BIT-ACT: An Ancient Chinese Translation System Using Data Augmentation

Li Zeng zengli@bit.edu.cn
Beijing Institute of Technology, Beijing, 100081, China
Yanzhi Tian tianyanzhi@bit.edu.cn
Beijing Institute of Technology, Beijing, 100081, China
Yingyu Shan shanyingyu@bit.edu.cn
Beijing Institute of Technology, Beijing, 100081, China
Yuhang Guo* guoyuhang@bit.edu.cn
Beijing Institute of Technology, Beijing, 100081, China

Abstract

This paper describes a translation model for ancient Chinese to modern Chinese and English for the Evahan 2023 competition, a subtask of the Ancient Language Translation 2023 challenge. During the training of our model, we applied various data augmentation techniques and used SiKu-RoBERTa as part of our model architecture. The results indicate that back translation improves the model's performance, but double back translation introduces noise and harms the model's performance. Fine-tuning on the original dataset can be helpful in solving the issue.

1 Introduction

Ancient Chinese translation is a Machine Translation task, aiming to translate ancient Chinese into modern Chinese or English, which is of great significance for the research and understanding of ancient Chinese. In the area of ancient Chinese translation, the presence of unique features in ancient Chinese poses challenges. Ancient Chinese has its own distinctive features, including a large number of rare characters, characters with different meanings in ancient and modern times. In the aspect of syntax, ancient Chinese are very different from modern Chinese, including lots of inverted or elliptical sentences, which increases the difficulty in translation. Maksym and Tetyana (2015) Additionally, due to the limited amount of data, it's necessary to apply data augmentation during training.

In our system, we use a transformer (Vaswani et al., 2017) architecture with adjusted parameters, to address the limited data, we applied various data augmentation techniques to generate additional data, achieved improved performance. Considering the unique grammatical structures of ancient Chinese, we tried a BERT Devlin et al. (2018) model pre-trained on ancient Chinese, we had analyzed different results from these approaches.

*Corresponding Author

Proceedings of ALT2023: First Workshop on Ancient Language Translation, pages 43–47
September 5, 2023, Macau SAR, China.

2 Method

In this section, we introduce our data processing and augmentation method. We also describe the architecture of our systems.

2.1 Data Processing

Training data for evaluation is excerpted from the Twenty-Four Histories(dynastic histories from remote antiquity till the Ming Dynasty), the Pre-Qin classics and "ZiZhi TongJian (Comprehensive Mirror in Aid of Governance)". The Twenty-Four Histories is the general name for the 24 official histories written by the various dynasties in ancient China. The Pre-Qin classics refer to historical materials from the Pre-Qin period (Paleolithic Period 221 B.C.), which play an important role in ancient books, including history books and subsidiary texts. "Zizhi Tongjian" is a multi-volume chronological history book compiled by Sima Guang, a historian of the Northern Song Dynasty. It covers 1,362 years of history of the sixteen dynasties.The Chinese ancient classic texts in the corpus exhibit diachronicity, spanning thousands of years and encompassing the four traditional types of Chinese canonical texts: Jing (Classics), Shi (Histories), Zi (Philosophical Works), and Ji (Literary Works)."

We conduct the following data processing method:

1. Because of the lack of segmentation for datasets, we apply segmentation with 'jieba' [1].
2. Apply 15K BPE (Byte-Pair-Encoding) (Sennrich et al., 2016) to the datasets.
3. Discard extremely long sentences (2048 tokens without BERT and 512 tokens with BERT) during training.
4. Randomly select validation sets: extract 2000 sentences from the Twenty-Four Histories and 40 sentences from Zizhi Tongjian.

After data processing, the quantities of the data are as follows:

Dataset	Sentences	Tokens
Ancient Chinese train set	311,352	7,912,087
Modern Chinese train set	311,352	9,495,032
Ancient Chinese valid set	2,040	51,875
Modern Chinese valid set	2,040	62,626

Table 1: Statistic of Datasets

2.2 Data Augmentation

Due to the limited size of the dataset and the constraints imposed by the closed track, where the use of external data is restricted, we employed various data augmentation techniques to augment the available data and enhance the capability of our model.

To apply data augmentation, we trained models for modern Chinese to ancient Chinese translation .Using these models, we translated the training set data, obtained new data in ancient Chinese, then mixed it with the original data, Then retrain a new model with mixed data. Which is commonly referred to as back translated(BT)Edunov et al. (2018). Notice that the vocabulary of the data may have changed after each back translation, we had apply BPE and preprocess on training data again. Details will be provided in selection 3.

[1] https://github.com/fxsjy/jieba

2.3 Model

We employed the transformer model provided by fairseq [2] Ott et al. (2019) as our architecture, based on the scale of the dataset, we conducted experiments using the following parameters.

Parameter	Value
Attention Heads	4
Number of Layers	6
Embedding Dimension	512
Feed-forward Hidden Size	1024

Table 2: Model Hyperparameters

Due to the limited amount of data, we also explored the use of pre-trained models. In our research, we employed SiKu-RoBERTa [3]Wang et al. (2022) as a pre-trained model. SiKu-RoBERTa is based on the RoBERTa model architecture and was trained on the complete text corpus of the "Si Ku Quan Shu" , a large-scale series of books compiled during the Qianlong period of the Qing Dynasty. We believe that SiKu-RoBERTa has the ability to capture rich knowledge of the ancient Chinese language.

3 Experiments

In this section, we delve into the training details and steps, including the utilization of hyperparameters. We also compare and analyze the results obtained from different models.

For the Model Configuration, we used the following set of Model Configuration as the default. We will point out if there are any modifications. We usd Adam Kingma and Ba (2014) as our optimizer, use the "inverse sqrt lr" scheduler with 4000 warm-up steps.

Configuration	Value
optimizer	Adam
lr-scheduler	inverse_sqrt
learning-rate	0.00005
dropout	0.2
weight-decay	0.0001

Table 3: Model Configuration

We conducted the training on two TITAN X GPUs, for each model, we trained for a total of 300,000 updates.

3.1 Ancient Chinese-Modern Chinese

In the task of translating ancient Chinese to modern Chinese, we experimented with a range of data augmentation strategies and evaluated the effectiveness of the RoBERTa model. Through extensive testing, we assessed the performance of these models and identified several factors that could lead to results.

First, we trained a baseline model using the given dataset and the configuration above.

[2] https://github.com/facebookresearch/fairseq

[3] https://github.com/hsc748NLP/SikuBERT-for-digital-humanities-and-classical-Chinese-information-processing

Then, we utilized the aforementioned data augmentation technique and trained a model for modern Chinese to ancient Chinese translation. Using this model, we performed one round of back-translation (BT) on the original dataset.

After retraining the model using the data augmented through back translation, we obtained a new model. We observed a significant improvement in the model's performance after the back translation process. Encouraged by these results, we performed the same back translation operation once again, to examine whether further improvement can be achieved. It was referred to as double back translation (double BT)

However, in reality, the double back translation didn't achieve the expected results. We suspect that this might due to the introduction of additional noise in the generated data. Considering this, we revert back to the original data, and fine-tune out model using the original dataset.

Finally, we tried to incorporate the BERT model. We used SiKu-RoBERTa as the embedding layer for the encoder.Zhu et al. (2020) After each training method, we evaluated the model's performance on valid set using the BLEU Papineni et al. (2002) score. The results are as follows.

Model	BLEU
baseline	32.4
BT	36.4
double-BT	35.2
double-BT and fine-tune	36.9
SiKu-RoBERTa	29.8

Table 4: BLEU Score of Ancient Chinese-Modern Chinese Model

3.2 Ancient Chinese-English

Considering the provided parallel text is limited, directly train the model on original text may not yield satisfactory results. Therefore, we only tested the method of training the model on the generated data.

First, we trained a baseline model using the provided ancient Chinese to English test. Due to data is limited, we were using dropout=0.4 Srivastava et al. (2014) to avoid overfitting.

Then, we generated data by decoding ancient Chinese in "Twenty-Four Histories"

Finally, we trained a new model using only the newly generated data. This model is served as our final model for the ancient Chinese to English translation task.

3.3 Official Evaluation Results

We used the best-performing model mentioned above to translate the official provided data and submitted it for testing. The results are as follows:

Translation Direction	BLEU
Ancient Chinese - Modern Chinese	21.95
Ancient Chinese - English	1.11

Table 5: Officially Evaluated BLEU Scores

The BLEU score obtained after submission showed a significant deviation from the scores we observed during local testing. We believe this discrepancy might be attributed to the model

overfitting to a certain extent or due to poor performance on the specific corpus likely caused by differences in the source of the data.

4 Conclusion

For the ancient Chinese to modern Chinese task, we found that back translation can help improve the model's performance with limited data. However, it's important to notice that back translation may introduce noise, and fine-tuning the model after back translation could potentially enhance its performance.

When using RoBERTa as embedding, the results were not as expected. One possible reason is that we did not train RoBERTa model on our data separately, leading to mismatch in vocabulary. Additionally, the large scale of RoBERTa might cause overfitting, We believe that more data will help address this issue.

For the ancient Chinese to English task, due to the limited dataset and significant linguistic differences between ancient Chinese and Endlish, we did not achieve satisfactory results.

References

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Edunov, S., Ott, M., Auli, M., and Grangier, D. (2018). Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Maksym, K. and Tetyana, S. (2015). Lexical difficulties in translation of ancient chinese texts into the ukrainian and english languages (case study of the chinese treatise "the art of war" and its translations into the ukrainian and english languages). *SECTION II. CROSS-CULTURAL COMMUNICATION IN CONTEMPORARY GEOPOLITICAL SPACE*, page 24.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). 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.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wang, D., Liu, C., Zhu, Z., Feng, J., Hu, H., Shen, S., and Li, B. (2022). Construction and application of pre-training model of "siku quanshu" oriented to digital humanities. *Library Tribune*, 42(6):31–43.

Zhu, J., Xia, Y., Wu, L., He, D., Qin, T., Zhou, W., Li, H., and Liu, T.-Y. (2020). Incorporating bert into neural machine translation. *arXiv preprint arXiv:2002.06823*.