Improving Asynchronous Interview System with Automatic Assessment and Follow-up Question Generation

Pooja Rao S. B.

Supervisor: Prof. Dinesh Babu Jayagopi

External Reviewer: Dr. Anutosh Maitra

Thesis Defense Examination

15th Oct 2019

Motivation

- Traditional interviews take place at the employer's site, in a quiet, distraction-free environment where presence is necessary

- Conventional hiring process is laden with challenges

- Recruitment is undergoing disruption by Artificial Intelligence

Challenges with Hiring

- Prolonged hiring

- Lack of interviewers

- Expensive labour costs

- Scheduling conflicts

- Not scalable

Alternatives

- Futuristic choices like social recruitment, online assessments, and video interviews

- Organisations adopting innovative methods like social media, proctored assessments, asynchronous or one-way interviews

- Online interviews are conducted using computer-mediated communication like instant messaging, email or video
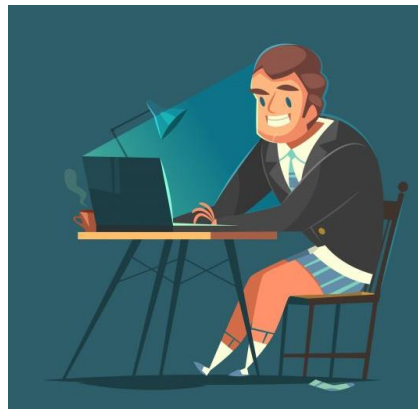
Online Interviews

Synchronous Interviews
Real-time simultaneous
communication exchange

Near-synchronous Interviews
near-immediate, on-going post
and response

Asynchronous Interviews
a time-lapse between
the communicating parties

Janet E. Salmons. Qualitative Online Interviews: Strategies, Design, and Skills. Sage Publications, Inc., Thousand Oaks, CA, USA, 2nd edition, 2014.

Preliminary Interviews Made Easy

Video Screening powered by Sonru is simply the best use of disruptive online video interviewing technology, designed specifically for the candidate selection process.
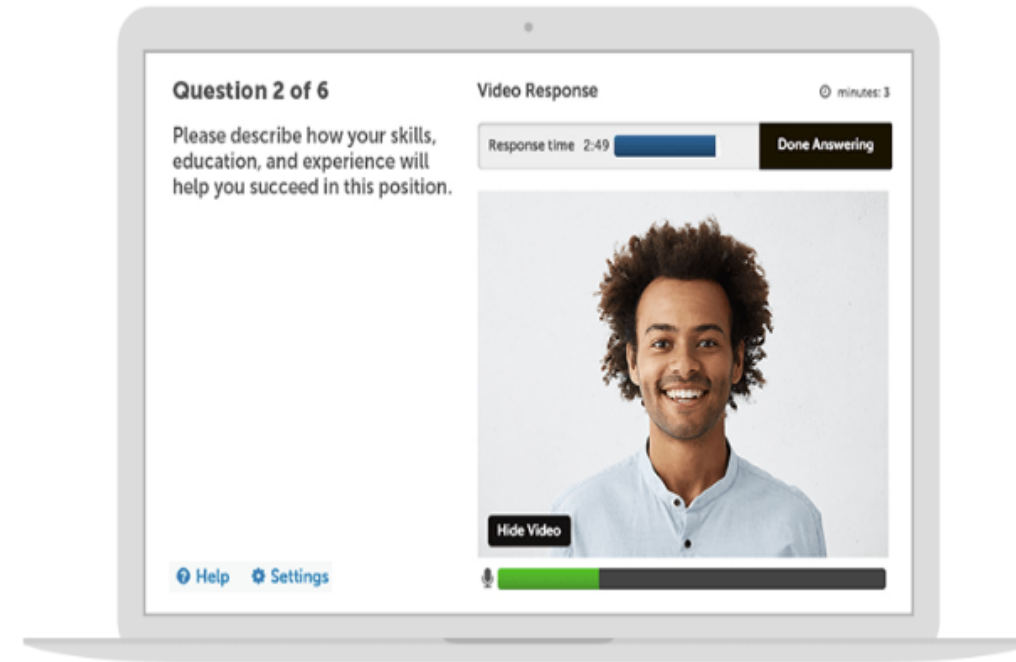
Sonru continually invests in our product design, user experience and best in class technology so you can focus on candidate selection. Even better, you will unlock substantial time and cost efficiencies.

HIREVUE VIDEO INTERVIEWING

WHERE STRUCTURED INTERVIEWS AND HIRING FLEXIBILITY MEET

SEE A DEMO

Question 2 of 6

Please describe how your skills, education, and experience will help you succeed in this position.

Video Response

⏱ minutes: 3

Response time 2:49

Done Answering

Hide Video

⦿ Help   ⚙ Settings

Attributes of Administration Medium

Transparency

Social Bandwidth

Interactivity

Surveillance

Denise Potosky. A conceptual framework for the role of the administration medium in the personnel assessment process. Academy of Management Review, 2008.

Attributes of Administration Medium

Feedback questionnaire and annotation from HR experts

Predictive model using multimodal features from audio, visual and lexical cues

Automatic follow-up question generation

Our Contributions

1. **Asynchronous Interview Dataset**

   ✓ Unique dataset of video and written interviews with short essays of same 100 participants

2. **Automated assessment of communication skill of participants**

   ✓ Automatic prediction and comparative analysis of non-conventional methods of employment interviews

3. **Automatic follow-up question generation**

   ✓ Approaches to generate follow-up questions given the interviewer question and the candidate response

Talk Outline

- Prior Work

- Automatic Communication Skill Prediction

- Comparative Analysis of Non-conventional Settings

- Automatic Follow-up Question Generation

- Future Directions

Talk Outline

- **Prior Work**

- Automatic Communication Skill Prediction

- Comparative Analysis of Non-conventional Settings

- Automatic Follow-up Question Generation

- Future Directions

Automated Short Answer Scoring

| Topic	Question
Science	Replicate an experiment based on the details of another experiment
Arts	Similarity between Pandas, Koalas and differences wrt. Pythons
Biology	Describe protein synthesis
English	Describe a character Mr. Leonard

Sample Questions from ASAP dataset

- C-rater, ETS 2009
- Louis Tandalla, Kaggle winner 2010
- Ramachandran, et al. Innovative Use of NLP for Building Educational Applications 2015
- Mohler, et al. ACL-HLT 2011
- Riordan, et al. Innovative Use of NLP for Building Educational Applications 2017
- Kumar, et al. AAAI 2019

Automated Essay Scoring

- Project Essay Grader, Page 1994

- The intelligent essay assessor, 2000

- Intellimetric, 2003

- E-rater, ETS 2006

- Alikaniotis, et al. ACL 2016

- Taghipour, et al. EMNLP 2016

- Jin, et al. ACL 2019

Automated Video Assessment

- Batrinca, et al. ICMI 2011

- Biel, et al. Facetube ICMI 2012

- Nguyen, et al. IEEE Transcations on Multimedia 2014

- Naim, et al. FG 2015

- Tanveer, et al. Rhema IUI 2015

- Fung, et al. Roc speak UbiComp 2015

- Rasipuram, et al. ICMI 2016
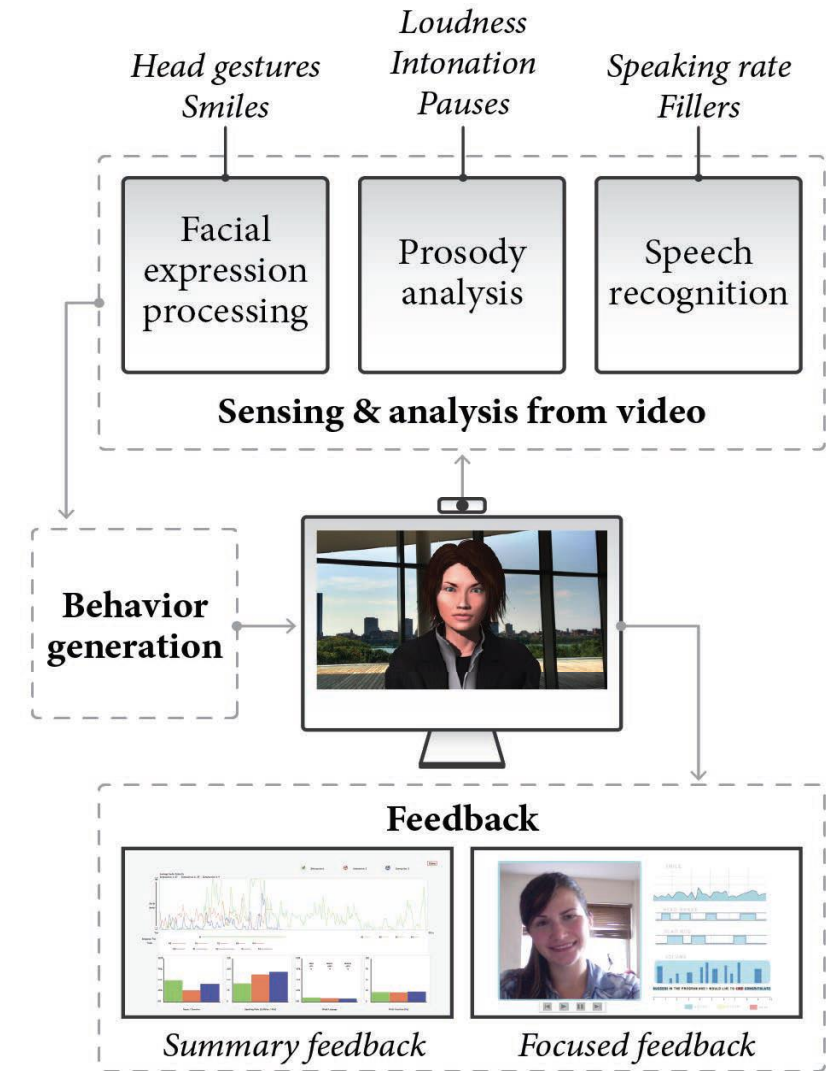
Mach: My automated conversation coach, UbiComp 2015

Question Generation

- Chali and Hasan. ACL 2015

- Serban, et al. ACL 2016

- Du, Shao and Cardie ACL 2017

- Yuan, et al. Rep4NLP@ACL 2017

- Wang, et al. L@S 2018

- Tang, et al. NAACL 2018

- Mrinmaya and Xing, NAACL 2018

- Su, et al. Interspeech 2018

Context: A healthy person in a family in which some members suffer from a recessive genetic disorder may want to know if he or she has the disease-causing gene and what risk exists of passing the disorder on to his or her offspring. Of course, doing **a test cross in humans** is unethical and impractical. Instead, geneticists use **pedigree analysis** to study the inheritance pattern of human genetic diseases.

QG-Net: What is unethical in humans?
QG-Net: What do geneticists use to study the inheritance pattern of human genetic diseases?
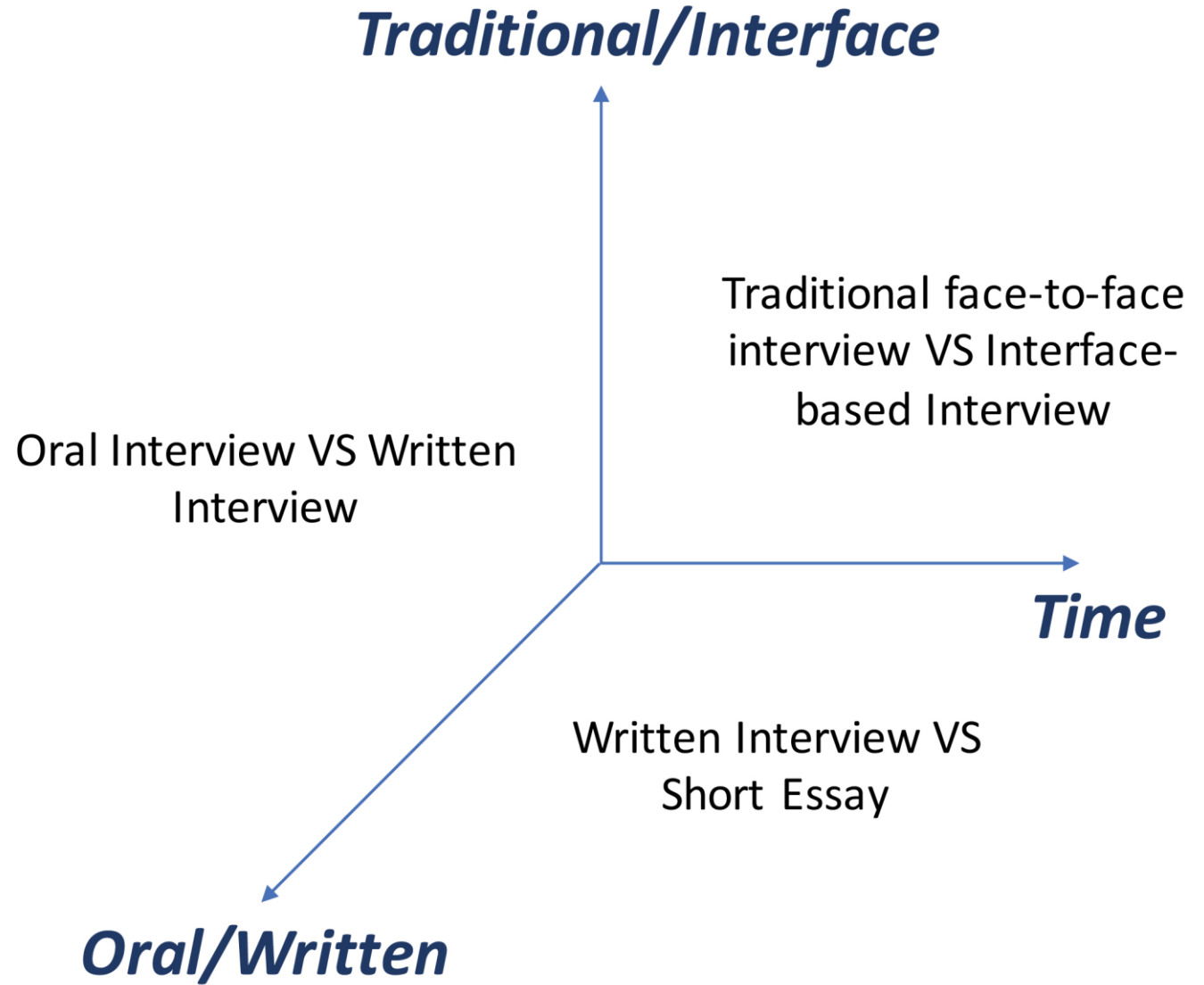
QG-Net, L@S 2018

Talk Outline

- Prior Work

- **Automatic Communication Skill Prediction**

- Comparative Analysis of Non-conventional Settings

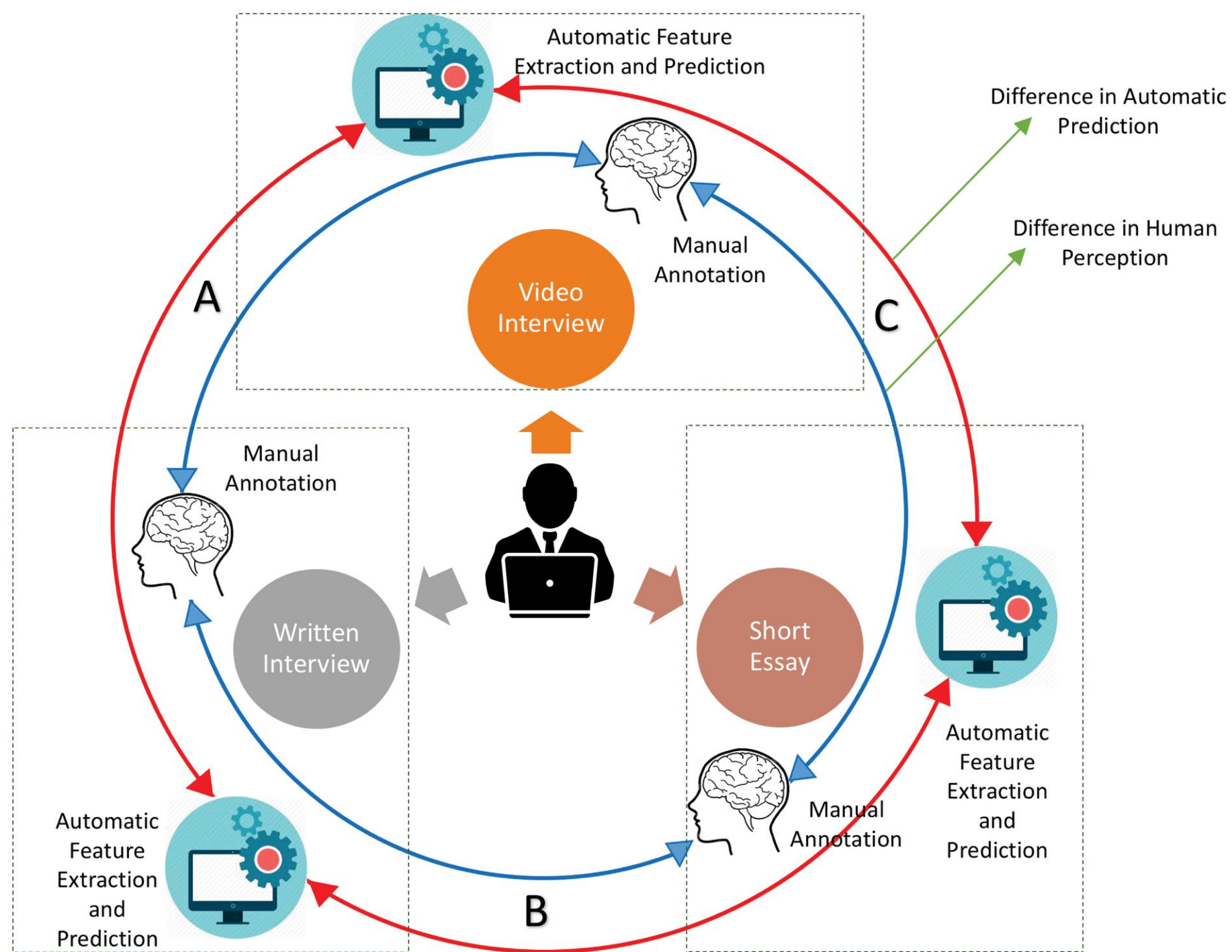- Automatic Follow-up Question Generation

- Future Directions

Non-Conventional Interview System

Traditional/Interface

Traditional face-to-face interview VS Interface-based Interview

Oral Interview VS Written Interview

Time

Written Interview VS Short Essay

Oral/Written

Data Collection

- Each participant is asked five behavioural questions

- Questions randomly selected from a pool of 75 commonly asked HR questions

- Each question is given 5 mins to answer in both settings

- Sixth question in the written phase is an essay topic

- Time limit to write essay is maximum of 30 mins

Data Statistics

AVERAGE RATING SCALE	1.5	2	2.5	3	3.5	4	4.5	5	Total
VIDEO INTERVIEW									100
WRITTEN INTERVIEW									118
SHORT ESSAY									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

- 5 mins per interview

- 70 words per answer

- 240 words per essay

- Two HR experts annotate all three interviews

- Scale of 1 to 5

- Average of two ratings

Rubric Analysis

- Indicates the necessity of lexical sophistication in video interviews

- Confidence and convincing attributes holds even in written interviews

- Ideas and content, organization substantial in essays

- Fluency, convincing and word usage essential in all three modes

RUBRICS ANALYSIS

Video Interview Rubrics	Correlation Coefficient r
Speaking fluency	0.88
Articulation	0.77
Use of eye contact	0.63
Facial expressiveness	0.71
Convincing	0.93
Confidence	0.86
Word usage/Vocabulary	0.86
Content relevance	0.85
Written Interview Rubrics	r
Writing fluency	0.86
Grammar	0.76
Conventions & Mechanics	0.74
Convincing	0.86
Confidence	0.8
Word usage/Vocabulary	0.84
Content relevance	0.8
Short Essay Rubrics	r
Writing fluency	0.89
Grammar	0.84
Conventions & Mechanics	0.81
Ideas and Content	0.95
Organisation	0.9
Word Usage/Vocabulary	0.87

Feature Extraction

Features For Written Interviews and Essay	Features for Video Interviews
Lexical Features	Prosody
LIWC	Speaking Activity
Sentiment Expression	Visual Features
Trait elements	Bag of Visual Words
	Text-based Features

Experimental Results

- XGBoost classifier is used to model the three systems separately on the binary classification task

- Baseline is 50% for the written interviews and essay and 59% for video interviews.

- Important feature groups:

  - Video – Energy, Lexical, Text-based features

  - Written - fusion of the lexical, sentiment expression and trait elements

  - Essay - count of word and tag collocations, count of subjectivity clues, exclusive words etc.

- More usage of word 'I' is a good personality indicator. First person singular in self introduction proved to be an important feature affirming the fact

Experimental Results

WRITTEN INTERVIEW	ACCURACY	F1 SCORE
Lexical	0.72	0.71
LIWC	0.7	0.68
Sentiment Expr.	0.6	0.59
Trait Elements	0.7	0.7
Lexical+Senti	0.73	0.73
Lexical+Senti+Traits	0.75	0.75
Boruta	0.71	0.7

Experimental Results

SHORT ESSAY	ACCURACY	F1 SCORE
Lexical	0.72	0.72
LIWC	0.73	0.72
Sentiment Expr	0.6	0.59
Lexical+Senti	0.71	0.7
Boruta	0.77	0.78

Experimental Results

VIDEO INTERVIEW	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	0.5	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

Feature Importance

Ranking	Feature Group	Fall in Accuracy
1	Best Model (Lex + Senti + Traits)	0
2	- Lexical	6%
3	- Traits	3%
4	- Sentiment	2%
5	+ LIWC	3%

| Important Features |
|--------------------|
| 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 |

Talk Outline

- Prior Work

- Automatic Communication Skill Prediction

- **Comparative Analysis of Non-conventional Settings**

- Automatic Follow-up Question Generation

- Future Directions

Comparative Analysis

- About 69 participants common across all the three settings

- Video and the written interviews differ in the mode of communication

- Written interview and the short essay differ on time dimension and the question type

- Majority of people who underperform in one mode of an interview are more likely to perform the same way in other modes

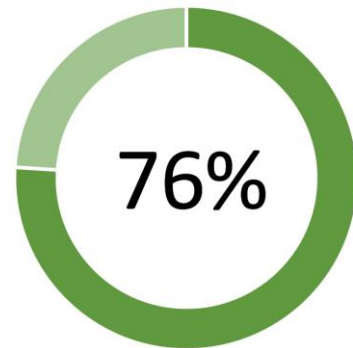- On average 83% of participants perform similarly in all three settings

Comparative Analysis

71%
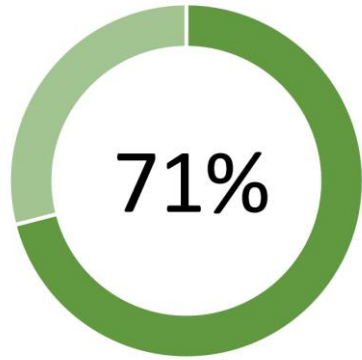
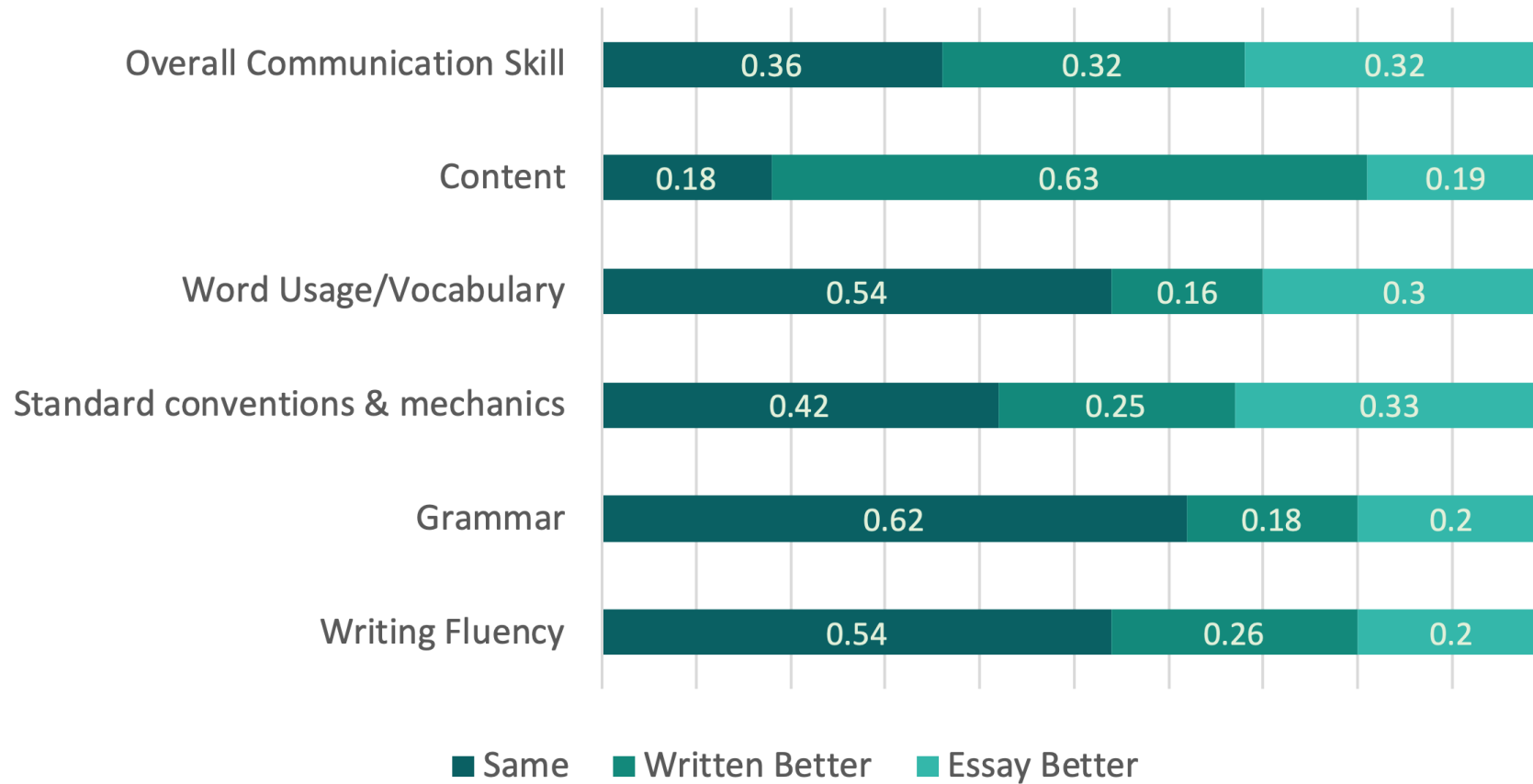Participants in class 1 of video interview assessment are in class 1 of written interview assessment

91%
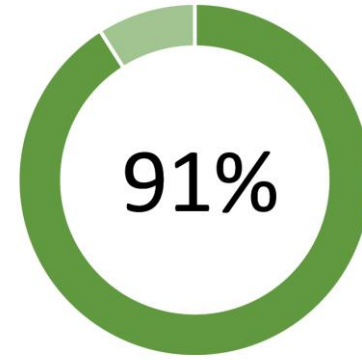
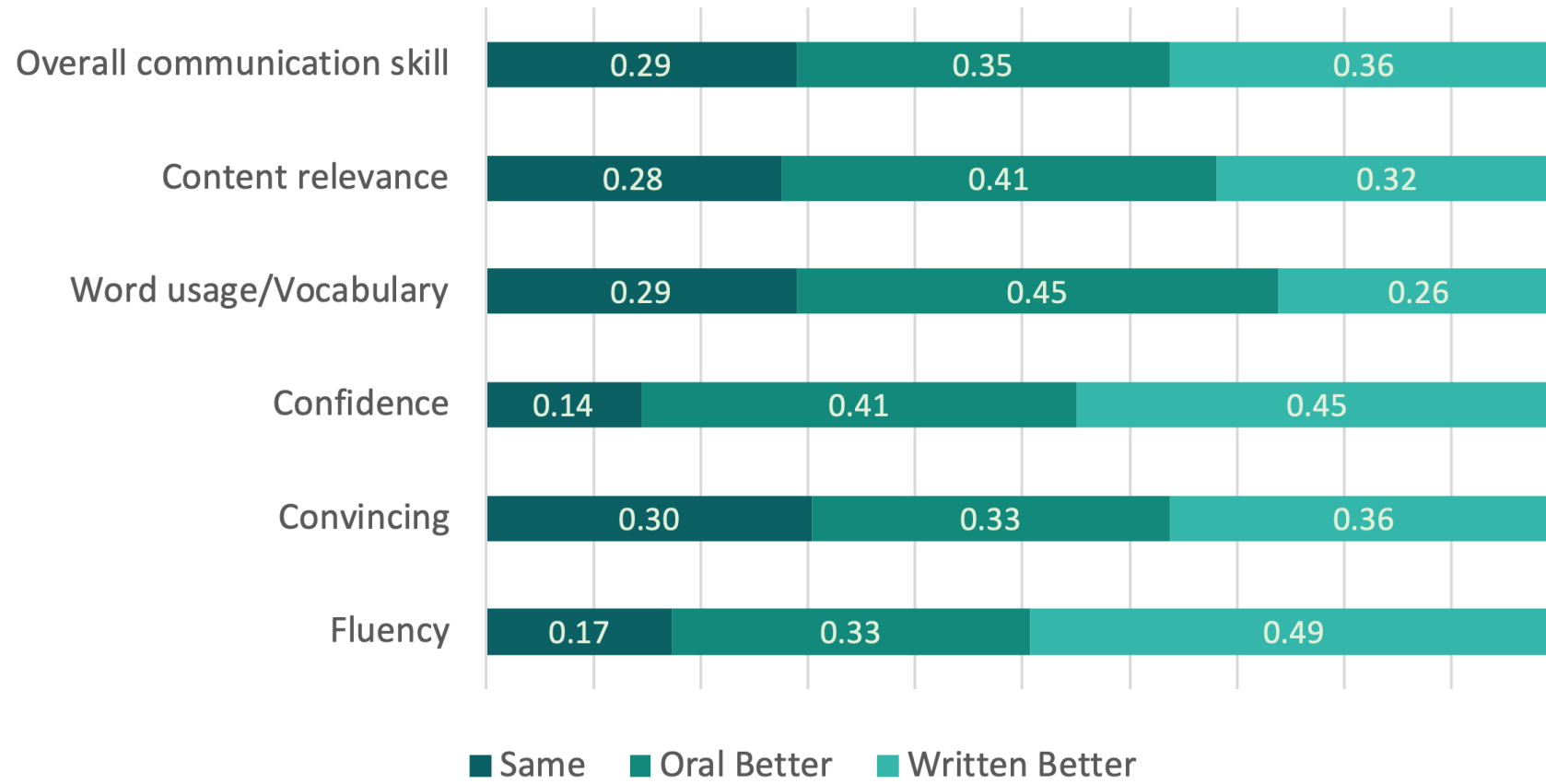Participants in class 1 of written interview assessment are in class 1 of essay assessment

76%

Participants in class 1 of essay assessment are in class 1 of video interview assessment

Comparison - Perceived Expert Annotations

	Same	Written Better	Essay Better
Overall Communication Skill	0.36	0.32	0.32
Content	0.18	0.63	0.19
Word Usage/Vocabulary	0.54	0.16	0.3
Standard conventions & mechanics	0.42	0.25	0.33
Grammar	0.62	0.18	0.2
Writing Fluency	0.54	0.26	0.2

■ Same   ■ Written Better   ■ Essay Better

Comparison - Perceived Expert Annotations

Feedback

- Questions about
  - Ease of usage and helpfulness of the interface – 97%
  - Recommendation of asynchronous systems to peers for interview assessment or training – 88%
  - Natural interaction – 63%
  - Comfortableness and expressiveness – 93%
- Preferences
  - Traditional face-to-face interviews – 51%
  - Interface-based video interviews – 27%
  - Interface-based written interviews – 22%

Talk Outline

- Prior Work

- Automatic Communication Skill Prediction

- Comparative Analysis of Non-conventional Settings

- **Automatic Follow-up Question Generation**

- Future Directions

Components of Asynchronous Interviews

- Asynchronous communication does not enable reciprocal exchange

- Adopt structured interviews

- Same questions/randomly selected from a fixed pool for each applicant

- Limited prompting and follow-up, and no elaboration on questions

- Reduce the impact of different biasing factors on interview ratings

Levashina, et al. The structured employment interview: Narrative and quantitative review of the research literature. Personnel Psychology, 2014
Schmidt, et al. The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 100 years. Fox School of Business Research Paper, 2016

Challenges

- With large scale adoption, might become repetitive and uninteresting for recruiters and candidates

- Increases predictability

- Reduces variability

- Makes them monotonous

- Crucial to find the right balance between structure and probing

Follow-up Questioning

Levashina et al. 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.

Follow-up Questioning

- Use of planned or limited probing may help

- Can collect more job-related information during the interview

- May lead to increased interview validity.

- Improves interaction

- Less predictable

Data Collection

A follow-up question dataset with 1089 triples

Task Formulation

- Training samples of {*q, r, f*}

  - q is the interviewer question

  - r is the candidate response

  - f is the follow-up question
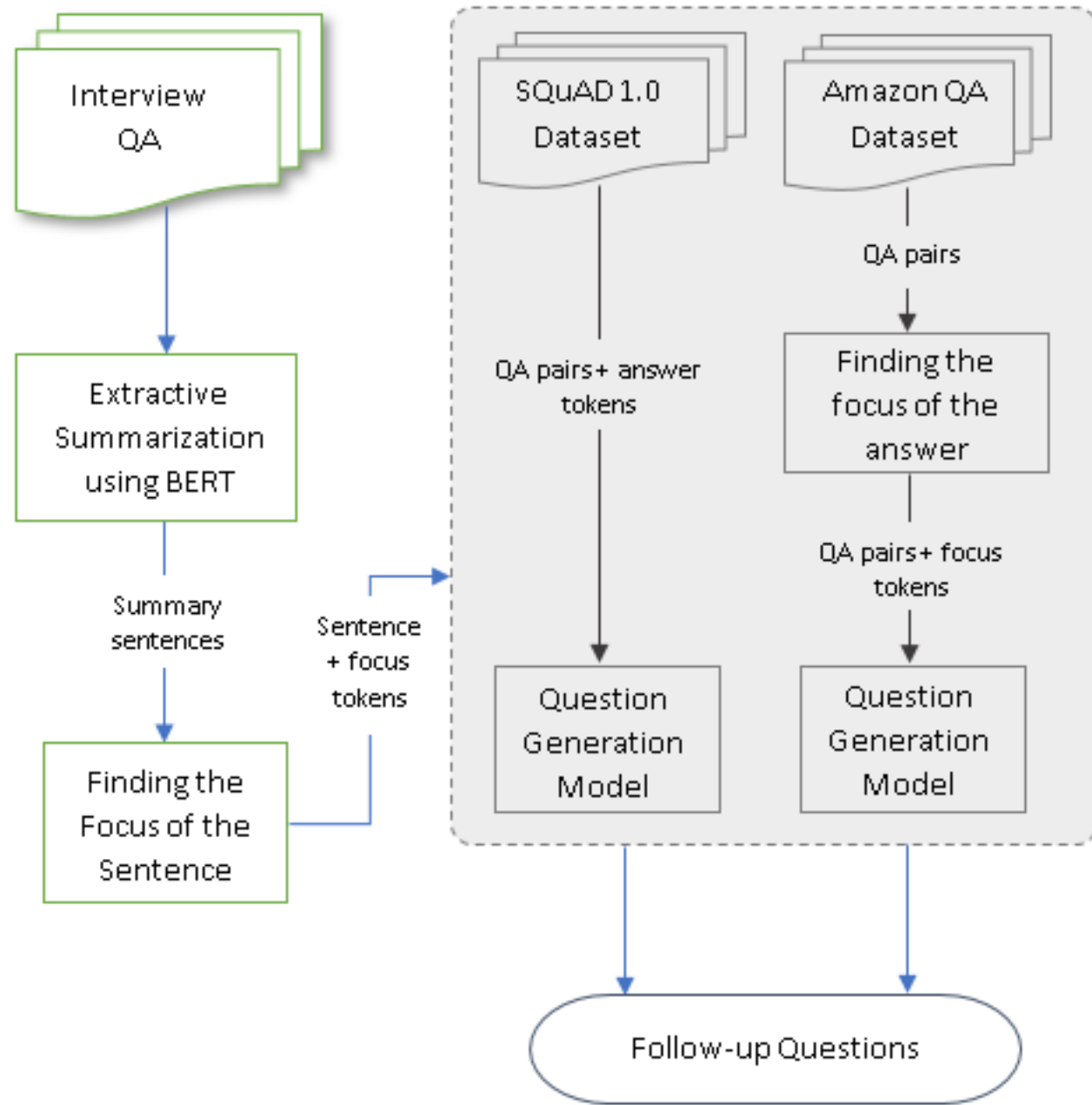
- Task is to generate f given q and r as inputs.

Automatic Follow-up Question Generation

- QG-Net based Follow-up Question Generation

- GPT-2 based Follow-up Question Generation

Automatic Follow-up Question Generation

- **QG-Net based Follow-up Question Generation**

- GPT-2 based Follow-up Question Generation

Approach

Finding the Focus of the Answer

- A and Q are represented as a sequence of tokens [a1;….;an] and [q1;….;qm]

- All the tokens in A as candidates for the topic and all the tokens in Q as voters polling for the candidates

- Final layer hidden state weights from the pre-trained language model BERT used as embeddings

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

Extractive Summarization using BERT

- Similar technique described to find the focus of each sentence employed

- R and S are two sentences from candidate response with their focus tokens represented as [f r1; ::::; f rp] and [f s1; ::::; f sq]

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

QG-Net Question Generation Model

- QG-Net is an encoder decoder (reader-generator) model

- Encoder is a bidirectional LSTM network

- Decoder is unidirectional LSTM incorporating pointer network

- Presence of focus tokens are encoded with each word as [0,1]

- Additional linguistic features like POS, NER and CAS are also encoded

- Words are all represented with d-dimensional GloVe vectors

- Trained on SQuAD and AQAD

Zichao Wang, Andrew S. Lan, Weili Nie, Andrew E. Waters, Phillip J. Grimaldi, and Richard G. Baraniuk. Qg-net: a data-driven question generation model for educational content. In L@S, 2018.

Qualitative Results

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 ?

Qualitative Results

Q: Are you an organised or a disorganised person, in what ways?

A: I think most of the time I am an organized person as I set goals for myself and ensure to complete them within stipulated time. I manage my time by making targets for myself and setting up deadlines and make sure to review my work on daily basis. I balance my schedule between academics, physical activities and social activities. Along with my career and academics, I pay proper attention towards my health and family.

Summary: I manage my time by making targets for myself and setting up deadlines and make sure to review my work on daily basis. I balance my schedule between academics, physical activities and social activities.

SQuAD
1. How do you feel about yourself ?
2. What is the main focus of your work ?

AQAD
1. Do you have to have a lot of time for it to work, or will it work on daily basis by making target ?
2. Can you balance the schedule between academics and social exercises ?

Qualitative Results

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 ?

Human Evaluation

- 76% of the questions score 2 or above in relevance, and 85% in grammar

- 70% and 85% of questions generated from SQuAD model score 2 or above in relevance and grammar respectively.

- 85% and 84% of questions generated from AQAD model score 2 or above in relevance and grammar respectively.
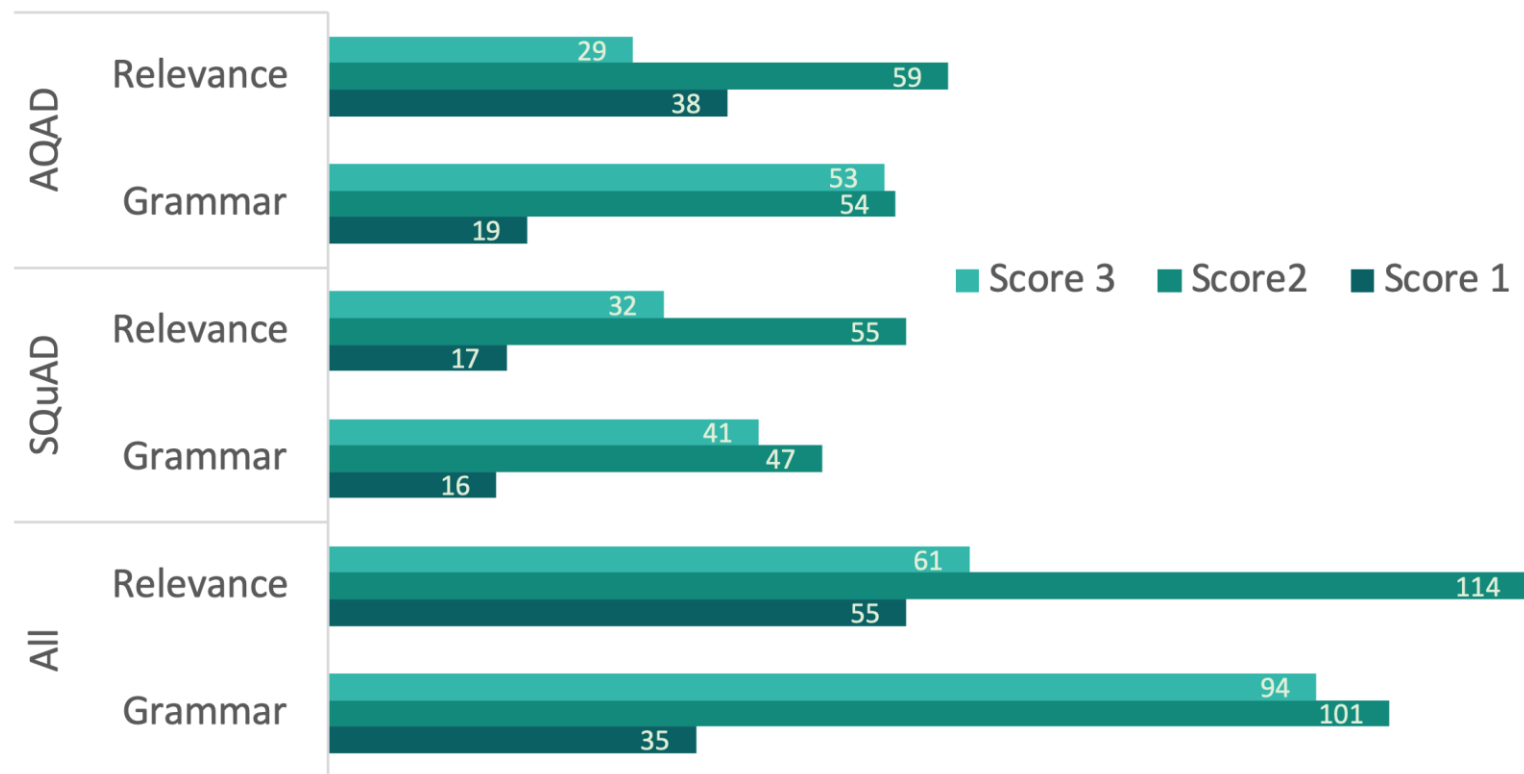
Automatic Follow-up Question Generation

- QG-Net based Follow-up Question Generation

- **GPT-2 based Follow-up Question Generation**

Approach

- Adaptation framework for generating follow-up questions using language models by fine-tuning it on $\{q, r, f\}$ triplets

- Data samples helps to learn the question structure and the relation between the triplets

- Knowledge from the language model pre-training produces novel questions

- A language model trained on a very large corpus of text will be able to generate long contiguous coherent text

GPT2 Language Model
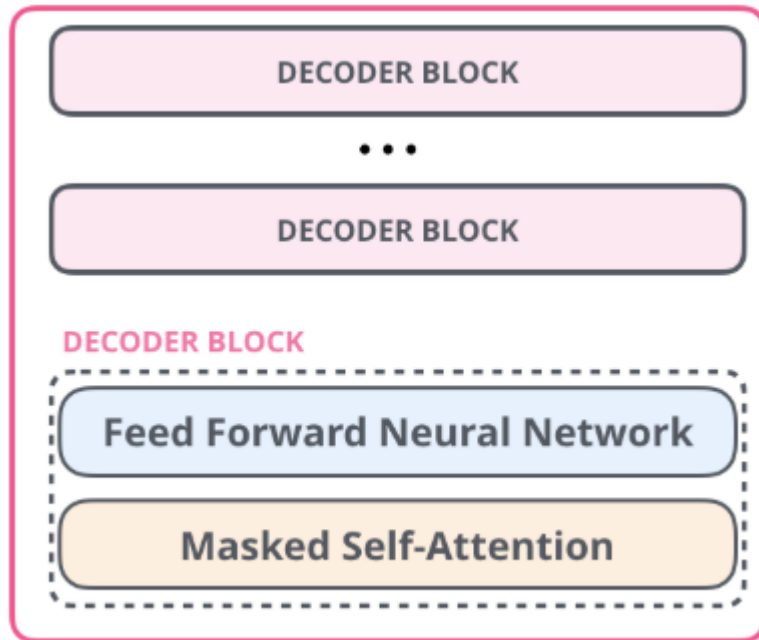
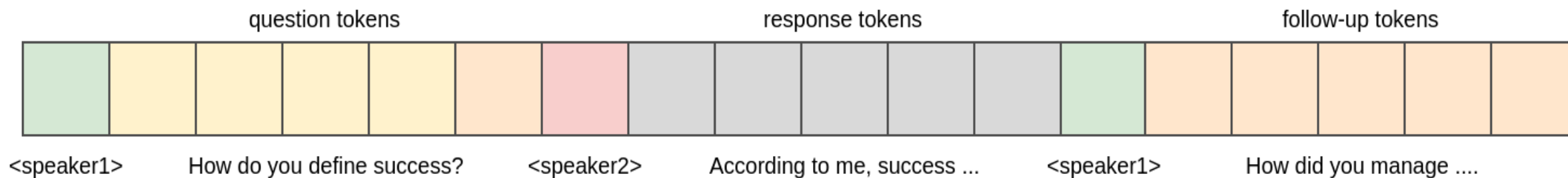Decoder-only transformer language model architecture

12-layer with self-attention heads containing 768 dimensional states and 12 attention heads.

Trained on the Web-Text dataset which contains the text of 45 million links from internet

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Input Representation

- Additional set of embeddings to indicate whether the token belongs to question, response or the follow-up

- These embeddings are learnt during the fine-tuning phase

question tokens — response tokens — follow-up tokens

<speaker1>   How do you define success?   <speaker2>   According to me, success ...   <speaker1>   How did you manage ....

Fine-tuning

- Trained for 3 epochs

- Only one level of history is considered

- Fine-tuning by optimizing two loss functions

  - Language modeling loss

  - Next-sentence classification loss

- Top-k random sampling strategy for decoding

Human Evaluation

- Almost all the time grammatically correct

- 77% of the questions are scored >= 2

- 89% of the questions score 2 or above taking rounded average

- 69% of questions generated on manual transcripts core 2 or above

- 64% of questions generated on ASR transcripts score 2 or above

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

Qualitative Results

	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?

Qualitative Results

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?

Talk Outline

- Prior Work

- Automatic Communication Skill Prediction

- Comparative Analysis of Non-conventional Settings

- Automatic Follow-up Question Generation

- **Future Directions**

Automatic Communication Skill Assessment

- A large and varied population of participants can contribute to the data

- Perform a domain adaptation using out-of-domain available datasets (eg. YouTube)

- Enhanced feature groups that capture the same aspects of the behavioural interviews across all the settings

- A computational formulation converging all the three models

Follow-up Question Generation

- Considering the complete history for questioning

- Augmenting knowledge base triples for generation

- Adversarial based training for to produce more diverse questions

- Incorporating the multimodal context of the audio and video

- Affect-aware conversational interviewing