# Fine-tuned T5 Models on FairytaleQA Chinese Dataset

**Sijie XIONG[a*], Haoling XIONG[b], Tao SUN[a], Haiqiao LIU[a], Fumiya OKUBO[a], Cheng TANG[a], Atsushi SHIMADA[a]**

[a]*Graduate School and Faculty of Information Science and Electrical Engineering, Kyushu University, Japan*

[b]*School of Business, ECUST, East China University of Science and Technology, China*

*xiong.sijie.630@s.kyushu-u.ac.jp

**Abstract:** Question answering (QA) training plays a significant educational role in boosting comprehension learning skills for both machines and young children. One representative of QA datasets is FairytaleQA, but it is limited to English. Therefore, in this project, we manually translate the essence of the dataset into Chinese and validate the utility of the FairytaleQA Chinese dataset based on five T5 models. The results demonstrate that the FairytaleQA Chinese dataset has acceptable utility.

**Keywords:** Question Answering, FairytaleQA, Chinese, Fine-tuned T5 Models

## 1. Introduction

Question answering (QA) training plays a pivot role in young children's education and machine reading systems (MRS) across the world, cultivating their abilities of language comprehension and reasoning (Peng, et al., 2023). The quality of such training processes, however, depends on the choice of QA datasets. One of such datasets is the FairytaleQA, which contains both explicit and implicit questions, and allows for Question generation (QG) tasks (Xu, et al., 2022). Although the FairytaleQA is appraised for its expertise and comprehensive contents, it also receives criticisms due to that there is English version only. However, the FairytaleQA has not yet been translated into Asian languages or hieroglyphics such as Chinese. Therefore, we manually translate essential training datasets, testing and validating datasets. In addition, to verify the utility of our Chinese FairytaleQA, we conduct QA, QG tasks based on five fine-tuned T5 models tailored for Chinese. The experimental results demonstrate that our dataset has acceptable utility for QA and QG tasks.

## 2. Machine Translation Issues

To extend the FairytaleQA into Chinese, we initially tried several machine learning models to automatically translate English into Chinese. The disappointed phenomena and the leading factors are demonstrated herein in detail. **Outdated Representation**: the machine translation models sometimes deliver a mixture of both Simplified Chinese and Traditional Chinese. This does not meet the end and deteriorates the readability for both machines and young children. For example, "忧郁" (simplified) and "憂鬱" (traditional) seem very different but they share the same meaning. **Authentic Flavo**r: for instance, "old women" are frequently translated as "老妇人" in long texts but as "老妪" in short texts, while the latter one rarely appears in the modern daily life. Given the listed issues, herein, we abandon a totally machine translated version. Instead, we eliminate these errors manually and we also re-order the positions of some sentences to comply with Chinese logics. We translate essential training dataset (10,379 rows), testing and validating datasets (4,636 rows and 4,132 rows).

## 3. Experiments

We employ five established T5-based models that have been specially trained, tested, validated for Chinese texts from Hugging Face. Herein, we number these models as T5-1, T5-2, T5-3, T5-4, and T5-5. To confirm the comprehensive performance of the translated dataset, we adopt five common tasks in QA fields (Q: Question, T: Text, A: Answer, G: Generation): AT-QG, QT-AG, T-QGAG, T-QG, T-AG. To be clear, for instance, T-AG means conducting **Answer Generation** tasks based on **Text**.

Table 1. *The Performance Measured by Cross-entropy Losses of Five Pre-trained T5 Models Across Five Subordinate Generation Tasks in QA with Random Seed at 45*

| Mode/Task | AT-QG | QT-AG | T-QGAG | T-QG | T-AG |
|---|---|---|---|---|---|
| T5-1 | 1.872 | 1.910 | 1.362 | 1.413 | 1.729 |
| T5-2 | 1.990 | 1.913 | 1.431 | 1.558 | 1.762 |
| T5-3 | 2.630 | 2.721 | 1.879 | 2.047 | 2.484 |
| T5-4 | 2.153 | 1.943 | 1.466 | 1.619 | 1.780 |
| T5-5 | 1.780 | 1.920 | 1.309 | 1.377 | 1.684 |

As shown in Table 1, the cross-entropy losses range between 1.300 to 2.800 for Chinese-oriented T5 models. This scope is larger than other FairytaleQA datasets in alphabet-based languages around 0.500 to 1.000. Three reasons contribute to this phenomenon. First, the tokenizer of Chinese-oriented T5 splits the input sentence into more tokens. Second, the most initial T5 models are trained on the C4 dataset, which is not capable of languages other than English. Although after being tuned on some Chinese datasets, the Chinese-oriented T5 models are still hybrid, failing to put full of their efforts to Chinese only. Third, Chinese contexts are more compacted and have stronger content-related independence, which is more difficult to extract precisely. This is also the contribution of our dataset, because we keep the key words in consistency to stand out such dependencies. Given acceptable gaps between five Chinese-oriented T5 model, we consider that our dataset is well-structured, and its utility is acceptable.

## 4. Conclusion

In this project, we manually translate it into the FairytaleQA Chinese dataset at a smaller scale. The abandon of machine translations is to avoid the five drawbacks: inconsistent translation, outdated representation, authentic flavor, failure translation, and resource copy. To validate the utility of the Chinese dataset, we conduct extensive experiments based on five fine-tuned T5 models for five generation tasks. The result demonstrates that the FairytaleQA Chinese dataset holds an acceptable utility and can be properly dealt by current fine-tuned models.
**Acknowledgements**: This work was supported by JST CREST Grant Number JPMJCR22D1 and JSPS KAKENHI Grant Number JP22H00551, Japan.

## References

Peng, W., Li, W., & Hu, Y. (2023, July). Leader-generator net: dividing skill and implicitness for conquering FairytaleQA. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 791-801).

Xu, Y., Wang, D., Yu, M., Ritchie, D., Yao, B., Wu, T., ... & Warschauer, M. (2022, April). 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)*.