System Report for CCL24-Eval Task 5: Multi-Model Classical Chinese Event Trigger Word Recognition Driven by Incremental Pre-training

Litao Lin[1], Mengcheng Wu[2], Xueying Shen[1], Jiaxin Zhou[1], Shiyan Ou[1,*],

School of Information Management, Nanjing University[1],

College of Information Management, Nanjing Agricultural University[2],

litaolin@smail.nju.edu.cn, wmc@stu.njau.edu.cn, sxy_77@smail.nju.edu.cn,

522023140121@smail.nju.edu.cn, oushiyan@nju.edu.cn

Abstract

This paper addresses the task of identifying and classifying historical event trigger words in Classical Chinese, utilizing both small-scale and large-scale language models. Specifically, we selected the small-scale language model GujiBERT for intelligent processing of classical texts, and the large-scale language model Xunzi-Qwen-14b. Both models underwent continued pre-training and fine-tuning, resulting in GujiBERT-CHED-mlm and Xunzi-Qwen-14b-CHED. For the small-scale language model, we used a BiLSTM as the feature extraction module and a CRF as the decoding module, employing a sequence labeling paradigm to complete the evaluation experiments. For the large-scale language model, we optimized the prompt templates and used a sequence-to-sequence paradigm for evaluation experiments. Our experiments revealed that GujiBERT-BiLSTM-CRF achieved the best performance across all tasks, ranking fourth in overall performance among all participating teams. The large-scale language model demonstrated good semantic understanding abilities, reaching a preliminary usable level. Future research should focus on enhancing its ability to produce standardized outputs.

1 Introduction

Ancient texts contain rich information on historical events, figures, geographical locations, and more, which are of significant value for historical research, cultural heritage preservation, and exploration of historical patterns. Events, as the smallest granular units describing historical knowledge, are critical for the detection and organization of information in ancient texts. This facilitates the enhancement of knowledge services for ancient literature from a digital humanities perspective. Due to the diverse content forms and complex language structures of ancient texts, information extraction typically relies on expert manual annotation and analysis, which is time-consuming and labor-intensive, making it difficult to scale for large volumes of ancient texts. In recent years, the rapid development of natural language processing (NLP), particularly the application of deep learning methods, has provided effective solutions for the automatic detection and identification of events in ancient texts.

The objective of this task is to evaluate and further improve algorithmic models for detecting historical events in ancient texts. Pre-training and fine-tuning are conducted on a prepared dataset of ancient

* Corresponding Author

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 172-184, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations

historical texts to develop an optimal performance model. This model should possess two key capabilities: first, accurately identifying the corresponding trigger words and their locations, meaning it should precisely pinpoint the words that best represent the occurrence of a historical event in ancient texts and clearly mark their positions within the text; second, correctly determining the event type associated with each trigger word according to the CCL-CHED official classification system for ancient text events, ensuring the accuracy and consistency of detection results.

Based on an extensive survey of research on Chinese ancient text event information extraction, we employed both natural language understanding models and natural language generation models to complete this evaluation task. Specifically, trigger word recognition and classification were treated as a sequence labeling task, with a domain-specific BERT model selected for continued pre-training and fine-tuning to construct the trigger word recognition and classification model. Additionally, to explore the applicability of large language models to this task, a base model tailored for intelligent processing of ancient Chinese texts was selected for continued pre-training and fine-tuning. Comparative testing was conducted to select the optimal prompts, enabling the large language model to generate character indices and trigger word categories for the given text. Figure 1 illustrates the overall technical approach employed in this study. Through a series of experiments, we discovered that smaller language models, exemplified by BERT, outperform in terms of prediction accuracy and the standardization of output. Although large language models have advantages in terms of ease of use, the standardization of their output is challenging to control, necessitating further manual processing. Overall, large language models are less suitable for tasks that require high precision in information extraction.

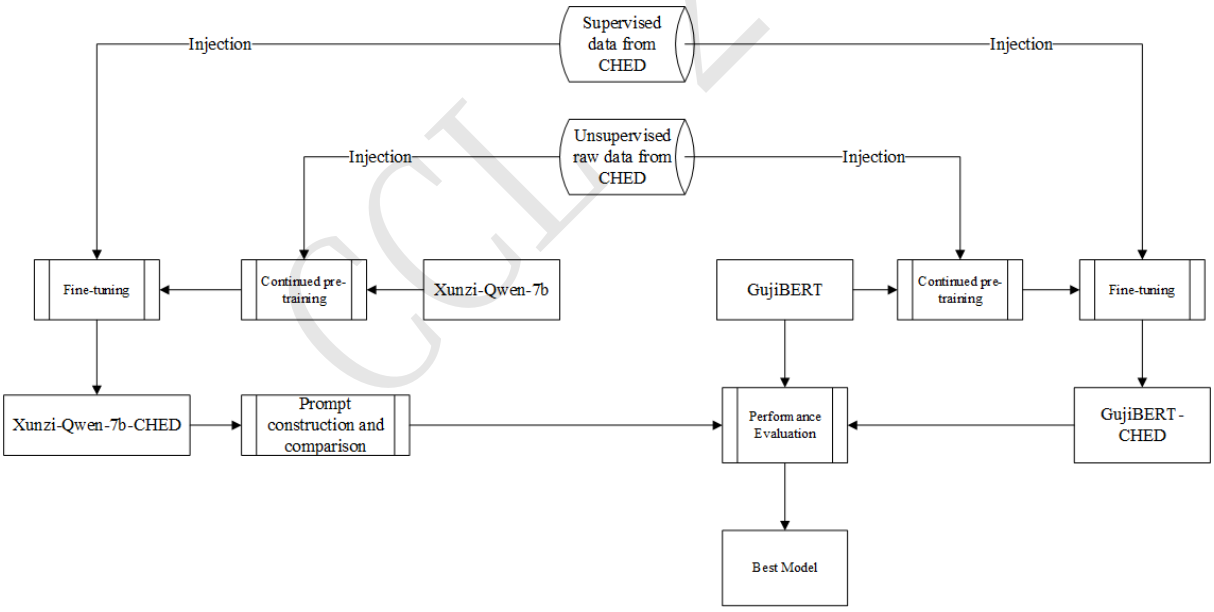

Figure 1: Technical Framework

2 Related Work

Currently, there are three main methods for event extraction from ancient Chinese texts: pattern-matching-based event extraction, statistical machine learning-based event extraction, and deep learning-

based event extraction. The pattern-matching method relies on expert-specified rules to parse sentences, identify trigger words, and categorize event types. The machine learning-based method converts the text into feature vectors and uses classifiers (such as support vector machines and maximum entropy models) for event identification and classification. For instance, Zhangchao Li et al.(2020) studied war events in the “Zuo Zhuan” (Chronicles of Zuo), exploring a combination of rules and machine learning. They used pattern matching to identify war sentences and then employed Conditional Random Fields (CRF) combined with feature models to identify and extract elements of war events. However, traditional statistical model-based methods rely heavily on manual feature engineering, making it difficult to fully extract deep semantic information from the text. With the significant advancements in natural language processing achieved through deep learning, constructing neural network models (such as CNN, RNN, LSTM, etc.) has enabled the automatic learning and extraction of textual features, substantially improving the accuracy and generalization capability of event extraction. Deep learning-based methods include various technical paradigms such as “sequence labeling”, “reading comprehension”, and “sequence-to-sequence”. Additionally, highly domain-specific pre-trained models are used to enhance the learning effectiveness on the text. Representative research based on the sequence labeling paradigm includes a study by Ma et al.(2021), they explored the multi-class automatic classification of canonical trigger words and event sentences using a Bi-LSTM model, achieving an accuracy of 95%. In their research focusing on “Records of the Grand Historian”, Zhongbao Liu et al.(2020) utilized a BERT-LSTM-CRF model to extract historical events from the corpus, achieving an F1 score of 82.3%. In a study by Xuehan Yu et al.(2021) focusing on war sentences from the “Zuo Zhuan” corpus, they compared the performance of five models: GuwenBERT-LSTM, BERT-LSTM, RoBERTa-LSTM, BERT-CRF, and RoBERTa-CRF, for event extraction. Among these models, RoBERTa-CRF achieved an F1 score of 82.1%. research based on the reading comprehension paradigm, such as that by Yu Xuehan et al.(2023), they integrated machine reading comprehension (MRC) into the neural network architecture to achieve event extraction by setting questions. Among these, RoBERTa-MRC_AC and RoBERTa-MRC_MC performed well in event type extraction, achieving F1 scores of 88.2% and 89.2%, respectively, significantly improving the performance. The sequence-to-sequence paradigm involves inputting the sentence to be processed into a generative model, which directly outputs the final result without the need for further manual processing. Representative research includes the study by Zhang et al.(2023) who proposed a generative approach using a knowledge graph-based event generation framework to extract war events from “Records of the Grand Historian”. The trigger word recognition achieved an effectiveness of 71.3%. Wang Y. et al.(2023) used “Records of the Three Kingdoms” as their research corpus and compared sequence labeling models with generative models. They explored the effectiveness of the BERT-BiLSTM-CRF fine-tuned BBCN-SG sequence model and the Stacking-TRN-SG model constructed by integrating three models—T5-SG, RoBERTa-SG, and NEZHA-SG—fine-tuned based on the T5 model, using stacking. The Stacking-TRN-SG model achieved a recall rate of 70.35% in event extraction.

Overall, pattern matching can achieve good results for specific types of texts, but its transferability is poor. Statistical machine learning methods heavily depend on the richness of training samples and perform poorly in parsing sentences that did not appear in the training corpus. Currently, deep learning methods have become mainstream. Among these, models based on the BERT architecture have achieved

good results under the sequence labeling paradigm. With the rise of transfer learning techniques, methods based on reading comprehension and generative models have also been applied to event information extraction from ancient texts, providing new perspectives for event extraction research. In addition to changing task paradigms and designing, comparing, and analyzing different architectures of deep learning models, many scholars also aim to provide the most basic and foundational optimizations for various natural language processing tasks by constructing domain-specific pre-trained language models (Dongbo Wang et al., 2022). By the end of 2022, the emergence of ChatGPT once again revolutionized the paradigm of natural language processing tasks. Subsequently, large language models continued to innovate, leading to the creation of models tailored for different vertical domains. This includes the "Xunzi" series of models designed for intelligent processing of Classical Chinese. However, as of now, there have been no studies observed that apply large language models to the extraction of trigger words in ancient texts.

Given the extensive research showing that domain-specific continued pre-training of deep pre-trained language models can effectively enhance the model's performance on downstream tasks (Gururangan et al., 2020). Related studies also indicate that sequence labeling models still maintain certain advantages in specific task scenarios. Furthermore, according to Wei et al. (2023), it is evident that within the sequence labeling task paradigm, the BERT-BiLSTM-CRF model achieves the best performance. Considering the current lack of relevant practices involving large language models in the extraction of trigger words from ancient texts, this paper proposes constructing incremental pre-training for BERT specifically for the evaluation task. It combines this with BERT-BiLSTM-CRF to enhance the effectiveness of event information extraction. Additionally, attempts are made to incrementally pre-train and fine-tune large language models, exploring their applicability in extracting trigger words from ancient texts.

3 Domain-adaptive continued pre-training of the model

3.1 Continuing Pre-training Data

This paper uses the contents of the "text" field from the training set provided by CCL-CHED as the continuing pre-training data for all models. Domain-adaptive pre-training is a form of self-supervised learning that only requires domain-specific text input to the model without any manual annotation. In this experiment, using a self-developed Python program, the entire contents of the "text" field from both the train.jsonl and validate.jsonl files in the CHED trigger word recognition task dataset were extracted as training and validation corpora. A total of 5650 valid training sentences were obtained, with an average sentence length of 21.67 characters. The minimum sentence length was 3 characters, and the maximum sentence length was 200 characters. The cumulative distribution of sentences with different numbers of characters is shown in Figure 2.

3.2 Selection and Continuing Pre-training of BERT Model

3.2.1 Selection of Base Model

The SikuBERT series models constructed based on the "Complete Collection of Four Repositories" corpus have laid the groundwork for intelligent processing of Chinese classical texts (Dongbo

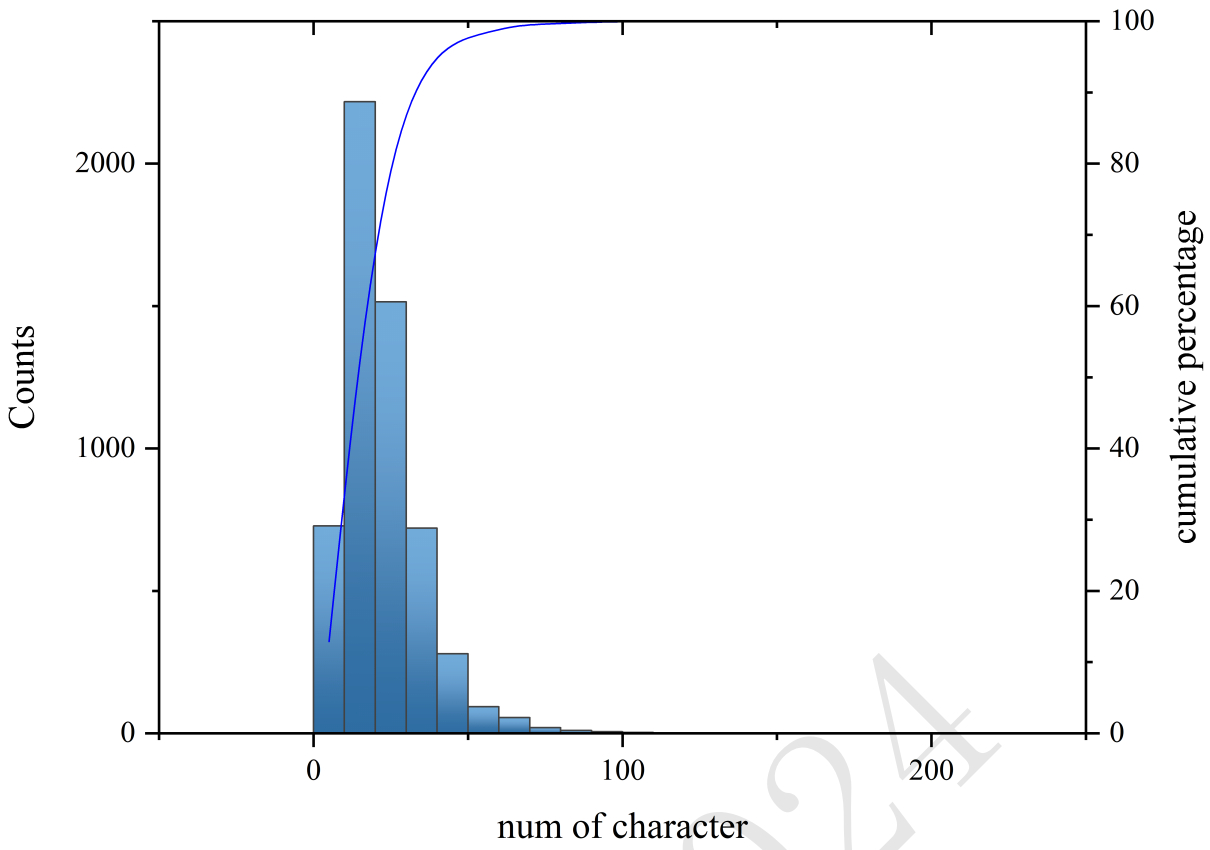

Figure 2: Cumulative distribution plot of the number of sentences with different numbers of characters in the continuing pre-training corpus

Wang et al., 2022). Subsequently, Professor Dongbo Wang's team from Nanjing Agricultural University further utilized approximately 170 million characters of ancient Chinese corpus from the Dazhige website to further pre-train the SikuBERT series models, leading to the development of the GujiBERT and GujiRoberta series models (D. Wang et al., 2023). According to the experimental data in the literature (D. Wang et al., 2023), it is evident that GujiBERT demonstrates relatively superior performance. Based on this, this paper selects GujiBERT as the baseline model for further pre-training.

3.2.2 Continued Pre-training Approach

This paper adopts Masked Language Modeling (MLM) (Devlin et al., 2019) as the task objective during the pre-training stage. MLM involves masking certain characters in the input sentence and requiring the model to predict these characters to learn bidirectional contextual relationships in language. During pre-training, BERT randomly masks 15% of the vocabulary in the input sequence. The masked words are replaced with a special [MASK] token. The objective of pre-training is for the model to predict the masked words based on context. In this way, the model needs to use information from other words in the sentence to infer the masked words, thereby learning deep semantic relationships and contextual dependencies between words. For example, for the sentence "遣内史王谊监六军，攻晋州城。" (Sending the Interior Minister Wang Yi to supervise the six armies to attack Jinzhou city.), after preprocessing by the model, it may become "遣内史王谊监六军，[MASK]晋州城。" (Sending the

Interior Minister Wang Yi to supervise the six armies, [MASK] Jinzhou city.), and the objective of MLM is to predict the character at the “[MASK]” position.

3.2.3 Training Process

During the training process, the preprocessed training data is inputted into the GujiBERT base model at the sentence level for continued training. In the forward pass, the predicted values of each masked word are computed, and the model parameters are adjusted through the calculation of the loss function and backpropagation, enabling the model to learn deep semantic relationships within the corpus. After multiple rounds of tuning and comparison, the final experimental hyperparameters are determined, as shown in Table 1. The model obtained from continued pre-training based on GujiBERT in this paper is named GujiBERT-CHED-mlm.

Experimental parameters	Value
Batch_size	8
Epochs	10
Max_sequence_length	512
Learning_rate	5e-5
optimizer	AdamW
Line_by_line	True

Table 1: Key Hyperparameters for Continued Pre-training for BERT model

3.3 Selection and Continued Pre-training of Large Language Models

3.3.1 Selection of Base Model

The Xunzi language model is a series of large language models tailored for intelligent processing of Classical Chinese texts. It is built upon models such as Qwen, Baichuan, and GLM, and has undergone extensive fine-tuning using a large amount of supervised ancient Chinese corpus data. This model series is capable of comprehensive tasks such as intelligent text indexing, information extraction, poetry generation, translation between classical and modern Chinese, reading comprehension, morphological analysis, and punctuation annotation[1] . In this paper, the Xunzi-Qwen-14b model[2] from this series is selected for continued pre-training and fine-tuning, aiming to construct a large language model suitable for the CCL2024-CHED task.

3.3.2 Continued Pre-training Approach

The Xunzi-Qwen-14b model is further pre-trained using Low-Rank Adaptation (LoRA) (Hu et al., 2021). LoRA fine-tuning is an efficient fine-tuning technique for large-scale language models. It adapts and updates specific parts of the model by introducing low-rank matrices, enabling fine-tuning for specific tasks while maintaining training efficiency and economical memory usage.

[1]https://github.com/Xunzi-LLM-of-Chinese-classics/XunziALLM

[2]https://www.modelscope.cn/models/Xunzillm4cc/Xunzi-Qwen1.5-14B/summary

3.3.3 Training Process

To further explore the effectiveness of large language models in this task, the study continues pre-training the Xunzi-Qwen-14b classical Chinese language model using the LLama-factory framework. This aims to enhance the efficiency and performance of event recognition in this task. During the training process, the corpus needs to be preprocessed into a specific format. Training corpus examples are shown in Table 2.

Xunzi-Qwen-14b Continued Pre-training Data Format

```
[
 {"text":"已酉，命习水战于新池。"}
 {"text":"九月，占城国王释利因德缓使蕃词散来。"}
]
```

Table 2: Xunzi-Qwen-14b Continued Pre-training Data Format

The hyperparameters set during the pre-training process of Xunzi-Qwen-14b are shown in Table 3. The model obtained after continuing pre-training on Xunzi-Qwen-14b is referred to as Xunzi-Qwen-14b-CHED in this paper. The perplexity of this model on the validation corpus is 21.99.

Experimental Parameters	Value
stage	pt
finetuning_type	lora
cutoff_len	1024
train_batch_size	1
learning_rate	5e-5
num_train_epochs	3

Table 3: Key Hyperparameters for Continued Pre-training for Xunzi-Qwen-14b

4 Fine-tuning Training and Evaluation of the Model

4.1 Fine-tuning of the BERT Model

4.1.1 Preprocessing of Fine-tuning Training Data

This paper utilizes the training and validation sets provided by CCL-CHED in the three task files for fine-tuning training and performance evaluation. Specifically, according to the event labeling patterns of different tasks, the corpora in each task dataset are converted to the input format required by the sequence labeling model. The B, I, E, S, O role tagging strategy is adopted, where B represents the starting character of a multi-syllable trigger word, E represents the ending character of a multi-syllable trigger word, I represents the remaining characters of a multi-syllable trigger word that are neither at the

start nor the end, S represents a single-syllable trigger word, and O represents characters in the sentence that do not form trigger words. Based on the B, I, E, S, O role tagging strategy, the number of characters with different role types for each task is shown in Table 4.

Task Name	**Number of Role Tags**
Trigger Word Recognition	5*1+1
Coarse Event Type Identification	5*9+1
Fine-Grained Event Type Identification	5*67+1

Table 4: Number of Character Role Tags for Different Tasks

4.1.2 Experimental Setup

The BERT-BiLSTM-CRF model is built using the PyTorch framework. The pre-trained weights for GujiBERT[3] are obtained from the Hugging Face official website. The computer system used for fine-tuning training of BERT-BiLSTM-CRF operates on Red Hat 4.8.5-44, with a CPU model of 48 Intel(R) Xeon(R) Silver 4214R CPU @ 2.40GHz, and a GPU model of Tesla P100. As shown in Figure 2, the longest sentence in the training corpus contains 200 characters, which is below the maximum sentence length that BERT can handle (512 characters). Therefore, the maximum sentence length for the model is set to the maximum sentence length in each batch of data. The key hyperparameters for fine-tuning training of BERT-BiLSTM-CRF are shown in Table 5.

Experimental Parameters	**Value**
Max sequence length	Pad to length
Batch size	32
Epoch	50
Learning rate	1e-5
LSTM hidden dim	256

Table 5: Key Hyperparameters for fine-tuning

In this experiment, after each training epoch, the model's precision (P), recall (R), and F1 score are computed on the validation set. Only when the F1 score of the new model is the highest, the weights of the model obtained in the current epoch are retained. Thus, in this experiment, the F1 score is used as the criterion for selecting the best model, rather than the lowest loss. The model with the lowest F1 score on the test set is considered the optimal model.

4.2 Fine-tuning of Large Language Models

4.2.1 Instructional Fine-tuning Template

Large language models are subject to various uncontrollable factors, particularly the design of prompts, which have a significant impact on the output results. Faced with various tasks, this study

[3]https://huggingface.co/hsc748NLP/GujiGPT_fan.

conducted multiple pre-experiments using different prompts. Finally, the prompts listed in Table 6 were determined for instructional fine-tuning and prediction.

Trigger Word Recognition	[{ "instruction": "将下面text古文中触发词和标签以及触发词的位置信息进行自动识别和抽取", "input": "sen_id: 5761, text: 己酉，命习水战于新池。", "output": "{\"sen_id\": 5761, \"text\": \"己酉，命习水战于新池。\", \"events\": [{\"trigger\": \"命\", \"start_offset\": 3, \"end_offset\": 4}]}" }]
Coarse Event Type Identification	[{ "instruction": "将下面text古文中触发词和标签以及触发词的位置信息进行自动识别和抽取", "input": "sen_id: 5761, text: 己酉，命习水战于新池。", "output": "{\"sen_id\": 5761, \"text\": \"己酉，命习水战于新池。\", \"events\": [{\"trigger\": \"命\", \"label\": \"交流\", \"start_offset\": 3, \"end_offset\": 4}]}" }]
Fine-Grained Event Type Identification	[{ "instruction": "将下面text古文中触发词和标签以及触发词的位置信息进行自动识别和抽取", "input": "sen_id: 5761, text: 己酉，命习水战于新池。", "output": "{\"sen_id\": 5761, \"text\": \"己酉，命习水战于新池。\", \"events\": [{\"trigger\": \"命\", \"label\": \"交流-个人交流-诏令-命令\", \"start_offset\": 3, \"end_offset\": 4}]}" }]

Table 6: Example Templates for Task-Specific Prompts

4.2.2 Experimental Setup

The experimental graphical display used is the RTX 8000, and the hyperparameters for instructional fine-tuning are shown in Table 7.

Experimental parameters	Value
stage	stf
finetuning_type	lora
cutoff_len	512
Learning_rate	5e-5
train_batch_size	2
learning_rate	5e-5
num_train_epochs	5

Table 7: Key Hyperparameters for Instructional Fine-tuning of Xunzi-Qwen-14b-CHED

5 Experimental Results and Analysis

5.1 Experimental Results

The experiment uses P (Precision), R (Recall), and F1 as evaluation metrics, with the final results being based on the Micro P, Micro R, and Micro F1 of each predicted label (Kelly, 2009) as the ultimate basis for assessing model performance. For the output results of large language models, this paper converts the outputs into character-level role sequences based on the B, I, E, S, O role labeling strategy, and then calculates Micro P, Micro R, and Micro F1(Kelly, 2009), thereby ensuring the comparability of evaluation results across different models.

5.1.1 Trigger Word Recognition

According to Table 8, the GujiBERT model, which did not utilize unsupervised data from CHED for continued pre-training, performed the best. The performance of large language models was the worst. Upon examining the output results of large language models, it was found that in the test samples with sen_id 8052 and 939, the character interval index of the last trigger word given by the large language models exceeded the sentence length range of the original sentence. Further inspection of the trigger word characters extracted by the large language models revealed that they matched the answers in the validation set. This indicates that large language models possess good semantic understanding capabilities, but their ability to produce standardized output needs improvement. Additionally, BERT based model achieved the maximum F1 score at training epoch 14.

Model Name	Micro P	Micro R	Micro F1
GujiBERT-BiLSTM-CRF	0.799	0.802	0.800
GujiBERT-CHED-mlm-BiLSTM-CRF	0.752	0.767	0.759
Xunzi-Qwen-14b-CHED	0.676	0.665	0.671

Table 8: Trigger Word Recognition Results of Various Models

5.1.2 Coarse Event Type Identification

According to Table 9, the GujiBERT model, which did not use unsupervised data from CHED for continued pre-training, performed the best, while the performance of large language models was

the worst. By examining the output results of the large language models, it was found that in the test sample with sen_id 917, the character interval index of the last trigger word given by the large language models exceeded the sentence length range of the original sentence. Further inspection of the trigger word characters identified by the large language models revealed that they matched the answers in the validation set. Additionally, BERT based model achieved the maximum F1 score at training epoch 67.

Model Name	Micro P	Micro R	Micro F1
GujiBERT-BiLSTM-CRF	0.837	0.820	0.828
GujiBERT-CHED-mlm-BiLSTM-CRF	0.832	0.779	0.805
Xunzi-Qwen-14b-CHED	0.653	0.650	0.651

Table 9: Trigger Word Recognition and Classification Results at the Coarse Granularity

5.1.3 Fine-Grained Event Type Identification

According to Table 10, the GujiBERT model, which did not use unsupervised data from CHED for continued pre-training, showed the best performance, while the performance of large language models was the worst. Upon examining the output results of the large language models in the fine-grained event extraction task, no issues with output format irregularities were found. Additionally, BERT based model achieved the maximum F1 score at training epoch 25.

Model Name	Micro P	Micro R	Micro F1
GujiBERT-BiLSTM-CRF	0.782	0.762	0.772
GujiBERT-CHED-mlm-BiLSTM-CRF	0.761	0.686	0.722
Xunzi-Qwen-14b-CHED	0.631	0.617	0.624

Table 10: Trigger Word Recognition and Classification Results at the Fine-Grained Granularity

5.2 Results Analysis and Discussion

The experiments found that using GujiBERT as the character encoder followed by a BiLSTM-CRF, the GujiBERT-BiLSTM-CRF model achieved the best performance in all tasks. We speculate that this may be due to the relatively small number of event trigger words in classical Chinese texts and their strong grammatical regularity, which explains why past pattern-matching methods have achieved good results. Additionally, the tasks of trigger word identification and classification are straightforward, not involving the recognition and classification of event arguments and their relationships. Therefore, context-aware sequence labeling models are better able to capture the position and type characteristics of event trigger words. On the other hand, the advantage of large language models lies in their understanding of the overall semantics of sentences, making them well-suited for complex tasks. However, for tasks with simple objectives but strict formatting requirements, these models do not hold as much of an advantage.The experiments also indicate that if the amount of continued pre-training data is too small, the model's performance may not improve and could even decline. This may be due to the insufficient amount of data used for fine-tuning, which is not enough to significantly optimize the large language model. Instead, it might introduce disturbances to the original model, resulting in unstable performance.

In the experiments, we also found that fewer epochs are needed to train fine-grained event trigger word recognition and classification models compared to training coarse-grained models. Furthermore, the accuracy of fine-grained event trigger word recognition and classification models was even higher than that of the trigger word recognition models. This suggests that for BERT-based sequence labeling models, the number of training epochs required to achieve the maximum F1 score is not positively correlated with the number of role labels. Additionally, more detailed event classification granularity seems to help the model better learn the semantic differences between different characters, thus improving the event extraction task performance.

6 Conclusion and future work

This paper constructs various models for event trigger word extraction and classification through domain-adaptive pre-training and fine-tuning of the small language model GujiBERT and the large language model Xunzi-Qwen-14b. Performance comparisons of different models were conducted through experimental testing. From our experiments, we conclude that small deep pre-trained language models like BERT, which follow a pre-training-fine-tuning approach, are more suitable for fine-grained information extraction and classification tasks.

There is still room for optimization in the sequence labeling paradigm used in this paper: we adopted a detailed role sequence labeling strategy of B, I, E, S, O in our experiments, but trigger words usually consist of only 1 to 2 characters. Thus, some role labels may not be used, which increases the computational load of the CRF module and potentially the classification difficulty. Future research could consider using the simplified B, I, O labeling strategy. Our experiments also show that large language models based on the pre-training-prompt paradigm have superior semantic understanding capabilities but cannot yet generate directly usable results when applied to information extraction tasks. The main challenges in applying large language models lie in the design of prompt templates and further processing of the output content.

References

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.),*Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Dongbo Wang, Chang Liu, Zihe Zhu, Jiangfeng Liu, Haotian Hu, Si Shen, & Bin Li. (2022). Construction and Application of Pre-trained Models of Siku Quanshu in Orientation to Digital Humanities. *Library Tribune, 42*(6), 31–43.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks.*Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8342–8360. https://doi.org/10.18653/v1/2020.acl-main.740

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). *LoRA: Low-Rank Adaptation of Large Language Models* (arXiv:2106.09685). arXiv. https://doi.org/10.48550/arXiv.2106.09685

Kelly, D. (2009). Methods for Evaluating Interactive Information Retrieval Systems with Users.*Foundations and Trends in Information Retrieval, 3*(1—2), 1–224. https://doi.org/10.1561/1500000012

Ma, X., He, L., Liu, J., Li, Z., & Gao, D. (2021). A Construction Method of the Classification System Oriented to Content Analysis of Ancient Books. *Journal of Library and Information Science in Agriculture, 33*(9), 27–36. https://doi.org/10.13998/j.cnki.issn1002-1248.21-0262

Wang, D., Liu, C., Zhao, Z., Shen, S., Liu, L., Li, B., Hu, H., Wu, M., Lin, L., Zhao, X., & Wang, X. (2023).*GujiBERT and GujiGPT: Construction of Intelligent Information Processing Foundation Language Models for Ancient Texts* (arXiv:2307.05354). arXiv. https://doi.org/10.48550/arXiv.2307.05354

Wang Y., Wang H., Zhu H., & Li X. (2023). Research on the Construction of an Event Recognition Model for Historical Antique Books Based on Text Generation Technology.*Library and Information Service, 67*(3), 119–130. https://doi.org/10.13266/j.issn.0252-3116.2023.03.011

Wei, C., Feng, Z., Huang, S., Li, W., & Shao, Y. (2023). CHED: A Cross-Historical Dataset with a Logical Event Schema for Classical Chinese Event Detection. In M. Sun, B. Qin, X. Qiu, J. Jing, X. Han, G. Rao, & Y. Chen (Eds.), *Chinese Computational Linguistics* (pp. 289–305). Springer Nature. https://doi.org/10.1007/978-981-99-6207-5_18

Xuehan Yu, Lin He, & Jian Xu. (2021). Extracting Events from Ancient Books Based on RoBERTa-CRF. *Data Analysis and Knowledge Discovery, 5*(7), 26–35.

Yu Xuehan, He Lin, & Wang Xianqi. (2023). Research on Event Extraction from Ancient Books Based on Machine Reading Comprehension. *Journal of The China Society For Scientific And Technical Information, 42*(3), 316–326.

Zhang, J., Wei, Y., Zhu, Y., & Wu, B. (2023). Self-adaptive Prompt-tuning for Event Extraction in Ancient Chinese Literature. *2023 International Joint Conference on Neural Networks (IJCNN)*, 1–8. https://doi.org/10.1109/IJCNN54540.2023.10191495

Zhangchao Li, Zhongkai Li, & Lin He. (2020). Study on the Extraction Method of War Events in Zuo Zhuan. *Library and Information Service, 64*(7), 20–29. https://doi.org/10.13266/j.issn.0252-3116.2020.07.003

Zhongbao Liu, Jianfei Dang, & Zhijian Zhang. (2020). Research on Automatic Extraction of Historical Events and Construction of Event Graph Based on Historical Records. *Library and Information Service, 64*(11), 116–124. https://doi.org/10.13266/j.issn.0252-3116.2020.11.013