Multi-Domain Ancient Chinese Named Entity Recognition Based on Attention-Enhanced Pre-trained Language Model

Qi Zhang, Zhiya Duan, Shijie Ma, Shengyu Liu, Zibo Yuan, Ruimin Ma
School of Economics and Management
Shanxi University, China
qi.zhang@sxu.edu.cn

Abstract

Recent advancements in digital humanities have intensified the demand for intelligent processing of ancient Chinese texts, particularly across specialized domains such as historical records and ancient medical literature. Among related research areas, Named Entity Recognition (NER) plays a crucial role, serving as the foundation for knowledge graph construction and deeper humanities computing studies. In this paper, we introduce a architecture specifically designed for multi-domain ancient Chinese NER tasks based on a pre-trained language model (PLM). Building upon the GujiRoberta backbone, we propose the GujiRoberta-BiLSTM-Attention-CRF model. Experimental results on three distinct domain-specific datasets demonstrate that our approach significantly outperforms the official baselines across all three datasets, highlighting the particular effectiveness of integrating an attention mechanism within our architecture.

Keywords: Named Entity Recognition, Ancient Chinese, Multi-Domain GujiRoberta-BiLSTM,-Attention-CRF.

1 Introduction

Thousands of years of Chinese civilization have been encapsulated within historical, political, economic, medical and various other types of ancient books. However, due to their vast quantity and significant deterioration over time, these invaluable resources have remained underexplored and underutilized. Recent rapid advancements in frontier technologies, such as big data and artificial intelligence, present unprecedented opportunities for the deep mining and revitalization of ancient texts. In particular, the integration of natural language processing (NLP) and knowledge graph technologies has rejuvenated research into ancient documents. Entities, serving as fundamental knowledge units within ancient texts, play a crucial role in humanities computing studies. Nevertheless, entity recognition from ancient Chinese texts remains significantly challenging, primarily due to the intrinsic complexity of ancient Chinese grammar, archaic vocabulary, semantic obscurity, and the domain-specific nature of texts.

To address these challenges, EVAHAN 2025 proposed a specialized NER task focused on ancient Chinese texts across multiple domains. Based on the PLM called SikuRoBERTa for ancient Chinese provided by EVAHAN 2025, we further propose the incorporation of a BiLSTM-Attention network for enhanced feature extraction, coupled with a CRF layer for decoding to improve the accuracy of entity label classification. Besides, through meticulous hyperparameter tuning, our model better accommodates domain-specific textual characteristics. Extensive experiments conducted on three provided datasets demonstrate the superior performance of our proposed model, significantly surpassing official benchmarks.

2 Related Research

Research on named entity recognition (NER) in ancient Chinese has gone through four technical evolution stages: rule-based templates, statistical modeling, neural networks and pre-trained models. Early template and statistical methods were gradually replaced by neural network learning frameworks due to their limited domain transferability (Huang et al., 2002; Li et al., 2023) . For instance, Huang et al. (2015) introduced the BiLSTM-CRF model, which captured long-distance syntactic dependencies in ancient texts through bidirectional long short-term memory networks and optimized label sequence prediction through Conditional Random Fields.

The subsequent emergence of pre-trained language models (PLMs) dramatically enhanced the processing efficiency and semantic comprehension capabilities for ancient Chinese

Proceedings of the Second Ancient Language Processing Workshop associated with NAACL 2025, pages 159–163
May 4, 2025.

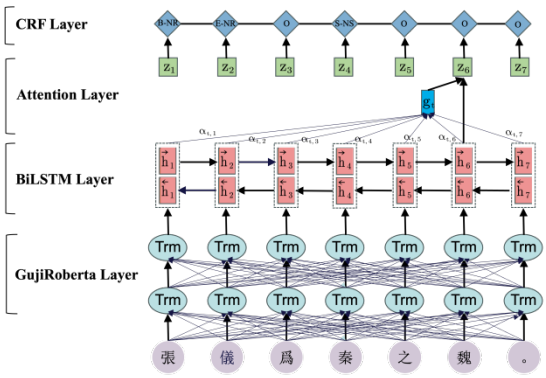

Figure 1: Model Architecture.

texts. The iterations of BERT architecture (Devlin et al., 2019) and the introduction of RoBERTa by (Liu et al., 2019) with dynamic masking mechanisms have redefined the pre-training paradigm. Wang et al.(2022) conducted incremental training using traditional Chinese text from the Complete Library in Four Sections to build SikuBERT and SikuRoBERTa pre-trained models. While generative large language models like GPT demonstrate considerable semantic understanding, they suffer from issues such as entity hallucination and boundary ambiguities, limiting their reliability for precise entity extraction tasks (Zhang et al., 2023) .Recent researches shows BERT-based methods still maintain significant advantages through domain-adaptive fine-tuning in IE tasks (Detroja et al., 2023; Diaz-Garcia and Lopez, 2024; Han et al., 2024).

However, most previous studies on ancient Chinese NER have primarily focused on general entities such as personal names, locations and dates. Recent advancements in digital humanities have broadened the scope, demanding sophisticated processing capabilities for specialized domains such as historical records and ancient medical literature, thus expanding annotation schemas from basic three-element frameworks to more comprehensive multi-element structures, including official titles, pathological terms, and cultural symbols (Zhang et al., 2023). Compared to standard named entity tasks, recognizing specialized domain entities poses greater challenges due to the need for enhanced contextual understanding and domain-specific adaptability.

3 Model Construction

3.1 Model Introduction

The GujiRoberta-BiLSTM-Attention-CRF model is a deep learning framework designed for the task of ancient text named entity recognition in ancient text. As illustrated in Table1, its process is divided into four stages: Firstly, through the pre-trained model GujiRoberta, context-aware word vectors are generated; Secondly, BiLSTM captures bidirectional long-distance semantic dependencies; Subsequently, the attention mechanism is utilized to enhance key features; Finally, the CRF layer is employed to achieve the global optimal label prediction.

The attention mechanism significantly enhances the model's ability to focus on key information. By introducing attention weights distribution in the output layer of BiLSTM, the importance scores of for each position are calculated through a learnable parameter matrix, and then normalized by softmax to generate a focusing vector. This mechanism can adaptively enhance entity-related features, such as the core verbs in disease descriptions, shile simultaneously suppressing irrelevant noise. It is particularly suitable for processing scattered entity expressions in ancient texts, thereby improving the model's sensitivity to key information.

In the named entity recognition task, CRF, as the decoding layer, solves the label conflict problem of traditional softmax decoding by modeling label transition probabilities. The global score function is constructed by defining emission scores (linear transformation of the BiLSTM-attention output) and transition scores (transition matrix between labels), and the optimal path is solved using the Viterbi algorithm. This design ensures that the output sequence conforms to the ancient text entity annotation specifications (such as the continuity constraint of the BIOES label system), effectively improving the recognition accuracy of entity boundaries.

3.2 Experimental Datasets

This competition involves three ancient text named entity recognition datasets: Dataset A is based on "Records of the Grand Historian" and labels personal names (NR), place names (NS), book titles (NB), official titles (NO), dynastic names (NG), and time (T). The complexity of this dataset arises from the evolution of historical naming conventions; Dataset B is selected from "Twenty-Four Histories" and focuses on personal names (NR), place names (NS), and time (T), requiring handling ambiguity issues caused by ancient language abbreviations; Dataset C originates from traditional Chinese medical classics and covers six categories of professional terms: traditional Chinese medicine diseases (ZD), syndromes (ZZ),

prescriptions (ZF), medicinal materials (ZP), symptoms (ZS), and acupoints (ZA). It faces the challenge of diverse term expressions.

Data processing adopts a multi-stage optimization strategy: Firstly, perform text data undergoes cleaning and standardization processing, which includes the removal of blank lines and the normalization of characters. Subsequently, dynamic segmentation is executed based on four sets of length thresholds (128/256/400/512) and locate the end symbols, such as periods and quotation marks, etc. through backtracking to ensure semantic integrity. This approach facilitates the model's capability to better learn the correlation information between ancient texts. After randomly shuffling process to eliminate sequence deviations, the dataset should be partitioned into training set and validation set in a 9:1 ratio. Finally, ensure the representativeness of each data subset to meet the multi-dimensional requirements of model training, hyperparameter optimization, and performance evaluation.

3.3 Evaluation Metrics

Precision, recall, and F1 score were used as the main metrics to evaluate model performance. Their calculation formulas are as follows:

$$P = \frac{TP}{TP+FP} \times 100\% \quad (1)$$

$$R = \frac{TP}{TP+FN} \times 100\% \quad (2)$$

$$F1 = \frac{2PR}{P+P} \times 100\% \quad (3)$$

Where TP = correctly identified entities, FP = incorrect identifications, and FN = missed entities.

3.4 Experimental Environment

The experimental setup utilized a Linux server equipped with an NVIDIA RTX 4090 GPU (24 GB of video memory), facilitating efficient large-scale deep learning model training. A 6-core Xeon Gold 6142 processor provided robust multitasking capabilities, while 64.4 GB of RAM and 420 GB of disk storage were sufficient to meet the computational and data storage requirements.

For the software environment, PyTorch 2.2.2 was chosen, which, although an older version, offered good compatibility and stability. Its dynamic computation graph, user-friendly APIs, and community support made it the preferred choice. Docker containerization technology was utilized to construct a standardized development environment, ensuring research reproducibility and consistency.

3.5 Model Training

(1) Loss Function

The objective of model training was to minimize negative log-likelihood loss, measuring prediction error by comparing the predicted label sequences with the true label sequences. The CRF layer calculated probabilities for all possible label sequences, ultimately selecting the most probable sequence as the final prediction.

(2) Optimizer

The AdamW optimizer was used, introducing weight decay to mitigate the risk of overfitting. The learning rate warm-up strategy was implemented to stabilize the initial gradient updates.

(3) Hyperparameter Tuning

This study employs a combined method of grid search and random search to optimize the key hyperparameters of the model. The search space for the learning rate is set from 8e-6 to 4e-5, while the batch size is dynamically adjusted within the range of 8 to 64. The dropout rate is explored within the range of 0 to 0.6, and the input text length is uniformly standardized to the range of 128 to 512 characters. We utilized several pre-trained model architectures, including bert, siku-roberta, GujiRoBERTa_jian_fan, roberta-classical-chinese-large-char, and employed some fine-tuning techniques. Through systematic verification of parameter combinations, the optimal configuration scheme of each model architecture was finally obtained, as shown in Table1.

Dataset	Dataset_A	Dataset_B	Dataset_C
Pretrained Model	GujiRoBERTa_jian_fan	GujiRoBERTa_jian_fan	GujiRoBERTa_jian_fan
Text Length	128	256	400
Learning Rate	0.00005	0.00002	0.00003
Batch Size	8	8	8
Dropout	0.4	0.4	0.6

Table 1: Hyperparameter Tuning.

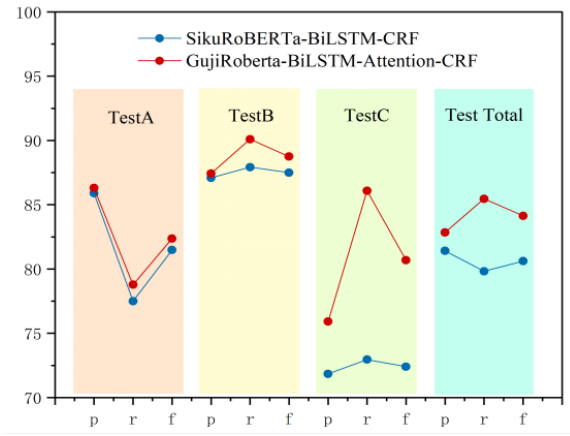

Figure 2: Model Effectiveness Comparison.

4 Experimental Results and Analysis

Prior to formal submission, we compared the GujiRoberta-BiLSTM-CRF and GujiRoberta-BiLSTM-Attention-CRF models on the validation set, with individual labels as the smallest unit to calculate P, R and F1-score. Experimental results confirmed that incorporating an attention mechanism consistently improved overall F1-scores. Specifically, under identical parameter configurations, the F1-score increased by over 1% for datasets A and C (the F1-score improved from 0.9103 to 0.9243 for dataset A and from 0.8613 to 0.8765 for dataset C). The attention mechanism significantly boosted global entity recognition by emphasizing critical feature information. Consequently, the GujiRoberta-BiLSTM-Attention-CRF model was selected for the final test

As shown in 错误!未找到引用源。 and Table 2, the test results further validated our model's effectiveness. EVAHAN 2025 adopted the classical SikuRoBERTa-BiLSTM-CRF architecture as its baseline. In the general dataset A, the F1 score of GujiRoberta reached 82.37, reflecting an increase of 0.89. Additionally, there was a significant improvement in the recall rate, which verifies the effectiveness of GujiRoberta's enhanced semantic understanding capabilities and its attention mechanism in capturing key information. In the professional historical dataset B, the F1 score increased to 88.74, representing an enhancement of 1.24, with the recall rate of 90.09. This indicates an advancement in the model's generalization ability concerning ancient terms and abbreviations. Furthermore, in the Chinese medicine classics dataset C, the F1 score improved by 8.28 percentage points to 80.68 compared to the baseline model, demonstrating a comprehensive ability to recognize professional terms. It is noteworthy that the accuracy rate of the three datasets were lower than the recall rates, reflecting that the model still has misidentification phenomena when dealing with complex historical entities, such as names and place names with omitted sentence patterns, as well as the diversity of TCM terms. These insights suggest meaningful directions for future research.

Dataset	Method	P	R	F
A	Ours	86.3	78.78	**82.37**
	Baseline	85.90	77.50	81.48
B	Ours	87.43	90.09	**88.74**
	Baseline	87.09	87.92	87.50
C	Ours	75.91	86.09	**80.68**
	Baseline	71.84	72.95	72.40
Total	Ours	82.84	85.46	**84.13**
	Baseline	81.41	79.82	80.61

Table 2: Model Effectiveness Comparison.

5 Conclusion & Future Directions

The experimental results indicate that the GujiRoberta-BiLSTM-Attention-CRF model proposed in this paper demonstrates a significant improvement over the official baseline on ancient book datasets in different fields such as history and medicine. These findings verify the effectiveness of the attention mechanism and the multi-module integration strategy employed in the model. By enhancing the parsing ability of ancient Chinese complex sentence patterns and long texts, the model significantly improves the entity recall rate, and provides a reliable solution for entity recognition in multi-domain ancient books. However, the accuracy of the model still remains potential for improvement in the face of fine-grained semantic contexts, such as the polysemy of words and variation of professional terms. Future research will focus on optimizing the dynamic attention allocation mechanism, enhancing semantic discrimination ability with domain adaptive pre-training, and further exploring the generalization of multi-domain feature adaptation modules.

References

Dongbo Wang, Chang Liu, Zihe Zhu, Jiangfeng Liu, Haotian Hu, Si Shen and Bin Li. 2022. *Construction and application of pre-training model of "Siku Quanshu" oriented to digital humanities*. *Library Tribune*, 42(06):31-43.

Jose A. Diaz-Garcia and Julio Amador Diaz Lopez. 2024. *A survey on cutting-edge relation extraction techniques based on language models*. *Computing Research Repository*, arXiv: 2411.18157.

Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. 2019. *BERT: pre-training of deep bidirectional transformers for language understanding. 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),* pages 4171-4186, Minneapolis, Minnesota. Association for Computational Linguistics. 10.18653/v1/N19-1423

Jianlong Li, Youren Yu, Xueyang Liu and Siwen Zhu. 2023. *System report for CCL23-Eval task 1;GuNER based on incremental pretraining and adversarial learning*. *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 23-33, Harbin, China.

Kartik Detroja, Ck Bhensdadia and Brijesh S. Bhatt. 2023. *A survey on r*elation *extraction. Intell. Syst. Appl.*, 19: 200244.

Liang Huang, Yinan Peng, Huan Wang and Zhenyu Wu. 2002. *PCFG parsing for restricted classical Chinese texts*. *Series PCFG Parsing for Restricted Classical Chinese Texts. (Volume 18)*, pages 1-6. https://doi.org/10.3115/1118824.1118830.

Ridong Han, Chaohao Yang, Tao Peng, Prayag Tiwari, Xiang Wan, Lu Liu and Benyou Wang. 2024. *An empirical study on information extraction using large language models. Computing Research Repository*, arXiv: 2305.14450.

Xinghua Zhang, Tianjun Liu, Wenyuan Zhang and Tingwen Liu. 2023. *System report for CCL23-Eval task 1: information theory constraint and paragraph based paragraph classical named entity recognition*. *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 1-13, Harbin, China.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer and Veselin Stoyanov. 2019. *RoBERTa: a robustly optimized BERT pretraining approach*. *Computing Research Repository*, arXiv: 1907.11692.

Zhiheng Huang, Wei Xu and Kai Yu. 2015. *Bidirectional LSTM-CRF models for sequence tagging*. *Computing Research Repository*, arXiv: 1508.01991.