Chinese Question Generation

Zekai Zhang
Peking University
2000013064@stu.pku.edu.cn

Abstract

In this project, I identified a helpful prompt structure that moderately improves the performance of the QG model. To guide the model's decoding process, I implemented a post-selection method that selects the generated texts with the highest similarity to the given answer. Upon analyzing the data, I discovered significant differences between the training and testing sets, so I applied a data augmentation technique to minimize these gaps. The combination of these methods resulted in a steady increase in content-related metrics, such as BLEU and ROUGE, and a significant improvement in QA-related metrics, such as QA_EM and QA_F1. Code is available at https://github.com/violets-blue/NLPDL_final.

1 Problem Description

Question generation is the process of creating natural language questions based on a given text or set of texts. Given a text T = $\{t_1, t_2, ..., t_n\}$ and a target answer phrase A, the task is to generate a natural language question Q such that Q is relevant to T and A, and Q is coherent and well-formed, as shown in figure 1.

文本：中国电信统一客服电话为10000，自助服务热线为 10001。

答案：10000
问题：中国电信客服电话

Figure 1: An example of Question Generation.

2 Baseline

I used the Randeng-BART-139M-QG-Chinese[1] model as a baseline and fine-tuned it on the provided dataset. The baseline model was pre-trained on the ChineseSQuAD dataset and already had the ability to perform Chinese question generation. I also conducted experiments in low-resource conditions to test the model's performance in these circumstances.

The results of the experiments showed a slow but steady improvement in performance with the increase in data size. This suggests that the model is mainly learning to adapt to the question form rather than learning the ability to generate questions. It is worth noting that the model only has 139M parameters, so its performance may be less desirable compared to models that use the mT5-base structure as their backbone.

Data	QA_EM	QA_F1	BLEU	ROUGE-L	BScore
5%	26.8	53.5	21.5	46.75	74.52
10%	27.0	53.9	23.59	48.74	75.06
20%	27.0	53.8	24.74	50.55	75.84
100%	29.1	56.0	27.99	54.1	77.63

Table 1: Experiment results of the baseline model under low-resource circumstances.

3 Case Analysis

I manually went through 100 randomly sampled model generations by the baseline model, and summarize the main error classes here.

Question Type Inconsistency: The baseline model often generates question types that do not match the ground truth. For example, when the answer is "可以" and the ground truth is "可以吗", the baseline model tends to generate other types of questions like "用什么".

Phrase Inconsistency: The baseline model sometimes changes the referred object when generating questions, such as "中国电信" becoming "上海电信".

Wrong Answer Location: In some cases, the answer appears multiple times in the context, leading to different possible questions for different an-

[1] https://huggingface.co/IDEA-CCNL/Randeng-BART-139M-QG-Chinese

Model	QA-EM	QA-F1	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	BScore
baseline	27.6	53.6	27.06	54.17	31.81	52.65	77.22
AS	29.0	56.0	27.99	55.62	32.79	54.1	77.63
AS+DA	35.0	60.9	29.53	55.28	32.66	53.85	77.86
AS+DA+PS	35.7	62.1	30.02	56.26	33.18	54.71	78.23

Table 2: Experiment results of all methods combined. AS, DA, PS stand for Answer Separation, Data Augmentation(Pseudo Question), and Post Selection respectively.

Model	QA-EM	QA-F1	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	BScore
baseline	27.6	53.6	27.06	54.17	31.81	52.65	77.22
AS	29.0	56.0	27.99	55.62	32.79	54.1	77.63

Table 3: Ablation study of Answer Separation.

swer locations. For example, in the text "汤姆克鲁斯的第一任妻子咪咪是山达基教教徒,而克鲁斯正是被妻子拉入教派的。 科幻作家罗恩·哈伯德于1954年创立了山达基教。山达基教公开反对现代医学、现代心理学,也限制信徒接受手术", the phrase "山达基教" appears multiple times, making it challenging to generate the correct question.

Analysis: Upon examining the provided datasets, I found significant differences between the training and testing datasets. For example, while the training dataset does not contain any questions of the type "可以吗", the testing dataset includes 149 such questions. This likely explains why the model has difficulty generating this type of question, as it has never encountered it during training. In conclusion, these errors are likely due to issues with the data rather than the model itself. The wrong answer location problem is a common issue in question generation and highlights the complexity of the task. The Phrase Consistency problems indicate that the baseline model lacks the ability to accurately retain content, suggesting room for potential improvements.

4 Proposed Methods

4.1 Answer Separation

According to the study by (Kim et al., 2019), replacing the target answer in a passage with a special token <ans> can improve the performance of QG model by allowing it to better utilize the information from both the passage and the target answer. In this project, I implemented this technique by separating the target answer from the rest of the passage, resulting in a boost in the model's performance.

4.2 Pseudo-Q Data Augmentation

In order to address the significant differences between the train and test data, I proposed to use data augmentation to minimize the gap. Specifically, I plan to generate pseudo questions using rule-based methods that incorporate question types not present in the train data. This will be achieved by searching the passages in the train data for specific keywords and then transforming those pieces into interrogative forms. For example, I will locate all sentence pieces that contain the keyword "吗" and add a "可以吗" to the end of each one, with the corresponding answer being "可以".

4.3 Post Selection

During the decoding phase, the model does not receive direct guidance from the the answer. This can sometimes lead to issues with inconsistency in the generated text. To address this, I adopted a sample and rank approach. This involves generating a set of potential outputs, evaluating them using BLEU score with the answer, and ranking them from the highest to the lowest. The output with the highest score is then chosen as the final result.

5 Experiments

5.1 Experimental settings

For training, I set batch_size, epoch, learning_rate as 64, 5, and 3e-04 respectively. All experiments are conducted on a single RTX2080Ti GPU. For more implementation details, please refer to my Github repository.

5.2 Comparison with baselines

The results of the proposed methods are presented in Table 2. It is evident that all the methods can

Model	QA-EM	QA-F1	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	BScore
AS	29.0	56.0	27.99	55.62	32.79	54.1	77.63
AS+PS	31.8	57.7	30.38	56.85	34.37	55.04	78.5
AS+DA	35.0	60.9	29.53	55.28	32.66	53.85	77.86
AS+DA+PS	35.7	62.1	30.02	56.26	33.18	54.71	78.23

Table 4: Ablation study of Post Selection.

Model	QA-EM	QA-F1	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	BScore
baseline	29.0	56.0	27.99	55.62	32.79	54.1	77.63
QA	26.3	52.6	26.65	54.36	31.89	52.96	76.9
SQuAD	30.2	56.3	26.74	53.74	31.18	52	76.98
Pseudo_Q	35.0	60.9	29.53	55.28	32.66	53.85	77.86
Purified	36.7	60.7	30.85	56.09	34.25	54.7	78.32

Table 5: Ablation study of different data augmentation methods.

enhance the model's performance. Among them, the Pseudo Question Data Augmentation method is particularly effective, resulting in a maximum gain of 5 points in QA-F1 and QA-EM. In general, all methods combined can improve the fluency and quality of the generated questions, as demonstrated by the boost in content-based metrics. Furthermore, the proposed methods increase the answerability of the generated questions, leading to notable enhancement in QA-related metrics.

5.3 Ablation Study

In this section, I perform ablation studies on each of the proposed methods to evaluate their individual contributions.

Effect of Answer separation: Based on Table 3, it can be inferred that the Answer Separation method leads to improvements on all metrics. It is surprising that such a small modification can result in overall improvements, highlighting the effectiveness of this method.

Effect of Pseudo-Q Data Augmentation: Table 4 presents the comprehensive results of the data augmentation methods, including the use of a pretrained QA model to extend the training dataset, the incorporation of Chinese SQuAD as an additional training set, and our own Pseudo Question method. We also evaluate the performance of the baseline model on a purified dataset, which has been manually cleaned to remove gaps between the training and test sets.

On the one hand, the SQuAD and QA methods do not bring significant benefits. One possible explanation is that we loaded the pretrained Chinese QG model's parameters at the beginning of training, meaning that our model already had the ability to generate questions from the start. As suggested by the finetuning process, most of the data is used to bridge the gap between question expression styles rather than to learn how to generate questions. However, the QA and SQuAD methods may introduce noise in the expression of questions, leading to a decline in performance.

On the other hand, our Pseudo Question method produces comparable results to the experiments on the purified dataset, indicating its effectiveness. By encountering pseudo questions during training, the model gains some knowledge about different types of questions, and is able to effectively bridge the gap between the training and test data.

Effect of Post Selection: To evaluate its effectiveness, I conducted experiments on the baseline and the baseline with data augmentation. As shown in Table 5, the performance steadily increases after adding post-selection in both experiments. As an improvement in coding strategy, post-selection can be easily combined with any previous improvement, making it highly useful.

6 Conclusion

In this project, my contributions are three-fold. Firstly, I implemented the Answer Separation method to enhance the model's performance. Secondly, I proposed a new data augmentation method called Pseudo Question to reduce the gap between the train and test data. Finally, I applied post-selection to better guide the model's decoding process. Overall, the refined model significantly outperformed the baseline model in QA-related met-

rics, demonstrating the success of the proposed methods.

References

Yanghoon Kim, Hwanhee Lee, Joongbo Shin, and Kyomin Jung. 2019. Improving neural question generation using answer separation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6602–6609.

7 Acknowledgements

I would like to express my sincere gratitude to the teacher assistants for their thorough planning and execution of this course. The design of Homework 4 and the final project was particularly effective, and I learned a great deal from completing them. These assignments allowed me to explore the capabilities of Huggingface and enhanced my skills in organizing my code and writing reports, which will be invaluable as I pursue scientific research. I am especially grateful to Professor Sujian Li, who generously offered to write a recommendation letter for me, even though I was only a student in her class. Her humble demeanor and dedication to her students played a significant role in the success of this flipped classroom experience.