Literature Review: Question Generation with Few Shot LLMs

Nitya Sampath
sampathnitya@gmail.com

1 Task definition

Question generation is a natural language task where a model is provided a context passage and an answer and trained to generate a question that asks for the answer. The task is most commonly used to improve question-answering (QA) systems by increasing the size of the training data for the model. Other use cases include providing suggested or recommended questions to users in a question answering system, improving the conversational skills of chatbots by allowing them to generate leading questions, and even for educational purposes by automatically generating questions from reading material for assessment.

A common approach to this problem is a sequence to sequence approach utilizing neural models, specifically recurrent neural networks (RNNs) such as long short-term memory (LSTMs) or gated recurrent units.

More recently, BERT and various derived models have been used for this task, finetuning BERT models on question-answering datasets and using them to generate questions.

Even more recently, with the emergence of large language models, GPT-3 has been used to accomplish this task by prompting the model with the context passage and the answer and asking it to generate a question for the answer.

2 Concise summaries of the articles

2.1 Leveraging Context Information for Natural Question Generation

Song et al. (2018) noted that previous work on question generation often ignored the target answer (Du et al., 2017) or provided hard-coded answer positions instead of the actual target answer (Zhou et al., 2017). Therefore they produced models that were unable to make use of the contextual information that could be provided when taking into account both the passage and the target answer. The authors decided to build on this work by providing the target answer and matching within the passage, in order to bring out the relevant context.

They used a sequence-to-sequence model, based on existing question generation work (Sutskever et al., 2014), as a baseline. This model used a birdirectional long short-term memory (BiLSTM) encoder to encode the passage and and an attentional LSTM decoder to generate the question. The baseline encoder includes information about the answer position, but no other contextual information is given.

To test their hypothesis that providing contextual information from the target answer would improve the generated question, the authors built a similar sequence-to-sequence model. However, unlike the baseline, their encoder used two separate BiLSTM encoders for the passage and the answer. Then they used the multi-perspective matching algorithm (Wang et al., 2017) to match the answer and the relevant context in the passage. The decoder was an attentional LSTM, similar to the baseline, but included this matching information.

The authors trained their model on the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016). They compared their model's performance with two state of the art models that both used the sequence to sequence model. One of these did not use answer information (Du et al., 2017), and the other used answer position features but no contextual information (Zhou et al., 2017). The authors found that their model outperformed both of the state of the art models they tested it against.

2.2 Question-type Driven Question Generation

Zhou et al. (2019) sought to improve state of the art results on this task by building a model that could more accurately predict question-type, allowing it to then better generate a question appropriate to the target answer. They were motivated by previous

work using models that would generate questions of the wrong type, as they could not properly predict what question type was most relevant to the target answer (Hu et al., 2018).

The authors used a feature-rich encoder to encode the context and the answer, using POS tags, NER tags and word case as features. They passed the embeddings generated from this to a BiLSTM. For the question type prediction step, they classified questions into 8 types and used a unidirectional LSTM layer to predict the question types. Then, they used an attention based decoder to generate the question. The model was trained on two datasets: SQuAD and MARCO. Th authors found that their model outperforms the baseline model significantly, with a BLEU-4 score of 16.31 on SQuAD.

2.3 A Recurrent BERT-based Model for Question Generation

Chan and Fan (2019) explores using BERT (Devlin et al., 2019) for the problem of question generation, building off of previous works which primary focused on sequence-to-sequence models, using LSTMs (Hochreiter and Schmidhuber, 1997) or gated recurrent units (GRU) (Chung et al., 2014). The authors note that because of the limitations on these recurrent neural network (RNN) models processing long sequences, most of the previous models use sentence-level information for generating questions and show worse performance on paragraph length sequences. The state of the art model that the authors cite maxout pointer and gated self attention meachanism, which enables it to process paragraphs (Zhao et al., 2018). The authors aim to improve on the state of the art models by using BERT, because the transformer architecture underlying BERT models allow for accurate processing of longer, paragraph level sequences.

Chan and Fan (2019) started with a naive BERT model finetuned for the question generation task. They called this initial model BERT-QG. However, they found this model had rather poor results, as a result of the fact that it does not take into account previous decoded results. So, to account for that limitation, the authors built second BERT model, a sequential generation model they called BERT-SQG. They found there were still shortcomings with their BERT-SQG model and sought to rectify issues with processing lengthy contexts and ambiguity that arises when the answer phrase appears multiple times in the context. The did this by defining a new token, [HL], that would be used to indicate the answer phrase in the context. They trained a new BERT model with contexts that used this token, which they called BERT-HLSQG.

The BERT models were again trained on the SQuAD dataset and the authors found that their models produced state of the art results in evaluation with BLEU 4. In particular, their BERT-SQG obtained a BLEU score of 21.04, over the previous state of the art score 16.85. Their BERT-HLSQG performed even better, achieving a BLEU score of 22.17.

2.4 Summary-Oriented Question Generation for Informational Queries

Yin et al. (2021) also uses a BERT based model to tackle the task of question generation. Specifically, they used a BERT-based Pointer Generator Network (BERTPGN) as the decoder to generate questions. To encode the context and answer, they used positional and type embeddings as the input for the BERT model. Their approach was motivated by their interest in generating suggested questions that could be recommended to users interacting with a QA system. The objective is to provide users with examples of complex questions the system is capable of answering. Because of this use case, the authors sought to generate questions that were *self-explanatory* and *introductory*, which do not require background knowledge to understand the generated question and that are answered by longer passages rather than just simple statements. These requirements motivated the use of the BERTPGN, as well as the use of a different dataset, the Natural Questions dataset (NQ) (Kwiatkowski et al., 2019). The NQ dataset consists of questions whose answers can be simple entites, short passages, or entire paragraph passages, which provides results that are more appropriate for this task's needs.

The model was evaluated on BLEU, METEOR and ROUGE and compared with the work of Mishra et al. (2020) as a baseline. Ultimately, they found that their model does outperform the baseline.

The authors also did an out of domain evaluation. For this, they used their model on out of domain news articles and compared the results to a baseline BERTPGN model trained on SQuAD data. They found that their model outperformed the baseline and conducted a human evaluation to confirm the evaluation results of this experiment.

2.5 Selecting Better Samples from Pre-trained LLMs: A Case Study on Question Generation

Yuan et al. (2023) take advantage of advancements in large language models (LLMs) for this task. They used GPT-3 and zero-shot prompting to generate questions based on context-answer pairs. They used two different datasets for training SQuAD and Fairytale QA (Xu et al., 2022). While SQuAD is a sentence-level question generation dataset and has each answer as substring of the context, Fairytale QA is paragraph-level and the answers are not always part of the context.

The authors first prompted a pretrained GPT-3 model to generate questions. Given the context and answer in the prompt, the model was asked to fill in the question. For evaluation, they used BLEU-4 for SQuAD and ROUGE for Fairytale QA to measure the generated questions against a ground truth question.

From these generated questions, the authors then continued with a question selecting step, in which they explored three methods for selecting the best quality questions. Their first method was using an n-gram similarity score in which the question was scored according to its similarity to the context as a measure of its relevance. The second method was called the round trip method, which measured the similarity of the generated question's answer to the original answer used to generate the question. This assumed that GPT-3 would answer a quality question in a way that was similar to the answer it was given to generate the question in the first place. The third method was to use prompting to ask GPT-3 to choose the highest quality questions. They found that for SQuAD, n-gram similarity was the best method for selecting high quality questions, while for Fairytale QA, prompt based scoring provided the best question selections.

3 Compare and contrast

The summarized articles represent an evolution in the strategies commonly used to accomplish the task of question generation. Both Song et al. (2018) and Zhou et al. (2019) followed the earlier tactics that involved sequence to sequence neural models. These two papers represent strategies that predated BERT and the transformer models. In addition, both Chan and Fan (2019) and Yin et al. (2021) used BERT based neural models for the task. On the other hand, Yuan et al. (2023) used an LLM, specifically GPT-3 and prompting to accomplish this task.

Most of the papers made use the the same dateset, SQuAD, a widely used dataset in the field of question answering that consists of questions on a number of Wikipedia articles (Rajpurkar et al., 2016). All of the answers in SQuAD are a span from the passage.

However, Yin et al. (2021) found SQuAD lacking for its specific use case and chose to use the Natural Questions dataset instead. In addition, most of the papers also used other datasets, notably MARCO and Fairytale QA in addition to SQuAD for their experiments.

For most of the papers, BLEU, METEOR and ROUGE were used as evaluation metrics. The papers most frequently cited scores with BLEU as comparison metrics and measures of state of the art performance, given the widespread use of this metric.

The papers also explore different subproblems and related problems to the main question answering task. A major part of Zhou et al. (2019)'s work was improving the model's ability to predict question types, in order to use that task to inform the nature of the question it generates. Yuan et al. (2023) additionally looked at methods for selecting questions of the highest quality from among the questions generated by the LLM. The papers approached the task of question generation from the perspective of different use cases, which informed the scope of related tasks their work also investigated.

4 Future work

Overall, these papers present a robust exploration of question generation as a natural language task. However, there are several avenues for further exploration. In particular, most of the papers use LSTMs or BERT in various ways, but now with LLMs increasing in popularity in the field, it would be interesting to expand the use of LLMs on this task. Yuan et al. (2023) uses GPT-3, but there are a number of other promising models that could generate comparable results, such as LLama2. In addition, making use of retrieval augmented generation to enhance the model's understanding of the context and answer may also be interesting to explore.

References

Ying-Hong Chan and Yao-Chung Fan. 2019. A recurrent BERT-based model for question generation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 154–162, Hong Kong, China. Association for Computational Linguistics.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352, Vancouver, Canada. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Wenpeng Hu, Bing Liu, Jinwen Ma, Dongyan Zhao, and Rui Yan. 2018. Aspect-based question generation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*. OpenReview.net.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Shlok Kumar Mishra, Pranav Goel, Abhishek Sharma, Abhyuday Jagannatha, David Jacobs, and Hal Daumé III au2. 2020. Towards automatic generation of questions from long answers.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Linfeng Song, Zhiguo Wang, Wael Hamza, Yue Zhang, and Daniel Gildea. 2018. Leveraging context information for natural question generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 569–574, New Orleans, Louisiana. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks.

Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral multi-perspective matching for natural language sentences. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 4144–4150.

Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Li, Nora Bradford, Branda Sun, Tran Hoang, Yisi Sang, Yufang Hou, Xiaojuan Ma, Diyi Yang, Nanyun Peng, Zhou Yu, and Mark Warschauer. 2022. Fantastic questions and where to find them: FairytaleQA – an authentic dataset for narrative comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 447–460, Dublin, Ireland. Association for Computational Linguistics.

Xusen Yin, Li Zhou, Kevin Small, and Jonathan May. 2021. Summary-oriented question generation for informational queries. In *Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering (DialDoc 2021)*, pages 81–97, Online. Association for Computational Linguistics.

Xingdi Yuan, Tong Wang, Yen-Hsiang Wang, Emery Fine, Rania Abdelghani, Hélène Sauzéon, and Pierre-Yves Oudeyer. 2023. Selecting better samples from pre-trained LLMs: A case study on question generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12952–12965, Toronto, Canada. Association for Computational Linguistics.

Yao Zhao, Xiaochuan Ni, Yuanyuan Ding, and Qifa Ke. 2018. Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3901–3910, Brussels, Belgium. Association for Computational Linguistics.

Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. 2017. Neural question generation from text: A preliminary study.

Wenjie Zhou, Minghua Zhang, and Yunfang Wu. 2019. Question-type driven question generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language*

Processing (EMNLP-IJCNLP), pages 6032–6037, Hong Kong, China. Association for Computational Linguistics.