completion [52] obtaining remarkable improvements over the models trained only with the in-domain data. Both the works use the transformer language model GPT for initialization. Our work builds on this to develop a Follow-up Question Generation model.

2.5 Conclusion

In this chapter, we have reviewed the literature in the fields of automated short answer scoring, automated essay scoring, automatic video assessment and question generation. The next chapter relies on these for automatic communication skill prediction in non-conventional interview settings like asynchronous written and video interviews.

CHAPTER 3

AUTOMATIC ASSESSMENT OF COMMUNICATION SKILL AND COMPARATIVE STUDY OF NON-CONVENTIONAL INTERVIEW SETTINGS

The experience of an asynchronous interview system will be seamless if it facilitates unobstructed communication and information exchange. In this chapter, we study the social bandwidth and transparency attributes. We build a predictive model by extracting multimodal features from the interviews to predict the communication skill of a participant.

The overall flow of the procedure is shown in Figure FC3.1. Our framework has three key phases.

1. Written Interview
2. Short Essay
3. Video Interview

All these settings qualify as asynchronous since the candidates can participate at any time and do not require the presence of an interviewer. The first part of this chapter (3.1, 3.2, 3.3) describes the automatic prediction of our variable of interest - effective communication in the interview settings, as mentioned above. We mainly concentrate

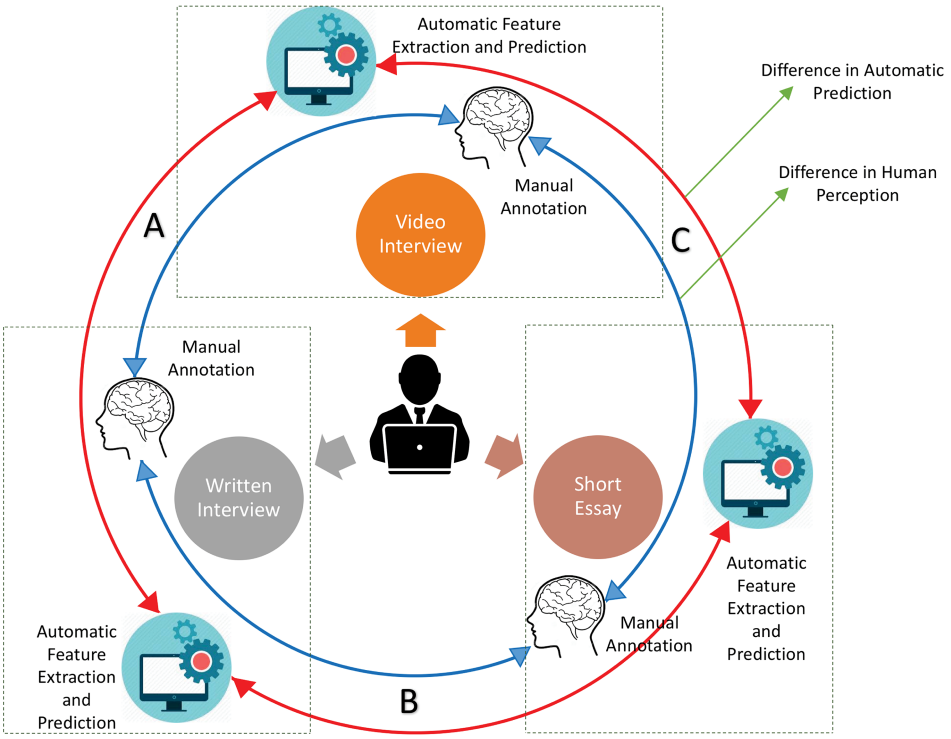

Figure FC3.1: Non-conventional Interview System. Blue lines - Differences in human perception, Red lines - Differences in automatic features and prediction, (A) Between the Video and Written interview (B) Between Written interview and Short Essay (C) Between Video interview and Short Essay

on written communication and compare it against the oral communication. The section 3.4 presents a comparative analysis of the three non-conventional settings [5].

The job profile of any knowledge worker demands a good command on one's communication. Written communication skill is the predictive variable in the case of written interviews and essays and oral or spoken communication skill in case of video interviews. We perform a comprehensive study on the written communication prediction and its comparison. This process takes place in three stages.

1. Data Collection and Labelling (3.1)
2. Feature Extraction (3.2)
3. Automatic Prediction (3.3)

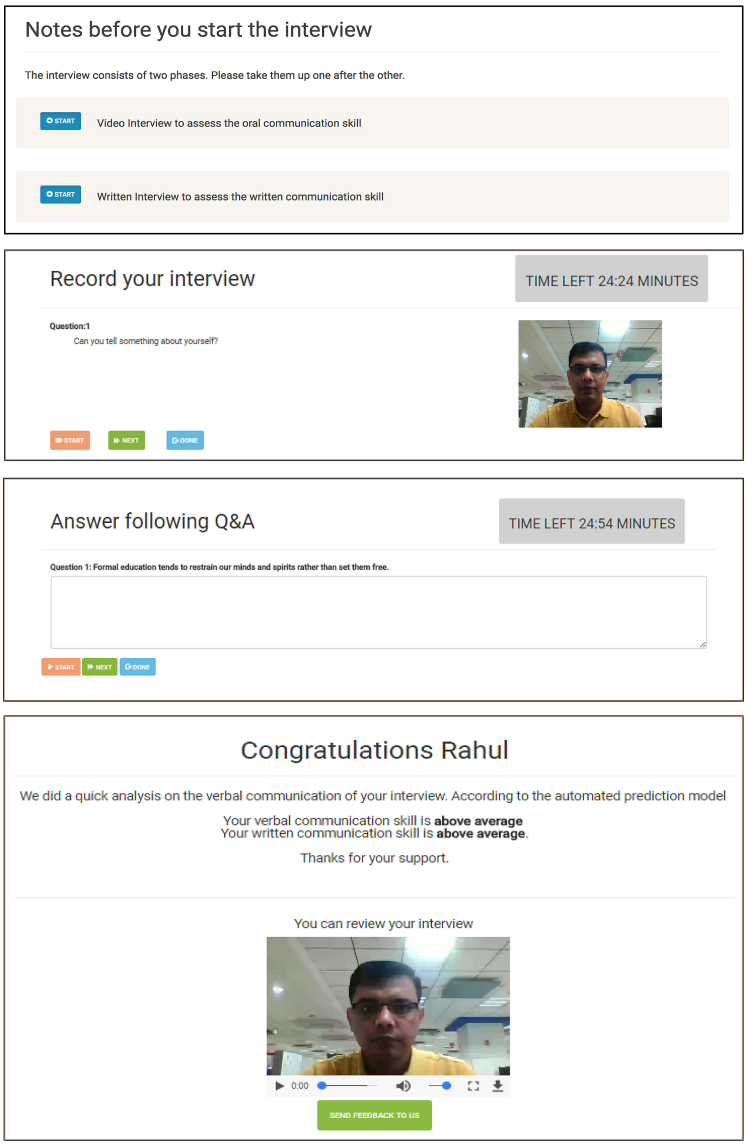

Figure FC3.2: Snapshots of the custom-built interface

3.1 Data Collection with Web Interface

Figure FC3.2 depicts the snapshots of the web interface scientifically used for data collection. The participant accesses the custom-built web application to start the interview process. He/she can start with either of the two options for the interview, as shown in the figure. The short essay is combined with the written phase. Each participant is asked five behavioural questions generally asked in any HR interview in both the video and the written modes. Participant views one question at a time and answers spontaneously before proceeding to the next question. The first question is always 'Tell me about yourself/Short self-presentation' (video and written), followed by four questions (totally five questions) randomly picked from a set of 75 behavioural questions. 'Describe a time when your workload was heavy and how you handled it.', 'What is the biggest challenge you have faced so far?', 'What do people most often criticise about you?' are some of the examples of the questions posed. Each of the video and the written interview phase is timed at a maximum of 25 minutes, considering 5 minutes for a question.

The sixth question in the written phase is an essay topic chosen randomly from a set of 10 topics. These essay questions are persuasive or argumentative type of essays where the participant is expected to write a short essay of fewer than 400 words by taking a stance and convincing the reader or making an appropriate argument. The time interval for this is a maximum of 30 minutes. At the end of the interview process, we present an automatically predicted result with three classes - below average, average and above average, as shown in the last snapshot of Figure FC3.2. After the automatic prediction, we ask a few feedback questions about the interface to the participants and their chosen preference among the three settings.

Though the format of the questions in video and written interviews are the same, the setting itself makes the difference. The mode of communication for conveying the

thoughts of the participant is different. The oral and writing skills of the participants are tested, whereas the written interview and short essay differ on the dimension of time and question type.

3.1.1 Data Labelling

We have collected 100 audio-visual recordings, 118 written interviews (118*5=590 answers), and 106 short essays. (The video interviews and the essays were also collected from more than 100 participants. Due to the technical quality issues of the videos and essays, the numbers are reduced). The total duration of the video interviews is around 512 minutes, with an average of approximately 5 minutes per interview. The written interview corpus has approximately 2,500 sentences and 41,000 words at an average of 70 words per answer. The short essay corpus has around 1,300 sentences and 25,000 words in total, with an average of 240 words per essay. All the participants are students from IIIT Bangalore in the age group of 18-23 years.

The data collected is submitted for manual annotation to two experts in the field of Human Resource and Development with experience in conducting interviews. Each expert rates the video, written interviews and the short essay on a scale of 1 to 5 based on rubrics specific to that setting. Table TC3.1 shows the rubrics and the inter-rater agreement, Cohen's Kappa k statistic between the two raters. The exact description of the questionnaire posed to the expert raters can be found in the appendix. To account for the subjectivity of the human ratings, we take the average of the two ratings.

Figure FC3.3 shows the data distribution across the three settings for the average ratings from two annotators. Since the data is skewed and there are no candidates fall in score 1, we consider the binary classification task of separating the capable and competent communicators from the average and below average class of communicators. To accomplish this, we combine the rating $\{1.5,2,2.5,3\}$ as poor and the average class

Average Rating Scale	1.5	2	2.5	3	3.5	4	4.5	5	Total
Video Interview	2	17	14	26	19	17	4	1	100
Written Interview	2	9	17	31	23	20	13	3	118
Short Essay	1	8	18	26	26	19	5	3	106

Class 1 Class 2

Binary Classification	Class 1	Class 2	Total
Video Interview	59	41	100
Written Interview	59	59	118
Short Essay	53	53	106

Figure FC3.3: Distribution of the data across the three settings.

and $\{3.5,4,4.5,5\}$ as the competent and capable class. Figure FC3.3 also shows the distribution of the binary classification.

3.1.2 Rubrics Analysis

Table TC3.1 shows the inter-rater agreement between the two experts for the three modes of the interview. It is evident from the table that they agree differently on different rubrics. The overall spoken and written communication ratings have a reasonable agreement of 0.4. This disagreement is effectively minimised by combining the ratings into binary classes. The rubrics on the essay has a relatively higher agreement with an agreement of 0.7 on effective communication which might indicate that the experts might agree more with the availability of the more content on a single subject to gauge the participants. The experts also agree more on other rubrics like the use of eye contact, speaking fluency, writing fluency and grammar. One note here is that the content relevance rubric has a negative value for both the video as well as the written interview, but the ideas and content has high Kappa in case of the essay. The behavioural questions in an interview are very subjective, and each expert might interpret the answer differently. Whereas in the case of essays, because of more information, the flow of ideas and content might be more straightforward to perceive.

Table TC3.1: Inter-rater agreement between the raters for different rubrics

Video Interview		Written Interview		Short Essay	
Rubrics	*Kappa*	*Rubrics*	*Kappa*	*Rubrics*	*Kappa*
Speaking fluency	0.45	Writing fluency	0.42	Writing Fluency	0.65
Articulation	0.28	Grammar	0.45	Grammar	0.58
Use of eye contact	0.56	Conventions and Mechanics	0.33	Conventions and mechanics	0.61
Facial expressiveness	0.14	Convincing	0.34	Ideas and Content	0.75
Convincing	0.32	Confidence	0.25	Organisation	0.63
Confidence	0.43	Word usage/Vocabulary	0.31	Word Usage/Vocabulary	0.58
Word usage/Vocabulary	0.01	Content relevance	-0.26	Effective Communication	0.71
Content relevance	-0.29	Overall written communication skill	0.41		
Overall spoken communication skill	0.40				

Table TC3.2 shows the pairwise correlation coefficients for each rubric with the overall communication skill in all three modes. The correlations are computed on the average ratings. The attributes convincing, confidence and speaking fluency show high correlation with the spoken communication skill. This confirms the fact that non-verbal communication or body language is of importance in spoken communication, especially in interviews. Vocabulary and content relevance are also vital, indicating the necessity of lexical sophistication.

Writing fluency, vocabulary and content relevance rubrics are outstanding in the case of written communication. Convincing and confidence in the written answers are also distinguishing attributes in the perception of a human annotator. This affirms that content and word usage are prominent, but the confidence and convincing required in an interview scenario holds even in written interviews.

In the case of short essays, the ideas and content and organisation are substantial to communicate the views effectively. It is closely followed by writing fluency and vocabulary. The main difference here is that along with the flow of the read, the ideas and organisation play a major role, given the increase in the duration to answer. It is important to note that the attributes fluency, convincing and word usage, be it spoken or written communication is essential in all the three modes.

Table TC3.2: Pearson's correlation coefficient of rubrics with overall communication skill

Video Interview		Written Interview		Short Essay	
Rubrics	*r*	*Rubrics*	*r*	*Rubrics*	*r*
Speaking fluency	**0.88**	Writing fluency	**0.86**	Writing Fluency	**0.89**
Articulation	0.77	Grammar	0.76	Grammar	0.84
Use of eye contact	0.63	Conventions and Mechanics	0.74	Conventions and mechanics	0.81
Facial expressiveness	0.71	Convincing	**0.86**	Ideas and Content	**0.95**
Convincing	**0.93**	Confidence	0.80	Organisation	**0.9**
Confidence	**0.86**	Word usage/Vocabulary	**0.84**	Word Usage/Vocabulary	0.87
Word usage/Vocabulary	**0.86**	Content relevance	0.80		
Content relevance	0.85				

3.1.3 Prediction using Rubrics

To further validate the importance of the rubrics for the communication skill, we perform automatic prediction on the binary classification task defined earlier. Figure FC3.3 shows the data distribution for the three settings. We use an average of the two expert ratings to train an XGBoost classifier model [53]. Table TC3.3 shows the prediction results for three modes of communication using manually annotated rubrics as features. The metrics are a result of 5-fold cross-validation. We obtain an accuracy of 96%, 92% and 88% for the spoken, written and essay communication skill respectively.

The rubrics convincing, speaking fluency and content relevance are shown to be critical for prediction of spoken communication skill. This prediction further validates the correlation analysis findings that, along with non-verbal communication, a good grip over the lexical information is necessary. Writing fluency, convincing and grammar are found to be distinctive features for the written communication skill. The main features for predicting effective communication in an essay are writing fluency, ideas & content and standard conventions & mechanics. As was the case in correlation analysis, the fluency rubric stands out in all the three settings.

Table TC3.3: Communication Skill rating prediction using rubrics

Mode of Interview	Accuracy	F1 score
Video	0.96	0.97
Written	0.92	0.93
Essay	0.88	0.89

3.2 Feature Extraction

This section describes the wide set of features used in each of the settings separately.

3.2.1 Features for Communication Skill Prediction in Written Interviews and Short Essays

Lexical Features

We utilise four types of lexical features for prediction. Length features like the average sentence length, average word length, number of sentences, the total number of words, the average length of each answer and longest sentence length are used. The vocabulary is an important rubric as we saw in Section 3.1.2. Vocabulary based features like type-to-token ratio, big words count (words with > 6 characters), number of spelling mistakes, a ratio of word types to the total number of word types in the vocabulary of the dataset, number of unique words, and number of difficult words capture this aspect. A list of 5000 words which frequently appear on SAT [1] (Scholastic Assessment Test) is considered difficult. We calculate periodical measures like complexity defined as the ratio of the number of syllables to the number of words. Grammatical features include the count of part of speech tag collocations and word collocations, part-of-speech tag densities of the pronouns, adjectives, verbs and adverbs.

[1] https://collegereadiness.collegeboard.org/sat

LIWC

We have also used the Linguistic Inquiry Word Count (LIWC) program [54] to extract lexical features. It provides a count of various word categories like negative emotions, positive emotions, content words, style or function words, words associated with the cognitive and perceptive processes. LIWC computes 81 features. Of the 81 features, redundant features like word count, words per sentence; zero-variance features; features that are null for more than 80% of samples, inappropriate categories like swear, death, religious words are all removed. These features are extracted from all the five written answers (referred to as LIWC features) and also only from the first answer, which represents the self-introduction (referred to as LIWC-intro features). We include this with the intuition that the LIWC toolkit will capture the emotions, the cognitive and perceptive features from the self-introduction where the participants will tend to express more about themselves.

Sentiment Expression

Intuitively, the answers to any behavioural question will involve some emotion and will clearly state the opinion of the writer. We consider the sentiment expression feature inspired by Farra et al. [55]. We use the MPQA subjectivity lexicon [56]. This lexicon provides a sentiment polarity along with the sentiment intensity for each word. We derive two features using this lexicon - the total count of the subjectivity clues in the answers for each interview and the sentiment intensity values.

Trait elements

In any behavioural interview, the interviewers look for few traits in any candidate like problem-solving, decision-making, creativity, trust, conflict resolution. We have

curated a list of 20 traits which is expected of the candidates concerning the 75 behavioural questions used for the interviews. One of the expert annotators further validated this. Using the WordNet lexical database [57], we create a dictionary of more than 8000 words corresponding to the traits defined. This dictionary is developed using the hyponyms and hypernyms for two levels obtained by parsing the WordNet tree. The trait elements feature is the count of each of such trait related words in the answer.

We use the same set of features used for the written interviews for the short essays except for the trait elements feature set.

3.2.2 Features for Spoken Communication Skill Prediction

We extract three types of features for every interview video. Prosody and Speaking Activity (non-verbal audio), Visual (non-verbal) and Lexical (verbal) features. Features computed are shown to be relevant for job interview scenarios [58] and are inspired by the previous works cited in the appropriate places. First, we describe non-verbal audio-visual feature extraction followed by verbal/lexical methods.

Prosody

Prosody captures information about the pattern, rhythm, intonation of speech. It conveys as much meaning as words they accompany. Prosodic features are shown to be effective for automatic analysis in social interactions [28] [59]. We extract a wide set of prosodic cues that include features from three open source tools OpenSmile [60], Praat [61], PyAudioAnalysis [62].

Features such as pitch, voicing probability, pcm loudness, spectral energy are computed at frame level using OpenSmile. We use statistical metrics such as mean, variance, maximum, minimum of these features for analysis. First three formant frequen-

cies are also used as a feature. Temporal-based features such as speaking rate (number of syllables/duration of speech signal), articulation rate (number of syllables/phonation time (excluding pauses)), average syllable duration (speaking time/number of syllables) are extracted using Praat. We also use PyAudioAnalysis library that generates a 68-dimensional vector of features such as spectral flux, centroid, MFCC's, Chroma coefficients. The first 34 dimensions correspond to mean of all features listed in [62], and the next 34 features correspond to the standard deviation of all features.

Speaking Activity (SA)

Speaking Activity helps us understand the aspects of pauses and speaking turns. We used Matlab speech processing library [63] to segment speech and silence segments of the audio signal. From the start and end time of every speech segment, we compute features described in [26] [29] such as mean pause, max pause, the number of speaking turns, average speaking turn duration, the variance of speaking turn duration. We also use a histogram of speaking turns, binned into ten values (number of segments shorter than 2sec, 2-3 sec, 3-4 sec, ..., $>$ 10sec) as a feature.

Visual Features

Non-verbal visual cues are important in social interactions. Features corresponding to facial expressions, head pose and head motion are automatically extracted using a commercially available software iMotions, which is a real-time face processing toolbox developed for the understanding of the facial expressions [64]. We compute features such as the amount of time each emotion is active in the video, the number of segments with active emotion (frame intensity values greater than 0.05 are considered as active segments) according to [25]. The variance of frame level emotion outputs is used as a feature. Along with frame level intensity values, iMotions also gives yaw, pitch and roll

angles of the head position. We measure the overall head movement in the horizontal and vertical directions from the bounding box position of each detected frame, at three predefined points corresponding to the eyes and mouth [26]. For displacements in both directions, statistics such as mean, maximum, minimum and variance are computed.

Bag of Visual Words

Initially used for applications such as image categorisation, scene understanding, it can also be applied to videos by treating every frame as an independent data point. It handles the unequal length videos by mapping to a fixed sized codebook. We have considered four different sizes of the codebook ranging from K=50,100,200,300. The histograms were computed on the frame-level intensity values from iMotions for seven basic expressions along with positive and negative intensity. We used one in every ten frames for histogram computation.

Text-based Features

The textual transcriptions of the videos are obtained automatically using an online Automatic Speech Recognition tool, VoiceBase [65] (WER=33%). We choose VoiceBase after comparing Word Error Rate (WER) with Koemei and SpeechLogger online tools on three randomly chosen videos. These transcriptions are used to extract many text-based features. These features are the same as that of the features used for prediction of written communication skill (lexical features, sentiment expression and trait elements) described in Section 3.2.1. Besides these, we also calculate the rate of speech feature measured as the number of syllables per second of the interview duration.

Table TC3.4: Writen Communication Skill rating prediction using different feature groups

Feature Groups	Accuracy	F1 score
Lexical	0.72	0.71
LIWC	0.70	0.68
Sentiment Expr.	0.60	0.59
Trait Elements	0.70	0.70
Lexical+Senti	0.73	0.73
Lexical+Senti+Traits	**0.75**	0.75
Boruta	0.71	0.70

Table TC3.5: Effective Communication Skill rating prediction for Short Essays using different feature groups

Feature Groups	Accuracy	F1 score
Lexical	0.72	0.72
LIWC	0.73	0.72
Sentiment Expr	0.60	0.59
Lexical+Senti	0.71	0.70
Boruta	**0.77**	0.78

3.3 Automatic Prediction

We model the three systems using the features described earlier (Section 3.2) with different learners from Scikit-learn toolkit [53] like SVM, Logistic Regression, Random Forests etc. We consider the binary classification task defined earlier where the competent and effective communicators $\{3.5,4,4.5,5\}$ are separated from the average class of communicators $\{1.5,2,2.5,3\}$ (Figure FC3.3). The baseline is 50% for the written interviews and essay and 59% for video interviews. We choose to represent the results with the Scikit-learn version of XGBoost classifier [53] since it is a faster, optimised implementation of the Gradient Boosting algorithm. Also, the results from other classifiers we tested were similar to the presented results. For our experiments, consistency is more important than especially good results, and so we choose to run the same classifier on all three systems rather than developing separate systems that require individual

Table TC3.6: Spoken Communication Skill rating prediction using different feature groups

Feature Groups	Accuracy	F1 score
openSmile	0.62	0.59
Praat	0.72	0.71
Energy	**0.75**	0.73
Pydata	0.67	0.66
Speaking Activity	0.72	0.61
Visual	0.57	0.55
BOVW	0.50	0.48
Lexical	**0.75**	0.74
Trait Elements	0.66	0.65
Praat+SA+Energy	0.74	0.71
Lexical+Senti+Energy	0.73	0.73

tuning. The results are shown in the Tables TC3.4, TC3.5, TC3.6. The metrics accuracy and macro-averaged F1 score reported are the cross-validated results of stratified 5-fold cross-validation averaged over 50 iterations.

The tables TC3.4, TC3.5, TC3.6 show results for each feature group and also the late fusion of well-performing feature groups. All results are above the baseline except for visual and bag of visual words. It is to be noted that the rubric facial expressiveness had a relatively lesser correlation with the spoken communication skill, and the Kappa statistic was also less for the same. The energy and the lexical features perform well with an accuracy of 75% and an F1 score of 0.73 and 0.74 respectively, in case of video interviews. The overall text-based features for the video interviews perform on par, if not better than the prosody and speaking activity showing that lexical sophistication is also important.

In the case of written interviews, the fusion of the lexical, sentiment expression and trait elements perform well with an accuracy of 75% and an F1 score of 0.75. We also perform feature selection using a method called Boruta [66]. The algorithm is designed as a wrapper around a Random Forest classification algorithm. It iteratively removes the features which are proved by a statistical test to be less relevant than ran-

dom probes. It gives a set of confirmed, tentative and rejected features with relevance. We use the confirmed and tentative set of features. The Boruta analysis is performed on the Lexical, LIWC and LIWC-intro features to understand which lexical features are crucial to predict communication skill. The algorithm selects ten features to be confirmed and tentative. They are average word length, count of word collocations, ratio of word types to the total number of word types in the whole vocabulary of the dataset, number of words in LIWC dictionary, count of unique words, count of articles in self-introduction, count of personal pronouns in self-introduction, complexity, word count, first-person singular in self introduction. More usage of the word 'I' is a good personality indicator [67]. As the self-introduction forms a critical first impression in an interview, the self-introduction features are useful.

The Boruta feature selection method selects a set of 14 features for the essay communication prediction. Features like the number of sentences, word count, count of word and tag collocations, count of subjectivity clues, exclusive words etc., are shown to be essential and perform well with an accuracy of 77% and an F1 score of 0.78. Automated essay scoring is a relatively well-studied area of research. We have touched upon only a few of the features for our model as the primary purpose is comparison.

3.4 Comparative Analysis

This section presents the cumulative results of comparisons across the three modes of interview. There are about 69 participants common to all the three settings and the analysis is carried out on those common participants. The video and the written interviews differ in the mode of communication, whereas the written interview and the short essay differ on time dimension and question type. The comparative analysis is carried out on the perceived behavioural changes of the candidates using expert annotation and on difference in participant behaviour with the help of automatically extracted features.

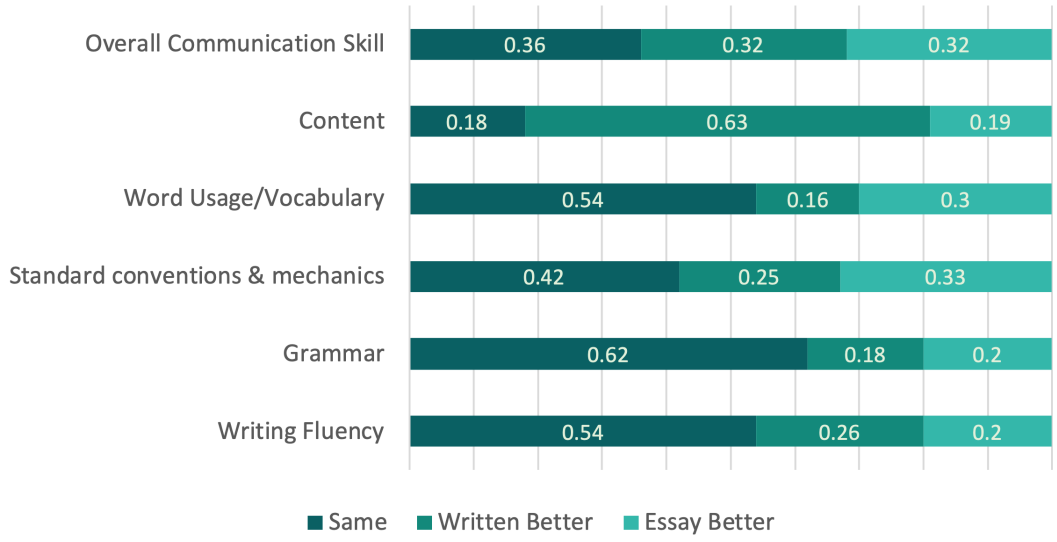

Figure FC3.4: Comparison of performance of common candidates across written interview and essays for all rubrics

.

3.4.1 Comparison of Perceived Expert Annotations

The following implications are based on the expert perception of the communication skill in each setting. We consider the average ratings from the two annotators for analysis.

Figure FC3.4 compares the annotations of all rubrics over the written interview and essay answers. 32% of candidates perform well in written interviews, 32% perform better in the essay, and 36% perform the same concerning communication skill. Considering the other rubrics like grammar, vocabulary, fluency and conventions, approximately 50% of the candidates perform equally. Both assessment media of writing facilitate good communication exchange. 63% of participants score better in content relevance in written interviews. Due to the increase in duration for the essay question and given the nature of the essay question, more might be expected of participants in the essay settings.

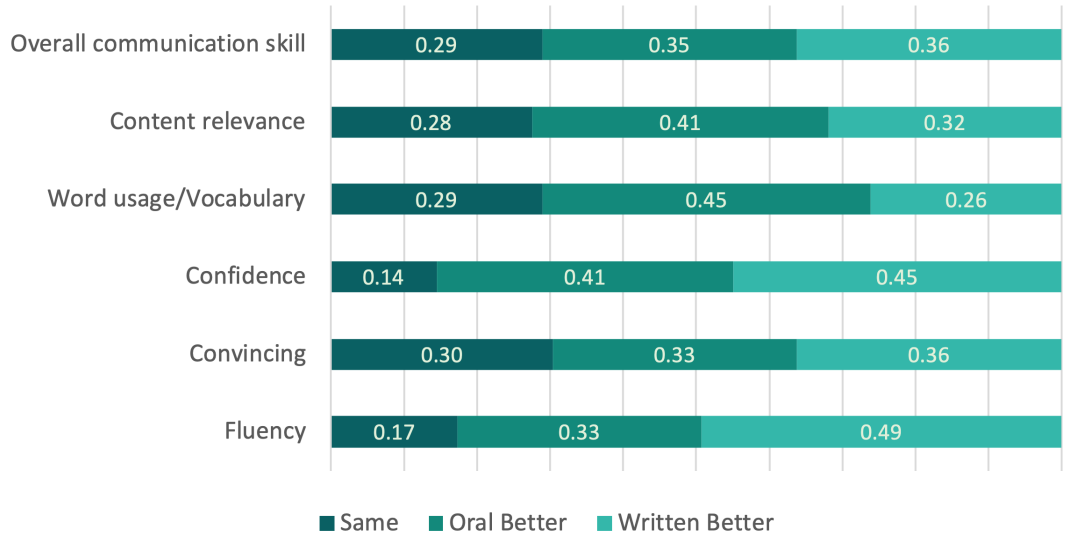

Figure FC3.5: Comparison of performance of common candidates across written and video interviews for all rubrics.

Comparing the video with the written interviews in Figure FC3.5, 36% of the participants performed better in written, 35% in video and 29% performed similarly in both the settings in communication skills. This might also be an indicator of the written interview being a plausible alternative. Rubrics content and vocabulary have better performance in oral interviews [2], whereas confidence, convincing and fluency are better perceived in written interviews. This invalidates the traditional belief that confidence and convincing characteristics cannot be well judged in writing. This further confirms our findings from the correlation analysis.

Considering video interview vs essay, 46% perform better in essay, 33% fare well in video interviews and 20% are the same (Figure FC3.6). Like in the case of written vs essay, content relevance is perceived well in oral for 54% of the candidates than in essays.

Of the participants who perform in the poor or average class (class 1) in spoken communication skill evaluation, 71% also perform poorly or average in written com-

[2] Video interviews and oral interviews are used interchangeably referring to Asynchronous Video Interviews.

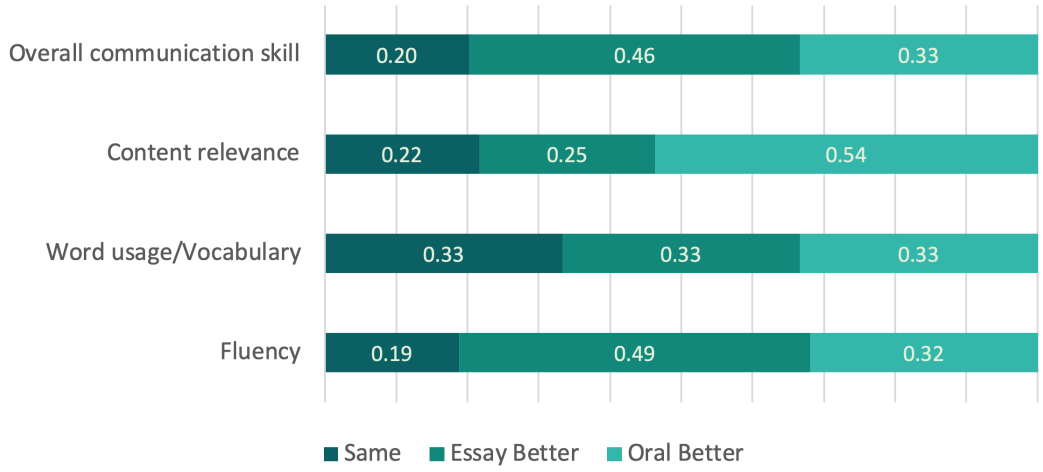

Figure FC3.6: Comparison of performance of common candidates across essays and video interviews for all rubrics.

munication skill evaluation. 91% of participants in class 1 of written communication skill are also in class 1 in the communication skill of the essay. Of the participants in class 1 of effective communication in the essay, 76% is also in class 1 of spoken communication. This indicates that majority of people who underperform in one mode of interview are more likely to perform the same in other modes. These facts affirm our hypothesis that the performance of the participants is independent of the setting.

3.4.2 Comparison of Automatically Extracted Features

We also perform the comparison of the overlapping features between the three settings for the 69 common participants. We consider the text-based features for comparison as it is shared across the three modes of interview. The average values of the selected few features are shown in the Figure FC3.7. Length features like the number of sentences, the total number of words, the average length of each answer are higher for oral interviews as the amount of content exchange is more while speaking than writing, although the information exchange might be the same. This also implies that word-based features like trait elements, uniques words also have higher values in video interviews.

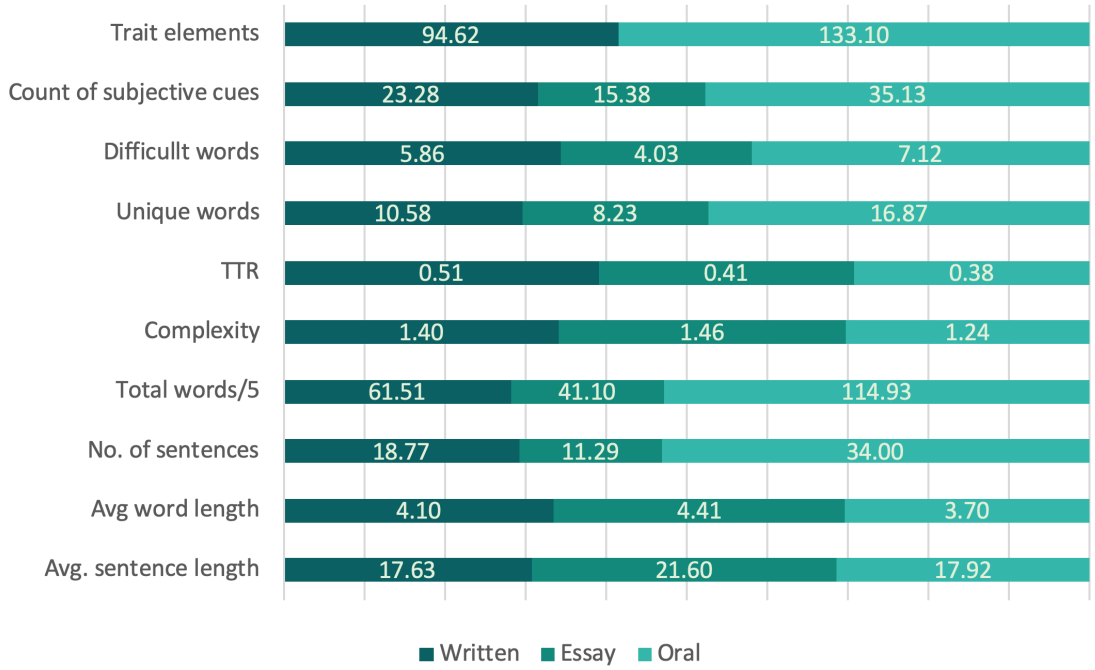

Figure FC3.7: Average feature values of common candidates across the three settings.

However, vocabulary-based features like type-to-token ratio, big words count, the ratio of word types to the total number of word types in the vocabulary of the dataset have higher values in written interviews illustrating rich word usage.

Feedback

At the end of the interview process, we presented a feedback questionnaire to the participants. Questions were about ease of usage, helpfulness of the interface, the recommendation of asynchronous systems to peers, natural interaction, comfortableness and their preference among traditional face-to-face interviews, interface-based video interviews, interface-based written interviews. The participants were requested to rate these parameters on a scale of 1 to 5, with 1 being the lowest, except the preference parameter.

97% of participants give a rating of 3 and above for the ease of use and helpfulness

of the interface, 87% rate 4 and above for the same. 88% of participants recommend asynchronous systems for interview assessment or training with a score 3 and above. Natural interaction with asynchronous systems when compared to traditional interview systems score 3 and above with 63% of the participants, score 4 and above with 29% of the participants. 93% rate 3 and above for the comfortableness and expressiveness parameter. All parameters get a good score of 3 and above except the natural interaction parameter. This supports the fact that asynchronous systems do not enable conversations. This inadequacy of interactivity in the medium is addressed in Chapter 4.

51% of the participants prefer traditional face-to-face, 27% prefer interface-based video interviews and 22% chose written interviews. It is a known fact that there is a general apprehension to accept the interface-based interviews, and these numbers confirm that fact. However, there is the other half of the population preferring the interface-based (video+written = 49%), which might be the reason for the advent of many automated talent assessment solutions.

Of the participants whose preference was traditional face-to-face interviews, 90% obtain almost similar human average ratings in all three settings with a difference of 0.5. Of the participants choosing the interface-based video, 80% perform similarly in the three settings with almost the same average ratings with 0.5 difference. 80% of the participants choosing the written mode of the interview are rated similarly with a difference of 0.5. These statistics further support our claim that participants' performance is independent of their preferred choice.

As a simple illustration, Figure FC3.8 shows the manual ratings in the three settings and their preferred choice among traditional face-to-face interviews, interface-based video interviews, interface-based written interviews taking three participants as an example.

In general, the participants found the interface comfortable and easy to use.

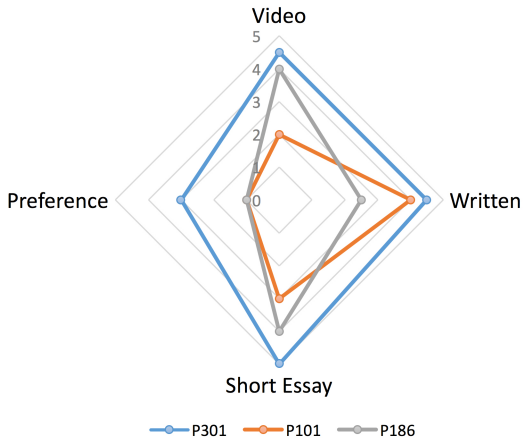

Figure FC3.8: Average Ratings of three participants P301, P101, P186 in Video, Written interviews and Short essay with their preference as 1-Traditional face-to-face, 2-Interface based video, 3-Interface based written interviews

3.5 Discussion

We present an empirical attempt to tackle the problem of automatic assessment of communication skill in non-conventional interview settings. Written interviews, video interviews and short essay are the three settings we consider. We propose a framework to predict the communication skill automatically and also compare the perceptions of the human experts and automatically extracted features. For this task, a dataset of more than 100 participants has been collected for each of the interview settings using a custom-built web-based interface in a controlled environment. Annotations are obtained from two experts in the field, on different rubrics specific to each mode of interview.

We make an extensive study of the rubrics with the help of the pairwise correlation analysis between each of the rubrics and the overall communication skill. It is revealed that along with the body language of the interviewee, lexical sophistication is also prominent with respect to the video interviews. Concerning the written interviews, the confidence and convincing required in answering the questions is also of impor-

tance. But in the case of essays, with the availability of more content, the ideas and the organisation are substantial to communicate the views.

We also perform automatic prediction of communication skill using various multimodal behavioural feature groups and highlight the important features in each of the settings. All the three models perform with accuracies well above the baseline with 75% being the best.

3.6 Conclusion

In this chapter, we analyse the transparency and social bandwidth attributes of the asynchronous interview medium. The expert perception of the various attributes of the candidates, manually annotated, signifies the transparency of the medium. It indicates the perceived measure of unobstructed communication. The feedback ratings from the participants also assess transparency. The ease of usage, helpfulness of the interface, recommendation of asynchronous systems to peers, comfortableness illustrates the extent to which the medium enabled the individuals to respond to questions as intended. All these parameters received high ratings from the participants.

The various multimodal cues derived from the participants' interviews like audio, visual and lexical quantifies the social bandwidth. These different kinds of social cues are included in the assessment of the candidates using the predictive model. The measure of these features across the three settings demonstrates the reasonable amount of data transfer regardless of the setting.

But the natural interaction parameter gets a relatively low from users. To make the asynchronous systems more natural and conversational, we need to address the interactivity attribute of the medium. The next chapter describes the investigation into this interactivity attribute through follow-up question generation.

CHAPTER 4

FOLLOW-UP QUESTION GENERATION

This chapter details the next contribution in this thesis, the automatic follow-up question generation (FQG). Asynchronous interviews, conventionally do not allow reciprocal or conversational exchange between the participant and the system. We address this attribute of interactivity in this chapter. In the scenario of an interview, our method generates a follow-up question from the current pair of interviewer question and the candidate's response. Section 4.1 details the procedure of data collection. Section 4.3 and section 4.4 proposes two seperate approaches to FQG. Human evaluation is conducted separately on both the models to validate the quality of the questions generated.

Structured interviews reduce the impact of different biasing factors on interview ratings. Limited prompting and follow-up, and no elaboration on questions is one of the components of structured interviews [68]. The current generation of asynchronous interview systems adopt structure and pose predefined or randomly selected questions from a relatively large set. However, with large scale adoption of these asynchronous interviews, it may eventually become repetitive and uninteresting for recruiters and candidates alike. The highly structured attribute of asynchronous video interviews (fixed questions, fixed time) increases their predictability, reduces the variability, and makes them monotonous [69]. Hence, it might be crucial to find the right balance between structure and probing in interviews. The adoption of planned or limited probing might

help interviewers collect additional information related to the job from the candidates, which may lead to increased interview validity [68].

Levashina et al. [68] 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. Integrating limited number of follow-up questions during the asynchronous interviews helps solve both aspects of the problem. A relevant follow-up question not only improves the interaction between the interviewer and the interviewee but also makes this interaction less predictable as the follow-up question is dynamic based on the interviewee's answer.

The follow-up questions generated by an FQG system should ideally

1. Take into account the candidate's response
2. Not be already answered in the response
3. Complement the original question.

A human interviewer is intelligent enough to scrutinize the candidate's answer concerning the question asked, keeping in mind the previous conversation with the candidate (previous pairs of question and answer). In this study, we only concentrate on the current question-answer pair in the conversation to generate a follow-up question(s). Considering the complete history of the conversation can be the focus of future work. Also, we consider one follow-up question as a proxy to planned or controlled probing. The results are evaluated on the single follow-up question generated on the current question-answer pair. However, the Follow-up Question Generator model is capable of generating multiple follow-up questions.

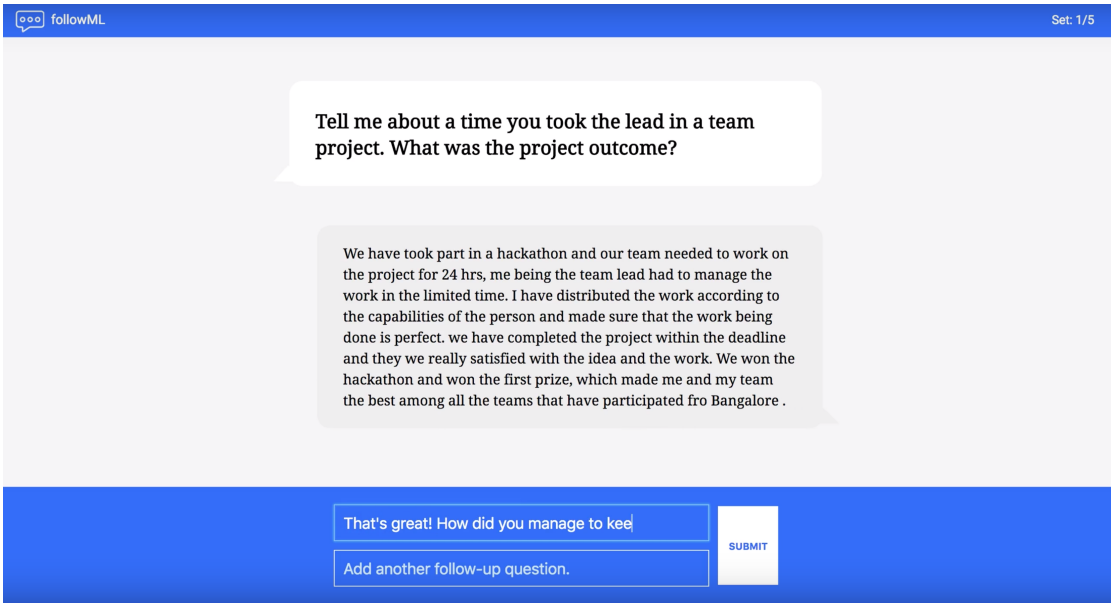

Figure FC4.1: Screenshot of the web application used for collection of follow-up questions.

4.1 Data

We conduct a restricted crowd-sourcing for this task. We instruct the volunteers to compose a follow-up question based on the presented snippet from the interview question answer on the screen. Figure FC4.1 shows the web application used for the data collection task.

These interview snippets were taken from the Asynchronous Interview dataset, specifically the written corpus. We call this the Asynchronous Written Interview dataset (AWI). An instruction or a demo video can be found here [1]. Thus, we obtain a follow-up question dataset with 1089 samples, each sample containing the triplet of a question, answer and a follow-up.

A text generation model requires large amounts of data to learn the nuances of a language and the task. This process of data collection is expensive to scale and time-consuming. To overcome this challenge, we use a generic and popular question-answer

[1] https://youtu.be/KbHF7_kMaA8

Table TC4.1: Number of QA pairs used for training

Dataset	No. of QA pairs
SQuAD	100,000
AQAD	311,678

datasets like Stanford Question Answer Dataset (SQuAD) [40], Amazon Question Answer Dataset (AQAD) [70] for training a question generation model or a large scale language model to augment the training process.

External Datasets

SQuAD [40] is an extensive, publicly available, general purpose reading comprehension dataset. It consists of 100,000 questions posed by human crowd workers on Wikipedia articles. It is a highly synthesized dataset with mostly factual questions. It has context paragraphs, questions and answers. The answer to each question is a span of text from the corresponding paragraph of the article.

AQAD [70], on the other hand, is a real-world dataset. It comprises of questions from the various product pages on the Amazon website. It has questions more natural and less syntactically structured. This dataset consists of two question categories, namely binary and open-ended questions. Binary questions are the ones where the answer amounts to either 'Yes' or 'No'. Open-ended questions or compound questions are more complex and are diversified question types like how-, why-, what- and so on. It consists of 135,000 products from Amazon, 808,000 questions and 3 million answers. Each question can have multiple answers from different users. We choose to use questions having a single answer from the various category of products like Electronics, Home and Kitchen, and so on. This accounts for around 300,000 question-answer pairs. We use this dataset as well to enhance the variety of questions generated. Table TC4.1 mentions the number of samples used for training the question generation model.

4.2 Task

The training samples 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, are assumed to be available. The task is to generate f given q and r as inputs.

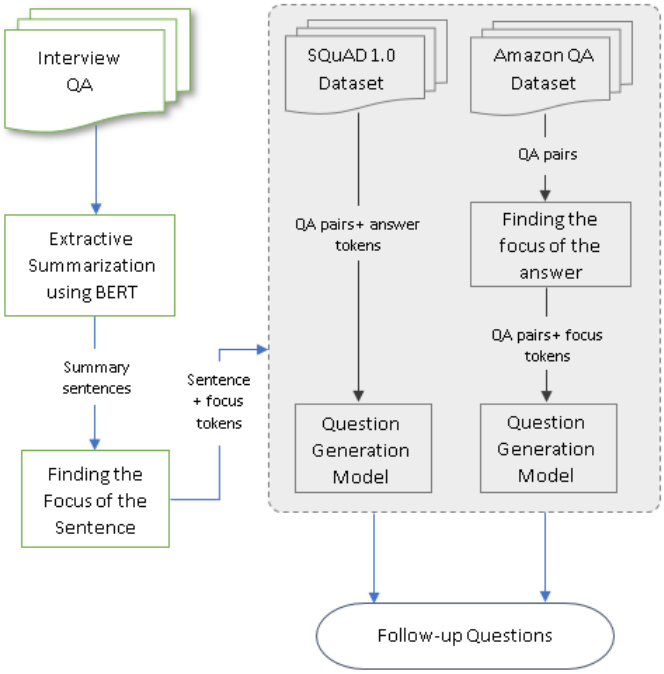

Figure FC4.2: Workflow describing the proposed approach to follow-up question generation.

4.3 QG-net based Follow-up Question Generation

In this approach, publicly available datasets – SQuAD and AQAD, are used to train the question generation model separately using the QG-net architecture. Follow-up questions corpus undergoes preparatory techniques before feeding into the question generation model as input. Figure FC4.2 illustrates our overall approach. The candidate's response is passed through an extractive summarization module and finding focus of the answer module to extract the important sentences in the response with their focus

Q: Give an example of how you worked in a team.

A: I once worked in a team of five individuals. Two of us were interested in working on the technicalities of the project while others were not. Hence, we divided the project work depending on our interests. Two of us worked on the technicalities of the project, one on implementation of those technicalities and while the rest helped in documentation.
We discussed our work during our meetings in order to ensure that everything is going in a right manner.

Summary: Two of us were interested in working on the technicalities of the project while others were not. Hence, we divided the project work depending on our interests.

FQ: What are the technicalities interested in working on?
FQ: What was the focus of the project ?

Figure FC4.3: An example interview question, answer and generated follow-up questions along with the summary and its focus tokens coloured respectively.

tokens. This is fed into the QG-net model to generate follow-up questions. QG-net [42] is an RNN-based encoder-decoder model with copy attention mechanism designed to generate questions from the educational content. Since this is closer to our domain of behavioural content, we adopt this architecture for our study. Figure FC4.3 shows an excerpt from the written interview dataset along with the follow-up questions generated using the workflow. Each module in our approach is explained below in detail.

- Finding the Focus of the Answer
- Extractive Summarization using BERT
- QG-net Question Generation Model

4.3.1 Finding the Focus of the Answer

QG-net model assumes that the answer tokens is a continuous segment within the source context. Its uses a binary valued indicator function as an added feature to indicate whether a word belongs to the answer. This compels the presence of answer tokens

similar to SQuAD, a rarity in real-world datasets. There exist overlapping tokens in the question (Q) and answer (A) pairs that can be seen as the topics shared between them. These are regarded as `focus tokens` of an answer. We employ a simple technique similar to [71] to automatically find those focus tokens. We allow non-contiguous and multiple focus tokens which aid in the generation of distinct follow-ups.

After removal of the stop words, A and Q are represented as a sequence of tokens $[a_1,..,a_n]$ and $[q_1,..,q_m]$ respectively. We consider all the tokens in A as candidates for the topic and all the tokens in Q as voters polling for the candidates. The final layer hidden state weights from the pre-trained language model BERT [45] 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

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

where $sim(a_i, q_j)$ is the cosine similarity [53] value between a_i and q_i and

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

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.

4.3.2 Extractive Summarisation using BERT

The input to the QG model should be a representative of the response and give information for a potential follow-up. We eliminate the sentence directly answering the question having the high cosine similarity value with the question above a threshold. We employ a simple extractive summarization technique on the remaining.

We use the technique 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 $[fr_1,...,fr_p]$ and $[fs_1,...,fs_q]$ respectively. The cumulative score for each focus token of R is calculated as

$$W_i = \sum_{j=1}^{q} p_{ij}.sim(fr_i, fs_j) \qquad\qquad N = \sum_{i=1}^{p} W_i$$

where $p_(ij)$ is the indicative variable which equals to 1 if cosine similarity value is greater than a threshold γ otherwise 0. 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* number of sentences with the highest frequency of similar sentences is considered.

This method ensures that the potential follow-up question is generated on the additional information from the candidate's response and not on the sentence directly answering the interview question if any. The sentence(s) thus selected is given as source input along with the focus of the sentence to the QG-net model at the inference time.

4.3.3 QG-net Question Generation Model

We use the QG-net [42] architecture to train a question generation model. QG-net is a Seq2Seq model (reader-generator). The context reader is a bi-directional long short-term memory (bi-LSTM) network [72] which processes each word in the input context and turns it into a fix-sized representation. The focus tokens are encoded with each word as an additional feature using one-hot encoding vector indicating that the word is a focus

token or not. The question generator is a uni-directional LSTM which generates the question word-by-word incorporating pointer network [73] on the generator vocabulary. This model design enables the generator to output questions that focus on specific parts of the input text. Additional linguistic features like the part of speech tag POS, named entity NER and word case CAS are also encoded. The words in the context, answer, and question are all encoded as d-dimensional GloVe [74] vectors. We refer the readers to the original paper for a detailed overview of the architecture [42].

4.3.4 Training Details

This model is trained on the two publicly available datasets, namely SQuAD and AQAD.

The SQuAD can be readily used with the QG-net model as it is in the required format. The context paragraph sentences are used as source input, questions as the target output and the explicit span of answer tokens are embedded into the source input as a binary indicator. Each paragraph is truncated to the single sentence containing the answer and is used as the source input.

AQAD dataset does not have the specific answer tokens mentioned within the answers. Since the questions are more complex than the factoid questions, there are no definite answer tokens. We employ the method mentioned in Section 4.3.1 to identify the focus tokens in the answer. These focus tokens are considered as answer tokens and embedded into the source input along with the linguistic features.

Two models are trained separately using the two datasets mentioned above. The first model is trained on the 100,000 answer-question pairs from SQuAD. The second model is trained on the 311,678 answer-question pairs from AQAD. The linguistic features are all derived using the SpaCy library [75] and embedded into the source.

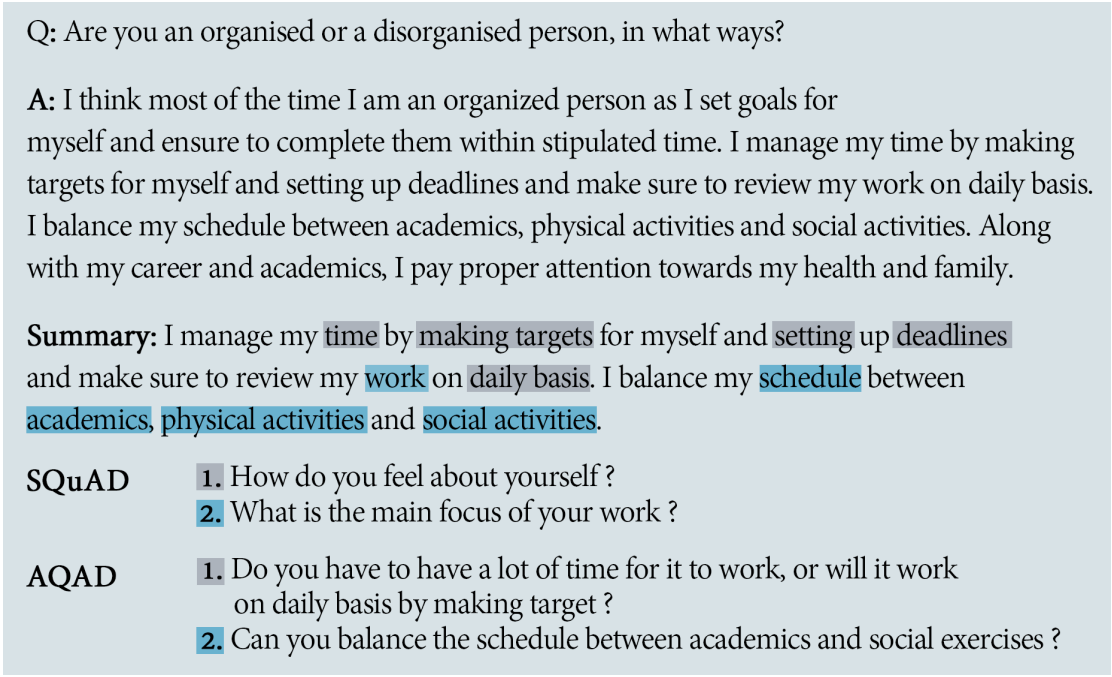

Figure FC4.4: Examples of follow-up questions generated from models trained on SQuAD and AQAD.

4.3.5 Results and Analysis

The asynchronous written interview data is used to conduct all experiments at the inference time. We prepare the data by passing it through the Extractive Summarization module and ”Finding the Focus of the sentence” module to obtain a sentence with its focus tokens as shown in Figure FC4.2. The sentence is further parsed using the SpaCy library to compute the linguistic features. The sentence tokens embedded with the focus tokens, linguistic features - POS tags, NER tags, CASE tags (tag indicative of whether a word starts with an upper or a lower case) are passed to the QG-net model to generate decoded question tokens.

Figure FC4.4 shows the results from both models. AQAD model generates complex and different types of questions like yes/no, why-, how- as compared to factoid questions from SQuAD model. They are less syntactically correct since the AQAD comprises of less grammatical QA pairs from real-world users.

Q: What are your views about further studies?

A: I believe that whether this option should be pursued depends on the person and what they wish to achieve in their life. As for me, I want to know more and more about things which I find interesting. I also think that I would love the interactions and intellectual stimulation that only an academic place can provide. I will definitely take up such an option if I get the chance.

Summary: I also think that I would love the interactions and intellectual stimulation that only an academic place can provide.

FQ w/o summarization: What is the difference between and intellectual
FQ: What is the difference between academic and intellectual?

Q. What obstacles or difficulties have you ever faced in communicating your ideas?

A. Sometimes it happens that I am unable to express what I want to tell because of demotivation from other parties. But still I don't lose hope and stick with my ideas and try to explain that in some innovative and real examples. One day, I and my project partners were discussing on project idea. At that time, my idea was good and new but problem was that I don't have any proof to prove that my idea will work very well. But after some days of research, I found actual proof that it is possible to implement and then we all would agreed on that topic. For the newer person, it is natural to have some difficulties in communicating but after some efforts you will surely get confidence to explain it in better way.

Summary: At that time, my idea was good and new but problem was that I don't have any proof to prove that my idea will work very well. But after some days of research, I found actual proof that it is possible to implement and then we all would agreed on that topic.

FQ w/o summarization: What did i and my project n't do ?
FQ w/o summarization: What did you state that it is possible to prove that it is possible ?
FQ: What did you believe was good and new about your idea ?
FQ: How did you view proof of research ?

Figure FC4.5: Examples of the questions generated with and without the summarization using the same answer and focus tokens.

To assess the advantage of the extractive summarization technique, we analyze the examples generated by the trained model with summarization and without. The experimental setup for both cases is the same except that the input is the entire answer for the former and the summary sentence for the latter. The focus tokens are the same for both, as shown in Figure FC4.5. We show that the questions generated are more syntactically correct in the latter case. As the length of the answer increases, the quality deteriorates.

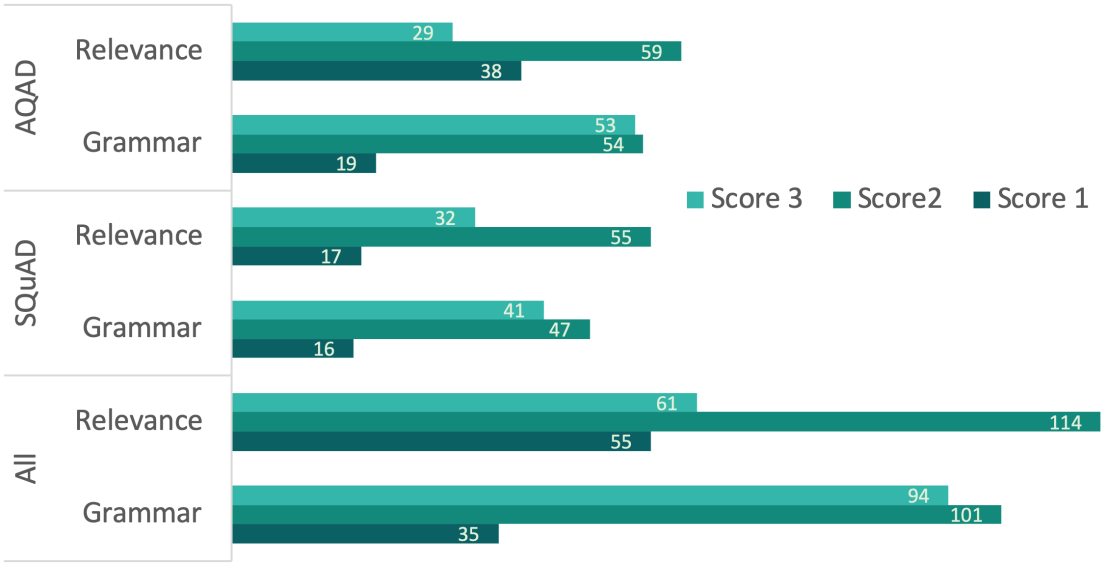

Figure FC4.6: Quantitative human evaluation on follow-up question generation.

4.3.6 Human Evaluation

To further evaluate the quality of the follow-up questions generated, we perform a human evaluation on the follow-ups generated from SQuAD and AQAD models. We choose a total of 230 questions randomly, 126 from SQuAD and 104 from AQAD for evaluation. The human evaluators are presented with original interview question, candidate response and generated follow-up question. They rate the follow-ups on two parameters, namely Grammar and Relevance on a scale of 1 to 3. A minimum of two people evaluate each question, and we consider the rounded average of the scores in the final evaluation. The results are depicted in Figure FC4.6.

The follow-up questions generated are relevant and adhere to grammar. Grammar is better than relevance. 76% of the questions score 2 or above in relevance, and 85% of the them are grammatically accurate with a score ≥ 2. 70% and 85% of questions generated from SQuAD model score 2 or above in relevance and grammar respectively. The questions generated from AQAD model seem to be more relevant than questions from SQuAD. They are 85% relevant and 84% grammatical with a score 2 or above.

4.4 GPT-2 based Follow-up Question Generation

In this approach, we use an adaptation framework for generating follow-up questions using language models by fine-tuning it on question, response and follow-up triples. A language model trained on very large corpus of text will be able to generate long contiguous coherent text. The fine-tuning with follow-questions data samples help FQG to learn the question structure and the relation between the triples, and the knowledge from the language model pre-training produces novel questions. We use the transformer language model architecture, Generative Pre-trained Transformer (GPT-2) introduced in Radford et al. [76]. This is very similar to the decoder part of the original transformer encoder-decoder model of Vaswani et al. [77]. It uses multiple transformer blocks each with multi-headed self-attention operation over the input context tokens followed by position-wise feed-forward layers to produce an output distribution over target tokens [46]. Our model is based on the recently published PyTorch adaptation of GPT-2. [2]

We initialize the Follow-up Question Generator model with 12-layer decoder-only transformer with self-attention heads containing 768 dimensional states and 12 attention heads. The parameters are initialized to the smallest version of the GPT-2 model weights open-sourced by Radford et al. 2019 [76]. The GPT-2 model is pre-trained on the WebText dataset which contains the text of 45 million links from internet [76].

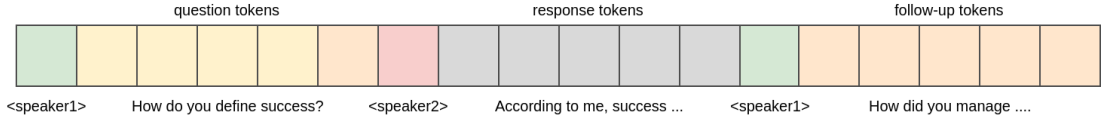

Figure FC4.7: Input representation for training Follow-up Question Generation model

[2] https://github.com/huggingface/pytorch-transformers

4.4.1 Fine-tuning

We fine-tune the pre-trained model using the follow-up questions dataset described in Section 4.1. 80% of the data is used for training and the rest is used for validation. The input sequence to the model constitutes of tokens from each of the $\{q, r, f\}$ in the triplet concatenated in a sequence. A set of input embeddings is constructed for this sequence. The word and position embeddings are learnt in the pre-training phase. We use an additional set of embeddings, speaker embeddings, similar to Wolf et al. [51] to indicate whether the token belongs to question, response or the follow-up. These embeddings are learnt during the fine-tuning phase. The input to the model is the sum of all three types – word, position and speaker embeddings for each token. Figure FC4.7 illustrates how the tokens in $\{q, r, f\}$ are organised to adapt the language model to the question generation task and form the speaker embeddings.

Following [51], [45], the fine-tuning is done by optimizing two loss functions – a language modelling loss, and a next-question classification loss. 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. The cross-entropy loss are used to score these probabilities where the human written follow-up question tokens are used as labels.

We append the dataset with randomly sampled questions from a pool of 100 questions, acting as distractors. A classifier is trained to recognize the correct next question appended to the input sequences. This trains the model to learn a sense of sentence ordering. The classifier is a linear layer applying a linear transformation to the last hidden state of the self-attention model to compute a value. A special token [CLS] is appended to the sentence at the end. The last hidden state of this special token is used for classification. Using the computed values from linear layer, a softmax layer obtains the classification probabilities. Then we apply a cross-entropy loss to correctly classify the

gold follow-up question among the distractors. We use $n = 2$ as the number of choices for classification. The language modelling loss and the next-question classification loss are optimized jointly by fine-tuning to maximize the log-probability of the correct label.

4.4.2 Decoding details

We use the top-k random sampling strategy for decoding [78]. At each timestep, the probability of each word in the vocabulary being the next likely word is given. The decoder randomly samples a word from the k most likely candidates. Here k is a hyperparameter determined to be k=10 experimentally.

4.4.3 Results

We report the results of the follow-up question generator model in terms of perplexity [79]. We also report the classification accuracy of next-question classification task. Perplexity is usually used to measure the quality of language models. It indicates how well the model predicts the next word correctly. Our model obtains an average validation perplexity of 20.6 and average validation accuracy of 63.1%. These values can be deemed reasonable considering the small size of the in-domain dataset used for fine-tuning. It may also be due to the fact that the questions generated are novel and relevant leveraging the knowledge from the pre-training step which may not be present in the human written follow-up questions.

4.4.4 Human Evaluation

To further evaluate the quality of the generated follow-up questions, we get human ratings for a small subset. We take 100 unseen question answer pairs from the AWI dataset and generate follow-up questions on the same. Three human annotators evaluate

Average Ratings	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

Figure FC4.8: Average Human Ratings of the follow-up questions

the quality of the questions on a scale of 1-3, 1 being the lowest. The annotators are instructed to rate the questions based on relevance and appropriateness of the question to the original base question and response. We consider the average ratings from three annotators for evaluation. Figure FC4.8 gives 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 FQG model generates quality follow-up questions and are scored well by humans. We also take rounded average of the ratings from the three annotators. Considering the rounded average, 89% of the questions are score 2 or more and 27% of them are rated a score of 3. Though not directly comparable with the QG-net based model due to the difference in the number of annotators, we can say that GPT-2 based model is superior in terms of relevance and grammar.

4.4.5 Qualitative Analysis

We present some examples generated by the GPT-2 based FQG model in Table TC4.2. As seen in the first part of the table, the questions generated on the validation samples are relevant and grammatically correct but differ from the human generated follow-up questions. This shows the richness and variety in the questions apart from the training data. 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.

Table TC4.2: 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?	Yes, 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 task is assigned to me, 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 from ASR	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?	

We also investigate the performance of the FQG model on transcripts generated by an automatic speech recognition engine (ASR). The aim of analysis is to check performance of the FQG model when incorporated in a real-time application like a virtual interviewing system. We use Web Speech API [80] to perform automatic speech recognition. The third part of the table shows the questions generated on the text from the ASR. As it can be seen, the questions are relevant and appropriate, inspite of the errors in the transcriptions. It is to be noted that the original base question is always correct and is not exposed to the errors of the ASR and can be a reliable input to the FQG model.

4.4.6 Robustness to Errors in Speech

Investigating the robustness of Follow-up Question Generator has important motivation. The model is trained on human-written triplets of (question, response, follow-up) whereas it will be inferred on the interviewee's answer obtained from ASR transcript in a real-world scenario. Hence, analyzing how follow-up question generation varies for ASR transcripts when compared with human transcripts helps to investigate the robustness of Follow-up Question Generator.

We use the asynchronous interface-based video interview dataset from Rasipuram et al. [29] for this purpose as they have manual transcriptions of the interviews. We randomly select 103 different candidates' responses to questions to form question answer pairs. We also obtain automatic transcriptions for the same pairs of 103 question answers using the Web Speech API [80]. We generate a follow-up question for each of this pair. This gives us 206 triplets of question, response, and follow-up questions, 103 each for manual and automatic transcripts.

Three human annotators evaluate the quality of the question on a scale of 1-3, 1 being the lowest. The annotators are displayed with the questions and answers from the

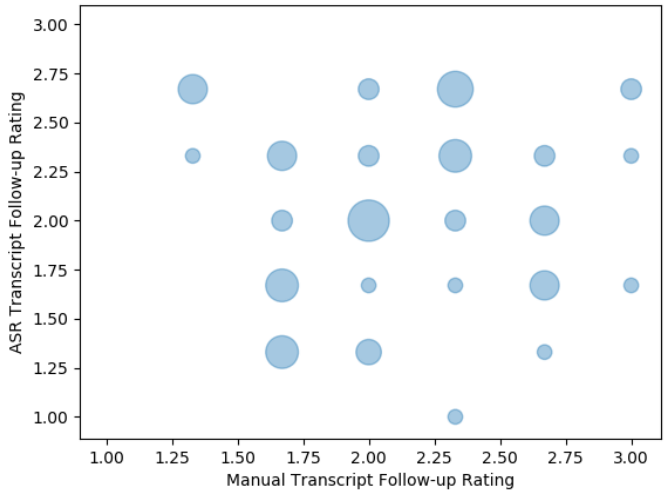

Figure FC4.9: Human Evaluation of the follow-up questions on manual and ASR transcripts. Bubble size depicts the count of the ratings

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.

4.4.7 Human Evaluation of FQG on ASR transcripts

Figure FC4.8 shows count of the average ratings for follow-up questions on manual and automatic transcripts. 69% of the questions generated on manual transcripts and 64% of the questions generated on ASR transcripts get a score of ≥ 2. This implies that the FQG model generates relatively good quality follow-up questions on both manual and automatic transcripts.

Figure FC4.9 depicts the average ratings of the two annotators for both manual and ASR generated follow-up questions with the bubble size depicting the count of the same. More points are concentrated along the diagonal of the chart indicating the ratings for both manual and ASR triplets are similar. As expected, there are also cases where the ratings are of manual transcripts are higher than the ASR counterparts. These ratings are further averaged over the 103 follow-up questions each for human and ASR transcripts.

The average rating for human transcript triplets is 2.15 out of 3, and the average rating for ASR transcript triplets is 2.03 out of 3. The minor difference between average ratings for human transcript triplets and ASR transcript triplets shows the robustness of the Follow-up Question Generator, which can generate nearly even relevant follow-up questions even with errored ASR transcripts.

4.5 Conclusion

This chapter presents two approaches to automatic follow-up question generation in an interview setting to enhance the interactions of a virtual agent/system. 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. This balances the structure of the interview as well as conversational flow between the system and candidate.

Using preparatory techniques like extractive summarization and finding the focus of the sentence before giving as an input to a question generation model enhances the quality of a follow-up. The FQs are rated well by human evaluators. 70% of the randomly chosen questions are relevant follow-ups. We also show that the summarization technique improves the quality of the questions generated.

The use of knowledge from 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. The model always generates grammatically correct questions and the quality of questions are enhanced. Though not directly comparable, the questions from GPT-2 based FQG model generates relevant questions 89% of the time as compared to 70% and 85% by the two QG-net based FQG models.

CHAPTER 5

COLLABORATIVE PROJECTS

In this chapter, we briefly summarise the other projects in collaboration carried out jointly with this thesis. The collaborative works are as follow:

1. *Maya*, a 3D virtual interactive interviewing system
2. Asynchronous Video Interviews vs Face-to-Face Interviews For Communication Skill Assessment

(1) was done in collaboration with Manish Agnihotri, research assistant from the multimodal perception lab. (2) was done in partnership with Sowmya Rasipuram, PhD student from the same lab.

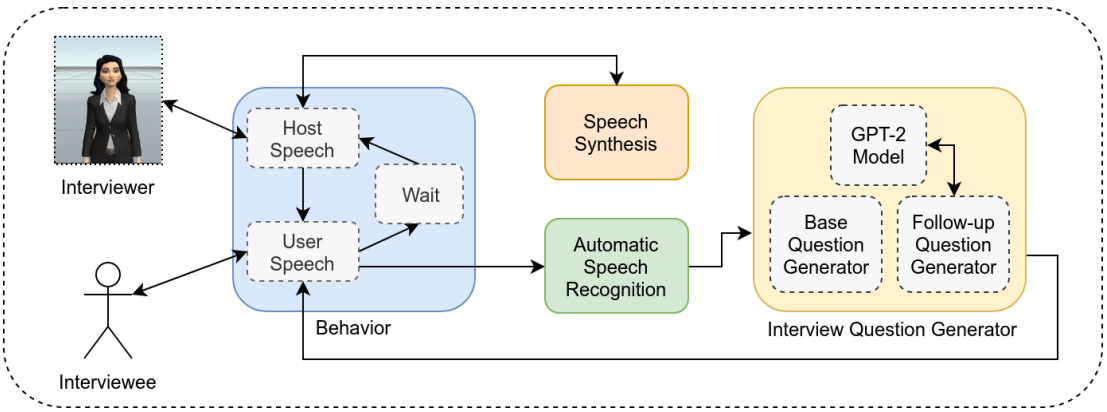

Figure FC5.1: *Maya* - Interactive Interviewing System

5.1 *Maya* - Interactive Interviewing System

The GPT-2 based follow-up question generation model (Section 4.4) is used to build a 3D virtual interviewing agent. Figure FC5.1 depicts the overview of *Maya*. Our interactive interviewing system, *Maya*, consists of two main components explained in detail in the below sections.

- 3D Virtual Interviewer
- Interview Question Generator

The first component is an Amazon Sumerian [81] based 3D virtual interviewing agent which asks questions and collects the interviewee's responses. We use Web Speech API [80] to perform Automatic Speech Recognition (ASR). This text data is fed to the question generator hosted on a server using an Application Programming Interface (API) call which returns a question. Using Amazon Polly [82] text-to-speech toolkit, the virtual agent communicates the generated question to the interviewee. This section describes these two components in detail.

5.1.1 3D Virtual Interviewer

The interviewing interface is a 3D scene hosted on Amazon Sumerian [81]. Amazon Sumerian is a managed service that enables developers to develop, host and deploy virtual reality (VR), augmented reality (AR), and 3D applications on WebGL compatible web browsers as well as on popular VR and AR hardware.

The interviewing interface consists of the following entities: default lights, default camera, and a host. A host is a digital character to narrate the scene, which in our case serves as the interviewer, as shown in Figure FC5.2. We have defined behavior

Figure FC5.2: The Virtual Interviewer

for the interviewer as a state machine component which makes the scene dynamic and interactive. The minimalist design of the interview interface makes it lightweight and is sized at less than 21 MB, making it suitable for even low-end devices.

5.1.2 Interviewer's Behavior

Behavior is a collection of actions that can be added to the state machine component. The state machine component defined for the host (interviewer) consists of 4 states: *Initialisation*, *Maya Response*, *User Response* and *Wait*. Each state consists of one or more actions that contain some logic. The state flow is described in Figure FC5.3

Initialisation, the first state, waits for the Amazon Web Services (AWS) SDK to get credentials before using features that call AWS services and signals that it's ready. When this occurs, it transitions to the next state *Maya* Response, which executes the HostSpeech script. 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 [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

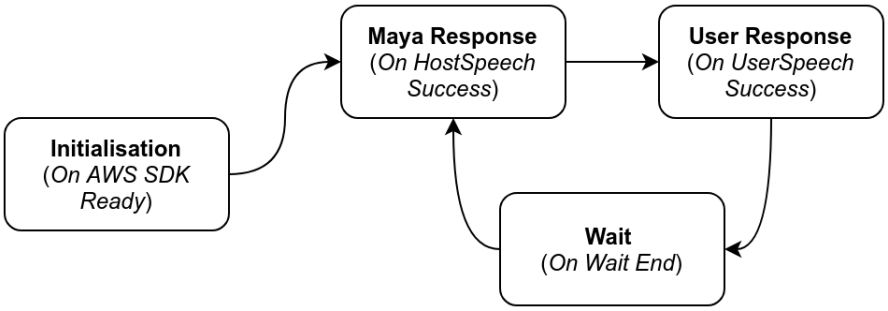

Figure FC5.3: *Maya's* Behavior: State Machine Diagram

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.

Upon successful completion of UserSpeech script, the current state transitions to the Wait state. In this state, the system waits for 2 seconds (empirically derived) before transitioning back to *Maya* Response state to give the interviewee some time before the audio for next question is played. This system works in a loop until the Dialogue System provides an end token which indicates the end of the interview.

5.1.3 Interview Question Generator

The Interview Question Generator component contains two modules which communicates with the 3D virtual interviewer in a request-response fashion – Base question selector and Follow-up question generator. We have a pool of 100 questions commonly asked in an HR interview, called Script of Questions (SoQ). At the start of the interview, this module selects a question randomly from SoQ and sends it to the virtual interviewer to be posed as the first base question.

The next question is a follow-up question based on the candidate response generated from follow-up question generator. The follow-up question generation is explained at length in Section 4.4. The next base question is again randomly selected when n number of follow-up questions are asked. 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. It is crucial to find the right balance between structure and probing in asynchronous interviews [68]. We consider one follow-up question as a proxy to planned or controlled probing. A demo video of the system is available here [1].

5.2 Asynchronous Video Interviews vs Face-to-Face Interviews For Communication Skill Assessment

In this section, we further compare the communication skill of participants in asynchronous interviews across another dimension of traditional face-to-face interviews.

We present a dual dataset of face-to-face interviews and asynchronous video interviews from the same participants. We perform a comparative study of the perceptual

[1] https://drive.google.com/open?id=1QbvRdI-K1khhJExhiXXI7s8VRKvYnC2-

changes of the candidates between both the settings. We also build a predictive model to assess the communication skill of the participants. The complete stack of features used, i.e. the speaking activity, prosody, visual and lexical features can be found in [29]

We achieve the best prediction accuracy of 83% in face-to-face interviews and 79% in asynchronous video interviews. In both cases, our experiments show that the automatic transcriptions from an ASR tool suffice for our prediction task without an absolute need for manual transcriptions.

CHAPTER 6

CONCLUSION

Throughout this thesis, we have made contributions to two different aspects of automating asynchronous interview systems. One is the automatic assessment of the communication skill (Chapter 3), and the other is the automatic generation of follow-up questions (Chapter 4). In this final chapter, we will recap the proposed methods and summarise our findings (6.1), and provide an outlook into the future (6.2).

6.1 Summary of Findings

Over the course of this thesis, we have presented methods and approaches to fulfil the objectives of this thesis. In this section, we recap how our methods addressed the objectives we laid out initially and summarise our contributions and findings.

Research Objectives

In chapter 1, we put forth the four attributes of the administration medium of an assessment to be studied in the case of the asynchronous interview medium.

- **Transparency** The asynchronous interview systems are transparent, allowing easy facilitation of communication exchange. In chapter 3, we measured this using

the automatic assessment of communication skill. We also presented the results from a feedback survey of the participants of their comfortableness with the asynchronous medium for interviews. The participants were reasonably comfortable with the setting. The annotations from the HR experts also confirmed the considerable exchange of communication.

- **Social bandwidth** Chapter 3 also discussed the various features that we extracted from the three different asynchronous interview settings. The lexical, audio and visual features are the different types of social cues included. Though it may seem like the written form of interviews have less social bandwidth, they can be assessed and predicted equally well (Chapter 3.3).

- **Interactivity** Asynchronous communication does not allow for any interactions. To improve on this aspect, Chapter 4 proposes two approaches to follow-up question generation. These follow-ups enable for a smooth conversational interview and assessment.

Contributions

Interview Dataset: We presented a dataset of 150 participants with written interviews, essays and video annotated independently for communication skill by two human expert annotators. This dataset is novel as the same participants have participated in the three modes of interview establishing a three-way dataset. This parallel corpus gave a unique opportunity to study the behaviour and performance of candidates. Section 3.1 detailed the procedure of the data collection and the web-interface specifically designed for it.

Automated assessment of communication skill and comparative analysis Chapter 3 presented a systematic study and automatic measurement of the communication skill of candidates in different modes of behavioural interviews. We demonstrated the

feasibility of automatic prediction of our variable of interest in written and oral modes. The computational models achieved an accuracy of 75% in written and oral, and 78% in the essay skill measurement. A comparative analysis of non-conventional methods of employment interviews in terms of behavioural perception and automated predictions was carried out (3.4). We show that the performance is independent of the administration medium since, on an average, 83% of the candidates perform similarly in all the three settings. This affirms the possibility of using low-infrastructure settings like written interviews instead of AVIs.

Automatic follow-up question generation In Chapter 4, we presented two powerful approaches to generate follow-up questions given the interviewer question and the candidate response in an asynchronous HR interview setting. We utilized external knowledge resources to train follow-up question generation model to surpass the need of a large specific dataset. We also saw that the GPT-2 based model (Section 4.4) generates more coherent and relevant follow-up questions and can be used in a virtual agent interviewing system.

6.2 Future Directions

In this section, we will give an outlook into the future for each of automatic skill assessment and follow-up question generation in particular and automation of asynchronous interviews in general.

Automatic Communication Skill Assessment

To indeed generalize the prediction model, a large and varied population of participants can contribute to the data. A more scalable option would be to use any out of domain available datasets and perform a domain adaptation for the classification mod-

els. With the availability of more data, various neural network architectures could be tested.

A computational formulation converging all three models can be attempted. Enhanced feature groups can be investigated that capture the same aspects of the behavioural interviews across all the settings. A more profound user study will strengthen the practicability of the asynchronous interview settings.

Follow-up Question Generation

There are several avenues for research in this area. A next step forward is considering the complete history of the interview to generate questions. Following Rao and Daumé, [83], we could use adversarial based training for the follow-up question generation to produce more diverse questions. Another avenue of further research could be to induce common-sense reasoning to question generation. A knowledge base could be of help in this case. A knowledge-aware generation model can be more robust and appropriate.

Another approach could be to use the multimodal context for the question generation. Mostafazadeh et al. [84] combine the test input with image input to generate more relevant questions. Including the multimodal input of the audio and video would be taking the generation system to top notch.

Bibliography

[1] W. Cascio and R. Montealegre, "How technology is changing work and organizations," *Annual Review of Organizational Psychology and Organizational Behavior*, vol. 3, pp. 349–375, March 2016. [Online]. Available: https://doi.org/10.1146/annurev-orgpsych-041015-062352

[2] talview.com, "Understanding recruitment troubles and trends," 2016. [Online]. Available: https://info.talview.com/understanding-recruitment-troubles-trends-research-2016. Accessed Oct 30, 2019.

[3] J. E. Salmons, *Qualitative Online Interviews: Strategies, Design, and Skills*, 2nd ed. Thousand Oaks, CA, USA: Sage Publications, Inc., 2014.

[4] D. Potosky, "A conceptual framework for the role of the administration medium in the personnel assessment process," *Academy of Management Review*, vol. 33, no. 3, pp. 629–648, 2008. [Online]. Available: https://doi.org/10.5465/AMR.2008.32465704

[5] P. Rao S. B, S. Rasipuram, R. Das, and 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*, ser. ICMI '17. New York, NY, USA: ACM, 2017, pp. 221–229. [Online]. Available: http://doi.acm.org/10.1145/3136755.3136756

[6] B. Spitzberg and T. Adams, *CSRS, the Conversational Skills Rating Scale: An Instructional Assessment of Interpersonal Competence*, ser. NCA diagnostic series. NCA, National Communication Association, 2007. [Online]. Available: https://books.google.co.in/books?id=TiZYnwEACAAJ

[7] E. B. Page, "The imminence of... grading essays by computer," *The Phi Delta Kappan*, vol. 47, no. 5, pp. 238–243, 1966. [Online]. Available: http://www.jstor.org/stable/20371545

[8] S. Burrows, I. Gurevych, and B. Stein, "The eras and trends of automatic short answer grading," *International Journal of Artificial Intelligence in Education*, vol. 25, no. 1, pp. 60–117, Mar 2015. [Online]. Available: https://doi.org/10.1007/s40593-014-0026-8

[9] J. Z. Sukkarieh and J. Blackmore, "c-rater: Automatic content scoring for short constructed responses." in *FLAIRS Conference*, 2009, pp. 290–295.

[10] J. Z. Sukkarieh and S. G. Pulman, "Information extraction and machine learning: Auto-marking short free text responses to science questions," in *Proceedings of the 2005 conference on artificial intelligence in education: Supporting learning through intelligent and socially informed technology*. IOS Press, 2005, pp. 629–637.

[11] L. Tandalla, "Scoring short answer essays," 2012. [Online]. Available: https://kaggle2.blob.core.windows.net/competitions/kaggle/2959/media/TechnicalMethodsPaper.pdf. Accessed Nov, 2017.

[12] L. Ramachandran, J. Cheng, and P. Foltz, "Identifying patterns for short answer scoring using graph-based lexico-semantic text matching," in *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*. Denver, Colorado: Association for Computational Linguistics, Jun. 2015, pp. 97–106. [Online]. Available: https://www.aclweb.org/anthology/W15-0612

[13] M. Mohler, R. Bunescu, and R. Mihalcea, "Learning to grade short answer questions using semantic similarity measures and dependency graph alignments," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, ser. HLT '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 752–762. [Online]. Available: http://dl.acm.org/citation.cfm?id=2002472.2002568

[14] D. Meurers, R. Ziai, N. Ott, and J. Kopp, "Evaluating answers to reading comprehension questions in context: Results for German and the role of information structure," in *Proceedings of the TextInfer 2011 Workshop on Textual Entailment*. Edinburgh, Scottland, UK: Association for Computational Linguistics, Jul. 2011, pp. 1–9. [Online]. Available: https://www.aclweb.org/anthology/W11-2401

[15] D. Higgins, C. Brew, M. Heilman, R. Ziai, L. Chen, A. Cahill, M. Flor, N. Madnani, J. R. Tetreault, D. Blanchard, D. Napolitano, C. M. Lee, and J. Blackmore, "Is getting the right answer just about choosing the right words? the role of syntactically-informed features in short answer scoring," *ArXiv*, vol. abs/1403.0801, 2014, accessed Oct 30, 2019.

[16] B. Riordan, A. Horbach, A. Cahill, T. Zesch, and C. M. Lee, "Investigating neural architectures for short answer scoring," in *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 159–168. [Online]. Available: https://www.aclweb.org/anthology/W17-5017

[17] E. B. Page, "Computer grading of student prose, using modern concepts and software," *The Journal of Experimental Education*, vol. 62, no. 2, pp. 127–142, 1994. [Online]. Available: https://doi.org/10.1080/00220973.1994.9943835

[18] Y. Attali and J. Burstein, "Automated essay scoring with e-rater® v.2," *The Journal of Technology, Learning and Assessment*, vol. 4, no. 3, Feb. 2006. [Online]. Available: https://ejournals.bc.edu/index.php/jtla/article/view/1650

[19] P. W. Foltz, D. Laham, and T. K. Landauer, "The intelligent essay assessor: Applications to educational technology," *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, vol. 1, no. 2, pp. 939–944, 1999.

[20] L. M. Rudner and T. Liang, "Automated essay scoring using bayes' theorem," *The Journal of Technology, Learning and Assessment*, vol. 1, no. 2, Jun. 2002. [Online]. Available: https://ejournals.bc.edu/index.php/jtla/article/view/1668

[21] D. Alikaniotis, H. Yannakoudakis, and M. Rei, "Automatic text scoring using neural networks," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 715–725. [Online]. Available: https://www.aclweb.org/anthology/P16-1068

[22] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th International Conference on Machine Learning*, ser. ICML '08. New York, NY, USA: ACM, 2008, pp. 160–167. [Online]. Available: http://doi.acm.org/10.1145/1390156.1390177

[23] K. Taghipour and H. T. Ng, "A neural approach to automated essay scoring," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 1882–1891. [Online]. Available: https://www.aclweb.org/anthology/D16-1193

[24] L. M. Batrinca, N. Mana, B. Lepri, F. Pianesi, and N. Sebe, "Please, tell me about yourself: Automatic personality assessment using short self-presentations,"

in *Proceedings of the 13th International Conference on Multimodal Interfaces*, ser. ICMI. New York, NY, USA: ACM, 2011, pp. 255–262. [Online]. Available: http://doi.acm.org/10.1145/2070481.2070528

[25] J.-I. Biel, L. Teijeiro-Mosquera, and D. Gatica-Perez, "Facetube: Predicting personality from facial expressions of emotion in online conversational video," in *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, ser. ICMI. New York, NY, USA: ACM, 2012, pp. 53–56. [Online]. Available: http://doi.acm.org/10.1145/2388676.2388689

[26] L. S. Nguyen, D. Frauendorfer, M. S. Mast, and D. Gatica-Perez, "Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior," *IEEE Transactions on Multimedia*, vol. 16, no. 4, pp. 1018–1031, June 2014. [Online]. Available: https://doi.org/10.1109/TMM.2014.2307169

[27] S. Rasipuram and D. B. Jayagopi, "Automatic assessment of communication skill in interface-based employment interviews using audio-visual cues," in *2016 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, July 2016, pp. 1–6. [Online]. Available: https://doi.org/10.1109/ICMEW.2016.7574733

[28] I. Naim, M. I. Tanveer, D. Gildea, and M. E. Hoque, "Automated prediction and analysis of job interview performance: The role of what you say and how you say it," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 1, May 2015, pp. 1–6. [Online]. Available: https://doi.org/10.1109/FG.2015.7163127

[29] S. Rasipuram, P. Rao S. B, and 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, ser. ICMI. New York, NY, USA: ACM, 2016, pp. 370–377. [Online]. Available: http://doi.acm.org/10.1145/2993148.2993183

[30] M. E. Hoque, M. Courgeon, J.-C. Martin, B. Mutlu, and R. W. Picard, "Mach: My automated conversation coach," in *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp '13. New York, NY, USA: ACM, 2013, pp. 697–706. [Online]. Available: http://doi.acm.org/10.1145/2493432.2493502

[31] M. I. Tanveer, E. Lin, and M. E. Hoque, "Rhema: A real-time in-situ intelligent interface to help people with public speaking," in *Proceedings of the 20th International Conference on Intelligent User Interfaces*, ser. IUI '15. New York, NY, USA: ACM, 2015, pp. 286–295. [Online]. Available: http://doi.acm.org/10.1145/2678025.2701386

[32] M. Fung, Y. Jin, R. Zhao, and M. E. Hoque, "Roc speak: Semi-automated personalized feedback on nonverbal behavior from recorded videos," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp '15. New York, NY, USA: ACM, 2015, pp. 1167–1178. [Online]. Available: http://doi.acm.org/10.1145/2750858.2804265

[33] M. I. Tanveer, R. Zhao, K. Chen, Z. Tiet, and M. E. Hoque, "Automanner: An automated interface for making public speakers aware of their mannerisms," in *Proceedings of the 21st International Conference on Intelligent User Interfaces*, ser. IUI '16. New York, NY, USA: ACM, 2016, pp. 385–396. [Online]. Available: http://doi.acm.org/10.1145/2856767.2856785

[34] H. Tanaka, S. Sakti, G. Neubig, T. Toda, H. Negoro, H. Iwasaka, and S. Nakamura, "Automated social skills trainer," in *Proceedings of the 20th International Conference on Intelligent User Interfaces*, ser. IUI

'15. New York, NY, USA: ACM, 2015, pp. 17–27. [Online]. Available: http://doi.acm.org/10.1145/2678025.2701368

[35] V. Rus, B. Wyse, P. Piwek, M. Lintean, S. Stoyanchev, and C. Moldovan, "The first question generation shared task evaluation challenge," in *Proceedings of the 6th International Natural Language Generation Conference*, 2010. [Online]. Available: https://www.aclweb.org/anthology/W10-4234

[36] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, pp. 3104–3112. [Online]. Available: http://dl.acm.org/citation.cfm?id=2969033.2969173

[37] I. V. Serban, A. García-Durán, C. Gulcehre, S. Ahn, S. Chandar, A. Courville, and Y. Bengio, "Generating factoid questions with recurrent neural networks: The 30M factoid question-answer corpus," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 588–598. [Online]. Available: https://www.aclweb.org/anthology/P16-1056

[38] X. Du, J. Shao, and C. Cardie, "Learning to ask: Neural question generation for reading comprehension," *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017. [Online]. Available: http://dx.doi.org/10.18653/v1/p17-1123

[39] Q. Zhou, N. Yang, F. Wei, C. Tan, H. Bao, and M. Zhou, "Neural question generation from text: A preliminary study," in *Natural Language Processing and Chinese Computing*, X. Huang, J. Jiang, D. Zhao, Y. Feng, and Y. Hong, Eds. Cham: Springer International Publishing, 2018, pp. 662–671. [Online]. Available: https://doi.org/10.1007/978-3-319-73618-1_56

[40] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2383–2392. [Online]. Available: https://www.aclweb.org/anthology/D16-1264

[41] X. Yuan, T. Wang, C. Gulcehre, A. Sordoni, P. Bachman, S. Zhang, S. Subramanian, and A. Trischler, "Machine comprehension by text-to-text neural question generation," in *Proceedings of the 2nd Workshop on Representation Learning for NLP*. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 15–25. [Online]. Available: https://www.aclweb.org/anthology/W17-2603

[42] Z. Wang, A. S. Lan, W. Nie, A. E. Waters, P. J. Grimaldi, and R. G. Baraniuk, "QG-net: A data-driven question generation model for educational content," in *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, ser. L@S '18. New York, NY, USA: ACM, 2018, pp. 7:1–7:10. [Online]. Available: http://doi.acm.org/10.1145/3231644.3231654

[43] M.-H. Su, C.-H. Wu, K.-Y. Huang, Q.-B. Hong, and 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. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1007

[44] X. Qiu and X. Huang, "Convolutional neural tensor network architecture for community-based question answering," in *Proceedings of the 24th International Conference on Artificial Intelligence*, ser. IJCAI'15. AAAI Press, 2015, pp. 1305–1311. [Online]. Available: http://dl.acm.org/citation.cfm?id=2832415.2832431

[45] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423

[46] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018. [Online]. Available: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf. Accessed Oct 30, 2019.

[47] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 328–339. [Online]. Available: https://www.aclweb.org/anthology/P18-1031

[48] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, 2018. [Online]. Available: http://dx.doi.org/10.18653/v1/n18-1202

[49] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," *ArXiv*, vol. abs/1508.01991, 2015, accessed Oct 30, 2019.

[50] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=rJ4km2R5t7. Accessed Oct 30, 2019.

[51] T. Wolf, V. Sanh, J. Chaumond, and C. Delangue, "Transfertransfo: A transfer learning approach for neural network based conversational agents," *ArXiv*, vol. abs/1901.08149, 2019, accessed Oct 30, 2019.

[52] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, and Y. Choi, "COMET: Commonsense transformers for automatic knowledge graph construction," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 4762–4779. [Online]. Available: https://www.aclweb.org/anthology/P19-1470

[53] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=1953048.2078195

[54] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: LIWC," *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.

[55] N. Farra, S. Somasundaran, and J. Burstein, "Scoring persuasive essays using opinions and their targets," in *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*. Denver,

Colorado: Association for Computational Linguistics, Jun. 2015, pp. 64–74. [Online]. Available: https://www.aclweb.org/anthology/W15-0608

[56] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, ser. HLT '05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 347–354. [Online]. Available: https://doi.org/10.3115/1220575.1220619

[57] G. A. Miller, "Wordnet: A lexical database for english," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995. [Online]. Available: http://doi.acm.org/10.1145/219717.219748

[58] A. I. Huffcutt, J. M. Conway, P. L. Roth, and N. J. Stone, "Identification and meta-analytic assessment of psychological constructs measured in employment interviews." *Journal of Applied Psychology*, vol. 86, no. 5, p. 897, 2001. [Online]. Available: http://dx.doi.org/10.1037/0021-9010.86.5.897

[59] T. DeGroot and J. Gooty, "Can nonverbal cues be used to make meaningful personality attributions in employment interviews?" *Journal of Business and Psychology*, vol. 24, no. 2, pp. 179–192, Jun 2009. [Online]. Available: https://doi.org/10.1007/s10869-009-9098-0

[60] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM International Conference on Multimedia*, ser. MM '13. New York, NY, USA: ACM, 2013, pp. 835–838. [Online]. Available: http://doi.acm.org/10.1145/2502081.2502224

[61] P. P. G. Boersma *et al.*, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, 2002. [Online]. Available: http://www.praat.org/. Accessed Oct 30, 2019.

[62] T. Giannakopoulos, "pyaudioanalysis: An open-source python library for audio signal analysis," *PLOS ONE*, vol. 10, no. 12, pp. 1–17, 12 2015. [Online]. Available: https://doi.org/10.1371/journal.pone.0144610

[63] T. Giannakopoulos and A. Pikrakis, *Introduction to Audio Analysis: A MATLAB Approach*, 1st ed. Orlando, FL, USA: Academic Press, Inc., 2014.

[64] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett, "The computer expression recognition toolbox (cert)," in *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011)*. IEEE, 2011, pp. 298–305.

[65] (2016) Discover big voice intelligence. https://www.voicebase.com/products-features/. Accessed Oct 30, 2019.

[66] M. Kursa and W. Rudnicki, "Feature selection with the boruta package," *Journal of Statistical Software, Articles*, vol. 36, no. 11, pp. 1–13, 2010. [Online]. Available: https://www.jstatsoft.org/v036/i11. Accessed Oct 30, 2019.

[67] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[68] J. Levashina, C. J. Hartwell, F. P. Morgeson, and 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. [Online]. Available: https://doi.org/10.1111/peps.12052

[69] F. L. Schmidt, I. Oh, and J. Shaffer, "The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings." *Fox School of Business Research Paper*, 1998. [Online]. Available: http://dx.doi.org/10.1037/0033-2909.124.2.262

[70] M. Wan and J. McAuley, "Modeling ambiguity, subjectivity, and diverging viewpoints in opinion question answering systems," *2016 IEEE 16th International Conference on Data Mining (ICDM)*, Dec 2016. [Online]. Available: http://dx.doi.org/10.1109/icdm.2016.0060

[71] W. Hu, B. Liu, J. Ma, D. Zhao, and R. Yan., "Aspect-based question generation," in *ICLR Workshop*, 2018. [Online]. Available: https://openreview.net/pdf?id=rkRR1ynIf. Accessed Oct 30, 2019.

[72] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," *Trans. Sig. Proc.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997. [Online]. Available: http://dx.doi.org/10.1109/78.650093

[73] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1073–1083. [Online]. Available: https://www.aclweb.org/anthology/P17-1099

[74] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. [Online]. Available: https://www.aclweb.org/anthology/D14-1162

[75] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing," *To appear*, 2019.

[76] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019. [Online]. Available: https://

d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf. Accessed Oct 30, 2019.

[77] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 5998–6008. [Online]. Available: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

[78] A. Fan, M. Lewis, and Y. Dauphin, "Hierarchical neural story generation," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 889–898. [Online]. Available: https://www.aclweb.org/anthology/P18-1082

[79] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, pp. 1137–1155, Mar. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944966

[80] G. Shires, "Web Speech API: Draft community group report," July 2019. [Online]. Available: https://w3c.github.io/speech-api/. Accessed Oct 30, 2019.

[81] 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/. Accessed Oct 30, 2019.

[82] J. Barr, "Amazon polly – text to speech in 47 voices and 24 languages," November 2016. [Online]. Available: https://aws.amazon.com/blogs/aws/polly-text-to-speech-in-47-voices-and-24-languages/. Accessed Oct 30, 2019.

[83] S. Rao and H. Daumé III, "Answer-based Adversarial Training for Generating Clarification Questions," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 143–155. [Online]. Available: https://www.aclweb.org/anthology/N19-1013

[84] N. Mostafazadeh, C. Brockett, B. Dolan, M. Galley, J. Gao, G. Spithourakis, and L. Vanderwende, "Image-grounded conversations: Multimodal context for natural question and response generation," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, Nov. 2017, pp. 462–472. [Online]. Available: https://www.aclweb.org/anthology/I17-1047

APPENDIX A

RATING CRITERIA FOR INTERVIEWS

Description for Rating Scale referenced by the expert HR annotators. Rating on a 1-5 scale :1 – Inadequate, 2 – Fair, 3 – Adequate, 4 – Good, 5 – Excellent

Written Communication

1. Writing fluency: Does the candidate produce written language rapidly, appropriately, creatively, and coherently?
    - 1 - Writing is difficult to follow or to read. Sentence structure that frequently obscures meaning. Sentences that are disjointed, confusing.
    - 2 - Writing tends to be either choppy or rambling. Sentence patterns are monotonous.
    - 3 - Writing tends to be mechanical. Sentence structure, length varied, with repetitive sentence patterns. Good control over simple sentence structure
    - 4 - Sentence patterns are somewhat varied, some repeated patterns of sentence structure, length. Strong control over simple sentence structure
    - 5 - Sentences show a high degree of consistency; makes oral reading expressive. Extensive variation in sentence structure, length, and beginnings
2. Grammar: Are the answers grammatically right?

- 1 - English so poor as to be barely understandable. Need for extensive editing.
- 2 - Problematic Sentence construction. Many grammar errors. Substantial need for editing.
- 3 - Sentence construction generally correct. Few grammar errors. Written style wordy.
- 4 - Written style clear and effective. Consistent use of standard grammar.
- 5 - No grammatical errors. Correct grammar and usage that contribute to clarity and style.

3. Standard conventions and mechanics: Spelling mistakes/ Punctuations/ Capitalization

    - 1 - Limited use of conventions. Basic punctuation misses; spelling of common words and capitalization inconsistent.
    - 2 - Little use of basic conventions. Many end-of-sentence punctuation errors; frequent spelling and capitalization errors.
    - 3 - Some control over basic conventions. End-of-sentence punctuation is usually correct; spelling and capitalization errors.
    - 4 - Correct end-of-sentence punctuation; internal punctuation inconsistent. Spelling usually correct, esp. on common words; correct capitalization.
    - 5 - Strong control of conventions; effective use of punctuation that guides the reader through the text. Correct spelling, even of difficult words; correct capitalization.

4. Convincing: Was the participant convincing in answering questions?

    - 1 – Not at all convincing
    - 2 – Tries to convince occasionally.

- 3 – Neither fully convincing nor fully contradicting
- 4 – Very convincing
- 5 – Fully convincing

5. Confidence: Did the participant exhibit good confidence (completely sure of the answers)?
    - 1 – Not confident
    - 2 – Confident while answering at least one question
    - 3 – Confident while answering few questions
    - 4 – Confident most of the times
    - 5 – Fully confident
6. Word usage/Vocabulary: Does the writing exhibit good diction (style of writing as dependent upon choice of words)?
    - 1 - Usage of totally unnecessary and inappropriate words. Extremely limited range of words
    - 2 - Unnecessary, imprecise word usage and repetition.
    - 3 - Used only the common words necessary to convey meaning, nothing more
    - 4 - Vocabulary is varied. Usage of commonly understandable words with few unique words and jargons
    - 5 - Vocabulary is striking, varied. Usage of commonly understandable, unique words, jargons naturally and judiciously avoiding overly complex words
7. Content relevance: How relevant is the content of the answer with respect to the question?
    - 1 - Unknown
    - 2 - Irrelevant

- 3 - Contradictory
- 4 - Partially correct/ Incomplete
- 5 – Relevant

8. Overall written communication skill: Does the answer convey a clear, concise and accurate meaning?

    - 1 - Vague, unclear and illogical answers with very poor presentation. Ambiguous and obscured message.
    - 2 - Uncertain and doubtful answers with poor presentation.
    - 3 - Conveys the answer in an acceptable way with fair logic and distinctness.
    - 4 - Consistent and logical answer with good presentation and articulation.
    - 5 - Clear, concise and coherent answer with professional presentation throughout. No ambiguity in the message to be conveyed.

Short Essay

1. Ideas and Content

    - 1 - Writing lacks a central idea or purpose, too short to develop the idea (if present)
    - 2 - Main ideas and purpose are somewhat unclear or development is attempted but minimal. Insufficient details.
    - 3 - Writing has an easily identifiable purpose and main idea(s), although they may be overly broad or simplistic. Supporting detail is often limited, insubstantial, or occasionally slightly off-topic.
    - 4 - Writing is clear and focused. The reader can easily understand the main ideas. Support is present, although it may be limited or rather general.

- 5 – Writing is exceptionally clear, focused, and interesting. It holds the reader's attention. Main ideas stand out and are developed by strong support and rich details suitable to purpose.

2. Organization

   - 1 - Writing lacks coherence; organization seems haphazard, disjointed. No identifiable beginning, body and/or ending.
   - 2 - Writing lacks a clear organizational structure. A missing or extremely undeveloped beginning, body, and/or ending.
   - 3 - An attempt at sequencing and paragraph breaks with a beginning and an ending, undeveloped.
   - 4 - Clear sequencing and paragraph breaks; developed beginning and a recognizable closure
   - 5 – Effective Writing with sequencing and paragraph breaks; inviting beginning and a strong closure

3. Effective Communication (More emphasis on structure and delivery)

   - 1 - The writing seems to lack a sense of involvement or commitment. Vague, unclear and illogical writing with very poor presentation. Ambiguous and obscured message.
   - 2 - The writing provides little sense of involvement or commitment. Uncertain and doubtful writing with poor presentation.
   - 3 - The writer's commitment to the topic seems inconsistent. Conveys the message in an acceptable way with fair logic and distinctness.
   - 4 - A writer's voice is present. Consistent and logical writing with good presentation and articulation.

- 5 - The writer has chosen a voice appropriate for the topic, purpose, and audience. Clear, concise and coherent writing with professional presentation throughout. Communicates the message most effectively with no ambiguity.

Description for the remaining questionnaire are same as described for the written communication.

Oral Communication

1. Speaking Fluency: Displays speech disturbances or dysfluencies such as stutters, omissions, repetitions or noticeable pause fillers (e.g., um, uh, er, ah, okay, like, you know, I mean, etc.) and maintenance of speech flow.
    - 1 - Displays almost constant use of dysfluencies, pauses, repetitions. Hardly keeps the speech flow smooth.
    - 2 - Displays frequent use of dysfluencies, pauses, repetitions. Tries to keep the speech flow going
    - 3 - Displays occasional use of dysfluencies, pauses, repetitions. Generally, maintains the speech flow smooth.
    - 4 - Displays few dysfluencies and keeps speech flow almost smooth
    - 5 - Displays no noticeable dysfluencies, pauses, repetitions. Maintains smooth speech flow consistently.
2. Articulation: Pronounces words such that they are understandable
    - 1 - Speaks with frequent errors, slurs, and/or incomprehensible utterances, resulting in unclear statements.
    - 2 - Speaks with occasional errors, slurs, and/or incomprehensible utterances, resulting in slightly unclear statements.

- 3 - Speaks with only a small number of errors, slurs, and/or incomprehensible utterances, resulting in no slightly clear statements.
- 4 - Speaks with no noticeable errors, slurs, and/or incomprehensible utterances, and clear statements.
- 5 - Speaks with clearly comprehensible utterances, but not with excessive "clip" or stilted pronunciation.

3. Use of eye contact: Does the participant maintain proper eye contact with the screen?
    - 1 - Completely avoids eye contact (looks down, corners of the room)
    - 2 - Frequently avoids eye contact
    - 3 - Provides occasional eye contact
    - 4 - Provides frequent eye contact of brief duration
    - 5 - Provides frequent eye contact
4. Facial Expressiveness: Facially displays range of affect, animation of facial musculature, and normative facial expressions compatible with verbal content.
    - 1 - Constantly displays blank, uninterested or hypnotic gaze
    - 2 - Frequently displays blank, uninterested or hypnotic gaze
    - 3 - Occasionally displays blank, uninterested or hypnotic gaze
    - 4 - Generally displays variation in facial affect consistent with subject matter
    - 5 - Consistently displays variation in facial affect consistent with subject matter
5. Confidence: Was the participant confident (completely sure) and do not display confused/blank expressions?
    - 1 – displays almost no confidence in answering the questions

- 2 – displays confidence occasionally
- 3 – displays confidence moderately
- 4 – displays confidence frequently
- 5 – answers all questions confidently

6. Convincing: Was the participant convincing in answering questions?
    - 1 – Not at all convincing
    - 2 – Tries to convince occasionally.
    - 3 – Neither fully convincing nor fully contradicting
    - 4 – Very convincing
    - 5 – Fully convincing
7. Word usage/Vocabulary: Does the participant exhibit good diction (choice of words)?
    - 1 - Usage of totally unnecessary and inappropriate words. Extremely limited range of words
    - 2 - Unnecessary, imprecise word usage and repetition.
    - 3 - Used only the common words necessary to convey meaning, nothing more
    - 4 - Vocabulary is varied. Usage of commonly understandable words with few unique words and jargons
    - 5 - Vocabulary is striking, varied. Usage of commonly understandable, unique words, jargons naturally and judiciously avoiding overly complex words
8. Content relevance: How relevant is the content of the answer with respect to the question?
    - 1 - Unknown
    - 2 - Irrelevant

- 3 - Contradictory
- 4 - Partially correct/ Incomplete
- 5 – Relevant

9. Overall oral communication skill: How clear, concise and understandable was the participant in conveying the answers?
    - 1 - Vague, unclear and illogical answers with very poor presentation. Ambiguous and obscured message.
    - 2 - Uncertain and doubtful answers with poor presentation.
    - 3 - Conveys the answer in an acceptable way with fair logic and distinctness.
    - 4 - Consistent and logical answer with good presentation and articulation.
    - 5 - Clear, concise and coherent answer with professional presentation throughout. No ambiguity in the message to be conveyed.