DAISI: the Deep Artificial Intelligence System for Interviews

Hyechan Jun
Department of Computer Science
Calvin University
Grand Rapids, MI 49546
hyechanjun@gmail.com

Ha-Ram Koo
Department of Computer Science
Calvin University
Grand Rapids, MI 49546
haramkoo@gmail.com

Advait Scaria
Department of Computer Science
Calvin University
Grand Rapids, MI 49546
ajs244@students.calvin.edu

Abstract

In this paper, we hope to harness the power of NLP by creating an AI interviewer capable of asking coherent, relevant, and useful interview questions, either on its own or as an aide for a human interviewer. Interviews are an integral part of gathering news and disseminating information and we hope to extend the power of interviewers by pairing them with an AI. To that end, we introduce "DAISI: the Deep Artificial Intelligence System for Interviews." By training on NPR interview data, we have created natural language models that can generate interview questions. We have also developed prototype metrics to score the performance of these models and an interactive web application for testing; however, there is yet much work to be done in order to fully realize this project.

1 Introduction

The goal of this project was to create an AI capable of generating interview questions, either to be used in tandem with a human interviewer or potentially even on its own when a human interviewer is not viable, such as in dangerous combat areas or situations requiring more manpower than a news company has available. This AI would ask relevant, meaningful interview questions that would achieve the journalistic goal of extracting information from a source [10]. Ideally, the AI would be able to carry on a conversation on its own and perhaps explore multiple paths a conversation could take and choose the most informative one, though that is beyond the scope of this paper.

More broadly, we also hoped that this project could serve as a launching point for further research into question-asking NLP. Currently, there is much more research being done on question-answering systems, the opposite of our task. Multiple datasets like SQuAD and QuAIL [9, 11] have been developed for use in question-answering tasks, but no similar dataset exists for question-asking. We found this to be a potential area of growth, given that question-asking could encompass tasks like generating questions for teachers to use or generating prompts for writers to explore creatively. By creating an interview AI, we hoped to contribute to a larger conversation on question-asking AI.

To achieve this task, we utilized models pre-trained on a task we saw as the most similar to question-asking: summarization. By fine-tuning freely available summarization models from the HuggingFace repository, we were able to create models that generate questions based on some context. We then developed some prototype metrics with which to evaluate these models, iterating through different

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

tests of models and metrics to continuously discover new issues, come up with solutions, and refine our methods.

Ultimately, we created 10 models, created a list of 6 qualitative metrics, and tested 3 different quantitative metrics. Our results are not astounding, nor are they truly complete. This project is still only in its early stages, and our contributions to the vast potential of question-asking AI are small; however, the models and metrics we created prove that question-asking AI is both viable and useful.

2 Background

To properly assess our models' performance, it was necessary to first determine what features made for good interview questions. To that end, we reached out to Professor Jesse Holcomb of the Calvin English Department, who teaches Journalism at a professional level. According to Professor Holcomb and the sources he provided, good interview questions have the following properties [10]:

1. Follow a set goal (whether that is obtaining a specific piece of information or more broadly learning about a subject)
2. Be open-ended
3. Result in the source talking more than the interviewer
4. Keep things on track toward the goal mentioned in 1.

In addition, we looked into current AI approaches in journalism, finding that most utilizations of NLP in the news field are for writing articles or engaging with users (e.g. through moderating comments) rather than asking questions [2]. These systems are designed to make the lives of journalists easier, but do not directly help them in the field.

The summarization models we utilized were from the HuggingFace model repository, specifically fine-tuned variations of BART [5]. This model has achieved state-of-the-art performance on multiple tasks, including SQuAD and GLUE, and was specifically recommended by HuggingFace in their example summarization task. The dataset we utilized to fine-tune our models was compiled from NPR transcripts by Majumder et al. [7], and consisted of labeled turns of conversation where each turn was tagged as either being a host's statement or a guest's statement.

To aid in our creation of metrics, we delved into prior work on dialogue bots, given that–though our task was explicitly unlike dialogue bots–we needed some example of what kind of metrics state-of-the-art systems were using to evaluate their models. To that end, we found that most chatbots are evaluated qualitatively based on user feedback [1, 8, 3], and furthermore are generally rated by how "human" they seem [4]. We mimicked these evaluations with our own set of qualitative metrics, though we also wished to develop more concrete, quantitative metrics in order to better quantify the performance of our models.

3 Approach

3.1 Data Preparation

For the majority of our models, we trained them on a processed version of the NPR dataset where the data was split into a "context" spoken by the guest (i.e. the one being interviewed) and a "question" asked by the host (the interviewer). To do so, we scraped the corpus and filtered every instance where the guest said something and the host responded with a question. Due to the fact that the original data was ordered by sentence-level utterances, this often meant combining several utterances into a single statement. These combined utterances formed single conversation turns; thus, no context was preserved from one data point to another. Moreover, any turn of conversation where the host did not ask a question was discarded. For example, the following exchange:

Guest (utterance 1): Good morning, Lulu.
Host (utterance 1): All right.
Host (utterance 2): What's the latest?

Would become:

Context: Good morning, Lulu.

Question: All right. What's the latest?

After scraping the dataset in this manner, we created a smaller dataset of around 80,000 context-question pairs. With a train-test split of 80:20, we fine-tuned the pre-trained BART model with various parameters, sometimes adding specific tags or removing certain characteristics. A full list of models and their attributes can be found in the Appendix, but the most useful models are presented in Table 1.

Table 1: Four Prominent Models

Model Name	**Characteristics**
Base Question Model	Model trained on the question dataset in its entirety
Blank Context Model	Model trained on data where the context was left blank but the question was still present
Generic Names Model	Model trained on data where named entities (e.g. people, places) were replaced by generic placeholders
Length Tagged Model	Model where tags were added to the context to force generation of certain length questions

Table 2 showcases some example output from our Base Question Model. The first is an example that we considered good; the second is one that we considered bad, for reasons that will be explored in the next section.

Table 2: Base Question Model Examples

	Good Example	**Bad Example**
Context	Well, I was born in South Africa, lived there until I was 17. Came to North America of my own accord, against my parent's wishes. And was in Canada for a few years. I started school there which is where I met my wife. Transferred down to the University of Pennsylvania and got a degree in physics, degree in business at Wharton. Came out to California with the intent of doing a PHD in the material science and physics [unintelligible] with an eye towards using that as an energy storage unit for electric vehicles. I ended up deferring that graduate work to start a couple to start a couple of area companies, one of which people have heard about, such as Pay Pal.	Well, a lot of things have happened in the interim. Two of the prosecutors who were assigned to the case became judges during this period, believe it or not. That's how long it's been, and they had to reconfigure the prosecution team. Phil Spector has had three different sets of lawyers, and now he has Bruce Cutler. And Bruce Cutler was the lawyer in New York who represented John Gotti; he's kind of known as a mob lawyer, a very enthusiastic, very flamboyant lawyer in the courtroom. He'll be something else to watch when you're watching on television.
Generated Question	Well, what are some of the things that people hear about the kind of ripple effects that have been seen in the auto industry?	What did you mean by that?

3.2 Evaluation

3.2.1 Qualitative Metrics

To evaluate our models, we began by creating six qualitative measures that we thought would make useful markers of good questions. We reasoned that these measures would be helpful because we noticed patterns in existing interview questions that fell into these categories.

1. **Appropriateness** is defined as the measure of how well the question flows in the conversation; that is, whether it breaks the conversation unnaturally or diverts the topic in a strange direction.
2. **Relevance** is defined as the measure of how closely related the question is to the context (in terms of ideas and themes).
3. **Specificity** is defined as the measure of how much detail is present in the question, whether that detail is pertinent to the context or not.
4. **Repetition** is defined as the measure of overlap of words between the question and the context.
5. **New Concepts** is defined as the measure of how much information was present in the question that was not in the context.
6. **Similarity to Original Question** is the measure of how similar the AI's response was to the host's original question, where we assume the host's question to be ground-truth and thus the best possible question (which, of course, may not always be the case).

Then, to provide a baseline, we randomly selected a set of 100 context-question pairs to use as a test set and scored the host's questions on that set using the first five metrics we devised (the sixth metric being useless because the similarity of the host's original question to the host's original question is obviously identical). These manual scores formed a general outline of what values we should expect from each metric to correspond with a good question. For example, a good interview question with the goal of extracting information should have high scores for appropriateness, relevance, and specificity, but a low score for new concepts and repetition.

From our manual scores, we found that length was a good predictor for several of our metrics. We found that–in general–questions that were longer were more specific, tended to introduce more new concepts, and were less relevant than shorter questions. Figure 1 depicts these relationships, from which we determined that questions of a particular length (<30 words) tended to make good interview questions.

3.2.2 Quantitative Metrics

Once we had defined our qualitative metrics, we began trying to find quantitative means of expressing them, with the goal of automating the evaluation of AI-generated questions. As mentioned previously, we found that length tended to be an acceptable predictor for several of our qualitative metrics, so we tried to see whether our model could be coerced into generating better questions by controlling its length. Unfortunately, our results were mixed; though shorter questions tended to be better overall than longer questions, we discovered that the appropriate length for a question was very context-dependent, meaning length was not as reliable of a metric as we had hoped. Questions that were too short (below 10 words) were often follow-up questions to previous ones, which posed a problem for our model because it operated on single turns only and thus lost all the context that made those short questions appropriate. Some long-form questions were similarly drawing from prior context by combining several things the guest had said previously, making the appropriate in context but inappropriate out of context. Given that our model could not handle more than a single turn of conversation, this meant length would fall short as a metric.

The second quantitative metric we considered was ROUGE [6], which measures n-gram overlap between two texts. We wanted to use ROUGE as a measure for repetition, which was a fairly obvious choice because ROUGE essentially measures how many words are repeated between two texts. After comparing ROUGE values to our manual metrics, we determined that a typically good question had a ROUGE F1 score between 0.1 and 0.25. The ROUGE output of our Base Question Model, when generating questions for the test set, is displayed in Figure 2.

Figure 1: Length as a Predictor of Three Metrics

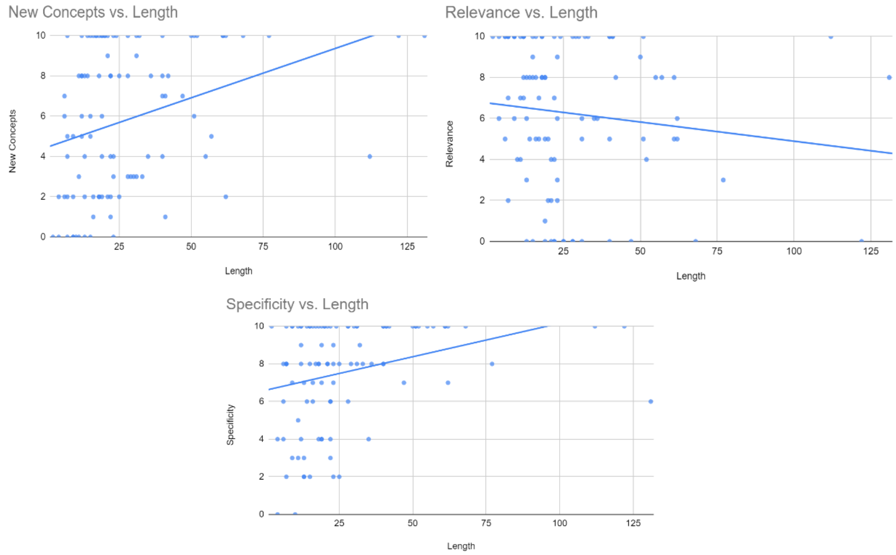
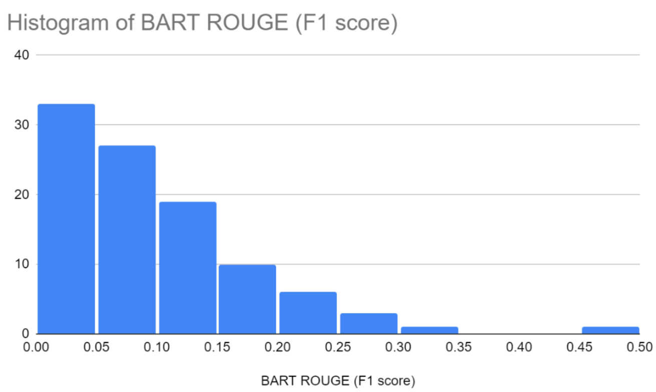


Figure 2: Histogram of Base Question Model ROUGE scores

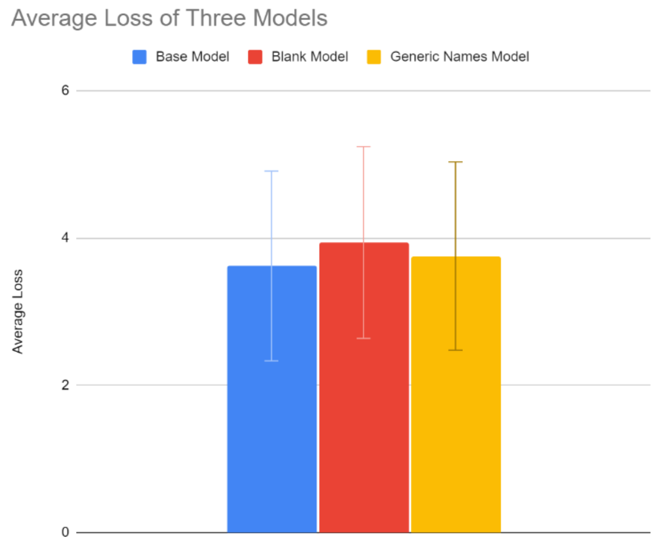


From the bounds we set, our Base model performed as we wished about 35% of the time, though we eventually realized that these bounds were rather arbitrary because, like length, the amount of repeated words in a question were dependent on the context. Sometimes it was appropriate to repeat a chunk of what was said in the context (e.g. if the context included a quote), and at other times it was better to not repeat any of the context (e.g. if the context had obvious information within it). We eventually determined that ROUGE, at least under our bounds, was not a robust enough metric to rely on.

The third quantitative metric we considered was the model's loss. The reasoning behind this was to use loss as a measure of relevance. We could train a model on blank contexts, thereby making an irrelevant model because it would simply have no context with which to ask questions. If we then compared the loss of our base model against this irrelevant model on our test set with labeled data, we could see if our base model was more relevant. To further experiment with loss, we also created a model trained on data where all named entities were made generic through the use of spaCy, replacing all the names of people with "person1" or "person2." This model would hopefully be more relevant than the blank model but less relevant than the base model. After running all three on the test set, we received the output in Figure 3.

As we had hoped, the average loss of the generic names model was lower than the blank model and the loss of the base model was lower still. However, the difference in loss tended to fluctuate greatly from question to question, and moreover was often negligible; as a result, though we had determined

Figure 3: Average Loss of Base, Blank, and Generic Name Models



that loss could be a viable metric, we unfortunately had to discard it as well because there was too much variability.

4 Discussion

As is evident from the previous section, our work has unfortunately not borne much progress. Though we have explored creating quantitative metrics from our qualitative baseline, our experiments have thus far only proven that we either need stricter bounds on what we define a good interview question to be or we need more rigid definitions of what constitutes relevance, appropriateness, or any other of our metrics. Our research has shown that there is yet more work that needs to be done in order to fully realize our goal of creating an interview AI; more quantitative metrics need to be uncovered, and potentially more qualitative metrics need to be devised as well.

However, that is not to say that our work was not useful. Our explorations and experiments have laid the foundation for future researchers to pick up where we left off, hopefully with more success than we have found. We have succeeded in creating prototype models and prototype metrics. We have begun the process of refining those models and metrics toward a more concrete idea of what makes a good question-asking AI. Most importantly, we have demonstrated that a question-asking AI model is viable, even if the technology is still in its infancy.

4.1 Future Work

One of the largest issues that plagued this project was the AI's inability to retain context for longer than a single turn of conversation. Our results were thus limited to single exchanges between interviewer and guest, which stripped away all the necessary context that could help inform our metrics. Our model was only good at responding to single statements at a time and incapable of holding a proper conversation, thereby lacking a key skill for an interviewer. We experimented a little with models like BigBird [12] to try and overcome this issue by elongating the context through prepending prior information, but our efforts did not lead very far. A large part of future work will be to make the system robust enough that it can hold natural conversations.

Another large part of future work will be crowd-sourced reviews of our model's performance. As we discovered when reading literature on chatbots, the current methods of evaluating this kind of conversational AI revolve around qualitative assessments made by a variety of people. We need to implement the same sort of approach to gather better qualitative assessments of our model, preferably on more than just 100 context-question pairs. To that end, we have created a web application where anyone can interact with our models; however, we have yet to implement a method to assess our models, which will need to be completed before any large-scale crowd-sourced assessment can take place.

Acknowledgements

We would like to thank Professor Kenneth Arnold for acting as our advisor during the course of this project and the Computer Science department at Calvin for providing the resources necessary to see this project through to the end.

References

[1] Alaa Abd-Alrazaq, Zeineb Safi, Mohannad Alajlani, Jim Warren, Mowafa Househ, and Kerstin Denecke. Technical metrics used to evaluate health care chatbots: Scoping review. *J Med Internet Res*, 22(6):e18301, Jun 2020.

[2] Meredith Broussard, Nicholas Diakopoulos, Andrea L Guzman, Rediet Abebe, Michel Dupagne, and Ching-Hua Chuan. Artificial intelligence and journalism. *Journalism & mass communication quarterly*, 96(3):673–695, 2019.

[3] Aaron Daniel Cohen, Adam Roberts, Alejandra Molina, Alena Butryna, Alicia Jin, Apoorv Kulshreshtha, Ben Hutchinson, Ben Zevenbergen, Blaise Hilary Aguera-Arcas, Chung-ching Chang, et al. Lamda: Language models for dialog applications. 2022.

[4] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1061–1068, 2014.

[5] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[6] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[7] Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. Interview: Large-scale modeling of media dialog with discourse patterns and knowledge grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8129–8141, 2020.

[8] Dijana Peras. Chatbot evaluation metrics. *Economic and Social Development: Book of Proceedings*, pages 89–97, 2018.

[9] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[10] Journalist's Resource. Interviewing a source: Tips, Dec 2020.

[11] Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. Getting closer to ai complete question answering: A set of prerequisite real tasks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8722–8731, 2020.

[12] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297, 2020.

Appendix: List of Models

Model Name	**Characteristics**
Base Question Model	Model trained on the question dataset in its entirety
Blank Context Model	Model trained on data where the context was left blank but the question was still present
Generic Names Model	Model trained on data where named entities (e.g. people, places) were made generic with the use of spaCy
Length Tagged Model	Model where tags were added to the context to force generation of certain length questions
Full Model	Model that was trained on the entirety of the NPR dataset, not just turns ending in a question mark
Length Percentile Model (Context)	Model that was trained on a dataset that stripped the lower and upper 5th percentile of context-question pairs by length of context (removing a set of the shortest and longest contexts)
Length Percentile Model (Question)	Model that was trained on a dataset that stripped the lower and upper 5th percentile of context-question pairs by length of question (removing a set of the shortest and longest questions)
Length Percentile Model (All)	Model that was trained on a dataset that stripped the lower and upper 5th percentile of context-question pairs by both length of context and length of question (removing a set of the shortest and longest context-question pairs overall)
BigBird Summarization Model	Model that was created in an attempt to allow longer contexts so that the AI could hold a proper conversation
BigBird Language Model	Model that was created for the same purpose as the BigBird Summarization Model, except with a Language Modeling head