Improving Question Generation with Fine-tuning of MT5: A Comparative Study of Pre-processing Techniques and Evaluation Metrics

Zihan Xie, Shengtong Jin, Xiangyuan Chi, Shengwen Jin, Cat Wang

Abstract

Traditionally generating questions from a text has been a time-consuming and laborious task for humans. NLP provides a solution by allowing us to train models that can automatically generate questions from a given text. We leverage a novel method using fine-tuned T5 model that can save a considerable amount of time and effort, especially when dealing with large amounts of text.

1 Introduction

Our project that uses a fine-tuned T5 model aims to address the problem of automating the process of generating questions from a pair of context and answer. Traditionally, generating questions from text has been a manual task performed by humans, and is often a time-consuming and labor-intensive process. We use NLP techniques to automate the task of generating questions from text at scale, and to ensure that the questions generated are accurate, relevant, and appropriate for the intended use case. Our project can solve a number of problems in various domains include:education,content creation, Q&A chatbots, and customer service,etc. Particularly, we focus on question generation in Chinese, which is a more complicated domain that needs a specialized way to process data. In all of these cases, the ability to automatically generate high-quality questions from text using a T5 model can help save time, reduce costs, and improve the overall quality of the output.

2 Related Work

2.1 T5-based models

There have been several works that have used T5 models for automated question generation. For example, the paper "How Much Knowledge Can You Pack Into the Parameters of a Language Model?" (Roberts et al., 2002) [1] showed that a T5 model can be fine-tuned for the task of question generation from context and answer.

2.2 Transformer-based models

Transformer-based models such as BERT and GPT have also been used for automated question generation. For example, the paper "Learning to Answer by Learning to Ask: Getting the Best of GPT-2 and BERT Worlds " (Klein and Nabi, 2019) [2] proposed a neural model that generates diverse and meaningful questions.

2.3 MT5

MT5 has been specifically designed for multilingual tasks and is pre-trained on 101 languages, which is significantly more than other state-of-the-art multilingual models. Its architecture includes language-specific components such as language embeddings and cross-lingual attention mechanisms that enable the model to better model the specific features of each language and transfer knowledge between languages more effectively.

2.4 Our Project

Despite so many prior works on question generation, study on question generation in Chinese has been relatively scarce. With this project, we aim to utilize the fine-tuned T5 model to experiment with Chinese question generation. We will benchmark the model with a Chinese Question-Ans dataset and compare its accuracy with English version. Moreover, based on the output, we will introduce some Chinese natural language preprocessing techniques to make the output more semantically and logically meaningful. We will also try to confine the context to TV shows and movies.

3 Problem Description

Question generation is a natural language processing task that involves generating a question from a given context and answer. The goal of this task is

to generate a question that is relevant to the context and answer, and that elicits information that is not explicitly stated in the context. This task is particularly challenging because it requires the model to understand the relationship between the context and answer and to generate a question that is both grammatically correct and semantically meaningful. Furthermore, the generated question should not simply repeat the context or answer but should instead probe deeper into the underlying meaning and implications of the text.

4 Methods

4.1 Data

- The Stanford Question Answering Dataset (SQuAD) is a popular benchmark dataset for machine reading comprehension tasks. It consists of a large collection of Wikipedia articles and a set of question-answer pairs, where each question corresponds to a span of text in the article. The goal is for a machine learning model to read the article and provide the correct answer to each question.

- The Delta Reading Comprehension Dataset (DRCD) is a Chinese language benchmark dataset for machine reading comprehension tasks. It consists of a large collection of news articles and web pages, with over 10,000 manually annotated questions and corresponding answers. DRCD was created to evaluate the ability of machine learning models to understand and reason over Chinese text, which is a significant challenge due to the complexity of the language and the lack of large-scale annotated datasets.

4.2 Preprocessing

The SQuAD dataset is a manually curated dataset, which means that it is carefully created and checked to ensure that the data is of high quality and free from errors. Each row of the SQuAD dataset, as found on Hugging Face, is already in the format of a single question-answer pair with its corresponding context. However, to use the data with a T5-based model, some pre-processing is required. The pre-processing involves reformatting each row to a format that the T5-based model can understand. The input format should be in the form of "ans:answer_text context:context</s>", where </s> is an end-of-sequence tag. This format enables the model to take in the answer text and the corresponding context and generate a relevant question.

The DRCD dataset is originally formatted such that each JSON object contains one paragraph and multiple question-answer pairs. However, we need to change the format to the SQuAD dataset format, where each JSON object contains only a single question-answer pair. To do this, we need to iterate through each paragraph in the DRCD dataset and extract each question-answer pair, along with its corresponding context. Then, we can create a new JSON object for each question-answer pair, with the context included in each object. This new format will allow us to use the DRCD dataset in a way that is compatible with existing SQuAD-based models and tools. In addition, each question-answer pair needs to be converted to the format of "ans: answer_text context: context_text</s>", where </s> represents the end-of-sequence tag. This conversion is necessary to ensure that the data is compatible with T5-based models, which require input in this specific format. By converting each question-answer pair to this format, the model will be able to more effectively analyze the context and generate relevant questions.

4.3 Model

In our study, the T5 model is imported from the transformers library as MT5ForConditionalGeneration and T5TokenizerFast. The model is specifically designed for generating questions based on a given context and answer. MT5ForConditionalGeneration is a variant of the T5 (Text-to-Text Transfer Transformer) model specifically designed for multilingual tasks, including tasks such as translation, summarization, and question generation. It extends the capabilities of the base T5 model by providing support for a wide range of languages. Like the original T5 model, MT5ForConditionalGeneration is built on the Transformer architecture, which utilizes self-attention mechanisms to process and generate sequences efficiently.

MT5ForConditionalGeneration consists of an encoder and a decoder, both of which are composed of a stack of Transformer layers. Each Transformer layer contains a multi-head self-attention mechanism, followed by a position-wise feed-forward neural network. These layers work together to process the input sequences and generate the output sequences.

The encoder processes the input text, which consists of tokenized words or subwords, and generates contextualized representations for each token in the sequence. These representations capture the meaning of the token within the context of the entire input sequence. The encoder uses self-attention to compute the relationships between tokens in the input sequence, allowing the model to understand and capture complex patterns, dependencies, and relationships.

The decoder is responsible for generating the output sequence, such as a translated sentence, a summary, or a question. It also consists of a stack of Transformer layers with self-attention mechanisms. The decoder takes the encoder's output representations as input, along with the target sequence (during training) or previously generated tokens (during inference). The decoder then generates the output sequence token by token, using the encoder's contextualized representations and its own self-attention mechanism to understand the input and generate the most relevant output.

MT5ForConditionalGeneration employs a conditional generation framework, where the model is trained to predict the target sequence given the input sequence. During training, the model learns to generate target sequences by minimizing the difference between its predictions and the ground truth target sequences. This enables the model to learn complex patterns and generate high-quality output sequences for a wide range of tasks and languages.

4.4 Training

In this study, we carried out a two-stage training process for question generation in English and Chinese. In the first stage, we trained a T5-based sequence-to-sequence model MT5ForConditionalGeneration on the SQuAD dataset in English. We fine-tuned the "google/mt5-base" pre-trained model using the Hugging Face Transformers library and adopted the T5TokenizerFast for tokenization. The model was trained for 3 epochs with a batch size of 5 on an NVIDIA A5000 GPU. After completing the training, we saved the model and tokenizer for further use.

In the second stage, we fine-tuned the saved English model on the DRCD dataset in Simplified Chinese to create a Chinese T-5 question generation model. The training procedure for the Chinese version closely resembled that of the English version, with minor modifications to accommodate the different dataset and starting point. The Chinese model was trained for 10 epochs using the same batch size, optimizer, learning rate scheduler, and data collator as in the English version. The fine-tuned model and tokenizer were saved upon completion.

Finally, we pushed the trained models and tokenizers for both English and Chinese question generation tasks to Hugging Face Model Hub, making them accessible for the broader research community and facilitating their use in various natural language processing applications. This two-stage training process aimed to achieve high-quality question generation performance in both languages, contributing to advancements in the field of natural language understanding and question-answering systems.

5 Experimental Results

The primary goal of this paper is to evaluate the performance of our fine-tuned models. The performance of is measured by the similarity between the questions generated by our models and the questions provided by the human-labeled datasets.

5.1 Metrics

We evaluated our model with four automated metrics:

Sentence Similarity (Hugging Face, 2021)[3] converts input texts into vectors (embeddings) that capture semantic information and calculate how close (similar) they are between them. This task is particularly useful for information retrieval and clustering/grouping.

Bleu (Papineni et al., 2002)[4] is a metric used to evaluate the quality of machine-generated text. It measures the overlap between the generated text and the reference text using n-grams and ranges from 0 to 1, with higher scores indicating better quality translations or summaries. However, it has some limitations and should be used in combination with other evaluation metrics and human judgement.

Rouge-L is a metric used to evaluate the quality of machine-generated summaries or translations by comparing them to human-written reference summaries. It measures the longest common subsequence (LCS) between the generated summary and the reference summary and places a higher weight on longer LCS, making it a "recall-oriented" met-

	SQuAD(EN)			CMRC2018(ZH)	
	Sentence Sim	Bleu	ROUGE-L	Sentence Sim	Bleu
GPT2-QG	74.09	24.38	46.74	NA	NA
MT5-TyDiQA	56.02	3.62	24.97	65.59	1.72
MT5-BilingualQG	**80.01**	**31.61**	**53.98**	**69.40**	**9.12**

Table 1: Score of considered metrics against human-generated questions from different datasets

ric.

5.2 Datasets

We evaluate our models on Squad_V2 and

Squad_V2 (Rajpurkar et al.,2018)[5] is a benchmark dataset for machine reading comprehension, consisting of over 100,000 questions designed to test a machine's ability to read and understand a passage of text and then answer questions about it. The key difference between Squad_v1 and Squad_v2 is that v2 includes unanswerable or ambiguous questions, challenging models to recognize and handle such cases.

CMRC2018 (Cui et al., 2019)[6] is a large-scale dataset for machine reading comprehension tasks in the Chinese language. It contains over 10,000 real-world documents and 30,000 manually annotated questions and answers. The dataset has been widely used for evaluating the performance of natural language processing models on Chinese language understanding and MRC tasks.

5.3 Results and Discussion

In table 1, we report the scores of considered metrics against human-generated questions from SQuAD and CMRC2018 for different question generation models. Since the GPT2-QG model doesn't support Chinese, the Sentence Similarity and Bleu score don't apply to it on the CMRC2018 dataset.

When evaluated with SQuAD, our fine-tuned MT5-BilingualQG model outperforms GPT2-QG and MT5-TyDiQA over all metrics(Sentence Similarity, Bleu, and ROUGE-L). This indicates that compared to the other two models, the questions generated by our model are closer to those generated by humans. The Bleu score of MT5-TyDiQA is abnormally lower than expected, which could be due to the fact that it was trained on a completely different dataset and supports over a hundred languages. It's important to take this into consideration when interpreting the results. Its Sentence Similarity score, which takes into account semantic similarity, and ROUGE-L score fall in the reasonable range, although still significantly lower compared to those of other models.

Our fine-tuned MT5 QG model also shows similar advantages over the other two models when generating questions in Chinese. The same problem can still be observed on MT5-TyDiQA, whose Bleu is abnormally lower.

6 Conclusions and future work

In this work, we presented a comparative study of different pre-processing techniques and evaluation metrics for question generation using fine-tuned MT5 models. The paper aimed to address the problem of automating the task of generating questions from a pair of context and answer, especially for Chinese language. The paper used two datasets, SQuAD and DRCD, to train and evaluate the models. The paper also introduced some Chinese natural language pre-processing techniques to improve the quality of the generated questions. The paper reported the results of various metrics, such as sentence similarity, word mover's distance, Bleu, and Rouge-L, to compare the performance of the models with human-generated questions. The paper concluded that the fine-tuned MT5-BilingualQG model outperformed other models on both English and Chinese question generation tasks.

Although the questions generated by our model are similar to those generated by humans, we did not evaluate the quality of the generated questions from other aspects, such as fluency and grammar. This suggests some future directions for research, in which we can evaluate the performance of the model using some other metrics for its fluency and grammar.

7 Division of labor

–Zihan Xie: Performed Datasets pre-processing and implemented model training process
–Shengtong Jin, Shengwen Jin: Model Evaluation and result analysis
–Cat Wang: Training model and debug.
– Xinagyuan Chi: Project presentation

8 Git Repository

group8-finalproject

References

[1] Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? *CoRR*, abs/2002.08910, 2020.

[2] Tassilo Klein and Moin Nabi. Learning to answer by learning to ask: Getting the best of GPT-2 and BERT worlds. *CoRR*, abs/1911.02365, 2019.

[3] Hugging Face. Sentence similarity. https://huggingface.co/tasks/sentence-similarity, 2021. Accessed: 2023-04-22.

[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.

[5] Pranav Rajpurkar, Robin Jia, and Percy Liang. Squad2.0: The stanford question answering dataset, 2018. Accessed on 2023-04-26.

[6] Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. A span-extraction dataset for Chinese machine reading comprehension. 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 5883–5889, Hong Kong, China, November 2019. Association for Computational Linguistics.