Sentence Segmentation and Punctuation for Ancient Books based on Supervised In-context Training

Shiquan Wang, Weiwei Fu, Mengxiang Li, Zhongjiang He,
Yongxiang Li, Ruiyu Fang, Li Guan, Shuangyong Song
China Telecom Corporation Ltd. AI Technology Company
{wangsq23, fuweiwei, hezj, liyx25, guanl, fangry, songshy}@chinatelecom.cn
limengx@126.com

Abstract

This paper describes the participation of team "TeleAI" in the third International ancient chinese Language Information Processing Evaluation (EvaHan2024). The competition comprises a joint task of sentence segmentation and punctuation, categorized into open and closed tracks based on the models and data used. In the final evaluation, our system achieved significantly better results than the baseline. Specifically, in the closed-track sentence segmentation task, we obtained an F1 score of 0.8885, while in the sentence punctuation task, we achieved an F1 score of 0.7129.

Keywords: sentence segmentation, sentence punctuation, in-context learning

1. Introduction

The tasks of sentence segmentation and punctuation in Chinese ancient texts are significant challenges and hold great importance in the field of Natural Language Processing (NLP). In ancient chinese texts, sentences are often written without explicit punctuation, making it difficult for modern readers and NLP systems to accurately interpret the text's structure and meaning. Sentence segmentation involves identifying boundaries between sentences, which is crucial for tasks such as text comprehension, information extraction, and machine translation. Furthermore, restoring missing or ambiguous punctuation marks is essential for improving the readability and understanding of ancient chinese texts. These tasks not only contribute to the preservation and analysis of historical texts but also serve as fundamental building blocks for various NLP applications, including language understanding, generation, and translation. Thus, addressing the challenges of sentence segmentation and punctuation restoration in ancient chinese is essential for advancing research in NLP and facilitating cross-temporal communication and understanding.

Our submitted system adopts a two-stage strategy to improve the performance of the model on ancient chinese sentence segmentation and punctuation restoration tasks. In the first stage, we enhance the performance of the XunziALLM base model through Supervised In-context Training. In the second stage, we improve the model's accuracy in discerning unreliable punctuation by employing a greedy character correction approach and a voting strategy. Our final submission achieved an F1 score of 0.8885 in the sentence segmentation task and 0.7129 in the joint task of sentence segmentation and punctuation.

2. Related Work

The absence of punctuation and sentence breaks in ancient chinese has been a longstanding cultural convention. However, the lack of sentence breaks poses a challenge for modern individuals in learning and utilizing ancient chinese. Manual sentence segmentation requires a clear understanding of semantics, grammar, rhythm, and indicative words, and consumes a significant amount of time and effort. To better understand and study the grammatical structure, logical relationships, and expression methods of ancient texts, as well as to grasp the semantics and rhythm of sentences, a large number of researchers are exploring the joint task of automatic sentence segmentation and punctuation restoration in ancient texts using natural language processing techniques.

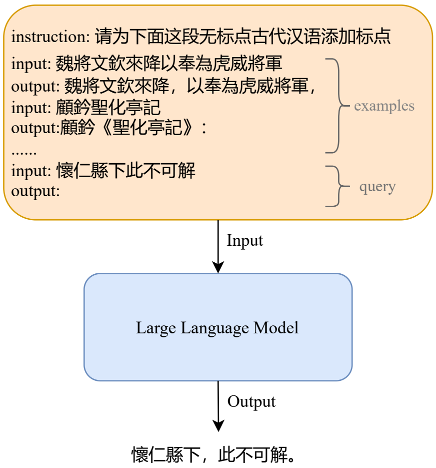
Early punctuation restoration tasks were predominantly based on rules along with LSTM, CNN, and other deep learning models. Tilk and Alumäe (2015) propose an LSTM-based punctuation restoration approach, first learning text features on a large text corpus, and then utilizing text features and prosodic features to predict punctuation marks on a small-scale corpus. Che et al. (2016) initially transform the text into a long word sequence, treating the punctuation restoration task as a sequence classification problem, and utilize Deep Neural Networks (DNN), Convolutional Neural Networks (CNN-A), and Double-layer Convolutional Neural Networks (CNN-2A) to predict punctuation marks. Cheng and Li (2020) proposed a method based on BiLSTM+CRF to achieve joint annotation of sentence segmentation and lexical analysis in Classical Chinese. They validated the effectiveness of this approach on four different test sets from different periods. Kim (2019) introduced a recurrent neural network model based on hierarchical multi-head attention. This model employs hierarchical attention to allow each layer to learn different contexts from various perspectives.

With the advancement of deep learning technologies, punctuation restoration methods based on transformers and pretrained language models have achieved significant success. Wang et al. (2018) framed punctuation restoration as a translation task, where the model takes unpunctuated sequences as input and produces sequences of punctuation marks and labels as output. This approach leverages Transformer networks based on self-attention mechanisms to extract hidden features. Wang et al. (2022) utilized validated high-quality corpora of the entire texts from the "Siku Quanshu" as the training set to construct SikuBERT and SikuRoBERTa for ancient chinese intelligent processing tasks. They validated the performance of these models across multiple ancient chinese tasks.

With the remarkable achievements of large language models (LLMs) in various fields of natural language processing, there has been a growing emphasis on integrating LLMs with classical literature processing to advance intelligent research on ancient texts. In this context, Nanjing Agricultural University has introduced the XunziALLM, aiming to facilitate the intelligent processing of classical texts. XunziALLM has demonstrated significant potential across multiple downstream tasks related to ancient texts.

3. Method

3.1. Supervised In-context Training



instruction: 请为下面这段无标点古代汉语添加标点
input: 魏將文欽來降以奉為虎威將軍
output: 魏將文欽來降，以奉為虎威將軍，
input: 顧鈐聖化亭記
output:顧鈐《聖化亭記》：
......
input: 懷仁縣下此不可解
output:

懷仁縣下，此不可解。

Figure 1: Illustration of in-context learning

With the scaling of model size and corpus size, large language models (LLMs) demonstrate an in-context learning (ICL) ability, wherein they learn from a few examples in the context. Numerous studies have indicated that LLMs can effectively perform a variety of complex tasks through ICL. Figure 1 provides an illustrative example depicting how language models make decisions using ICL. In essence, the model estimates the likelihood of potential answers conditioned on the demonstration, leveraging a well-trained language model.

Formally, given a query input text x and a set of candidate answers $Y = \{y_1, \ldots, y_m\}$, a pretrained language model M selects the candidate answer with the maximum score as the prediction, conditioning on a demonstration set C. The set C comprises an optional task instruction I and K demonstration examples, thus $C = \{I, s(x_1, y_1), \ldots, s(x_k, y_k)\}$ or $C = \{s(x_1, y_1), \ldots, s(x_k, y_k)\}$, where $s(x_k, y_k, I)$ represents an example written in natural language texts according to the task. The likelihood of a candidate answer y_j can be represented by a scoring function f of the entire input sequence with the model M :

$$P(y_j|x) \triangleq f_M(y_j, C, x) \tag{1}$$

The final predicted label $\hat{y}$ is the candidate answer with the highest probability:

$$y = \arg\max_{y_j \in Y} P(y_j|x) \tag{2}$$

While LLMs have demonstrated promising ICL capability, several studies also suggest that this capability can be further enhanced through a continual training stage between pretraining and ICL inference (Wei et al., 2023; Chen et al., 2022). Therefore, we enhance the ICL capability of LLMs by constructing context-learning instruction training data and eliminating the gap between pretraining tasks and downstream ICL tasks through supervised instruction fine-tuning. Specifically, we utilize a differential selection method to choose example data in ICL and construct supervised ICL training data, followed by training XunziALLM based on the supervised ICL data.

3.2. Character Correction and Voting Strategy

During the inference process, we observed discrepancies between some predictions of the Large Language Model (LLM) and the input text at the character level. To ensure consistency between the model's predictions and the original text, we propose a greedy character correction algorithm, as shown in Algorithm 1. This algorithm sequentially examines the characters in the predicted and original texts. If a character is a punctuation mark, it is directly appended to the result string. Otherwise, each character in the original and predicted texts is compared, and corresponding operations, such

Algorithm 1 Greedy Character Correction Algorithm

```
Require: Original text $original\_text$, Predicted text $predicted\_text$
Ensure: Text with restored punctuation $res$
Initialize empty string $res$
Initialize indices $i \leftarrow 0, j \leftarrow 0, max\_try \leftarrow 0$
while $i < $ length$(original\_text)$ and $j <$ length$(predicted\_text)$ do
    $max\_try \leftarrow max\_try + 1$
    if $max\_try > 100000$ then
        break
    end if
    if $predicted\_text[j]$ is a punctuation mark then
        Append $predicted\_text[j]$ to $res$
        $j \leftarrow j + 1$
        continue
    end if
    if $original\_text[i] = predicted\_text[j]$ then
        Append $original\_text[i]$ to $res$
        $i \leftarrow i + 1$
        $j \leftarrow j + 1$
        continue
    end if
    $k \leftarrow j$
    while $predicted\_text[k+1]$ is a punctuation mark and $(k+1) <$ length$(predicted\_text)$ do
        $k \leftarrow k + 1$
    end while
    if $original\_text[i+1] = predicted\_text[k+1]$ then
        Append $original\_text[i]$ to $res$
        $i \leftarrow i + 1$
        $j \leftarrow j + 1$
        continue
    end if
    if $original\_text[i+1] = predicted\_text[j]$ then
        Append $original\_text[i]$ to $res$
        $i \leftarrow i + 1$
        continue
    end if
    if $original\_text[i] = predicted\_text[k+1]$ then
        $j \leftarrow j + 1$
        continue
    end if
end while
if $j <$ length$(predicted\_text)$ then
    Append remaining characters of $predicted\_text$ to $res$
end if
return $res$
```

as replacement, deletion, or addition of characters, are performed based on their equality or inequality.

Simultaneously, we observed discrepancies in the predictions of different models for the same input, reflecting variations in the models' confidence levels regarding candidate entities. To leverage the advantages of different models, mitigate the limitations of individual models, and enhance overall predictive performance, we initially retained predictions from multiple models across different iterations. Subsequently, we employed a voting method to obtain the final prediction result.

4. Experiments

This section will introduce the experimental aspects involved in our participation in this evaluation task, primarily encompassing three parts: data preprocessing, experimental parameter settings, and experimental results and analysis.

4.1. Data Preprocessing

The EvaHan2024 dataset comprises texts sourced from classical literature, especially the Siku Quanshu (Four Treasuries) and other historical texts. Constructed through initial label predictions by models and subsequent human expert corrections, the original training set consists of 254,360 data points, with 412 data points in the test set. Through rule-based filtering, we selected 126,372 high-quality data points from the training set, which were then redivided into training and validation sets in a 9:1 ratio. All subsequent comparative experiments were conducted based on this redivided training and validation set.

The evaluation in this assessment involves two tasks: sentence segmentation and sentence Punctuation. Sentence segmentation is the process of converting Chinese text into a sequence of sentences, with each sentence separated by a single space. Additionally, sentence punctuation involves correctly placing punctuation marks at the end of each sentence. In many Chinese language processing systems, these two tasks, sentence segmentation and punctuation, are typically addressed together. Therefore, we developed a set of evaluation scripts for offline assessment of this joint task, calculating precision, recall, and F1 scores. Based on the offline evaluation results, we selected the optimal outcome as the final submission version. The scores on the validation sets in subsections 4.2 and 4.3 are all computed based on this offline evaluation script.

4.2. Experiment1: Supervised Fine-tuning

In pursuit of identifying the most suitable XunziALLM base model for handling sentence segmentation and punctuation restoration tasks in Classical Chinese, we conducted experiments on the re-partitioned EvaHan2024 dataset. The experimental findings are presented below.

Table 1 showcases the performance of various XunziALLM base models on the EvaHan2024 dataset. Precision, Recall, and F1 metrics denote

Model	Precision	Recall	F1	ER
Xunzi-Qwen	0.7517	0.6961	0.7228	0.19
Xunzi-Qwen-CHAT	0.7573	0.7079	0.7318	0.32
Xunzi-GLM	0.7548	0.7365	0.7455	0.06
Xunzi-Baichuan	**0.7759**	**0.7588**	**0.7672**	**0.04**

Table 1: Experimental Results of XunziALLM Models on the EvaHan2024 Dataset

the highest scores achieved by different XunziALLM base models in the joint task of sentence segmentation and punctuation restoration, while ER represents the proportion of character inconsistencies between model predictions and the original texts. The experimental results reveal that Xunzi-Baichuan attained the highest F1 score on the EvaHan2024 dataset, accompanied by the lowest proportion of character discrepancies between predicted results and the original texts. Consequently, for subsequent experiments, we elected to utilize this model as the primary base model.

4.3. Experiment2: Supervised In-context Training

While pre-trained language models have demonstrated initial capabilities in In Context Learning (ICL), there remains a certain gap between their pretrained objectives and downstream ICL tasks. To fully harness the potential of XunziALLM in context learning, we employed a differential selection approach to curate sample data suitable for ICL, thereby constructing a supervised ICL training dataset. Subsequently, we trained XunziALLM based on this supervised ICL dataset. The experimental results are shown in Table 2.

Model	Precision	Recall	F1	ER
Xunzi-Qwen	0.7640	0.7250	0.7465	0.09
Xunzi-Qwen-CHAT	0.7687	0.7399	0.7540	0.35
Xunzi-GLM	0.7862	0.7639	0.7749	**0.04**
Xunzi-Baichuan	**0.8013**	**0.7892**	**0.7952**	**0.04**

Table 2: Experimental Results of XunziALLM Models on the EvaHan2024$_{\mathrm{ICL}}$ Dataset

Table 2 illustrates the performance of XunziALLM on the EvaHan2024 dataset after supervised ICL training. The results indicate that supervised fine-tuning with ICL supervision enhances the ability of LLMs to learn from context during inference, thereby improving XunziALLM's performance on sentence segmentation and punctuation restoration tasks in classical chinese. Simultaneously, it can be observed that supervised in-context training outperforms direct supervised fine-tuning.

4.4. Experiment3: Online Submission

Table 3 presents the experimental results of our system compared with the baseline model (Xunzi-Qwen-7B-Chat) on Test Set A, while Table 4 presents the experimental results of our system compared with the baseline model on Test Set B. In the joint task of sentence segmentation and punctuation, our system achieved a relative improvement of 16.75% on Test Set A and 9.13% on Test Set B compared to the baseline model. These experimental results demonstrate the effectiveness of our proposed method, indicating that supervised in-context training can enhance the performance of models in sentence segmentation and punctuation tasks for ancient texts.

Model	Task	Precision	Recall	F1
Xunzi-Qwen-7B-Chat	Seg	0.9053	0.6612	0.7642
	Punc	0.7352	0.5222	0.6106
Our System	Seg	**0.9170**	**0.8671**	**0.8885**
	Punc	**0.7433**	**0.6848**	**0.7129**

Table 3: Experimental Results on the Test Set A Compared with Baseline Model.

Model	Task	Precision	Recall	F1
Xunzi-Qwen-7B-Chat	Seg	0.9528	0.8717	0.79104
	Punc	0.7925	0.7209	0.7550
Our System	Seg	**0.9632**	**0.9146**	**0.9383**
	Punc	**0.8599**	**0.7910**	**0.8240**

Table 4: Experimental Results on the Test Set B Compared with Baseline Model.

5. Conclusion

In this paper, we describe our submission system for the EvaHan2024 shared task. We present our solution in two stages: (a) Supervised In-context Training and (b) Character Correction and Voting. In the final evaluation, our system achieved outstanding results in the closed track, with a final F1 score of 0.7129 on Test Set A and 0.8240 on Test Set B.

6. References

Xiaoyin Che, Cheng Wang, Haojin Yang, and Christoph Meinel. 2016. Punctuation prediction for unsegmented transcript based on word vector. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 654–658.

Mingda Chen, Jingfei Du, Ramakanth Pasunuru, Todor Mihaylov, Srini Iyer, Veselin Stoyanov, and Zornitsa Kozareva. 2022. Improving in-context few-shot learning via self-supervised training. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3558–3573.

Qian Chen, Wen Wang, Mengzhe Chen, and Qinglin Zhang. 2021. Discriminative self-training for punctuation prediction. *arXiv preprint arXiv:2104.10339*.

Ning Cheng and Bin Li. 2020. A joint model of automatic sentence segmentation and lexical analysis for ancient chinese based on bilstm-crf model.

Maury Courtland, Adam Faulkner, and Gayle McElvain. 2020. Efficient automatic punctuation restoration using bidirectional transformers with robust inference. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 272–279.

Seokhwan Kim. 2019. Deep recurrent neural networks with layer-wise multi-head attentions for punctuation restoration. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7280–7284. IEEE.

Ottokar Tilk and Tanel Alumäe. 2015. Lstm for punctuation restoration in speech transcripts. In *Interspeech*, pages 683–687.

Dongbo Wang, Chang Liu, Zihe Zhu, Jiangfeng Liu, Haotian Hu, Si Shen, and Bin Li. 2022. Construction and application of pre-trained models of siku quanshu in orientation to digital humanities. *Library tribune*, 42(06):31–43.

Feng Wang, Wei Chen, Zhen Yang, and Bo Xu. 2018. Self-attention based network for punctuation restoration. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2803–2808. IEEE.

Jerry Wei, Le Hou, Andrew Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma, et al. 2023. Symbol tuning improves in-context learning in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 968–979.