Pre-trained Model In Ancient-Chinese-to-Modern-Chinese Machine Translation

Jiahui Wang wangjiahui@smail.nju.edu.cn
Xuqin Zhang 2580334082@qq.com
Kuangyaming Honor School, Nanjing University, Nanjing, 210023, China

Jiahuan Li lijh@smail.nju.edu.cn
Shujian Huang huangsj@nju.edu.cn
National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, 210023, China

Abstract

Neural Machine Translation (NMT) has emerged as a powerful approach for language translation, with the Transformer model revolutionizing the field. One key aspect that has propelled the Transformer's success is the utilization of pre-training techniques. This paper presents an analysis of the pre-trained Transformer model NMT for the Ancient-Chinese-to-Modern-Chinese machine translation task.

1 Introduction

The Transformer model (Vaswani et al., 2017) has demonstrated exceptional performance in various natural language processing tasks, including machine translation. One key aspect that has propelled the Transformer's success is the utilization of pre-training techniques, such as the popular BERT (Devlin et al., 2018) model. By pre-training on large-scale corpora, BERT captures rich linguistic representations and context, allowing for more effective transfer learning.

In this study, we incorporate a pre-trained Transformer model, Chinese-RoBERTa-wwm-ext (Cui et al., 2021) into our NMT system, enabling it to leverage the wealth of linguistic knowledge encoded in the pre-training process. By fine-tuning the pre-trained model on translation-specific data, we aim to exploit the benefits of both pre-training and task-specific learning.

2 Related Work

The currently most commonly used BERT model (Devlin et al., 2018) is pre-trained on general-domain text using universal language representation. While the model exhibits strong generality, its performance is easily constrained when applied to natural language processing tasks involving domain-specific texts. Due to the inherent differences in grammar, semantics, and pragmatics between Ancient Chinese and other languages, there are significant deviations in features. It is challenging to achieve the same performance level as in general corpora. Therefore, the direct application of BERT in projects related to Ancient Chinese does not yield ideal results.

Proceedings of ALT2023: First Workshop on Ancient Language Translation, pages 23–28
September 5, 2023, Macau SAR, China.

AnchiBERT (Tian et al., 2021) is a pre-trained model specifically designed for Ancient Chinese texts. It "reads" a total of 39.5 million characters of Ancient Chinese, including historical records, prose, ancient poetry, and couplets, spanning thousands of years. The downstream tasks of AnchiBERT include comprehension and generation of Ancient Chinese texts. The paper suggests that to use AnchiBERT for text generation in Ancient Chinese, a framework based on the Transformer model can be adopted. The encoder part of the framework utilizes AnchiBERT, while the decoder part uses the original Transformer model's decoder with randomly initialized parameters.

In the same year, Dongbo Wang et al. employed high-quality, validated full-text corpus from the Qing Dynasty's Qianlong period edition of the extensive series "Siku Quanshu" as an unsupervised training set. They continued training a BERT model based on the BERT structure, incorporating a large amount of Ancient Chinese texts. This led to the development of the Siku BERT (Wang et al., 2022) pre-trained language model specifically tailored for intelligent processing tasks related to Ancient Chinese. By directly using the pre-trained model as initialization parameters, not only did the model possess stronger generalization capabilities and faster convergence speed, but it also required only a small amount of labeled data for fine-tuning, significantly improving the performance of natural language processing tasks while avoiding overfitting.

The Siku BERT pre-trained language model and AnchiBERT pre-trained model introduced the idea of transfer learning in low-resource machine translation research. This can provide a theoretical and practical foundation for the text generation project in Ancient Chinese.

3 Approach

We employed a standard Transformer-based NMT architecture with the pre-trained model, Chinese-RoBERTa-wwm-ext (Cui et al., 2021) as the encoder and a randomly initialized decoder, depicted in Figure 1 .Chinese-RoBERTa-wwm-ext(Cui et al., 2021), as an advanced language model specifically designed for processing and understanding the Chinese language, has demonstrated exceptional performance on a wide range of Chinese NLP benchmarks, surpassing previous state-of-the-art models in tasks such as text classification, sentiment analysis, and natural language understanding. During the training process, we employed a strategy where the encoder parameters were frozen, and only the decoder parameters were updated. At the same time, we adopted the technology of joined dictionary. Thanks to the similarity between Ancient Chinese and Modern Chinese, joined vocabulary simplifies the process of aligning words between Ancient Chinese and Modern Chinese. By using a joined dictionary, identical words only require one embedding vector, reducing the number of model parameters, memory consumption, and computational overhead, thereby enhancing training efficiency.

4 Experinental Setup

Data: The source of the training data includes the Ancient-Chinese-to-Modern-Chinese parallel texts of China Twenty-four Histories, with 9,583,749 characters for the original Ancient Chinese texts as source data and 12,763,534 characters modern Chinese translation as target data.

Training Details: The model was implemented on the top of fairseq toolkit[1]. The dropout rate was set to 0.3. We set weight decay to $1e-4$ to overcome over-fitting. We used Adam (Kingma and Ba, 2014) to optimize the model parameters, with $\beta_1 = 0.9$ and $\beta_2 = 0.98$.

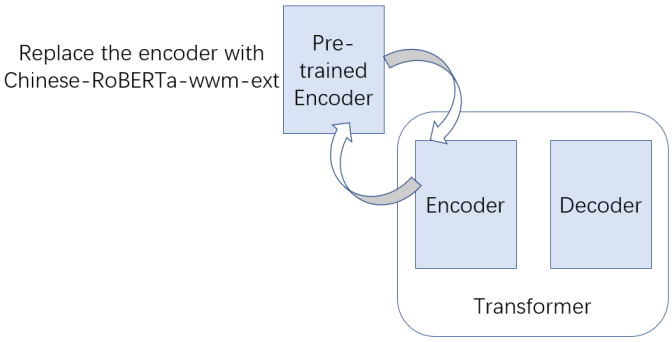
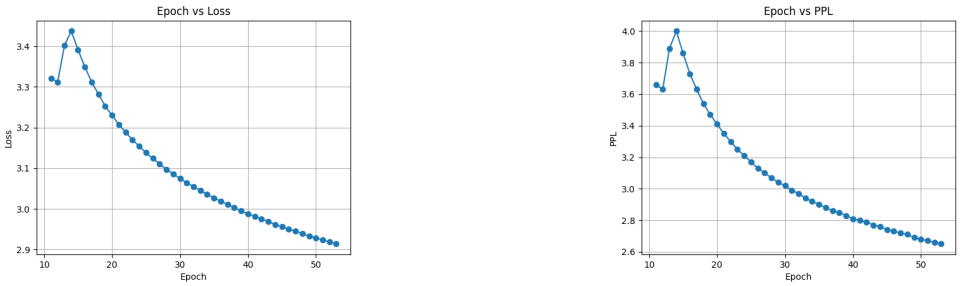


Figure 1: Model Architecture



(a) Loss

(b) PPL

Figure 2: Loss and PPL

5 Results

During local test, we get the following results. Figure 1 shows the loss during training and the change of perplexity(PPL), both of which decrease steadily. Figure 2 shows the changes in the BLEU (Papineni et al., 2002) score of the validation set, showing an upward trend. The performance on the valid set of the best checkpoint was evaluated with the following metrics: Loss: The overall loss achieved on the valid set was 3.192. Loss represents the discrepancy between the predicted output and the ground truth and is minimized during training. Negative Log-Likelihood (NLL) Loss: The NLL loss was calculated as 1.61. It measures the average negative log probability of the correct target tokens given the model's predictions. Perplexity (PPL): The perplexity value obtained was 3.05. PPL is a measure of how well the model predicts the next token in the sequence. Lower perplexity indicates better predictive performance. BLEU Score: The BLEU score achieved on the valid set was 35.84. BLEU is a widely used metric to evaluate the quality of machine translation outputs. A higher BLEU score indicates better translation quality.

By integrating our translation examples, we have observed that our model can produce relatively smooth translations for short sentences. While some individual words may not be translated directly into modern Chinese, this does not hinder the conveyance of meaning, as depicted in Figure 4. However, in Figure 5, the model struggles to accurately analyze pronouns

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

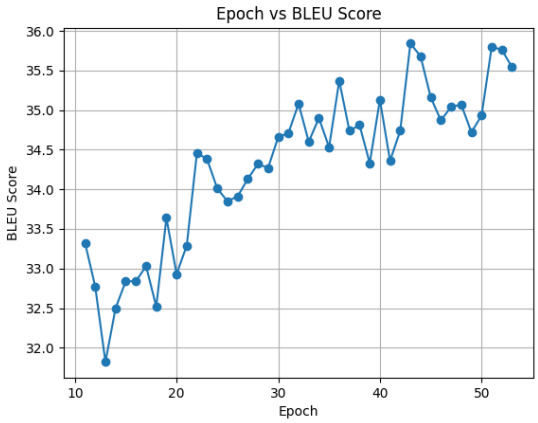


Figure 3: BLEU on valid set

Ancient:復勑罷江西巡撫韓光祐。（简体：复勑罢江西巡抚韩光祐）

Reference:又揭發罷免江西巡撫韓光祐。（简体：又揭发罢免江西巡抚韩光祐）

Hypothesis:又彈劾罷免江西巡撫韓光祐。（简体：又弹劾罢免江西巡抚韩光祐）

Figure 4: Short example1

that require expansion based on context. Although this does not significantly impact the overall comprehension of the translated sentences, it does emphasize the importance of incorporating longer contextual information. For short sentences, relying solely on the source-to-target sentence pairs may not suffice to effectively translate referential information.

For longer sentences, as shown in Figure 6, we have found that our model is generally able to maintain consistency in conveying meaning. Although the accuracy of the translation may slightly decrease with increasing sentence length, there are instances where certain vocabulary may not be rendered with utmost precision. Nonetheless, overall comprehension of the sentences can still be achieved. On some vocabulary, our model translates more detail. Also, the model demonstrates a high level of proficiency in accurately recognizing and classifying proper nouns. Its advanced language processing capabilities enable it to effectively identify and distinguish names of specific people, places, organizations, and other entities.

In the official EvaHan2023[2] test set, the best model achieves the BLEU score of 22.05.

[2]https://github.com/GoThereGit/EvaHan

Ancient: 其子孫年幼者戚配流嶺外，誅其親黨數百餘家。(简体：其子孙年幼者戚配流岭外，诛其亲党数百馀家。)

Reference: 他們的子孫年幼的都流放嶺外，誅殺他們的親黨幾百家。（简体：他们的子孙年幼的都流放岭外，诛杀他们的亲党几百家）

Hypothesis: 其子孫年幼的都流放到嶺外，誅殺其親黨數百多家。（简体：其子孙年幼的都流放到岭外，诛杀其亲党数百多家。）

Figure 5: Short example2

(a)

(b)

Figure 6: Translation examples sampled from validation set

6 Conclusion

In this study, we explored the application of pre-trained Transformer models in Ancient-Chinese-to-Modern-Chinese machine translation. By incorporating the Chinese-RoBERTa-wwm-ext model as the encoder in our NMT system, we aimed to leverage the rich linguistic representations and contextual knowledge captured through pre-training. Our findings and experimental results shed light on the effectiveness of pre-training techniques for improving translation quality in this specific language pair. By leveraging pre-training techniques and adopting a joined dictionary approach, we achieved moderately satisfactory results, exploring the way for Ancient-Chinese-to-Modern-Chinese machine translation.

References

Cui, Y., Che, W., Liu, T., Qin, B., and Yang, Z. (2021). Pre-training with whole word masking for chinese bert.

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*.

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

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.

Tian, H., Yang, K., Liu, D., and Lv, J. (2021). Anchibert: A pre-trained model for ancient chinese language understanding and generation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, D., Liu, C., Zhu, Z., Liu, J., Hu, H., Shen, S., and Li, B. (2022). Sikubert and sikuroberta: Construction and application research of pretrained models for digital humanities in siku quanshu. *Library Tribune*, 42(6):31–43.