AnchiLm: An Effective Classical-to-Modern Chinese Translation Model Leveraging bpe-drop and SikuRoBERTa

Jiahui Zhu miugod0126@gmail.com
School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai, 200093, China

Sizhou Chen jjyaoao@126.com
Blockchain Industry College, Chengdu University of Information Technology, Chengdu, 610225, China

Abstract

In this paper, we present our submitted model for translating ancient to modern texts, which ranked sixth in the closed track of ancient Chinese in the 2nd International Review of Automatic Analysis of Ancient Chinese (EvaHan). Specifically, we employed two strategies to improve the translation from ancient to modern texts. First, we used bpe-drop to enhance the parallel corpus. Second, we use SikuRoBERTa to simultaneously initialize the translation model's codec and reconstruct the bpe word list. In our experiments, we compare the baseline model, rdrop, pre-trained model, and parameter initialization methods. The experimental results show that the parameter initialization method in this paper significantly outperforms the baseline model in terms of performance, and its BLEU score reaches **21.75**.

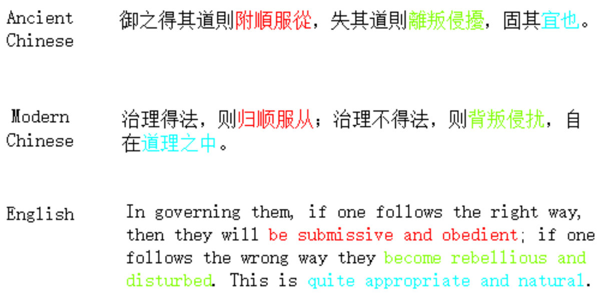
1 Introduction

Ancient Chinese historical texts are not only key parts of Chinese civilization but treasures of global culture. Understanding these works is challenging for modern people, hence, translation is vital to bridge this gap. Traditional manual translation of these texts is time-consuming and challenging. With the advancement of computer science, Machine Translation (MT) provides a new solution to this problem. Among various MT methods, Neural Machine Translation (NMT) is representative.

However, applying NMT to ancient text translation faces significant challenges. The main issue is the relatively small corpus for ancient text translation, making it hard for models to learn and capture the complex grammar and rich semantics from limited data. Further, as shown in **Figure 1**, the conversion between modern and ancient Chinese is highly complicated. Consequently, the results of ancient text translation often fall short of expectations.

In recent years, the "pre-training + fine-tuning" paradigm has become a powerful technique in the field of Neural Machine Translation (NMT). Researchers have attempted to leverage this paradigm to enhance translation models, such as XLM Conneau and Lample (2019) , MASS-Song et al. (2019) , mBARTLiu et al. (2020) , and mRASP2Pan et al. (2021) , which have performed excellently in machine translation tasks. Meanwhile, there have also been some excellent solutions in the field of ancient text translation. For instance, a Transformer model was

Proceedings of ALT2023: First Workshop on Ancient Language Translation, pages 55–60
September 5, 2023, Macau SAR, China.



Ancient Chinese	御之得其道則附順服從，失其道則離叛侵擾，固其宜也。
Modern Chinese	治理得法，则归顺服从；治理不得法，则背叛侵扰，自在道理之中。
English	In governing them, if one follows the right way, then they will be submissive and obedient; if one follows the wrong way they become rebellious and disturbed. This is quite appropriate and natural.

Figure 1: Changes in Linguistic Features between Modern Chinese and Ancient Chinese

trained to translate ancient Chinese into modern Chinese , AnchiBERTLiu et al. (2019) pre-trained BERTDevlin et al. (2018) on ancient texts and then initialized the encoder of the NMT model, or the Bert-FusedZhu et al. (2020) method fused the BERT output into every layer of the encoder-decoder using the attention mechanism...

Existing research provides a series of valuable insights. In this paper, we aim to further advance research in this field by proposing **AnchiLM** to deal with specific problems and challenges more effectively. It has the following main features:

1) We initialize the encoder with the SikuRoberta model pre-trained on the Siku Quanshu, and introduce the method of alternating initialization from Deltalm[9] to initialize the decoder parameters.

2) We noticed that when using the SikuRoberta model, tokenizing at the character level could lead to excessively long translation sequences. To address this issue, we encode sentences in a mixed character-word manner and initialize the new vocabulary embedding with the topn character embedding vectors.

3) We introduce bpe-drop, a simple and effective subword regularization method. It randomly disrupts the BPE segmentation process, resulting in multiple segmentations within the same fixed BPE framework.

2 Methods

In this section, we will detail the methodology of our approach, which consists of three main components: the encoder, the decoder, and the word embedding. Each of these components plays a crucial role in our translation model.

2.1 Encoder

Based on the idea of domain adaptation training, SikuRoBERTa continues to train on the basis of the RoBERTa structure combined with a large amount of ancient Chinese corpus from ”Siku Quanshu”, so as to obtain a pre-training model for the ancient Chinese field. It can provide rich semantic information of ancient texts. We take SikuRoBERTa to initialize the encoder parameters.

2.2 Decoder

Since the Transformer decoder adds a cross-attention layer to each layer of Bert to capture the correlation between the source language and the target language, this paper uses the alternate initialization method proposed in deltalm to self-attention in each layer of the decoder. A feed-forward layer is inserted in the middle of cross-attention, so that the two layers of bert initialize one layer of the decoder, thus fully initializing the six-layer decoder.

2.3 Word Embedding

In response to the problem that the word-level tokenization of Chinese SikuRoBERTa leads to excessively long sentences, low encoding efficiency, and difficulty in training convergence, this paper rebuilds the vocabulary at the subword level. Specifically, the ancient text and modern text are merged and jointly trained with BPE to build a unified vocabulary. Then for each token in the new vocabulary, the SikuRoBERTa word embedding vector of its first character is taken to initialize the vocabulary.

In summary, our approach combines a pre-trained encoder, a specially initialized decoder, and a reconstructed vocabulary to effectively handle the challenges of ancient text translation. The detailed structure of our model is illustrated in **Figure 2**.

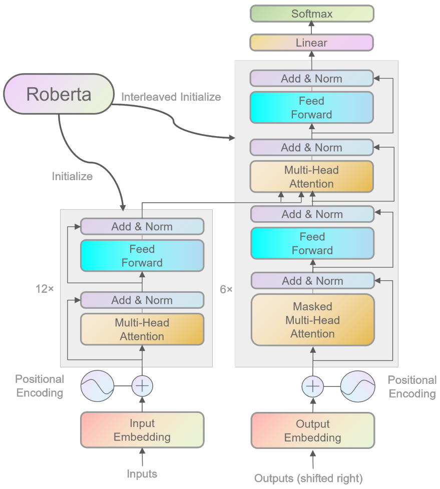


Figure 2: AnchiLm model structure diagram

3 Experiment

For the introduction of the experimental data, we put it in the **Appendix A** part, and then start to explain the experimental part from the data processing.

3.1 Data Processing

Our experiments mainly focus on the translation from Classical Chinese to Simplified Chinese characters. In the process, the Classical Chinese text is segmented using the Jiayan Classical Chinese word segmenter, while the Modern Chinese text is segmented using the Jieba segmenter and then converted to Traditional Chinese. The Classical Chinese and Modern Chinese texts are combined and trained with 12,000 BPE subword merge operations.A discussion of the vocabulary is in **Table 3** in Appendix B.1.

The data of all subsequent experiments were expanded by 3 times using bpe-drop with a drop probability of 0.1. The results of bpe-drop are shown in **Table 4** of Appendix B.2.

3.2 Evaluation Metrics

In this paper, BLEU-4 is used as the evaluation metric for the validation set. First, the bpe is removed from the translation and candidates, then the Traditional Chinese is converted to Simplified Chinese and segmented with Jieba, and finally, the BLEU score is calculated using the multi-bleu.perl script. For the test set, the official score is used.

3.3 Experimental Settings

Our experiments are based on the open-source machine translation framework fairseq. The training parameters are: each batch contains up to 8192 subwords, the update round is 60,000, the learning rate is 0.0005, the inverse square root learning rate adjustment strategy is used, the warmup step is 4000, and the Adam optimizer is used with parameters 1=0.9, 2=0.98. During decoding, beam search with a beam width of 5 is used. The models compared are as follows:

Baseline: The baseline uses the transformer base model, both the encoder and decoder are 6 layers, the embedding dimension is 512, the feed-forward layer dimension is 2048, 8-head attention is used, and dropout is 0.2.

R-drop enhancement: R-Drop minimizes the bidirectional KL divergence between the output distributions of two sub-models sampled by dropout, thereby producing more robust output.

Span pre-training: Span corruption is to reconstruct the text spans based on the masked input document. The probability of corruption is 0.15, and the average length of spans is 3.

DAE pre-training: The denoising autoencoding task proposed in BART, which improves the performance of the model by reconstructing the text from the noised text. The probability of token mask is 0.35.

Sikuroberta parameter initialization: The model used in this paper, the encoder is 12 layers, the decoder is 6 layers, the embedding dimension is 768, the feed-forward layer dimension is 3072, 12-head attention is used, and dropout is 0.1.

Among them, 1 to 5 all use the same transformer base architecture. For the pre-training tasks of 3 and 4, they are first trained for 30,000 rounds, and then the translation task is learned.

3.4 Results

We report the validation and test set BLEU for all compared methods, as shown in **Table 1**. Compared with the baseline, rdrop improves by 0.74, while the pre-training method using span or dae improves by 1.49 and 1.26, respectively. However, we use the SikuRoBERTa initialization method to increase the BLEU score by 2.1, and the final test set score is **21.75**. Moreover, if the model of the same scale is directly trained without SikuRoBERTa initialization, the training will not converge.The display of the training effect is in **Table 5** of Appendix B.3.

Model	Valid BLEU	Test BLEU
Baseline	28.31	-
+R-Drop	29.0	-
+Span	29.80	-
+DAE	29.57	-
SikuRoBERTa-init	**30.41**	**21.75**

Table 1: EvaHan2023 Ancient Chinese Translation Closed Task (All models were trained using BPE-drop to augment data by a factor of 3)

4 Conclusion

This paper describes our submission to the 2nd International Evaluation of Automatic Analysis of Ancient Chinese (EvaHan) closed track for ancient Chinese translation. Our ancient text-modern text translation model includes two parts: bpe-drop data enhancement and parameter initialization. Future work is to combine parameter initialization and pre-training tasks simultaneously.

References

Conneau, A. and Lample, G. (2019). Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

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*.

Liu, D., Yang, K., Qu, Q., and Lv, J. (2019). Ancient–modern chinese translation with a new large training dataset. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(1):1–13.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Pan, X., Wang, M., Wu, L., and Li, L. (2021). Contrastive learning for many-to-many multilingual neural machine translation. *arXiv preprint arXiv:2105.09501*.

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T.-Y. (2019). Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*.

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*.

Appendix

A Dataset

The training datasets are respectively the parallel corpora of ancient Chinese to modern Chinese from the Twenty-Four Histories of China, the pre-Qin classics, and the ”Zizhi Tongjian”. The statistical data is as follows in Table X:

Data	Total volume of ancient texts	Total translations
Parallel Corpus of Twenty-Four Histories of Ancient Bai	9,583,749 Character	12,763,534 Character
Pre-Qin classics and ancient English parallel corpus of ”Zi Zhi Tong Jian”	618,083 Character	838,321 Words

Table 2: EvaHan2023 training data details

Among them, the parallel corpus of ancient and modern Chinese in the Twenty-Four Histories has 307,494 lines, and the pre-Qin and ”Zizhi Tongjian” have 5,899 lines. 250 pairs, a total of 500, are taken from the two corpora as the validation set.

B Some More Extra Material

B.1 Vocabulary Experiment

In order to explore the impact of word segmentation tools on model performance in ancient text translation, we used jieba and jiayan to conduct three sets of experiments. Table x shows that using jiayan and jieba word segmentation works best for ancient texts and modern texts, respectively.

No	Method	Bleu
1	Jieba_Jiaba	27.08
2	Jieba_Jiaba	27.17
3	Jiayan_Jieba	**27.89**

Table 3: Vocabulary Experiment Details

B.2 Bpe Drop

bpe drop is a data processing method that can increase data granularity and generate multivariate data, specifically as shown below.

No.	**Augmented text**
1	三@@ 者 同@@ 时 发生 而 又 出现 黄河 的 水@@ 清 。
2	三@@ 者 同时 发@@ 生 而 又 出现 黄河 的 水@@ 清 。
3	三@@ 者 同时 发生 而 又 出现 黄河 的 水@@ 清 。

Table 4: An example of BPE-Dropout result(Factor=3, Drop=0.1).

B.3 Model training effect

The following is the training effect of AnchiLm on the test data Id 1 and Id 100:

ID	**Language**	**Sentence**
1	An-CH	契丹侵澶，公相真宗北伐，騙河未渡。
	Mo-CH	契丹 侵犯 澶 州 ， 萊 公相 真宗 北伐 ， 臨近 黃河 沒有 渡過 黃河 。
98	An-CH	所著《索蘊》，乃其學也。
	Mo-CH	所著 的 《 索蘊 》 ， 就是 他 的 學問 。

Table 5: Case Study.