InterviewBot: Real-Time End-to-End Dialogue System to Interview Students for College Admission

Anonymous ACL submission

Abstract

We present the InterviewBot that dynamically integrates conversation history and customized topics into a coherent embedding space to conduct 10 mins hybrid-domain (open and closed) conversations with foreign students applying to U.S. colleges for assessing their academic and cultural readiness. To build a neural-based end-to-end dialogue model, 7,361 audio recordings of human-to-human interviews are automatically transcribed, where 440 are manually corrected for finetuning and evaluation. To overcome the input/output size limit of a transformer-based encoder-decoder model, two new methods are proposed, context attention and topic storing, allowing the model to make relevant and consistent interactions. Our final model is tested both statistically by comparing its responses to the interview data and dynamically by inviting professional interviewers and various students to interact with it in real-time, finding it highly satisfactory in fluency and context awareness.

1 Introduction

With the latest advancement of Conversational AI, end-to-end dialogue systems have been extensively studied (Zhang et al., 2020; Adiwardana et al., 2020; Roller et al., 2021). One critical requirement is context awareness; robust dialogue systems must consider relevant parts in conversation history to generate pertinent responses (Serban et al., 2017; Mehri et al., 2019; Bao et al., 2020; Zhou et al., 2021; Xu et al., 2022). However, these systems still suffer from issues such as hallucination, inconsistency, or lacking commonsense (Bao et al., 2021), hindering them to take a place in real applications.

Numerous admission interviews are given every year to students located in 100+ countries applying to colleges in the U.S., where the interviews are often conducted online. Those interviews are usually unscripted with an emphasis on asking the applicants thought-provoking questions based on their interests and experiences. The main objective is to provide decision makers (e.g., admissions officers, faculty members) with an unfiltered look at those students in a daily academic environment.

Building an interview chatbot, called InterviewBot, will save time and effort of the interviewers, and provide foreign students a cost-efficient way of practicing interviews when native speakers are not available. Nonetheless, there are a few hurdles for developing an end-to-end InterviewBot. First, it is hard to collect a sufficient amount of data covering dialogues crossing open & closed domains (Sec. 3). Second, most transformer-based encoder-decoder models adapted by current state-of-the-art systems are not designed to handle long context; thus, they often repeat or forget previously discussed topics (Sec. 4). Third, it is demanding to find appropriate people to interactively test such a dialogue system with the professional objective (Sec. 5).

This paper presents an end-to-end dialogue system that interacts with international applicants to U.S. colleges. The system questions on critical perspectives, follows up on interviewee’s responses for in-depth discussions, and makes natural transitions from one topic to another until the interview ends, which lasts about 30 turns (5 mins for text-based, 10 mins for spoken dialogues). To the best of our knowledge, it is the first real-time system using a neural model, completely unscripted, conducting such long conversations for admission interviews.

2 Related Work

Dialogue systems can be categorized into closed- and open-domain systems (Ilievski et al., 2018). Closed-domain systems require efficient access to domain knowledge (Lian et al., 2019) and serve specific professions such as education (Cunningham-Nelson et al., 2019), healthcare (Fan et al., 2021; Amiri and Karahanna, 2022), or customer service (Baier et al., 2018; Nichifor et al., 2021). Open-domain systems converse across multiple domains.
with natural transitions (Adiwardana et al., 2020) and conduct interactions in a broader horizon (Ahmadvand et al., 2018; Wang et al., 2017; Finch et al., 2020). For admission interviews, however, the conversation is often a mixture of closed (job-related questions) and open domains (general aspects of the applicant) dialogues, which makes it more challenging to build an end-to-end system.

Several dialogue systems have been developed to communicate with humans for information exchange or elicitation (Finch et al., 2020; Li et al., 2017; Kim et al., 2019) across multiple domains (Safi et al., 2020; Khoa, 2021; Okonkwo and Adeibijola, 2021). Conversational agents for interviews have been experimented with for law enforcement (Minhas et al., 2022), healthcare (Ni et al., 2017), job application (Xiao et al., 2019), and psychology (Siddig and Hines, 2019), among which most are proof of concept. A few interviewbots have been developed on commercial platforms such as Google Dialogflow and IBM Watson Assistant, with the limitation of pre-scripted interviews; thus, they cannot proactively follow up to the user contents.

3 Interview Dataset

Audio recordings of 7,361 interviews are automatically transcribed with speaker identification by the online tool RevAI,1 where 440 are manually corrected on speaker ID assignment for finetuning and evaluation of our models (Table 1). Each recording contains an average of a ≈15-min long dialogue between an interviewer and an interviewee. The interviews were conducted by 67 professionals in 2018 - 2022. The largest age group of interviewees is 18-years-old with 59.3%, followed by 17-years-old with 29.4%. The male-to-female ratio is 1.2:1. The major country of origin is China with 81.4% followed by Belgium with 10.5%, alongside 37 other countries. Appendix A.1 provides detailed demographics of the interviewees.

4 Dialogue Generation

Figure 1 depicts an overview of our dialogue generation model. Since inputs to the encoder \( \mathcal{E} \) and the decoder \( \mathcal{D} \) are limited by the total number of tokens that the pretrained language model accepts, sliding window (Sec. 4.1) and context attention (Sec. 4.2) are proposed to handle long utterances and contexts in the previous utterances, respectively. In addition, topic storing is used to remember user-oriented topics brought up during the interview (Sec. 4.3). The input to \( \mathcal{E} \) and output of \( \mathcal{D} \) include the speaker ID \( S_1 \) or \( S_2 \) as the first token followed by an utterance from the interviewer or interviewee, respectively. Hyperparameters are finetuned by cross validations.

### Table 1: Distributions of our data

<table>
<thead>
<tr>
<th>D</th>
<th>U</th>
<th>S1</th>
<th>S2</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRN</td>
<td>140</td>
<td>43.8</td>
<td>39.3</td>
</tr>
<tr>
<td>DEV</td>
<td>150</td>
<td>45.0</td>
<td>36.2</td>
</tr>
<tr>
<td>TST</td>
<td>150</td>
<td>44.3</td>
<td>37.8</td>
</tr>
<tr>
<td>RAW</td>
<td>6,921</td>
<td>40.4</td>
<td>41.5</td>
</tr>
</tbody>
</table>

1https://www.rev.ai
matrix $A \in \mathbb{R}^{d \times n}$ such that $C^T \cdot A \rightarrow S^T \in \mathbb{R}^{d \times n}$. Thus, $S \in \mathbb{R}^{n \times d}$ represents the context summary of $U_{i-k}, \ldots, U_{i-1}$, which is fed into the decoder $D$.

### 4.3 Topic Storing

Even with the context attention, the model still has no memory of contexts prior to $U_{i-k}$, leading it to repeat the same topics that it has already initiated. To overcome this issue, topic storage is introduced to remember key topics derived by the interviewer. Every interview in our data came with 8-16 questions by the interviewer, who used those during the interview and thought they led to assess crucial aspects of the interviewee. Our final model considers these questions the “key topics” and dynamically stores them as the dialogue progresses.

Let $Q = \{q_1, \ldots, q_h\}$ be the topical question set. During training, $D$ learns to generate $Q$ instead of $S$ as the first token of the interviewer’s utterance that contains any $q_i \in Q$. In addition, it generates B/E if the interviewer begins/ends the current dialogue with that utterance (Table 3). Any utterance starting with $Q$ is encoded by $E$ that creates the utterance embedding $v_i \in \mathbb{R}^{1 \times d}$. These embeddings get stacked as the interview goes to create the topic matrix $V \in \mathbb{R}^{h \times d}$. If $|Q| < h$, then zero-padding is used to create $V$ (in our case, $h = 16$). Finally, $V$ is stacked with the context matrix $C$ (Sec. 4.2), and $(V \oplus C)^T \in \mathbb{R}^{d \times (h+\ell)}$ is multiplied by the attention matrix $A \in \mathbb{R}^{(h+\ell) \times n}$ to create the transpose of the context summary matrix $S \in \mathbb{R}^{n \times d}$.

### 5 Experiments

For our experiments, the encoder and the decoder in BlenderBot 1.0 (Roller et al., 2021) are used. Three models are developed as follows:

- **BB: Blender Baseline** that takes only up to 128 tokens in $U_{i-1}$ as context.
- **SW: BB + Sliding Window (Section 4.1)**, taking all tokens in $U_{i-1}$ as context.
- **CT: SW + Context Attention (Section 4.2) + Topic Storing (Section 4.3)**, taking all tokens in $\{U_{i-k}, \ldots, U_{i-1}\}$ as context.

All models are first trained on raw and finetuned on TRN in Table 1. To assess real-life performance, 10 interviews are conducted per model, where each interview consists of exactly 30 turns. A qualitative analysis is performed on the top-3 most frequently occurring errors as follows:

- **Repetitions**: how often it repeats topics already covered in the previous utterances.
- **Early Ending (EE)**: implies to end the interview without covering a sufficient amount of topics.
- **Off Topic (OT)**: how often it makes utterances that are not relevant to the current topic.

Table 2 shows the error analysis results. The repetition rates are significantly reduced as the model gets more advanced. Compared to the baseline, the CT model conducts 3.5 times longer conversations before it attempts to end the interview, while generating fewer fewer off-topic utterances, which is very promising. Examples of these error types are provided in Appendix A.2.

<table>
<thead>
<tr>
<th>Model</th>
<th>Repetitions (R)</th>
<th>Early Ending (EE)</th>
<th>Off Topic (OT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB</td>
<td>30.0</td>
<td>13.3</td>
<td>20.0</td>
</tr>
<tr>
<td>SW</td>
<td>16.7</td>
<td>23.3</td>
<td>26.7</td>
</tr>
<tr>
<td>CT</td>
<td>6.7</td>
<td>46.7</td>
<td>10.0</td>
</tr>
</tbody>
</table>

Table 2: The error analysis of all generation models. R: avg-% of repeated topics, EE: avg-% of the interview conducted before the model attempts to end (higher is better), OT: avg-% of off-topic utterances.

### 5.1 Static Evaluation

Following previous work, static evaluation is performed on the CT model, where the input is every
Hello, and thank you so much for coming today. Let us start by having you just to state your full legal name and also the school that you are currently attending?

Sure, my name is David, and I go to Shanghai high school in China.

Great. um, yeah, let’s talk about leadership experiences. Can you give me some specific examples?

Sure, when I was a junior, I was elected to be the Vice President in the Student Union and I organized many activities to help students to enrich their life on campus.

Okay, great. Thank you so much for sharing that. so, um, tell me a little bit about your leadership style. What do you think you bring to the student union?

Sure, I am very open-minded, and with some of my colleagues, we together established a lot of new programs on top of our existing ones. Many of those programs became very popular.

Ok, great. and so, um, what, uh, what would you say has been the most challenging aspect of being a leader?

Sure, I took Math, English, PE, Chemistry, Art, and so on. I like Math most.

Okay. so, um, tell me a little bit about your, your interest in mathematics.

Okay, great. well, thank you very much for sharing that with me. um, and thank you for taking the time out of your day to interview with us.

Table 3: An interview dialogue conducted by our best model (CT in Section 5). S1/S2: interviewer/interviewee (chatbot/human), B/E: beginning/ending utterance (chatbot), Q: topical question (chatbot).

<table>
<thead>
<tr>
<th>Rating Score</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>∑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluator Count</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>15</td>
</tr>
</tbody>
</table>

Table 4: The rating distribution of the InterviewBot conversations for real-time evaluation. 5: very satisfied, 4: satisfied, 3: neutral, 2: unsatisfied, 1: very unsatisfied.

In the future work, the main focus is to enrich the follow-up discussions on topics or sub-topics during interactions by training the InterviewBot with more structured topic-flow materials. This task would indirectly alleviate the early ending issue by deepening the discussions on certain topics.
7 Ethics Statement and Broader Impact

This research project is conducted under full compliance of research ethics in the ACL Code of Ethics. The data revealed in this paper are either generalized or generated from models that have no identifying or private information released. The interviewers and interviewees involved in the conversational data used in this paper have agreed to the usage of the data by the data collecting party and its collaborators for research purposes.

The future application of the Interviewbot will impact thousands of students applying to U.S. colleges. While institutions emphasizing on the research environment, the Interviewbot could help students understand the basics of research conduct of ethics, and even admission committees to make decisions on students’ readiness in compliance of code of ethics in research, through interview conversations.

However, constant observations of its performance may be enforced to prevent negative impact due to malfunction, biased behaving or unbalanced demographic distributions in learning.

References


A Appendix

A.1 Interviewee Demographics

We summarize the demographics of the interviewees in this section. Table 2 shows the distribution of the ages of applicants. Most interviewees are between 17 to 19, which is an accurate reflection of the ages of high school students applying to colleges.

![Figure 2: The interviewee age demographics.](image)

The distribution of countries of origin of applicants is shown in Figure 3. There are 38 countries in total. The majority of applicants come from China. Other major countries are Belgium, Bangladesh, Canada, India and Belarus.

![Figure 3: The interviewee country demographics.](image)

The gender distribution of applicants is shown in Figure 4. The numbers of male and female applicants are close, with exclusion of applicants not providing gender information.

![Figure 4: The interviewee gender demographics.](image)

A.2 Examples of Generation Limitations

The following are examples for each type of generation errors (underlined). Early ending prevents the Interviewbot from covering critical aspects of applicants and further discussion. The Interviewbot cannot name mentions properly, either skipping over or generating wrong entities in the conversations. Random Generation generates random words and punctuation, and concatenates them incorrectly. Not following up on previous topics hinders the Interviewbot from becoming proactively engaged with interviewees.

- Early Ending
  * InterviewBot: Could you please state your name?
  * Interviewee: Sure. My name is David.
  * ... 4 more turns ...
  * InterviewBot: Thank you and best wishes to your applications.

- Repetition
  * InterviewBot: Now why, why do you think receiving an education is important?
  * ... more turns ...
  * InterviewBot: Why do you think people value education so much?

- Off Topic
  * Interviewee: Sure, I really like Mathematics.
  * InterviewBot: What do you do for fun?