CSE 582 FINAL PROJECT REPORT

Generating Targeted Novelty Questions with AskMeAnything QA Session Data

Ehsanul Kabir
May 5, 2023

1 INTRODUCTION

Automatic question generation is an important task in natural language processing. It has many applications in education, information retrieval, and question answering. In this project, I focus on generating targeted questions on a given topic and context on the questionee that is not covered in the context or not common knowledge about the questionee. These specific type of questions I refer to as novelty questions. I use IAmA subreddit's AskMeAnything (AMA) sessions as the source of context and questions. I train a T5 seq2seq model on the AMA data to generate novelty questions. I also finetune a BERT model on the AMA data to rank generated questions based on their context relevance. The latter is used to filter out questions that are not relevant to the context. I evaluate the generated questions on metrics such as BLEU, ROUGE, and METEOR and a human evaluation.

2 DATA COLLECTION AND PREPROCESSING

I use the IAmA subreddit's AskMeAnything (AMA) sessions as the source of context and questions. I use a combination of PRAW and PMAW python libraries to collect the data. I picked the years from 2014 to 2021 to collect the data. There were in total 498 AMA sessions in this time period with more than 1000 level-0 comments (level-0 comments are the comments directly under the original post and has a high probability of being a question to the original poster). The cutoff of 1000 comments was chosen to ensure that these are the popular AMA sessions. Even after that there were some popular AMA request posts that got into the data which were filtered out. I picked the sessions from 2015 to 2021 to be used as the training and

	IAmA Subreddit Original Post Titles			
	I am **Jason Steele**, creator of Charlie the Unicorn, Llamas with Hats, and other internet videos. Ask me anything!	I'm a retired **bank robber**. AMA!	I'm **Lynda Carter**, Wonder Woman actress and singer. Ask Me Anything!	Sup muthafuckas! **Ron Perlman** here. Don't be shy- Ask Me Anything!
Comments	What caused you to take the Llamas with Hats series in the direction that you did, from hilariously morbid to actually somewhat sad?	You say that you're retired. But I know a guy who is looking to put together a crew for a major job. This is the one you've been dreaming of all these years. Are you up for one last job?	Hello! Thanks for doing this AMA. Were you ever asked to reprise your role as Wonder Woman for another show, even a cameo?	Why doesn't anybody ever bring up how fucking great City of Lost Children was?
	Hi there! What was your inspiration for Charlie the Unicorn?	How many pounds of shit, would you say, were in your pants while walking out the door?	Thank you for doing this AMA. I just wanted to know, how did it feel, starring on modern superhero shows like Smallville and Supergirl, as opposed to the Wonder Woman show of the 70's?	Now that photos have been released, how do you feel about the 'new' Hellboy?
	How does it feel having created some of the most memorable animations on the Internet?	"Why, is that mister Clay coming in with the gun? Well, goshdarnit, is it thursday already?"	Hi Lynda! Thanks for doing this AMA! How did you like working on Fallout 4? Magnolia was such a great character!	Who would you want to play you in a Sons of Anarchy prequel?

Table 2.1: Sample of AMA context and questions

Question	Topic
Hello! Thanks for doing this AMA. Were you ever asked to reprise your role as Wonder Woman for another show, even a cameo?	Wonder Woman reprisal
Thank you for doing this AMA. I just wanted to know, how did it feel, starring on modern superhero shows like Smallville and Supergirl, as opposed to the Wonder Woman show of the 70's?	Acting in superhero shows
Hi Lynda! Thanks for doing this AMA! How did you like working on Fallout 4? Magnolia was such a great character!	Fallout 4 and Magnolia
Hi, I just want to know one thing. Where did you park your plane?	Airplane Parking
What was your most funny moment behind the scenes of Wonder woman?	Funny moments behind the scenes

Table 2.2: Sample of generated topics using GPT3 API

the validation set. I picked the sessions from 2014 to be used as the test set. Further selection criteria were applied to the comments to reduce the training data size. For each AMA session, I only kept the level-0 comments that had at least 10 upvotes. This helped pick better quality comments. After the above selection criteria, I ended up with $22,246$ AMA questions with their context as the original post. The test set had 2284 questions and the training and validation set had $11,123$ questions each. Table 2.1 shows a sample of the collected data.

ADDING TOPIC

The topic of the question is not explicitly mentioned in the context. However, it can be inferred from the question. As it would take a lot of time to manually annotate the topic of each question, I use GPT3 API to generate the topic of the question. I use the following prompt to generate the topic: *Describe the topic of the question below in a single word/phrase. Q: <question> A: ?* Table 2.2 shows a sample of the generated topics. From the sample, we can see that the generated topics are relevant to the question.

PADDING CONTEXT WITH WIKIPEDIA SUMMARY

The AMA posts contained title and body. The title is meant to be a brief Introduction to the questionee/OP. However, it proved to be too short in a lot of cases. For example, well-known TV/movie actors and actresses just wrote their name in the title which is enough for the readers to know who they are but not enough for the model to learn the context. The post body did not contain relevant information in most cases. Hence it was necessary to append context of named entities to the original post. However, this proved to be challenging to be done automatically. The questionees ranged from celebrities to various form of organizations. Often times there were multiple entities in the title and adding context of all of them would be too much for the model to learn. Therefore, only one entity was picked as the main entity and this task was done manually. I also had to make sure that the entity also has a Wikipedia page. There were some instances the OP was not a celebrity or a well-known organization and did not have a Wikipedia page. In those cases, OP actually provided sufficient details about

Title	Summary
Hello, I'm Lorin. I make music called Bassnectar - lately i've been working nonstop to remix my record collection into new versions for an experiment called the Freestyle Sessions in Colorado! What's on your mind? An me asky-thing...	Lorin Gabriel Ashton, better known under his stage name Bassnectar (born February 16, 1978), is an American DJ and record producer.
I Work at a Costco in Oregon dealing with this corona virus craze, AMA!	Costco Wholesale Corporation (doing business as Costco Wholesale and also known simply as Costco) is an American multinational corporation which operates a chain of membership-only big-box retail stores (warehouse club). As of 2022, Costco is the fifth largest retailer in the world and is the world's largest retailer of choice and prime beef, organic foods, rotisserie chicken, and wine as of 2016.
I'm a 23 year old woman living with Addison's Disease, Lupus, Psoriasis and Psoriatic Arthritis. Basically a walking autoimmune disease. Ask me anything!	Addison's disease, also known as primary adrenal insufficiency, is a rare long-term endocrine disorder characterized by inadequate production of the steroid hormones cortisol and aldosterone by the two outer layers of the cells of the adrenal glands (adrenal cortex), causing adrenal insufficiency. Symptoms generally come on slowly and insidiously and may include abdominal pain and gastrointestinal abnormalities, weakness, and weight loss.
Hey Reddit, it's Kris, Rob and Dave from Cyanide & Happiness! Let's talk about things!	Cyanide & Happiness (C&H) is a webcomic created by Rob DenBleyker, Kris Wilson, Dave McElfatrick and Matt Melvin. The comic has been running since 2005 and is published on the website explosm.net along with animated shorts in the same style. Matt Melvin left C&H in 2014, and several other people have contributed to the comic and to the animated shorts.

Table 2.3: Sample of generated topics using GPT3 API

themselves in the post title. For titles with at least one renowned entity, I used the Wikipedia API to get the summary of the entity's Wikipedia page. However, the summaries had varied length. I picked as many sentences as needed to have a total of 50 tokens in the summary. I appended the summary to the original post title which was then used as the context for the question. For the questionees that did not have a Wikipedia page, I used the original post title as the context. Table 2.3 shows a sample of the generated summaries. From the samples, we can see that the summaries are relevant to the title and the appending is seamless.

Additional Pre-processing of Title

The titles contained a lot of special characters and emojis. They also contained a lot of variations of the text 'Ask Me Anything' which needed to be filtered. I used various regular expressions to filter out the special characters and emojis along with the variations of 'Ask Me Anything'.

Qualitative Analysis of Data

The above selection criteria and pre-processing steps helped clean the data to a great extent. Specifically, I did not notice any post among the selected posts that were not an AMA session. However, there were some questions that seemed to be more of a comment than a question and some questions that seemed like a follow-up question. I did not remove those questions as they were not a significant portion of the data and the task of identifying them would be challenging and out-of-scope for this project.

Below I have listed some issues with the data that could not be solved within the limited time frame of this project:

- Some reddit questions were informal and contained a lot of abbreviations and slang. (Not necessarily an issue but could be a challenge for the model to learn)
- Some questions were actually comments or follow-up questions which is not expected from a question generator model without knowledge of the prior conversation.
- The appended context could not be expected to cover all the information about the questionee and as a result, some questions are relevant to the questionee but it is not obvious from the context. The model would need more familiarity with the questionee to be expected to generate such questions which may only be possible with a large LLM model like GPT-3 or GPT-4.

3 Methodology

For the task of question generation, I used one question generator model and one question ranker model. The question generator model's goal is to generate as many well-constructed questions as possible on the given topic to the questionee. The question ranker model's goal is to rank the generated questions based on their quality. Afterwards, the top ranked question is selected as the final output.

3.1 Question Generator

The following steps were taken to train the question generator model:

- The T5 seq2seq Transformer was picked as the architecture for the question generator model.
- I picked a pre-trained T5 model from HuggingFace's model hub which was trained on an accumulation of various well-known QA datasets (SQuAD, RACE, CoQA, and MSMARCO).
- The pre-trained model was fine-tuned on the AMA dataset that I prepared above.
- The fine-tuning is implemented as a conditional generation (seq2seq) task that maximizes the log-likelihood of the target sequence given the source sequence.
- The source sequence is a <topic> token followed by the topic of the question followed by a <context> token followed by the context of the questionee. The <topic> and <context> tokens are used to help the model learn the boundaries of the topic and the context and they were added to the original T5 tokenizer's vocabulary.
- The target sequence is the question asked by the reddit user on the given topic to the questionee.

3.2 Question Ranker

The following steps were taken to train the question ranker model:

- The BERT Transformer was picked as the architecture for the question ranker model.
- The Transformer was fine-tuned on the AMA ranking dataset of positive and negative examples that I mentioned above.
- The fine-tuning is implemented as a binary classification task that maximizes the log-likelihood of the target label given the source sequence.
- The source sequence is the question and the context separated by <sep> token and the target label is 1 if the question is a relevant question to be asked to the questionee and 0 otherwise.

4 Experiments

Decoding Method

Selecting the right decoding method is critical to ensure good output quality of the question generator model. Beam search and nucleus sampling are two popular decoding methods for Transformer models to generate questions. Nucleus sampling is a stochastic decoding method that samples from the smallest possible set of words whose cumulative probability exceeds a threshold p. Since AMA questions are typically elaborate and long, I used beam search in my experiments. The hyperparameters picked for beam search are as follows:

- Number of beams: 20
- Number of beam groups: 10
- Maximum length of the generated question: 50
- Number of question candidates to generate: 5
- Diversity penalty: 0.5 (to encourage diversity among the generated questions)

Evaluation Metrics

Qustion Generation task shares similarity with the summarization task and as such, the same evaluation metrics (ROUGE [2], BLEU [4], METEOR [1], etc.) are typically used to evaluate the quality of the generated questions. However, [3] showed that the standard evaluation metrics are not suitable for evaluating the quality of NLG tasks such as question generation. Therefore, manual evaluation is necessary to evaluate the quality of the generated questions. There was no baselines considered for this project considering the limited time and resources.

Hyperparameter	Question Generator	Question Ranker
Batch size	8	64
Learning rate	0.0001	0.0001
Number of epochs	20	20

Table 4.1: Training Hyperparameters

Test Set	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE
Familiar Questionee	16.47	6.26	3.82	2.82	15.68	13.57
Unfamiliar Questionee	17.24	7.20	4.45	3.26	16.73	14.02

Table 5.1: Performance of the question generator model on the familiar and unfamiliar questionee test sets. (The scores are in percentage.)

Hyperparameters

The hyperparameters used for the question generator model are as follows:

Both model was trained on a single NVIDIA A100 GPU. The code was written in a combination of PyToch and HuggingFace's Transformers library. Code from the github repository `https://github.com/AMontgomerie/question_generator.git` was used as a reference for the implementation of the question generator and the question ranker models. The generator model was trained for 20 epochs and took about 6 hours to train. The ranker model was trained for 20 epochs and took about 1 hours to train. The trained question generator model is available at `https://huggingface.co/ehsanul007/IAmA-question-generator` and the trained question ranker model is available at `https://huggingface.co/ehsanul007/IAmA-question-ranker`.

5 Results and Analysis

Automatic Evaluation

Two test sets were considered in automatic evaluation of the question generator model. The unfamiliar questionee test set consists of questions asked by reddit users to questionees that were not present in the training set i.e. from the AMA sessions in 2014. The familiar questionee test set consists of questions asked by reddit users to questionees that were present in the training set and they were randomly sampled from the validation set. The motivation behind using two test sets is to evaluate the generalization ability of the question generator model and whether it is affected by exposure to the questionee context during training.

Table 5.1 shows the evaluation metrics for the question generator model on the familiar and unfamiliar questionee test sets. The performance of the question generator model is slightly better on the unfamiliar questionee test set than the familiar questionee test set. However, the difference is negligible. One explanation for this is that the context available to the question generator model is not sufficient to generate inference capability about the questionee which could have helped the model to generate better questions. Thus, the model tries to generate similar questions for the same questionee to what it has seen during training even though topics are different. However, the above is just a hypothesis and further investigation is required to confirm it. Overall, the automatic evaluation results show that the question generator model is not able to generate questions that are similar to the questions asked by reddit users in the AMA sessions.

Qualitative Analysis

References

[1] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

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

[3] Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. Why we need new evaluation metrics for nlg. *arXiv preprint arXiv:1707.06875*, 2017.

[4] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.