Ancient Chinese Sentence Segmentation and Punctuation on Xunzi LLM

Shitu Huo, Wenhui Chen
Beijing Normal University, Xiamen University
Beijing, Fujian, China
saitofok@gmail.com, 793994571@qq.com

Abstract

This paper describes the system submitted for the EvaHan2024 Task on ancient Chinese sentence segmentation and punctuation. Our study utillizes the Xunzi large language model as the base model to evaluate the overall performance and the performance by record type. The applied methodologies and the prompts utilized in our study have shown to be helpful and effective in aiding the model's performance evaluation.

Keywords: Natural Language Processing, Ancient Chinese, Sentence Segmentation, Punctuation

1. Introduction

Throughout history, Chinese civilization has given birth to countless invaluable classics, imbued with rich philosophical thought, historical records, and literary enlightenment. These ancient texts are not only the crystallization of the Chinese nation's precious wisdom but also an integral part of the common heritage of human civilization. However, due to the significant differences between ancient Chinese and modern Chinese in terms of grammar, vocabulary, and semantics, the digitalization and automatic computational understanding of these ancient texts pose tremendous challenges.

In the process of digitizing ancient texts, accurate sentence segmentation and punctuation are crucial steps. Reasonable sentence segmentation can enhance the reading experience and lay the groundwork for subsequent semantic analysis. However, because ancient texts contain a large number of unique grammatical constructions and rhetorical devices, traditional sentence segmentation and punctuation often rely on manual processing by experts, which is time-consuming and laborious. Therefore, developing automated evaluation models and algorithms is an urgent need to improve efficiency and quality.

EvaHan2024 is an international evaluation currently focusing on automatic sentence parsing and punctuation assessment tasks in Classical Chinese. This research proposes to utilize the Xunzi large language model and tailor prompt engineering strategies specifically on sentence segmentation and punctuation for ancient Chinese. As a result, we have a relatively higher performance than baseline with effective prompts.

2. Related Study

2.1 Study on Statistical Machine Learning and Deep Learning Methods for Segmentation and Punctuation for Ancient Chinese

Segmentation mainly divides into rule-based methods and statistical methods. Rule-based methods are typically formulated by experts in ancient Chinese, using common linguistic knowledge to help construct a system for sentence segmentation. For example, segmentation can be based on antonymous compound words, book citation markers, numerals, reduplicated words, and verb-noun structures (Huang & Hou, 2008). However, actual sentence segmentation is very complex, with a word having multiple meanings and combinations, making it impossible to segment based solely on a single word or combination. Rule-based methods cannot cover all situations, leading to scenarios akin to Gödel's incompleteness theorems.

Statistical methods were subsequently widely used. Early experiments could use n-grams (Chen et al., 2007), Conditional Random Fields (Zhang et al., 2009), and the relationship features between adjacency collocation intensities (Xu, 2011) for judgment. Later, scholars increasingly turned to deep learning methods, with BERT being one of the most widely utilized models. BERT (Bidirectional Encoder Representation Transformers) shows excellent performance at language inference and other NLP tasks (Devin et al., 2018). Yu et al. (2019) used BERT for ancient Chinese sentence segmentation research, achieving better results than the BiLSTM+CRF model. Wei (2020) fine-tuned the BERT model, achieving F1 scores of 70.40% for punctuation and 91.67% for segmentation on a large-scale composite corpus. Hu et al. (2021) compared the sequence labeling methods of BERT+FCL, BERT+CRF, and BERT+CNN on the task of ancient Chinese sentence segmentation, finding that BERT+CNN had the best automatic sentence segmentation performance in the three literary forms of poetry, ci, and ancient prose, reaching F1 scores of 99%, 95%, and 92%, respectively. Tang et al. (2023) used a large-scale traditional ancient Chinese corpus to incrementally train the BERT Chinese model, achieving automatic sentence segmentation F1 scores of 95.03% and 99.53% for ancient prose and poetry,

LT4HALA 2024@LREC-COLING-2024, pages 242–245
25 May, 2024.

respectively, and automatic punctuation F1 scores of 80.18% and 98.91%, respectively.

In conclusion, the evolution of methodologies in ancient Chinese sentence segmentation has shown a clear trajectory from rule-based approaches towards the adoption of deep learning techniques.

2.2 Study on Large Language Models for Ancient Chinese

With the successful implementation of scaling laws on large language models, these models have been able to grasp the deep semantics and grammatical rules of languages. Several studies have recently focused on evaluating the capabilities of large language models (LLMs) in comprehending ancient languages, with a particular emphasis on ancient Chinese. One notable contribution in this area is the work by Zhang and Li (2023), who introduced ACLUE, an evaluation benchmark designed specifically to assess LLMs' language abilities in relation to ancient Chinese. ACLUE comprises 15 tasks covering various linguistic skills, including phonetic, lexical, syntactic, semantic, inference, and knowledge. Notably, ChatGLM2 exhibited the highest performance level among the evaluated models, achieving an average accuracy of 37.45%.

Currently, the existing ancient Chinese large models include AI Jiusi, AI Taiyan, and the Xunzi model, which are mainly based on existing pre-trained models and fine-tuned on ancient Chinese datasets.

AI Jiusi is a large model fine-tuned by Huazhong University of Science and Technology based on the Alibaba Cloud Tongyi Qianwen as the base model[1].

AI Taiyan is a large language model specifically designed for understanding Classical Chinese texts, developed by the Digital Humanities Department at Beijing Normal University[2].

The Xunzi large language model includes versions fine-tuned on Qwen-7B, GLM-6B, Baichuan-7B for Classical Chinese[3]. In summary, the above models have not yet undergone comprehensive evaluation on segmentation and punctuation benchmarks, necessitating further exploration.

3. Employed Model

Qwen-7B is a large language model based on the Transformer architecture, trained on an extensive pre-training dataset (Bai et al., 2023).

Xunzi large model is fully fine-tuned based on Qwen-7B. The training of this model utilized the Zero2 technology in the DeepSpeed framework for memory optimization, distributing the model's state parameters and gradients across 8 A800 model GPUs. The fine-tuning dataset comprised approximately 5GB of ancient text corpora mixed with modern Chinese texts, command data, and other types of corpora, thus creating a mixed dataset containing 4 billion Chinese characters.

This study employs the Xunzi large model and deploy it to a server and conduct large-scale evaluations on various text datasets to assess its practical utility and scalability.

4. Experiment

4.1 Experimental Environment

The NVIDIA card is configured in Table 1:

CUDA Version	GPU	Memory
12.1	NVIDIA GeForce RTX 4090	24GB

Table 1: The Nvidia Info

4.2 Prompt Engineering

This study explores how carefully designed prompts can guide Xunzi model to generate more accurate and rich text content. The method employed in this study adopts a two-round prompt design strategy, aimed at refining and optimizing the final output through the text generated initially.

In the first round, we designed an initial prompt: "Please think boldly, and as diversely and richly as possible, punctuate the following text, and respond in traditional characters: {text}." The goal of this prompt is to guide the model to process a complex sentence, making the meaning of the text clearer by adding appropriate punctuation, while maintaining the diversity and richness of sentence structure. In this stage, the model was run five times, generating five different punctuated results.

Following this, in the second round, another prompt was adopted: "Please consider and integrate the optimal sentence breaking scheme from the following five sentences: {response}." This step requires the model to select and integrate the best sentence breaking scheme from the five punctuated sentences generated in the first round[4]. In the end, we collect all the best sentences from the test set.

1 https://mp.weixin.qq.com/s/c-NeKg4z4dMgBSFUbYDtbg
2 https://mp.weixin.qq.com/s/Cp5NOSOcjvBt9qzcVZ9igQ
3 https://github.com/Xunzi-LLM-of-Chinese-classics/XunziALLM
4 The actual prompt is in Chinese. The first prompt is: "请大胆思考，尽可能多样、丰富地给下面的文本打上标点符号，请使用繁体字回答：{text}." The second prompt is : "请从下列五个句子中，思考并整合出最优的断句结果：{response}."

4.3 Temperature Setting

We systematically varied the parameter known as 'temperature', a scaling factor applied to the logits of the model before sampling. The temperature setting is shown in Table 2. During the initial round of interaction, the temperature was set to 0.95, promoting a diverse and creative output by allowing for a broader probability distribution of potential responses. Subsequently, in the second round, the temperature was reduced substantially to 0.1, significantly narrowing the scope of variability in the model's output. This reduction in temperature typically yields more deterministic and possibly repetitive results, given the higher likelihood of sampling the most probable outcomes. This methodological adjustment of the temperature parameter is critical in fine-tuning the model's performance to align with the desired level of creativity and variability in the generated content.

Temperature	Value
first round	0.95
second round	0.1

Table 2: The Parameter of the Model

5. Results

5.1 Overall Performance

Test sets include Test A and Test B. Test A refers to the data released the first time while Test B refers to the Zuozhuan data released the second time.

The performance metrics in Test A presented in Table 3 underscore the relative strengths and weaknesses of different models in handling segmentation and punctuation tasks for text analysis. The segmentation task, as evident from the data, benefits from higher accuracy across all models when compared to punctuation.

Our prompt engineering method based on Xunzi-Qwen-7B model outstrips its predecessors, achieving a precision of 90.70% in segmentation, indicating exceptional reliability in predicting segment boundaries. However, a recall of 71.54% suggests room for improvement in identifying all true segment boundaries. The F1-score, at 79.99%, represents a favorable balance between precision and recall, underscoring a robust segmentation model. In comparison, GPT-3.5 and Xunzi-Qwen-7B demonstrate precision rates of 83.81% and 90.53%, respectively, with the latter nearly matching our model. However, both models fall short in recall, and consequently, F1-scores, with GPT-3.5 at 59.85% and 69.83%, and Xunzi-Qwen-7B at 66.12% and 76.42%, respectively.

For the punctuation task, the results indicate more challenges across the board. Our method achieves a precision of 73.63%, suggesting a correct prediction in approximately three out of four instances. Yet, the recall of 56.86% reveals that the model fails to detect a significant number of true punctuation marks. This is reflected in the F1-score, which at 64.17%, points to moderate overall effectiveness. GPT-3.5's punctuation capability is weaker still, with precision and recall scores of 63.90% and 43.88%, respectively, and an F1-score of 52.03%. Xunzi-Qwen-7B presents comparable results to our model in precision at 73.52% but lags in recall at 52.22%, culminating in an F1-score of 61.06%.

Method	Task	Precision	Recall	F1-score
Ours	Seg	**90.70%**	**71.54%**	**79.99%**
GPT-3.5		83.81%	59.85%	69.83%
Baseline (Xunzi-Qianwen-7B-CHAT)		90.53%	66.12%	76.42%
Ours	Punc	**73.63%**	**56.86%**	**64.17%**
GPT-3.5		63.90%	43.88%	52.03%
Baseline (Xunzi-Qianwen-7B-CHAT)		73.52%	52.22%	61.06%

Table 3: Experiment Results on Test A of Ours(our prompt engineering methods on Xunzi-Qwen-7B), GPT-3.5, Baseline (Xunzi-Qianwen-7B-CHAT).

The performance metrics in Test B, as shown in Table 4, highlight our method's competitiveness against the baseline. Our segmentation model achieved a precision slightly lower than the baseline but exhibited higher recall, indicating a trade-off between precision and recall. Similarly, for the punctuation task, our model demonstrated a balanced trade-off between precision and recall compared to the baseline, suggesting comparable performance between the two models.

Method	Task	Precision	Recall	F1-score
Ours	Seg	95.25%	**88.15%**	**91.57%**
Baseline (Xunzi-Qianwen-7B-CHAT)		**95.28%**	87.17%	91.04%
Ours	Punc	79.06%	**73.66%**	**76.26%**
Baseline (Xunzi-Qianwen-7B-CHAT)		**79.25%**	72.09%	75.50%

Table 4: Experiment Results on Test B of Ours(our prompt engineering methods on Xunzi-Qwen-7B) and Baseline (Xunzi-Qianwen-7B-CHAT).

5.2 Performance by Specific Record Type

The results further show the model's performance on four different types of records (Table 5): Products in Local Products in Local Chronicles(方志物产), County Annals(县志), Buddhist Sutra(佛经), and Academy Records(书院志).

For Products in Local Chronicles, the model achieved a high segmentation precision of 94.78% and a moderate recall of 69.81%, demonstrating high reliability in detecting segments, yet missing some. The punctuation precision was decent at 78.09%, outperforming the recall at 55.22%.

County Annals saw slightly lower segmentation precision but a higher recall, indicating a more balanced performance, and also led the records with the highest punctuation F1-score.

However, Buddhist Sutra presented considerable challenges, with the lowest performance metrics including a segmentation recall of just 46.72%, suggesting the model frequently missed segment points, and the punctuation F1-score fell to 47.23%.

Lastly, Academy Records achieved relatively high segmentation scores and better punctuation performances, although still not surpassing the punctuation results of County Annals. This analysis indicates that while the model shows competency in segmentation, its performance in punctuation is less consistent and requires targeted improvements, particularly within the more complex texts like Buddhist Scriptures.

		Precision	Recall	F1-score
Products in Local Chronicles	Seg	94.78%	69.81%	80.4%
	Punc	78.09%	55.22%	64.7%
County Annals	Seg	89.61%	81.32%	**85.26%**
	Punc	72.01%	63.66%	67.58%
Buddhist Sutra	Seg	89.02%	46.72%	61.28%
	Punc	68.92%	35.92%	47.23%
Academy Records	Seg	90.78%	78.78%	84.36%
	Punc	77.24%	67.18%	71.86%

Table 5: Our Method's Performance by Record Type

6. Conclusions

Our method is generally more effective at segmenting than punctuating, indicating the need for further training or a different approach for punctuation.

The notably lower performance on Buddhist Scriptures could be due to various factors such as language complexity, formatting, or the presence of Sanskrit or Pali words. Tailored solutions, like adding more scriptural training data or using a specialized tokenization approach.

Improving punctuation accuracy through contextual understanding integration could significantly enhance the model's performance, particularly in ancient texts. Thorough error analysis can uncover specific challenges, while targeted improvements address discrepancies, enhancing consistency and accuracy.

7. References

Bai, J., Bai, S., Chu, Y., et al. (2023). Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Chen, T. Y., Chen, R., Pan, L. L., et al. (2007). Archaic Chinese punctuating sentences based on context n-gram Model. *Computer Engineering, 33*(03), 192-193.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

Hu, R. F., Li, S., & Zhu, Y. C. (2021). Knowledge representation and sentences segmentation of ancient Chinese based on deep language models. *Journal of Chinese Information Processing*, 35(04), 8-15.

Huang, J. N., & Hou, H. Q. (2008). On sentence segmentation and punctuation model for ancient books on agriculture. *Journal of Chinese Information Processing, 22*(04), 31-38.

Tang, X. M., Su, Q., Wang, J., et al. (2023). Automatic traditional ancient Chinese texts segmentation and punctuation based on pre-trained language model. *Journal of Chinese Information Processing, 37*(08), 159-168.

Wei, Y. (2020). *Research on Automatic Texts Segmentation and Word Segmentation For Ancient Chinese Texts.* Beijing: Master Dissertation, Peking University.

Xu, J. Y. (2011). *Research on Automatic Sentence Reading of Ancient Chinese Texts*. Beijing: Ph.D. Dissertation, Peking University.

Yu, J. S., Wei, Y., & Zhang, Y. W. (2019). Automatic ancient Chinese texts segmentation based on BERT. *Journal of Chinese Information Processing, 33*(11), 57-63.

Zhang, K. X., Xia, Y. Q., & Yu, H. (2009). CRF-based approach to sentence segmentation and punctuation for ancient Chinese prose. *Journal of Tsinghua University (Science and Technology), 49*(10), 1733-1736.

Zhang, Y., & Li, H. (2023). Can Large Language Model Comprehend Ancient Chinese? A Preliminary Test on ACLUE. In *Proceedings of the Ancient Language Processing Workshop* (pp. 80–87). Varna, Bulgaria: INCOMA Ltd., Shoumen, Bulgaria.