Few-Shot Question Generation with GPT-J

Sukrit Rao
Center for Data Science
New York University
str8775@nyu.edu

Rohith Bollineni
Center for Data Science
New York University
rb4987@nyu.edu

Hasan Khan
Courant Institute
New York University
hk3550@nyu.edu

Abstract

Neural Question Generation (QG) systems aim to automate the process of question construction by generating novel questions given a particular context, thus reducing time and costs of question generation for educators and test developers. We propose Question Generation using GPT-J in a few-shot setting. Generating questions in this manner reduces time and resource cost required to construct datasets and fine-tune increasingly complex models like GPT-J, thereby increasing usage for educational purposes such as adaptive education. We compare our results against a GPT-J model fine-tuned on the task.

1 Introduction

Asking relevant questions of varying difficulty forms an essential part of assessing a student's grasp of concepts. Instructors are required to spend considerable amounts of time constructing exam and assignment questions that assess students on material taught in the classroom. In this process of question creation, instructors must make sure that the questions cannot all be of similar difficulty and must range from easy to difficult to ensure optimum learning outcomes. Additionally, questions often need to be replaced continually as content is revised to reflect the latest updates in the domain or because the questions become publicly available. Another aspect that affects learning outcomes is the student's aptitude and pre-existing knowledge. In classroom settings, learning takes place in groups in which some students would be stronger in the concepts the material tests while others would not and would require additional practice and/or instruction. As a result, applying universal pedagogy in such settings is not ideal. (Liu et al., 2020) Adaptive education systems have great potential to improve learning outcomes by increasing accessibility (Srivastava and Goodman, 2021).

Neural Question Generation (QG) (Pan et al., 2019) systems aim to automate the process of question construction by generating novel questions given a particular context, thus reducing time and costs of question generation for educators and test developers. Advanced QG systems with configurable parameters could help offer students custom material based on their individual ability, and act as a foundation for adaptive testing and learning.

Recent work in QG has focused on generating quiz-style questions (Lelkes et al., 2021), with particular focus on generating questions of selected difficulty levels (Gao et al., 2018). However, these techniques have relied on fine-tuning a language model on a task-specific dataset such as SQuAD (squ) or RACE (Lai et al., 2017). As a result, these models are limited in their domain of use. Moreover, constructing such datasets that contain thousands of examples, specific to that task, is time-consuming and costly, and thus not a viable means for widespread adoption. In this paper, we propose using GPT-J in a few-shot setting to produce questions that are fluent in linguistic construction, relevant to the input context, and appropriately difficult as desired. We compare our work against a GPT-J model fine-tuned on the task.

In the context of reading comprehension, we look at two types of QG variants. In answer-focused QG, a reference passage and an answer are passed as inputs into the system, resulting in the generation of questions relevant to the input answer. In general QG, only a context passage is passed as input, resulting in the generation of unmapped questions relevant to the context.

2 Related Work

2.1 Difficulty Controllable Question Generation

DQG discusses a framework to generate questions given a difficulty level along with the context and answer. Bidirectional LSTMs are used to encode

the input. For the decoder, another LSTM is used along with a global difficulty control variable trained on ground truth difficulty labels. The authors use R-Net and BiDAF to assess the difficulty of a question; if both models guess correctly then it is labeled easy; if both models fail then it is labeled hard.

2.2 Quiz Style Question Generation

As mentioned above, (Lelkes et al., 2021) focuses on quiz style questions as part of the NewsQuizQA dataset. Unlike prior datasets like SQuAD, prior context does not have a strong impact on the NewsQuizQA dataset with regards to generating questions which reference source text. NewsQuizQA implements Minimum Reference Loss to judge its models, allowing for multiple correct outputs based on how well results match with QA pairs. This metric style allows for some open ended behavior from the model.

2.3 Evaluation of Text Generation: A Survey

This paper (Celikyilmaz et al., 2020) serves as a framework for evaluation metrics used to judge machine generated text. The authors delineate three categories of evaluation. Human-centric evaluation involves humans (often experts) judging the quality of generated text, usually through manual review. Untrained automatic evaluation focuses on comparing text generated by models with human written text (such as BLEU, BLEURT and ROUGE). Machine-learned evaluation involved a machine learning model acting in place of a human judge to determine similarities in human vs. machine (or machine vs. machine) generated texts. All three categories of evaluation are used in our paper, discussed further in the evalatuation section.

3 Methodology

3.1 Data

We use RACE (Lai et al., 2017), a compilation of reading comprehension questions from middle and high school English exams administered to Chinese students, to fine-tune and prompt our GPT-J model. We choose RACE since "the difficulty of RACE questions mostly comes from the understanding of the story but not from the way how the question is asked." Gao et al. (2018) (p.2), which is what one finds in the education domain. Moreover, RACE categorizes questions into either *middle* or *high*, based on their relative difficulty. As a result, no manual labeling is required to obtain difficulty labels for questions.

Questions that are generic (i.e less than 5 words, non-specific to the context, etc.) are dropped in order to prevent the model from generating irrelevant questions.

For the answer-focused setting, training inputs to the QG model are composed of a question, an answer, a context and a desired difficulty of the generated question. We determine difficulty by mapping middle school questions to *easy* and high school questions to *hard* difficulties. 31051 data points are used for fine-tuning, and 1695 for testing. We only use context passages that have at least two associated questions, in order to allow for the prompt to contain at least one example for every test context. A similar input is used for the general setting, except no answer is included.

We experiment with various prompts for the few-shot setting (insert future prompt variant info here). An example of the input prompt used in the few-shot setting is shown below:

> **Context**: "Family" is of course an elastic word. But when British people say that their society is based on family life, they are thinking of "family" in its narrow, peculiarly European sense of mother ,father and children living together alone in their own house as an economic and social unit. Thus, every British marriage indicates the beginning of a new and independent family–hence the tremendous importance of marriage in British life. Readers of novels like Jane Austen's Pride and Prejudice will know that in former times marriage among wealthy families were arranged by the girl's parents, that is, it was the parents' duty to find a suitable husband for their daughter, preferably a rich one, and by skillful encouragement to lead him eventually to ask their permission to marry her.
>
> **Difficulty**: Hard.
>
> **Answer**: It gives quite some idea of English social life in the past.
>
> **Question**: What is true concerning the book Pride and Prejudice?
>
> **Difficulty**: Hard.
>
> **Answer**: Different definitions could be given to the word.

Question:

In this instance, the true reference question is "What does the author mean by "Family is of course an elastic word"?"

3.2 Model & Tools

For our experiments, we select GPT-J-6B (Wang and Komatsuzaki, 2021) pulled from HuggingFace. We choose this model as it performs well in a zero-shot setting (Wang and Komatsuzaki, 2021) and for its manageable parameter size (6 billion) and open-source code, allowing us to compare performance between a fine-tuned model and a model used in a few-shot setting. During decoding, we use a **temperature** value of 1.0, **top-k** of 50, and **top-p** of 1.0. We set a **length penalty** of 0.8 to favor shorter sequences

For difficulty classification, we use RoBERTa (Liu et al., 2019). The model is trained with a classification head, taking in a combined input of context, question, and answer with the difficulty labels of 'hard' and 'easy'. Training is performed for 3781 examples, evaluation on 222 examples, and testing on 197 samples. The data is preprocessed to follow a similar ruleset to Gao et al. (2018), with some adjustments.

We perform model training on NYU HPC. All code can be found in our Github repository.

3.3 Evaluation

Generated questions are evaluated along three dimensions: fluency, relevancy, and difficulty. Generally, we use a mix of manual human-centric evaluation, automatic metrics and Machine-Learned metrics as described above.

Fluency determines weather a generated question is easy to read and understand, without taking the source (passage or reference question) into account, and is manually evaluated on a 0-2 Likert scale (ranging from poor to excellent). **Relevancy** determines whether the generated questions and input reference questions are topically related (Dudy et al., 2021), and is measured using manual evaluation on a 0-1 Likert scale (either relevant or irrelevant) as well as with evaluation metrics including BLEU (Lin and Och, 2004), ROUGE-L (Lin, 2004) and BLEURT, (Sellam et al., 2020) a learned evaluation metric that captures the semantic similarities between two pieces of text. These untrained automatic metrics are computed between the reference question from RACE and the question by GPT-J.

Difficulty is determined by checking if the generated question's difficulty matches that of the input difficulty fed to the model, and is measured as a Machine-Learned metric using a RoBERTa (Liu et al., 2019) classifier trained on RACE to classify questions as being either easy or hard. We compare the classifier's output difficulty with the desired difficulty from the input, and evaluate the F1 score.

4 Results

Metric	Finetuned	One-Shot
Fluency		
Manual	-	1.27
Relevancy		
Manual	-	0.18
BLEU-1	-	1.40
BLEU-2	-	0.00
BLEU-3	-	0.00
BLEU-4	-	0.00
ROUGE-L	-	16.0
BLEURT	-	32.31
Difficulty		
RoBERTa	-	82.00

Table 1: Results from GPT-J. Manual fluency and relevancy are averages calculated over manual human evaluation scores. Further untrained metrics are included for relevancy. Difficulty is the F1 score achieved by the difficulty classifier

Results on all three evaluation metrics are included above. See Appendix on page 5 for example generated questions.

5 Collaboration

Credit to Sukrit Rao for formulating the project question. He developed and executed the inference scripts for the GPT-J question generating models, prompt design, and developed scripts for the automated evaluation metrics. Credit to Rohith Bollineni for writing the training/evaluation script, building the difficulty classification model. Credit to Hasan Khan for question preprocessing, a majority of the paper writing, and pulling references together. All team members worked on manual evaluations of the generated questions.

Acknowledgements

We would like to thank our Professor Sam Bowman and TA Richard Pang for their assistance and guidance through the development of this paper.

References

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey.

Shiran Dudy, Steven Bedrick, and Bonnie Webber. 2021. Refocusing on relevance: Personalization in nlg.

Yifan Gao, Jianan Wang, Lidong Bing, Irwin King, and Michael R. Lyu. 2018. Difficulty controllable question generation for reading comprehension. *CoRR*, abs/1807.03586.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. 2017. RACE: large-scale reading comprehension dataset from examinations. *CoRR*, abs/1704.04683.

Ádám D. Lelkes, Vinh Q. Tran, and Cong Yu. 2021. Quiz-style question generation for news stories. *CoRR*, abs/2102.09094.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chin-Yew Lin and Franz Josef Och. 2004. ORANGE: a method for evaluating automatic evaluation metrics for machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 501–507, Geneva, Switzerland. COLING.

Junhua Liu, Lionell Loh, Ernest Ng, Yijia Chen, Kristin L. Wood, and Kwan Hui Lim. 2020. Self-evolving adaptive learning for personalized education. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*. ACM.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Liangming Pan, Wenqiang Lei, Tat-Seng Chua, and Min-Yen Kan. 2019. Recent advances in neural question generation.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *ACL*.

Megha Srivastava and Noah Goodman. 2021. Question generation for adaptive education. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 692–701, Online. Association for Computational Linguistics.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

A Appendix

Desired Difficulty	Reference Question	Generated Question	Answer
Hard	What is special about the pens produced in 1234 PENS?	Why would you choose 1234 PENS?	Their material is less harmful to the earth
Hard	What would be the best title for the passage?	How is it to climb from the Tibetan side of Mt. Everest?	California teen becomes youngest to climb Mt. Everest
Easy	What can be the best title for the passage ?	What do you think is the moral of the story?	Love Saves All

Table 2: Questions generated by GPT-J in a one-shot setting