Some Trials on Ancient Modern Chinese Translation

Li Lin efsotr_l@stu.pku.edu.cn.Peking University.China
Xinyu Hu huxinyu@pku.edu.cn.Peking University.China

Abstract

In this study, we explored various neural machine translation techniques for the task of translating ancient Chinese into modern Chinese. Our aim was to find an effective method for achieving accurate and reliable translation results. After experimenting with different approaches, we discovered that the method of concatenating adjacent sentences yielded the best performance among all the methods tested.

1 Introduction

Chinese characters are the writing system of Chinese and are considered one of the oldest written languages in the world. According to verifiable records, over 3000 years ago, Chinese characters had already developed a mature writing system, including oracle bone inscriptions. While some characters have been retained throughout the subsequent development process, the expression forms and meanings of ancient Chinese and modern Chinese differ significantly. Ancient Chinese often features rare words that are not commonly found in modern Chinese, and the grammar structures also vary. Consequently, reading ancient Chinese poses difficulties for modern individuals, often necessitating the expertise of professionals to translate it into modern Chinese.

Neural Machine Translation has already demonstrated remarkable performance in various bilingual translation tasks. However, there has been limited exploration of the existing advanced Neural Machine Translation technology in the domain of monolingual translation from ancient Chinese to modern Chinese. This relatively specialized field has received little attention in terms of developing Neural Machine Translation technology. In this Evahan 2023 competition, we are participating in the task of translating ancient Chinese into modern Chinese. The training data of this task is extracted from the Twenty-Four Histories (recording the history from the pre Qin period to the Ming Dynasty), which was finished by the research group of the National Social Science Foundation of China major project "Research on the Construction and Application of Cross-language Knowledge Base of Ancient Chinese Classics" (project No. :21&ZD331)

In this paper, we begin by introducing the methods employed in our study, including data augmentation, fine-tuning of pre-trained models, and attention mechanisms such as group attention. Next, we provide details on our training setting, including the use of pre-trained models, data segmentation, vocabulary construction, and the division of training and validation sets. We also present the experimental results, including the performance of different methods on evaluation metrics such as BLEU and ChrF. Lastly, we discuss the outcomes of our attempts, highlighting the effectiveness of certain techniques, the limitations of others, and the potential for further improvements in low-resource translation tasks.

Proceedings of ALT2023: First Workshop on Ancient Language Translation, pages 29–33
September 5, 2023, Macau SAR, China.

2 Method

In this section, we will describe some methods that we have tried.

Data augmentation Data augmentation is an essential technique in machine learning tasks to increase the size and diversity of the training data. In our case, we have observed that adjacent sentences in the training data are typically in the same article. Therefore, a straightforward data augmentation method we employ is concatenating adjacent sentences to form longer sentences. We denote the concatenation of no more than k sentences as k-cat.

Tune Training translation models can be seen as training classification models, as both tasks involve categorizing text. However, the challenge arises from the presence of numerous categories, leading to a long tail problem. To address this issue, we draw inspiration from the approach described in (Menon et al., 2020).

In our methodology, we augment the trained model by adding $\log P(x)$ to the bias term of the final classification layer. Here, $P(x)$ represents the prior distribution of token x, which is derived from the distribution observed in the training set. Subsequently, we continue training the model, while keeping the non-classification layers frozen.

Group attention Group attention techniques (Bao et al., 2021), such as group attention and combine attention, were applied to data augmented with concatenated sentences. Group attention focuses exclusively on information within the same sentence during attention calculations, while combine attention combines group attention with traditional global attention.

Finetune pre-trained model To further enhance the model's performance, we conducted fine-tuning using the Masked LM technique inspired by BERT (Devlin et al., 2018). The pre-trained model was trained for 50 epochs on the training dataset. Subsequently, the fine-tuned model was loaded and trained on the 3-cat and 4-cat data.

3 Experiment

3.1 Baseline setting

For the translation model, we utilize the encoder-decoder Transformer architecture (Vaswani et al., 2017) and employ the EncoderDecoderModel which is provided by the Hugging Face library [1] to build our system. The basic model loads the first 6 layers of SikuBERT (Wang et al., 2022) (resp. Chinese-BERT-wwm (Cui et al., 2020)) parameters for encoder (resp. decoder) part. In the embedding layer of encoder (resp. decoder), the corresponding parameters are loaded for the original tokens (resp. the original simplified tokens), and if unavailable, the [UNK] parameters are used. The remaining parameters in the model are initialized randomly.

3.2 Training detail

For all experiments, we adopt the AdamW optimizer (Loshchilov and Hutter, 2017). The hyper parameters of AdamW optimizer set as follows : $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e - 8, \lambda = 0.01$. The learning rate is scheduled using inverse_sqrt with a maximum learning rate of 1.5e-4 and warmup steps of 30000. We set the label smoothing as 0.1 and the batch size as 32. For the Tune method, we set the maximum learning rate to 4e-5 and the warmup steps to 10000. We use beam search with beam size 5 for decoding and report the BLEU score (Papineni et al., 2002) and the ChrF score (Popović, 2015) on validation set.

During our experiments, we observed that after training for 4 to 7 epochs, the validation set loss would start to increase. However, during this time, the BLEU and ChrF scores of

[1] `https://huggingface.co/`

the validation set continued to improve. So, our stopping strategy is based on the validation set's BLEU score: if it does not exceed the highest point for five consecutive epochs, we stop training. After stopping, we select the checkpoint with the highest BLEU score among the last five epochs or their average as the final training result.

3.3 Data preprocess and vocabulary

Brief overview of training data The training data comprises around 0.3 million parallel corpora, encompassing both ancient Chinese and modern Chinese texts. In the training data, ancient Chinese has been supplemented with modern punctuation marks. An important point to note is that both ancient Chinese and modern Chinese are written using traditional Chinese characters.

Data preprocess The training data undergoes several processing steps. To handle sentences that exceed the length of 256 tokens, we first segment them based on the ending punctuation marks. In order to align the segmented portions, we utilize a similarity metric based on 1-gram, 2-gram, and 3-gram comparisons. To determine the optimal alignment scheme, we employ dynamic programming techniques. This helps us align the segments in a way that maximizes the similarity between the original and translated texts while ensuring that the segments do not exceed the maximum length limit of 256 tokens. However, some segments may still surpass the 256-length limit or exhibit significant length discrepancies between the original and translated texts. These segments, representing either non-translatable sections or incorrect training data alignment, are discarded.

The training set and validation set are randomly divided in a 9:1 ratio. Then, based on the training set, we generate k-cat data by concatenating adjacent sentences, as described in section 2. The statistics of the training, k-cat, and validation sets, including the number of sentences and characters, are presented in Table 1.

	baseline	2-cat	3-cat	4-cat	valid
#sents	320K	584K	801K	980K	36K
#src chars	8M	21M	36M	53M	0.89M
#tgt chars	10M	27M	48M	69M	1.2M

Table 1: Number of sentences and characters in training, k-cat, and validation sets.

Vocabulary The word segmentation method used in this task is based on single characters. When constructing the vocabulary, we consider only those words that appear more than 5 times in the training set. The dictionary sizes for the encoder and decoder can be seen in table 2.

	#vocab(occur ≥ 6)	#vocab
src	6,367	9,425
tgt	5,997	8,548

Table 2: Dictionary Sizes for Ancient and Modern Chinese

4 Result & Discussion

The results of the experiments are presented in table 3. It can be observed that as the number of concatenated sentences increases, the performance of the model improves. However, there is a diminishing return effect, and the performance essentially plateaus at 4-cat.

Method	baseline	2-cat	3-cat	4-cat	3-cat(tune)	4-cat(tune)
BLEU	61.451*	63.386	63.728	63.838*	63.824	**63.941**
ChrF	58.986*	61.186	61.561	61.696*	61.651	**61.774**

Table 3: Performance comparison of baseline and k-cat models, where k ranges from 2 to 4. The table also includes the performance of 3-cat and 4-cat models that were tuned without modifying the bias. Note: * indicates that the result is the average of multiple checkpoints.

We have also experimented with alternative initial embedding parameter methods, such as integrating the embedding of ancient and modern parts and incorporating additional dictionary definitions. However, these methods did not yield significant improvements in performance. In fact, in some cases, these alternative methods even resulted in worse results. As a result, we concluded that the current initial embedding parameter approach, as described earlier, is the most suitable for our task.

For the tune method, we have found that the performance of the model remains comparable even after removing these modifications and training the model using the same settings. Furthermore, after applying the tune method, there is a slight improvement in BLEU score of approximately 0.1. Similarly, the ChrF metric shows a slight improvement of less than 0.1. These improvements indicate that the tuning process has a positive impact on the model's translation quality, albeit with modest gains.

For the group attention method, our experiments involved testing different learning rates. However, the results consistently showed that this method resulted in significantly worse performance compared to the standard attention method. In this particular setting, this indicates that either using group attention limits the model's capability or that training the model with this method requires careful adjustment of learning rates across all model components. It is possible that the introduction of group attention affects the gradients differently, making the learning process more sensitive and challenging. Therefore, further investigation and fine-tuning of the model's learning rates would be necessary to achieve better performance with the group attention method.

For the fine-tune pre-trained model method, the results showed an improvement of approximately 0.06 in BLEU and 0.07 in ChrF for the 3-cat data. However, there was no improvement observed for the 4-cat data, which can be attributed to the lower performance after the average checkpoint. The reason for the limited improvement in this method is likely due to the small size of the fine-tuning data. With a small amount of data available for fine-tuning, the model may not have sufficient exposure to the specific characteristics and patterns of the task at hand.

5 Conclusion

In conclusion, among all the methods explored in our experiments, the data augmentation technique of directly concatenating sentences proved to be the most effective. However, as the number of concatenated sentences increased, the improvement in performance became less significant. This suggests that simply adding more repetitive data does not necessarily lead to better results. It also indicates that the potential of the model may not be fully utilized with only 0.3M parallel sentences. Therefore, for low-resource translation tasks such as translating from ancient Chinese to modern Chinese, data augmentation methods should be the primary approach to consider.

References

Bao, G., Zhang, Y., Teng, Z., Chen, B., and Luo, W. (2021). G-transformer for document-level machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3442–3455, Online. Association for Computational Linguistics.

Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., and Hu, G. (2020). Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online. Association for Computational Linguistics.

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*.

Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Menon, A. K., Jayasumana, S., Rawat, A. S., Jain, H., Veit, A., and Kumar, S. (2020). Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*.

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.

Popović, M. (2015). chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

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). Construction and application of pre-trained models of siku quanshu in orientation to digital humanities [j]. *Library Tribune*, 42(06):31–43.