Translating Ancient Chinese to Modern Chinese at Scale: A Large Language Model-based Approach

Jiahuan Cao jiahuanc@foxmail.com
Dezhi Peng pengdzscut@foxmail.com
Yongxin Shi yongxin_shi@foxmail.com
Zongyuan Jiang eejiangzongyuan@mail.scut.edu.cn
Lianwen Jin eelwjin@scut.edu.cn
School of Electronic and Information Engineering, South China University of Technology, Guangzhou, 510641, China

Abstract

Recently, the emergence of large language models (LLMs) has provided powerful foundation models for a wide range of natural language processing (NLP) tasks. However, the vast majority of the pre-training corpus for most existing LLMs is in English, resulting in their Chinese proficiency falling far behind that of English. Furthermore, ancient Chinese has a much larger vocabulary and less available corpus than modern Chinese, which significantly challenges the generalization capacity of existing LLMs. In this paper, we investigate the Ancient-Chinese-to-Modern-Chinese (A2M) translation using LLMs including LLaMA and Ziya. Specifically, to improve the understanding of Chinese texts, we explore the vocabulary extension and incremental pre-training methods based on existing pre-trained LLMs. Subsequently, a large-scale A2M translation dataset with 4M pairs is utilized to fine-tune the LLMs. Experimental results demonstrate the effectiveness of the proposed method, especially with Ziya-13B, in translating ancient Chinese to modern Chinese. Moreover, we deeply analyze the performance of various LLMs with different strategies, which we believe can benefit further research on LLM-based A2M approaches.

1 Introduction

Ancient Chinese plays a crucial role in carrying the invaluable heritage of traditional Chinese culture. However, ancient Chinese expresses in a significantly

Proceedings of ALT2023: First Workshop on Ancient Language Translation, pages 61–69
September 5, 2023, Macau SAR, China.

different way compared with modern Chinese, which hinders the understanding of ancient Chinese books by non-experts. Therefore, automatic Ancient-Chinese-to-Modern-Chinese (A2M) translation is essential to the preservation of traditional Chinese culture.

Existing neural machine translation methods mainly adopted a sequence-to-sequence paradigm, evolving from architectures based on recurrent neural networks [1, 2], to convolutional neural networks [3], and to Transformer [4]. Although great industrial and academic success has been achieved in the neural machine translation area, the A2M translation [5, 6, 7] is still quite under-explored. With the emergence of large language models (LLMs) [8, 9], they have rapidly been applied to a wide variety of natural language processing (NLP) tasks, exhibiting high generalization and reasoning capacities. Although there have been studies [10] that attempt to use LLMs for ancient Chinese, their model sizes are limited.

To this end, we propose to solve the A2M translation problem using LLMs with large-scale parameters and datasets. Specifically, the model architecture is based on LLaMA [8] and its variants (*e.g.,* Ziya [11]). Furthermore, owing to the lack of Chinese texts in the pre-training corpus of LLaMA, we align the Chinese understanding ability of LLaMA with English using vocabulary extension and incremental pre-training following recent works [12, 13, 14]. After that, a large-scale A2M translation dataset with 4M pairs is employed to fine-tune the LLM, so as to transfer its general capacity to the specific A2M translation task. The experimental results demonstrate the effectiveness of our method, exhibiting 29.68% BLEU-4 on the testing set of the EvaHan2023 competition dataset [15].

2 Methodology

In this section, we present our approach for transferring the knowledge of a pre-trained LLM (*e.g.*, LLaMA [8]) using English pre-training corpus to the task of translating ancient Chinese to modern Chinese. As depicted in Figure 1, our approach involves three main steps, *i.e.*, vocabulary extension, incremental pre-training, and large-scale finetuning. In the following sections, we will give detailed descriptions of these steps.

2.1 Vocabulary Extension

As the original vocabulary of LLaMA lacks sufficient Chinese characters, the encoding of a single Chinese character commonly requires multiple tokens, resulting in low efficiency and unsatisfactory performance. Therefore, based on the training

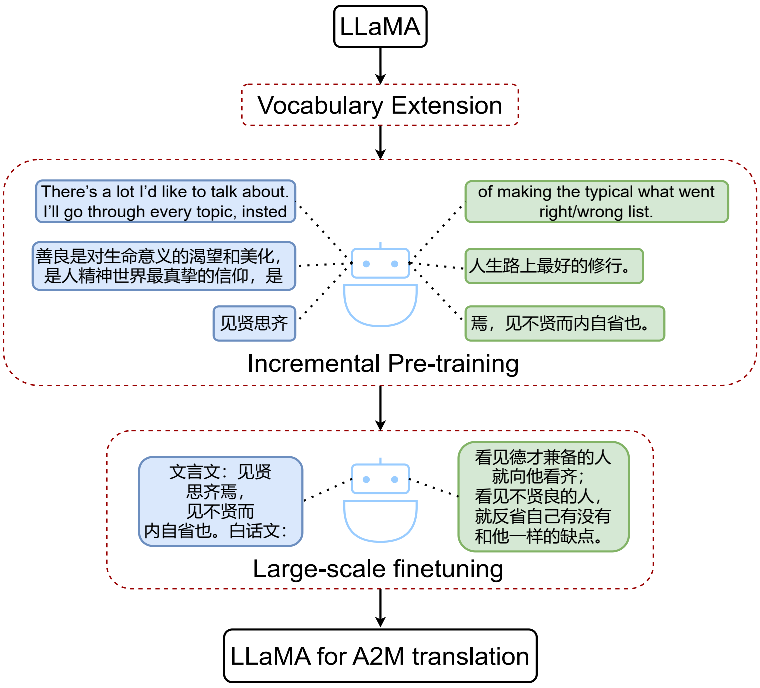

Figure 1. Overview of our method.

data of EvaHan2023, we extend 9,180 common characters and words of ancient and modern Chinese in addition to the original LLaMA vocabulary.

Furthermore, to fully utilize the knowledge of the pre-trained LLaMA, we propose DecompInit which initializes the embedding of the extended characters and words by token decomposition. As shown in Fig. 2, instead of using the popular random initialization, we initialize the embedding of a new character/word by averaging the embeddings of the tokens that it can be decomposed into. Specifically, we first denote the original LLaMA vocabulary and corresponding embeddings as $\{w_i\}_{i=1}^m$ and $\{E_i\}_{i=1}^m$, respectively, where m is the vocabulary size. Then the embedding E_{m+1} of a new word w_{m+1} that can be decomposed to $\{w_{a_i} | 1 \leq a_i \leq m\}_{i=1}^n$ is initialized as

$$E_{m+1} = \frac{1}{n} \sum_{i=1}^{n} E_{a_i}. \tag{1}$$

Empirical experiments demonstrate that this novel initialization strategy can help the model converge faster in the early stages. Moreover, the proposed DecompInit can better preserve the capacity of pre-trained LLaMA. As shown in Tab. 1, the DecompInit enables the LLaMA with extended vocabulary to generate a reasonable response to the input “你好” while the random initialization results in meaningless

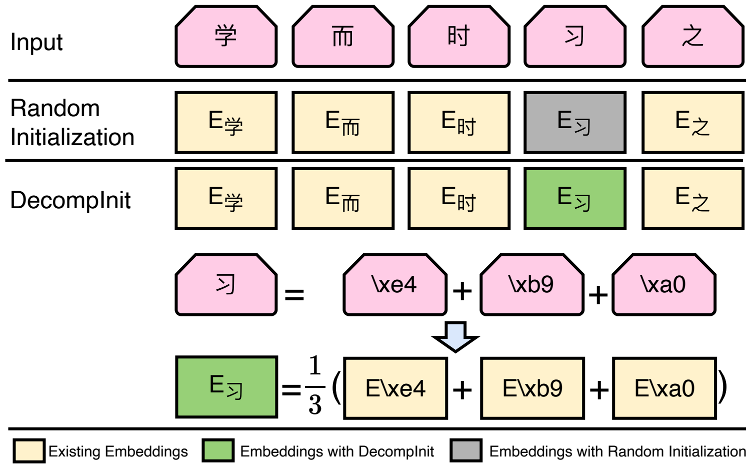

Figure 2. Illustration of the proposed DecompInit for embedding initialization.

Method	Response
Random Initialization	你 好 踐 踐\ue4bd逻 F 逻 逻 逻\n\ue4bd佾 佾 逻\n\ue4bd佾佾\n疋\n\n\n
DecompInit (Ours)	你好，我是杨晓莉，欢迎来到我的博客。\n我是一个杭州大学的学生，我的专业是计算机科学。\n我的博客是一个带有许多资料的博客，我希望能够给大家带来很多的资料。

Table 1. Comparison of model responses with random initialization and DecompInit.

outputs.

2.2 Incremental Pre-training

To further enhance the Chinese understanding capability, we conduct incremental pre-training based on LLaMA with extended vocabulary. Existing approaches to incremental pre-training are based on Chinese-only corpora [14] or mixed Chinese and English corpora [16, 11]. In this study, we validate the effectiveness of these two types of corpora. Specifically, for the Chinese-only corpora, we use Dazhige[1] and Wudao [17] for ancient Chinese and modern Chinese, respectively, while for the mixed Chinese and English corpora, we additionally incorporate Com-

[1] https://github.com/garychowcmu/daizhigev20

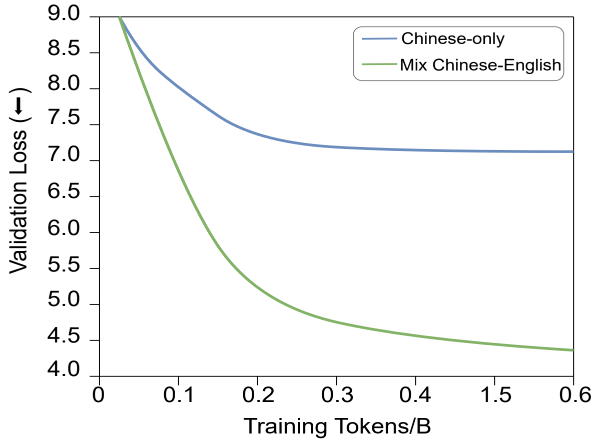

Figure 3. Validation loss on ancient Chinese.

monCrawl[2] for English. The validation loss curves on ancient Chinese of different corpora are shown in Fig. 3, which demonstrate that incorporating a mixture of Chinese and English corpora during incremental pre-training can accelerate convergence compared with using Chinese-only corpora.

2.3 Large-scale Finetuning

In order to enhance the capability of our language model in translating Ancient Chinese to Modern Chinese, we conduct large-scale finetuning using three base models, including LLaMA-7B with extended vocabulary (LLaMA-7B-EXT), LLaMA-7B with extended vocabulary and incremental pre-training (LLaMA-7B-EXT-INC), and Ziya-13B [11] that is a variant of LLaMA-13B with vocabulary extension and 110B-token incremental pre-training.

Finetuning Data. EvaHan2023 [15] originally provides 307,494 A2M translation pairs for training. We randomly sample 10,000 pairs for validation while the remaining 297,494 pairs are used for training. Moreover, we additionally use 972,467 A2M translation pairs provided by NiuTrans[3] and 2,800,000 in-house A2M translation pairs, finally yielding a large-scale finetuning dataset with 4,056,223 pairs in total.

Translation Prompts. Previous studies [18] have shown that a well-designed prompt can fully unleash the potential of large models. In our experiments, the prompt for A2M translation is "文言文：[文言文] 白话文：[白话文]", where "[文言

[2] `https://commoncrawl.org`

[3] `https://github.com/NiuTrans/Classical-Modern`

文]” represents the ancient Chinese text to translate and “[白话文]” indicates the corresponding translation in modern Chinese.

Optimization During training, the models are optimized to minimize the cross entropy loss for the tokens corresponding to the “[白话文]” part without considering the tokens of other parts, which ensures the model is focused on the translated modern Chinese text.

Inference During inference, we fill the ancient Chinese text that requires to be translated in the “[文言文]” position, yielding a translation prompt formatted as “文言文：[文言文] 白话文：”. Based on this prompt, the model predicts the “[白话文]” part which is the translation result in modern Chinese.

3 Experiments

3.1 Setting

The 7B-sized models (*i.e.*, Vanilla LLaMA-7B, LLaMA-7b-EXT, and LLaMA-7B-EXT-INC) are fine-tuned with a learning rate of 2e-5, while the 13B-sized model (*i.e.*, Ziya-13B) is fine-tuned with a learning rate of 1e-5. Other experimental settings follow Vicuna[4]. We utilize the BLEU-4 [19] and CHRF-2 [20] metrics to evaluate the performance. All experiments are conducted using 8 A100 GPUs with 80GB memory.

3.2 Ablation Study on Base Model

The ablation experiments on different base models are conducted using the 297,494 training pairs from EvaHan2023. The performances on the validation set are presented in Table 2. It can be seen that the vocabulary extension and incremental pre-training contribute to significant improvement in terms of BLEU-4 and CHRF-2. Furthermore, the Ziya-13B with much more parameters than the other three base models achieves the best A2M translation performance.

3.3 Final Results

Based on the ablation results in Section 3.2, we choose Ziya-13B as the final base model. To produce the final results of the Evahan2023 competition, we fine-tune the Ziya-13B using all available data comprising 4,056,223 pairs (Section 2.3) for 5 epochs to obtain the Ziya-13B-FT1 model, and then further fine-tune the Ziya-13B-FT1 using the total EvaHan2023 competition data with 307,494 pairs for 1 epoch to obtain the Ziya-13B-FT2 model. After performing inference on the

[4]`https://github.com/lm-sys/FastChat`

Method	BLEU-4	CHRF-2
Vanilla LLaMA-7B [8]	59.66	56.38
LLaMA-7B-EXT	60.15	56.85
LLaMA-7B-EXT-INC	<u>60.60</u>	<u>57.49</u>
Ziya-13B [11]	**61.41**	**58.22**

Table 2. Ablation study on different base models. The performances on the validation set are reported. The bold and underline indicate the best and the second best, respectively.

test set of the Evahan2023 competition using the Ziya-13B-FT1 and Ziya-13B-FT2 models, we get two sets of final results as shown in Table 3.

Method	BLEU-4
Ziya-13B-FT1	29.54
Ziya-13B-FT2	**29.68**

Table 3. Final results on the test set of the Evahan2023 competition.

4 Conclusion

In this paper, we propose a novel approach to address the Ancient-Chinese-to-Modern-Chinese (A2M) translation task using large language models (LLMs). Specifically, based on existing pre-trained LLMs, the proposed method involves vocabulary extension, incremental pre-training, and large-scale finetuning. The experimental results demonstrate the effectiveness of our method on the A2M translation task. Moreover, the ablation study highlights the importance of vocabulary extension and incremental pre-training for LLMs to improve their understanding of low-resource languages. We believe that our findings can benefit further research on LLM-based A2M approaches and contribute to the preservation of traditional Chinese culture.

Acknowledgement

This search is supported in part by NSFC (Grant No.: 61936003).

References

[1] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Neural Information Processing Systems*, pages 1–9, 2014.

[2] Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*, pages 1–15, 2015.

[3] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *International Conference on Machine Learning*, pages 1243–1252, 2017.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, pages 1–11, 2017.

[5] Dayiheng Liu, Kexin Yang, Qian Qu, and Jiancheng Lv. Ancient–modern Chinese translation with a new large training dataset. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 19(1):1–13, 2019.

[6] Hongyang Zhang, Muyun Yang, and Tiejun Zhao. Exploring hybrid character-words representational unit in classical-to-modern Chinese machine translation. In *International Conference on Asian Language Processing*, pages 33–36, 2015.

[7] Zhiyuan Zhang, Wei Li, and Qi Su. Automatic translating between ancient Chinese and contemporary Chinese with limited aligned corpora. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 157–167, 2019.

[8] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[9] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[10] Liu Chang, Wang Dongbo, Zhao Zhixiao, Hu Die, Wu Mengcheng, Lin Litao, Shen Si, Li Bin, Liu Jiangfeng, Zhang Hai, et al. SikuGPT: A generative pre-trained model for intelligent information processing of ancient texts from the perspective of digital humanities. *arXiv preprint arXiv:2304.07778*, 2023.

[11] Jiaxing Zhang, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, and Chongpei Chen. Fengshenbang 1.0: Being the foundation of Chinese cognitive intelligence. *arXiv preprint arXiv:2209.02970*, 2022.

[12] Yunjie Ji, Yan Gong, Yong Deng, Yiping Peng, Qiang Niu, Baochang Ma, and Xiangang Li. Towards better instruction following language models for Chinese: Investigating the impact of training data and evaluation. *arXiv preprint arXiv:2304.07854*, 2023.

[13] Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. Huatuo: Tuning LLaMA model with Chinese medical knowledge. *arXiv preprint arXiv:2304.06975*, 2023.

[14] Yiming Cui, Ziqing Yang, and Xin Yao. Efficient and effective text encoding for Chinese LLaMA and Alpaca. *arXiv preprint arXiv:2304.08177*, 2023.

[15] Dongbo Wang, Si Shen, Minxuan Feng, Chao Xu, Lianzhen Zhao, Wenlong Sun, Bin Li, Liu Liu, and Wenhao Ye. Evahan2023. `https://github.com/GoThereGit/EvaHan`, 2023.

[16] Zhongli Li. Billa: A bilingual LLaMA with enhanced reasoning ability. `https://github.com/Neutralzz/BiLLa`, 2023.

[17] Sha Yuan, Hanyu Zhao, Zhengxiao Du, Ming Ding, Xiao Liu, Yukuo Cen, Xu Zou, Zhilin Yang, and Jie Tang. WuDaoCorpora: A super large-scale Chinese corpora for pre-training language models. *AI Open*, 2:65–68, 2021.

[18] Daniel Khashabi, Shane Lyu, Sewon Min, Lianhui Qin, Kyle Richardson, Sean Welleck, Hannaneh Hajishirzi, Tushar Khot, Ashish Sabharwal, Sameer Singh, et al. Prompt waywardness: The curious case of discretized interpretation of continuous prompts. *arXiv preprint arXiv:2112.08348*, 2021.

[19] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Association for Computational Linguistics*, pages 311–318, 2002.

[20] Maja Popović. CHRF: Character n-gram F-score for automatic MT evaluation. In *Workshop on Statistical Machine Translation*, pages 392–395, 2015.