Sentence Segmentation and Sentence Punctuation based on XunziALLM

Zihong Chen
Nanjing University, China
chenzihong_gavin@foxmail.com

Abstract

In ancient Chinese books, punctuation marks are typically absent in engraved texts. Sentence segmentation and punctuation heavily rely on the meticulous efforts of experts and scholars. Therefore, the work of automatic punctuation and sentence segmentation plays a very important role in promoting ancient books, as well as the inheritance of Chinese culture. In this paper, we present a method for fine-tuning downstream tasks for large language model using the LoRA approach, leveraging the EvaHan2024 dataset. This method ensures robust output and high accuracy while inheriting the knowledge from the large pre-trained language model Xunzi.

Keywords: sentence segmentation, sentence punctuation, ancient Chinese information processing

1. Introduction

Chinese classical texts hold tremendous value as sources of literature. The compilation and organization of these ancient works not only serve as a bridge between the present and the past but also contribute to the scholarly exploration of cultural heritage. However, ancient Chinese writings generally lacked punctuation marks. Consequently, many surviving classical texts lack proper sentence segmentation and punctuation. This poses a significant challenge for readers seeking to comprehend these texts, as well as for scholars engaged in their analysis and interpretation.

Sentence segmentation refers to the process of converting continuous text into a sequence of sentences, where each sentence is separated by a single space. Furthermore, sentence punctuation involves placing the correct punctuation marks at the end of each sentence. However, in classical Chinese texts, sentence punctuation serves the function of sentence segmentation itself, as punctuation marks inherently possess the ability to separate sentences.

Given this situation, an effective automated algorithm needs to be proposed for batch Chinese text segmentation and punctuation tasks. In this paper, we describe the method we used in EvaHan2024. Our system is based on XunziALLM, which is a large pre-trained language base model for ancient Chinese processing. We executed extra training on the fixed provided dataset from classical sources, notably Siku Quanshu, along with other historical texts. The effectiveness of our method is demonstrated by the experimental results obtained from two test sets. Our results reveal performance gains compared to the baselines employed in the evaluation. Our findings not only showcase the adaptability of the fine-tuned model on this downstream task but also demonstrate the generalization capabilities of the Xunzi model in the domain of ancient Chinese text processing.

2. Related Work

2.1. Sentence Segmentation and Sentence Punctuation

Methods of Chinese sentence segmentation can be primarily classified into rule-based, sequence labeling model-based, and neural network language model-based approaches.

Rule-based methods are not suitable for large-scale processing of ancient texts. In recent years, research has often treated sentence segmentation in ancient Chinese texts as a sequence labeling problem similar to word segmentation. To address the issue of sentence segmentation in ancient texts, researchers have employed Conditional Random Fields (CRF) (Lafferty et al., 2001) for modeling purposes. Also, the combination of LSTM and CRF models (Wang et al., 2019) often yields better results. Wang et al. propose a sentence segmentation method for ancient Chinese texts based on neural network language models (Wang et al., 2016).

Sentence punctuation has a wide range of application scenarios in the field of speech recognition, as the textual sequences generated after recognition often lack punctuation. While neural network methods have achieved considerable success in restoring punctuation in English text, there have been relatively few efforts made to apply these techniques to Chinese punctuation restoration (Zhang et al., 2020), let alone ancient Chinese texts.

LT4HALA 2024@LREC-COLING-2024, pages 246–250

2.2. Pre-trained Language Model

Currently, large language models based on the Transformers architecture, such as GPT, T5, and BERT, have achieved state-of-the-art (SOTA) results in various natural language processing tasks. Fine-tuning pre-trained language models on downstream tasks has become a paradigm for handling NLP tasks. Compared to using out-of-the-box pre-trained LLMs (e.g., zero-shot inference), fine-tuning these pretrained LLMs on downstream datasets yields significant performance improvements. The idea behind Domain-Adaptive Pre-training is to adapt the model to a particular domain by exposing it to domain-specific language patterns, terminology, and characteristics during the pre-training phase.

However, as models grow larger, performing full parameter fine-tuning on consumer-grade hardware becomes infeasible. Additionally, storing and deploying individually fine-tuned models for each downstream task becomes highly expensive due to the comparable size of fine-tuned models to the original pre-trained models. Consequently, in recent years, researchers have proposed various parameter-efficient transfer learning methods (Lialin et al., 2023). These methods involve fixing the majority of parameters in the pre-trained model and only adjusting a small subset of parameters to achieve similar effects as full fine-tuning. The adjusted parameters can include both inherent model parameters and additional ones introduced.

3. Method

3.1. Pre-processing

We performed pre-processing on the raw data. Firstly, we detected duplicate sentences in the training data. Most of these are short sentences, such as "其二 (the second)" appearing 349 times and "宋史 (history of the Song dynasty)"appearing 174 times. As these duplicates do not contribute to performance improvement in model training, we retained only one instance of each sentence. In addition, within the training data, there are some texts lacking punctuation annotations, possibly due to annotation oversights or missing original historical records, such as "和君擊築吟請君側耳聽不是更容貌誰能知姓名主人莫稱善坐客何須驚酒酣欲罷奏壯心難自平 (With you, I strike the zither and sing, please listen closely, as appearances may be deceiving and the name of the master remains unknown, so let the seated guest not be startled, for with wine in hand and the music about to end, a strong heart finds it hard to be at peace)". In order to maintain a high standard of quality in the training data, we decided to remove these sentences which have a length of over 30.

Unlike the conventional paradigm used by previous expert models, the current LLMs primarily employ the "training + context" learning paradigm. As a result, it is necessary to select appropriate prompt templates for each downstream task to help the model recall the knowledge it acquired during training, thus achieving alignment between the downstream and pre-training tasks. The training data is partitioned into fixed-length segments, where the input consists of text sequences with designated punctuation removed, and the output is the original text. The instruction specifies, "Please add punctuation to the following unpunctuated classical Chinese passage without any additional output." To optimize context token length, no examples are included in the prompt.

3.2. Model

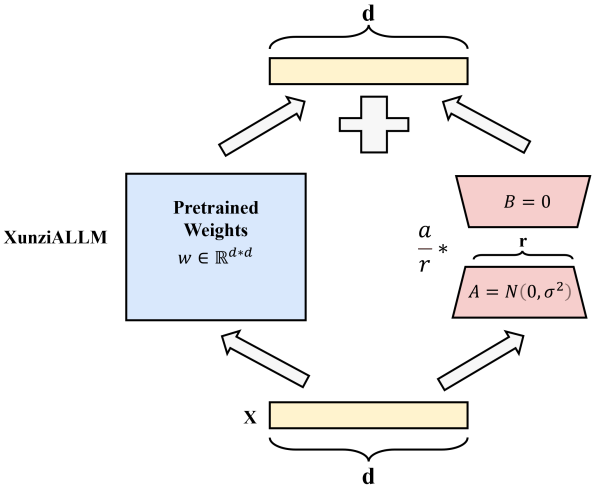

Figure 1: The structure of LoRA model.

We utilized the large language model XunziALLM, which is built upon the Qwen-7B (Bai et al., 2023) model and further pre-trained using corpora consisting of classical Chinese texts. Consequently, XunziALLM possesses extensive knowledge of classical Chinese and various capabilities in processing classical texts. To avoid making full parameter modifications to the original large-scale model, we employed the LoRA method for efficient parameter fine-tuning and supervised training, known as SFT (Supervised Fine-Tuning).

The principle behind the LoRA model involves approximating the incremental updates with low-rank matrices A and B, which are placed alongside the original pre-training matrix. This approximation is used to perform parameter updates efficiently. A large-scale model processes data by mapping it into a high-dimensional space. In fact, when deal-

ing with a specific and narrow task, it may not be necessary to employ such a complex large-scale model. Instead, it might be sufficient to focus on a sub-space range to address the task. We can define the intrinsic rank of the parameter matrix in the sub-space as the rank that achieves a certain level of performance comparable to optimizing the full parameters for the specific problem at hand.

$$W_0 + \Delta W = W_0 + BA$$
$$B \in \mathbb{R}^{d\times r}, A \in \mathbb{R}^{r\times k}, r \ll \min(d, k) \quad (1)$$

Figure 1 shows the structure of LoRA model. As can be seen in Formula 1, the pre-trained weight matrix W_0 can be approximated using a low-rank decomposition to represent the parameter update ΔW. During the training process, the parameters of W_0 are frozen, and only the parameters in A and B are trained. For $h = W_0 x$, the forward propagation process is modified as follows:

$$\begin{aligned} h &= W_0 x + \Delta W x \\ &= W_0 x + BAx \end{aligned} \quad (2)$$

During the training process, the low-rank adaptation matrix amplifies the useful features for downstream tasks, enabling the large-scale model to adapt to sentence punctuation tasks in classical texts.

3.3. Post-processing

Accuracy is a crucial aspect in sentence segmentation and punctuation tasks. Due to the greedy sampling approach employed by large models and the variations in input tokenization, the direct output of the model often contains mistakes. For example, there might be instances where the model overlooks a particular character from the original text or introduces an extra character. To address these issues, we have identified and abstracted most of the possible scenarios and implemented a post-processing step for refining the output generated by the Xunzi model.

As can be seen in Algorithm 1, this post-processing step aims to rectify inaccuracies and inconsistencies in the punctuation predictions by considering the specific context and linguistic rules. As a result, we enhance the reliability and coherence of the model's output, ensuring that it aligns with the intended punctuation patterns in practical usage scenarios.

Algorithm 1 Post-process

Input: original sentence s_1 and sentence after punctuation s_2
$s_3 \leftarrow move\_punctuation(s_2)$
if s_1 == s_3 **then**
  return s_2
else
  if $len(s_3)$ == $len(s_1)$ **then**
    $IDs \leftarrow s_3[id] \neq s_1[id]$
    for id in IDs **do**
      $s_3[id] \leftarrow s_1[id]$
    end for
  else
    $b_{s_1} \leftarrow 0, e_{s_1} \leftarrow len(s_1) - 1$
    $b_{s_3} \leftarrow 0, e_{s_3} \leftarrow len(s_3) - 1$
    while $s_1[b_{s_1}] == s_3[b_{s_3}]$ **do**
      $b_{s_1} \leftarrow b_{s_1} + 1$
      $b_{s_3} \leftarrow b_{s_3} + 1$
    end while
    while $s_1[e_{s_1}] == s_3[e_{s_3}]$ **do**
      $e_{s_1} \leftarrow e_{s_1} - 1$
      $e_{s_3} \leftarrow e_{s_3} - 1$
    end while
    $s_3[b_{s_3} : e_{s_3}] \leftarrow s_1[b_{s_1} : e_{s_1}]$
    $s_2 \leftarrow restore\_punction(s_3)$
  end if
end if
Output: s_2

4. Experiments

4.1. Dataset

We used the training dataset released by EvaHan2024, which consists of texts from classical sources. The corpus of ancient Chinese classical texts demonstrates a diachronic nature, encompassing a vast time span of thousands of years and encompassing the four traditional categories of Chinese canonical texts, namely *Jing (经)*, *Shi (史)*, *Zi (子)*, and *Ji (集)*. We conducted a statistical analysis on the occurrence of punctuation marks in the training text, as shown in Table 2. It is worth mentioning that the corner brackets 【 】appeared 53818 times. Since they are not within the scope of sentence segmentation and punctuation in this context, we can treat them as two special Chinese characters.

There are two test datasets. Test A includes approximately 50000 characters of Ancient Chinese texts and comes from different sources. Test B mainly comes from the book *Zuo Zhuan*.

4.2. Metric

Precision (P), Recall (R), and F1 Score are employed as evaluation metrics for all experiments, with the results being expressed in percentages.

Task (Test A)	Seg			Punc		
	Precision	Recall	F1	Precision	Recall	F1
Xunzi-Qianwen-7B-CHAT	90.53	66.12	76.42	73.52	52.22	61.06
ChatGPT 3.5	83.81	59.85	69.83	63.90	43.88	52.03
Our system	**90.80**	**76.34**	**82.94**	**77.75**	**63.85**	**70.12**
Task (Test B)	**Seg**			**Punc**		
	Precision	Recall	F1	Precision	Recall	F1
Xunzi-Qianwen-7B-CHAT	95.28	87.17	91.04	79.25	72.09	75.50
Our system	**95.98**	**90.54**	**93.18**	**85.08**	**79.93**	**82.43**

Table 1: Experimental results on two tests.

Punctuation	Name	Count
，	Comma	1879220
。	Period	954948
、	Slight-pause	126394
：	Colon	163968
；	Semicolon	55256
？	Question	73067
！	Exclamation	45623
“ ”	Double Quotes	240176
‘ ’	Single Quotes	10036
《 》	Book Title Mark	120558

Table 2: Statistics of punctuation marks in the training text

4.3. Implementation Details

We utilized the latest XunziALLM as the base model. A complete three-round training using LoRA was conducted on a device with a Nvidia A100 40G. Each training session, which lasted approximately 20 hours, was conducted on LLaMA Factory (Zheng et al., 2024), an integrated large-scale model training platform. During training, the learning rate is set to $4e-5$, and the LR scheduler is $cosine$. The other important hyperparameters are listed in Table 3.

Hyperparameters	Value
Learning rate	4e-5
LR scheduler	cosine
Warmup steps	200
LoRA rank	8
LoRA Alpha	32
LoRA modules	all

Table 3: Hyperparameters of training

4.4. Results

The results are shown in Table 1. Compared to baselines, our approach achieved comprehensive improvements in sentence segmentation and punctuation tasks. In terms of different test tasks, our method performs relatively well on test B, specifically the *Zuo Zhuan* dataset. Possibly because it has been specifically tailored to understand the stylistic consistency and historical context of classical Chinese texts, with refined preprocessing and post-processing steps that effectively capture the unique linguistic patterns and nuances present in this historical narrative. This validates the effectiveness of the additional layers we designed and demonstrates the advantages of the Xunzi model in processing classical Chinese texts.

5. Conclusion

In this paper, we described the method for tasks in EvaHan2024 using the LoRA approach in the context of ancient Chinese text processing. Our focus has been on sentence segmentation and punctuation. By leveraging the training dataset and building upon Xunzi model, we demonstrated significant improvements over baselines in these tasks. Our experimental results on the test sets, particularly the Zuo Zhuan dataset (test B), validate the effectiveness of our method and showcases its robustness, accuracy, and generalization capabilities.

However, the method may have some limitations. For instance, it relies on a series of predefined rules to correct the model's output, which may not cover all types of errors and may not be flexible enough when dealing with complex or atypical texts.

Automated sentence segmentation and punctuation play a vital role in promoting the study and preservation of ancient books, as well as the inheritance of Chinese culture. With further advancements and refinements in this area, we can contribute to the broader accessibility and understand-

ing of classical Chinese literature for scholars and readers worldwide.

6. References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

John Lafferty, Andrew McCallum, Fernando Pereira, et al. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Icml*, volume 1, page 3. Williamstown, MA.

Vladislav Lialin, Vijeta Deshpande, and Anna Rumshisky. 2023. Scaling down to scale up: A guide to parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.15647*.

Boli Wang, Xiaodong Shi, Zhixing Tan, Yidong Chen, and Weili Wang. 2016. A sentence segmentation method for ancient chinese texts based on nnlm. In *Chinese Lexical Semantics*, pages 387–396, Cham. Springer International Publishing.

Hongbin Wang, Haibing Wei, Jianyi Guo, and Liang Cheng. 2019. Ancient chinese sentence segmentation based on bidirectional lstm+ crf model. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 23(4):719–725.

Zhe Zhang, Jie Liu, Lihua Chi, and Xinhai Chen. 2020. Word-level bert-cnn-rnn model for chinese punctuation restoration. In *2020 IEEE 6th International Conference on Computer and Communications (ICCC)*, pages 1629–1633.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.