Generating Targeted Novelty Questions with AskMeAnything QA Session Data

Ehsanul Kabir
May 5, 2023

1 INTRODUCTION

Automatic question generation is a fascinating and vital aspect of natural language processing, which holds the potential to revolutionize fields such as education, information retrieval, and question answering systems. In this particular project, I endeavor to develop a method for generating unique and targeted questions that focus on a specific topic and context about a person, an organization, or any similar entity of interest whom we can ask questions to. This latter entity will be referred to as the "questionee" throughout this report. However, the question should use the context as a knowledge base to ask questions that are not explicitly mentioned in the context but a rational agent would be curious to know about the questionee. For example, if the questionee is a famous actor, the context is a brief biography of the actor, the topic is the actor's career, then a good question would be *"What was the most challenging role you have ever played?"* which is highly relevant to the context and topic, but not explicitly mentioned in the context. These questions should also pertain to the questionee's knowledge or experiences that aren't covered in the provided context or aren't common knowledge about the individual. I refer to these types of questions as "novelty questions" as they spark curiosity and provide deeper insights.

For this project, I decided to harness the power of the IAmA subreddit's AskMeAnything (AMA) sessions. AMA sessions are a treasure trove of information, presenting a rich variety of topics and contexts to explore. People from diverse backgrounds participate in these sessions, answering questions from the Reddit community about their personal experiences, expertise, or unique perspectives. This dataset provides an excellent opportunity to create and refine a model that generates novelty questions that are both engaging and relevant.

To achieve this goal, I employed a two-step approach. First, I trained a T5 seq2seq model on the AMA dataset. The T5 model is known for its strong performance in various natural language processing tasks, and by fine-tuning it on the AMA data, it would ideally be able to generate novelty questions that are coherent, contextually relevant, and interesting.

However, I wanted to take this project a step further by filtering out questions that might not be as relevant to the context. To accomplish this, I fine-tuned a BERT model on the AMA data as well, which served as a ranking mechanism. The BERT model assessed the generated questions based on their relevance to the context, ensuring that only the most appropriate and intriguing questions made it through the filtering process.

Once the models were trained and the question generation process was established, it was time to evaluate the effectiveness of the generated novelty questions. To do this, I used several well-known evaluation metrics, such as BLEU, ROUGE, and METEOR. However, these metrics are not equipped to assess the novelty of the generated questions. Therefore, I also conducted a human evaluation, seeking input from real people to assess whether the questions were engaging, contextually relevant, and genuinely novel.

In summary, this project is dedicated to generating novelty questions on specific topics and contexts using a T5 seq2seq model trained on IAmA subreddit's AMA sessions.[1] Furthermore, a fine-tuned BERT model is employed to rank and filter out questions based on their contextual relevance. The generated questions are evaluated using a combination of standard metrics and human evaluation, ensuring that the questions are both engaging and insightful. The trained question generator is available as a web application at HuggingFace[2] and can be accessed by anyone. The datasets are also available at HuggingFace[3] and can be accessed by anyone.

2 Data Collection and Preprocessing

I chose to use the IAmA subreddit's AskMeAnything (AMA) sessions as my primary source of context and questions. To collect the data, I utilized both the PRAW and PMAW Python libraries, which made the process much smoother. I decided to focus on AMA sessions from the years 2014 to 2021 for a comprehensive dataset.

During this time period, there were a total of 498 AMA sessions that garnered more than 1000 level-0 comments each. Level-0 comments are those that appear directly beneath the original post and are highly likely to be questions aimed at the original poster. I opted for a cutoff of 1000 comments to ensure that the dataset would primarily consist of popular AMA sessions. However, some popular AMA request posts still slipped through the cracks and needed to be filtered out afterward.

I then split the dataset, selecting sessions from 2015 to 2021 for training and validation purposes, while using the 2014 sessions as the test set. To further refine the data and reduce the training set size, I applied additional selection criteria to the comments. For each AMA

[1] `https://www.reddit.com/r/IAmA/`

[2] `https://huggingface.co/spaces/ehsanul007/IAmA-question-generator`

[3] Question Generation Dataset: `https://huggingface.co/datasets/ehsanul007/IAmA-question-generator`
Question Ranking Dataset: `https://huggingface.co/datasets/ehsanul007/IAmA-question-ranking`

	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!
	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?
Comments	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 post 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

session, I kept only level-0 comments with at least 10 upvotes, which helped to highlight better-quality comments.

After applying these selection criteria, I ended up with a total of 22,246 AMA questions, along with their respective original posts as context. The test set contained 2284 questions, while the training and validation sets each comprised 11,123 questions. Table 2.1 shows a sample of the collected data.

ADDING TOPIC

The topic of each question is not explicitly stated in the context. Nevertheless, it can often be deduced from the question itself. Manually annotating the topic for every question would be a time-consuming process, so I opted to use the GPT-3 API to generate the topics instead.

To do this, I employed the following prompt: *Describe the topic of the question below in a single word/phrase. Q: <question> A: ?*. This prompt allowed GPT-3 to generate a topic based on the input question. Table 2.2 presents a sample of the topics produced by the API. As we can see from this sample, the generated topics are relevant and closely related to the questions themselves.

PADDING CONTEXT WITH WIKIPEDIA SUMMARY

AMA posts are composed of a title and a body. The title serves as a brief introduction to the questionee or the original poster (OP). However, in many cases, the titles proved to be too short to provide adequate context for the model. For instance, famous TV or movie actors and actresses often simply included their name in the title, which was enough for readers to identify them but insufficient for the model to learn the context. Furthermore, the post body usually didn't contain much relevant information.

Consequently, it became necessary to append additional context from named entities to the original post. Automating this task was quite challenging, as the questionees ranged from celebrities to various organizations. Titles frequently featured multiple entities, and including

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 appended context from Wikipedia

context for all of them would overload the model. Thus, I manually selected a single main entity from each title, ensuring that the entity had a Wikipedia page.

Occasionally, the OP was not a well-known figure or organization and didn't have a Wikipedia page. In these cases, the OP typically provided enough information about themselves in the title. For titles containing at least one notable entity, I employed the Wikipedia API to retrieve summaries from the entity's Wikipedia page. Since the summaries varied in length, I chose a sufficient number of sentences to achieve a total of 50 tokens in each summary.

I then appended these summaries to the original post title, which served as the context for the questions. For questionees without a Wikipedia page, I used the original post title as context. Table 2.3 displays a sample of the generated summaries. As we can see from these samples, the summaries are relevant to the title and blend seamlessly into the appended context.

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.
- The AMA sessions were from a few years back but the appended context was from the current Wikipedia page of the questionee. This could be an issue for the model to learn the context.

Quantitative Analysis of Data

	mean	min	max	std
Context	68.63	23	132	17.86
Question	18.20	1	48	10.18
Topic	2.61	1	5	0.95

Table 2.4: Word count statistics of the data

I have listed some statistics about the data below:

- Among the 22,246 posts, there are 14,313 unique topics. Top 5 most frequent topics are: Time Management, Clarification, Confusion, Identity verification, Miscommunication.
- The number of unique questionees is 16027. Top 3 most frequent questionees are: *'What is your favorite dinosaur?'*asked 4 times, *'What is your favorite flavor of ice cream?'* and *'Is a hotdog a sandwich?'* asked 3 times each.
- Of 428 questionees, 134 of them were random/anonymous persons hosting the AMA session with title hooks such as *'I am a retired bank robber, AMA!'*.
- 24 of the host reappeared in multiple AMA sessions.
- Table 2.4 shows the word count statistics of the data.

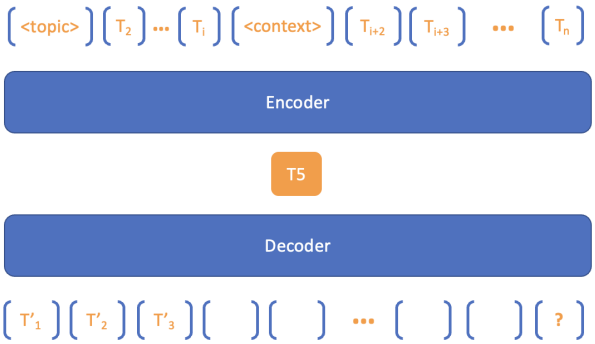

Figure 3.1: Question Generator Architecture

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 [8] 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 [9], RACE [3], CoQA [10], and MSMARCO [5]).
- 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 [2] 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 [4], BLEU [7], METEOR [1], etc.) are typically used to evaluate the quality of the generated questions. However, [6] 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. BLEU-1, BLEU-2, BLEU-3, BLEU-4 refer to BLEU scores for n-grams of length 1, 2, 3, 4 respectively.)

HYPERPARAMETERS

The hyperparameters used for the question generator model are given in Table 4.1.

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 HuggingFace[4] and the trained question ranker model is available at HuggingFace[5].

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

[4] `https://huggingface.co/ehsanul007/IAmA-question-generator`
[5] `https://huggingface.co/ehsanul007/IAmA-question-ranker`

Semantic Similarity with Actual Question	Grammatical Correctness	Degree of Novelty	Relevance to Questionee
0.83	2.5	2.1	1.37

Table 5.2: Human evaluation results for the generated questions. (The scores are out of 3)

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.

HUMAN EVALUATION

Table 5.2 shows the results of the human evaluation of the generated questions on various aspects. As this was a one-person project, the human evaluation was done by the author of this report. The human evaluation was done on a sample of 100 generated questions.

The results show that the generated questions are not semantically similar to the actual questions asked by reddit users in the AMA sessions. However, that is neither surprise nor unwelcome in this case. Novelty questions are not expected to be semantically similar to the actual questions.

The generated questions have a high degree of grammatical correctness. This is impressive considering that the questions generated by the question generator model are long and complex.

The generated questions have a high degree of novelty evaluated by the human evaluator. This means that the goal of generating novel questions is moderately achieved. However, the generated questions are not relevant to the questionee. The novelty comes at the cost of losing connection with the context.

CASE STUDY

Table 5.5 shows some examples of generated questions by the question generator model. Some interesting observations are listed below:

- The generated question likes to *address the questionee by name*. Example 2 shows one interesting such case where an inference is made about the questionee's name.
- Except few patterns such as finding the name of the questionee, not much attempt is made to infer about the questionee. The evidence of this is the fact that *most questions could be generated by looking at the topic alone*. Not much information about the questionee is integrated into the generated question.
- The effort to generate an interesting/novel question is evident in the generated questions. Instead of asking dull factoid questions, the generated *questions are more opinionated and eager to use superlatives* such as *best, craziest, favorite* etc.

Questionee Context	Topic	Context	Actual Question	Generated Question
Familiar	Work experiences	I am a Video and Audio Forensic Expert who has consulted on cases like Trayvon Martin, Malaysia Airlines Flight 307, and the JFK Tapes!	In your 34 years of work, what it the weirdest or strangest thing you learned in your time?	What is the **craziest** thing to ever happened to your job?
Familiar	Inter-coastal	We are the staff of Giant Bomb, a site about video games, and our decades of video game coverage experience are here to serve you today. *Giant Bomb is an American video game website and wiki that includes personality-driven gaming videos, commentary, news, and reviews, created by former GameSpot editors* ***Jeff*** *Gerstmann and Ryan Davis. The website was voted by Time magazine as one of the Top 50 websites of 2011.*	With how busy you guys are on different coasts, how much content on the site that you're not directly involved in do you consume? For example, do you listen to each others podcasts?	Jeff, what are your thoughts on intercoastal waters?
Familiar	Movie Rankings/Voting	We are Jeremy & Barrett of CinemaSins, MusicVideoSins, & the SinCast podcast./AMA	Once you've reached 2016 on SinCast, will you continue to vote on your favourite movies for each year? What other topics do you have planned for then?	What's your favorite topic of movies that you have discussed on the **Sincast**? And which movie(s) that you have mentioned that you have the most fun times watching? Also, which movies/shows are you watching should
Familiar	Unconventional viewing habits	We are the Pornhub team.	Odd? I browse Pornhub exclusively for the Halo montages!	What is the **craziest** thing you've seen someone do on **Pornhub**?
Familiar	Clothing style comparison.	Hi, Im Hiroshi Lockheimer, here at Google with the team that build Nexus 5X & 6P.	Who wins the best shirt award on the Android team? Matias Duarte with his colorful dress shirts or Dave Burke with his graphic Ts?	How does the Nexus 6P differ from the other android 6P?
Unfamiliar	Opinion of Victoria	I am Cyndi Lauper. Whaddya wanna know?	Is Victoria really as cool as everyone says she is?	What do you really think of Victoria?
Unfamiliar	Actor Collaboration	I am Cyndi Lauper. Whaddya wanna know?	Was it more fun to work with Christopher Walken, or Paul Rubens?	Who was your **favorite** actor to work with?

Table 5.3: Examples of generated questions by the question generator model. Appended context is omitted for brevity unless it is necessary to understand the question.

Topic	Context	Generated Question
Size Comparison	Sarah McLachlan here on reddit.	Would you rather fight 100 duck-sized Nate Silver?
	Jason Bateman here	Would you rather fight 100 duck-sized Nate Silver?
	Make it Work. Tim Gunn here from Project Runway	Would you rather fight 100 duck sized horses or 100 duck sized horses?
	Sean Schemmel and Christopher Sabat here, the voices of Goku and Vegeta from Dragon Ball Super.	Would you rather fight a syllable duck, or 100 duck-sized Vegetas?
	I'm Cheech	On a scale of mouse to giraffe how high are you right now?

Table 5.4: Examples of generated questions with the topic "Size Comparison" but to different questionees.

Topic	Context	Generated Question
Time Management	Im Astead W. Herndon, a national political reporter for The New York Times. I spent 3 months reporting on the Sunrise Movement, a group of young climate activists trying to push Democrats to the left ahead of the 2020 election.	Why do you think you have so little time?
	I am Aisha Tyler and I have all the jobs!	Hi, Aisha. How do you spend most of your time?
	I am a 14yr old Ebola survivor in remote Liberia	What do you do all day?

Table 5.5: Examples of generated questions with the topic "Time Management" but to different questionees.

- If there are multiple words in the topic, the *words that are not relevant to the questionee are ignored*. In example 5, the topic is *Clothing style comparison*. The words *Clothing style* are not relevant to the questionee. So, the generated question is about the word *comparison*. This observation is in contrast to the observation made in the second bullet point.
- No difference is visible in the generated questions when the questionee is familiar or unfamiliar to the question generator model.

TOPIC DEPENDENCY OF GENERATED QUESTIONS

Table 5.5 shows some examples of generated questions with the topic *Size Comparison* but to different questionees. The generated questions are not very different from each other. Size comparison is a very generic topic and the generated questions are not only generic but also repetitive, potentially memorized by the model. The high variation of topics seen in the training data is probably the reason for this. Since most topics only occur once in the training data, the model is not able to learn the nuances of the topic and generate interesting questions.

Table 5.5 presents a different scenario. The topic is *Time Management* and the generated questions are very different from each other. There is some connection between the generated questions and the questionee especially in the 3rd example. This could be because the topic *Time Management* occurs multiple (23) times in the training data and the model is able to learn the nuances of the topic. Future attempts to improve the model could focus on increasing the number of occurrences of each topic in the training data as a first step.

6 Lessons Learned and Potential Future Directions

- Human-like targeted novelty question generation is a hard problem. The model is not able to generate questions that are interesting and novel with respect to the questionee.
- Automatic evaluation metrics are not useful at all for this task. Even in summarization or closed-domain question generation, automatic evaluation metrics have some merit as the distance from the reference is a good proxy for the quality of the generated text. But in this task, the generated question could be very different from the reference question and still be a good question on the given topic to the given questionee. A new automatic evaluation metric is needed for this task.
- The approach in this project did not utilize the reply given by the OP in the AMA session at all. That could be a potential area of improvement or a new direction to explore, possibly an attempt to build dialogue generation models.
- Another potential direction could be to use context to generate a knowledge graph and use that knowledge graph to generate questions. This could be a way to integrate information about the questionee into the generated question.

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] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*, 2017.

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

[5] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. Ms marco: A human generated machine reading comprehension dataset. *choice*, 2640:660, 2016.

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

[7] 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.

[8] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[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] Siva Reddy, Danqi Chen, and Christopher D Manning. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, 2019.