Make Good Use of GujiRoBERTa to Identify Entities in Ancient Chinese

Lihan Lin, Yiming Wang, Jiachen Li, Huan Ouyang, Si Li*
School of Artificial Intelligence
Beijing University of Posts and Telecommunications, China
{linlihan, wym2001, jiachen-li, ouyanghuan, lisi}@bupt.edu.cn

Abstract

This report describes our model submitted for the EvaHan 2025 shared task on named entity recognition for ancient Chinese literary works. Since we participated in the task of closed modality, our method is based on the appointed pre-trained language model GujiRoBERTa-jian-fan and we used appointed datasets. We carried out experiments on decoding strategies and schedulers to verify the effect of our method. In the final test, our method outperformed the official baseline, demonstrating its effectiveness. In the end, for the results, this report gives an analysis from the perspective of data composition.

1 Introduction

Named Entity Recognition (NER) is a cornerstone task in Natural Language Processing (NLP), which involves identifying and classifying named entities such as person names, locations, and organizations within text. These entities carry significant semantic information and are crucial for various NLP applications, including information extraction (Nasar et al., 2021), machine translation (Yang et al., 2017), and historical text analysis (Won et al., 2018). The complexity of ancient Chinese texts, characterized by classical grammar, lack of punctuation, and evolving vocabulary, presents unique challenges for NER tasks.

Previous research (Yu and Wang, 2020) on ancient Chinese NER has largely framed the problem as a sequence labeling task, leveraging pre-trained language models to achieve notable performance improvements. However, most existing pre-trained language models are pre-trained on modern Chinese or multilingual corpora, which may not adequately capture the linguistic nuances of ancient Chinese. To address this gap, recent efforts have focused on developing specialized Pre-trained Language Models, such as GujiBERT and GujiGPT (Wang et al., 2023), which are specifically pre-trained on ancient Chinese corpora to better support NER tasks in this domain.

Building on these advancements, EvaHan 2025 has been launched as the fourth International Evaluation of Ancient Chinese Information Processing. This competition focuses on NER tasks using large language models and provides a benchmark for evaluating the performance of different approaches to ancient Chinese texts. The datasets used in EvaHan 2025 include historical texts from sources like the *Shiji* and the *Twenty-Four Histories*, as well as medical texts from Traditional Chinese Medicine Classics. These datasets have been carefully annotated by experts to ensure high-quality training materials and gold-standard texts. Using these high-quality datasets, we can further explore how to better perform NER tasks in ancient Chinese.

This report introduces our NER system for EvaHan 2025 and its performance on testing datasets.

2 Related Work

2.1 Named Entity Recognition

NER for ancient Chinese is a more specialized and challenging task due to the unique linguistic characteristics of classical texts, such as archaic grammar, lack of punctuation, and lexical evolution. Early studies on ancient Chinese NER adopted rule-based methods and statistical models (Liu et al., 2018), but these approaches struggled with the complexity and variability of historical texts. Recent advancements have shifted toward deep learning and pre-trained language models (Tian et al., 2020), with researchers developing models tailored to ancient Chinese. The introduc-

*Corresponding author

Proceedings of the Second Ancient Language Processing Workshop associated with NAACL 2025, pages 136–140
May 4, 2025.

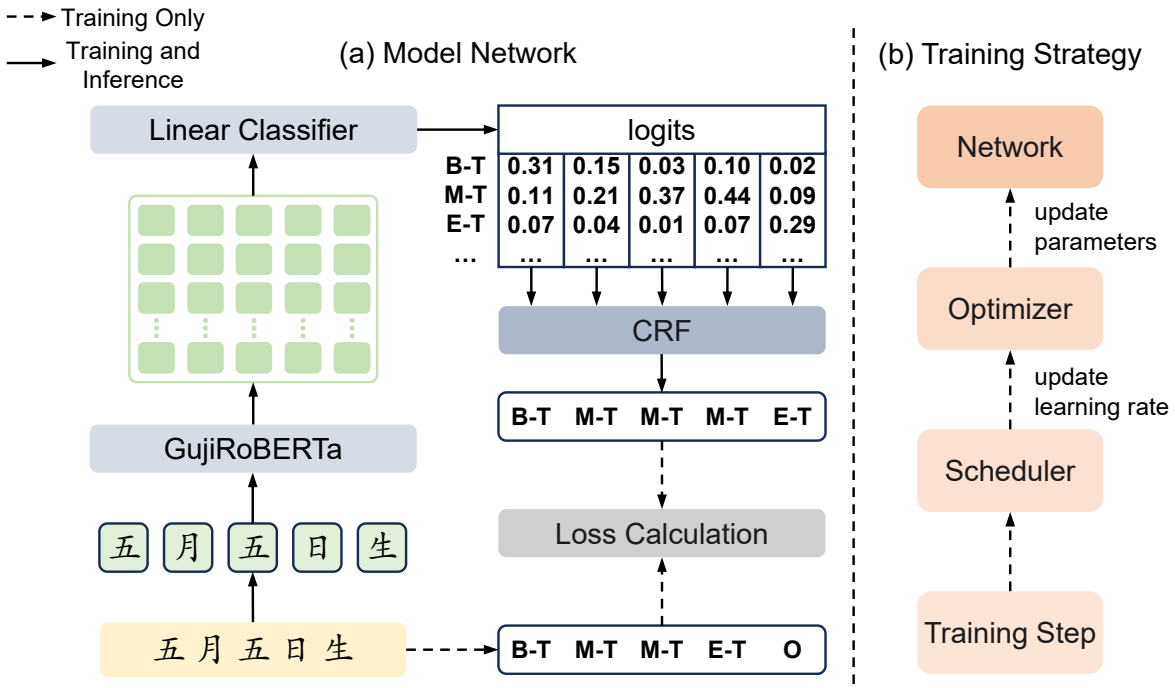

Figure 1: Our model, where (a) is model network (actions of training and reasoning are distinguished by arrows) and (b) is training strategy that is used only at training time to update parameters of (a).

tion of pre-trained language models pre-trained on ancient Chinese corpora, such as SIKU-BERT (Wang et al., 2021) and SIKU-RoBERTa (Wang et al., 2021), has further enhanced NER accuracy by capturing the linguistic nuances of classical texts. Some other researchers (Liu et al., 2021) propose a Chinese NER method for historical and cultural texts using a BERT-BiLSTM-CRF model, which significantly improves the accuracy and efficiency of entity extraction in ancient Chinese documents by leveraging contextualized embeddings and sequence tagging. These models leverage large-scale ancient Chinese datasets, including historical documents like the *Shiji* and *Hanshu*, to address the limitations of modern pre-trained language models. Additionally, the EvaHan 2022 competition has provided a benchmark for evaluating NER systems on ancient Chinese, fostering innovations in this domain. Despite these advances, challenges such as data scarcity, entity ambiguity, and cross-era vocabulary variations remain, driving ongoing research in ancient Chinese NER.

2.2 Pre-trained Language Model

In the domain of Named Entity Recognition (NER), Pre-trained Language Models have been pivotal, with BERT (Devlin et al., 2019) being widely recognized. However, the application of these models to ancient Chinese texts presents unique challenges due to the significant linguistic differences compared to modern Chinese (Sun et al., 2019). To address this, specialized models such as SIKU-RoBERTa and GujiRoBERTa (Wang et al., 2023) have been developed, specifically pre-trained on ancient Chinese corpora to enhance NER performance for historical documents.

3 Method

3.1 Model

The model is shown in Figure 1, including model network and training strategy.

In model network, input text is first tokenized into single tokens, forming the input sequence $S = \{c_1, c_2, ..., c_n\}$. If in training stage, the truth entity annotation TE of the text input is also entered, as shown in the figure 'B-T M-T M-T E-T O'. The input sequence S is then passed through the GujiRoBERTa, a multi-layer Transformer structure model. In the l-th layer of Transformer, the hidden representation H_l is calculated as following:

$$H_l = LayerNorm(H_{l-1} + Attention(H_{l-1})), \quad (1)$$

$$H_{l+1} = LayerNorm(\hat{H}_{l+1} + FFN(\hat{H}_{l+1})), \quad (2)$$

where H_0 is S, *LayerNorm* is the layer-wise normalization layer, and *Attention* is the multi-head

Combination	Dataset A	Dataset B	Dataset C
LinearScheduleWithWarmup+Softmax	**92.21**	87.24	80.67
CosineSchedulerWithWarmup+Softmax	91.98	85.68	79.50
ConstantSchedule+Softmax	88.03	83.96	77.71
LinearScheduleWithWarmup+CRF	92.14	**89.03**	**83.11**
CosineSchedulerWithWarmup+CRF	92.01	85.70	80.73
ConstantSchedule+CRF	88.61	83.22	76.20

Table 1: F1 Score comparison of different combinations in training(%)

attention layer. We initialize the model using pretrained GujiRoBERTa. After obtaining the encoding representation H from RoBERTa, these embeddings are passed through a linear classification layer to produce logits:

$$R = MLP(H), \tag{3}$$

which represent the raw, unnormalized scores indicating the models confidence for each possible class. Finally, we apply the Conditional Random Field (CRF) to decode the logits into tags:

$$PE = CRF(R). \tag{4}$$

where PE is entity labels predicted by model network, as shown by 'B-T M-T M-T M-T E-T' in Figure 1. During training, PE and TE are used to calculate loss. In inference, PE is the output of the NER task.

In training strategy, we use scheduler, specifically, LinearScheduleWithWarmup, which receives training step and outputs an updates learning rate. The new learning rate is used by optimizer to update parameters of model network.

3.2 Decoding Strategy

We employ a Conditional Random Field (CRF) layer as the decoding mechanism. The CRF layer explicitly models sequential dependencies between output tags by incorporating both emission scores (token-level label confidences from the encoder) and transition scores (learnable inter-tag relationships). The CRF layer jointly optimizes these two components to ensure globally coherent predictions. The model computes the most likely tag sequence by maximizing the conditional probability:

$$A = \sum_{i=1}^{T} \psi_{\text{emission}}(x_i, y_i), \tag{5}$$

$$B = \sum_{i=2}^{T} \psi_{\text{transition}}(y_{i-1}, y_i), \tag{6}$$

$$P(y|x) = \frac{1}{Z(x)} \exp(A + B). \tag{7}$$

where $Z(x)$ is the partition function, T is the sequence length, x_i is the hidden state of the i-th token, y_i is the tag at position i, ψ_{emission} is the emission score from the encoder, and $\psi_{\text{transition}}$ is the transition score between tags.

3.3 Scheduler

As described in 3.1, during training, we employ LinearScheduleWithWarmup as scheduler, updating learning rate based on training step:

$$\text{lr}_{\text{warmup}}(t) = \text{lr}_{\text{base}} \cdot \frac{t}{t_{\text{warmup}}}, \tag{8}$$

$$\text{lr}_{\text{decay}}(t) = \text{lr}_{\text{base}} \cdot \left(1 - \frac{t - t_{\text{warmup}}}{t_{\text{max}} - t_{\text{warmup}}}\right), \tag{9}$$

$$\text{lr}(t) = \begin{cases} \text{lr}_{\text{warmup}}(t), & \text{if } t < t_{\text{warmup}}, \\ \text{lr}_{\text{decay}}(t), & \text{otherwise.} \end{cases} \tag{10}$$

where lr_{base} is preset base learning rate, t_{warmup} is preset warmup timestep, t is training step, and $\text{lr}(t)$ is updated learning rate.

3.4 Solution for Long Sentences

The testing datasets contain some long length sentences, which are beyond the maximum length processed by model. Considering this situation, we split these long sentences into some short subsentences. We try to keep all sub-sentences semantically complete thus we split the long sentence according to punctuation instead of the maximum length. Then we revert sentences from the output file of system and obtain our final submisson.

	TestA			TestB			TestC			Test Total		
	P	R	F	P	R	F	P	R	F	P	R	F
Baseline	85.90	77.50	81.48	87.09	87.52	87.50	71.84	72.95	72.40	81.41	79.82	80.61
Ours	**88.16**	76.38	**81.84**	86.87	**90.09**	**88.45**	**75.57**	**85.50**	**80.23**	**82.92**	**84.56**	**83.74**

Table 2: Baseline and testing results of our model(%)

4 Experiments

4.1 Dataset

Given the closed modality competition we participated in, our experiments were limited to the datasets provided by EvaHAN 2025, including three different training datasets and their corresponding three test datasets. Dataset A comes from *Shiji*; Dataset B is extracted from *Twenty-Four Histories*; Dataset C consists of texts on Traditional Chinese Medicine Classics.

Models trained on training datasets A, B, and C are then used to test on the corresponding test datasets A, B, and C.

4.2 Metric

According to the requirements of EvaHAN 2025, Precision, Recall and F1 Score are selected as metrics, which are simply denoted as P, R and F in tables of this report. The results are presented in percentages (%).

4.3 Setting

When training model, we set some hyperparameters. Importantly, we set base learning rate to 5e-5, dropout ratio to 0.1, weight decay to 0.01, and training epoch to 50.

4.4 Training

We divided labeled training dataset into training data and validation data in a ratio of 0.95: 0.05. Specifically, in our experiment during training, we used Softmax and CRF for decoding strategies. And we selected LinearScheduleWithWarmUp, CosineScheduleWithWarmup, ConstantSchedule as candidate scheduler respectively. Based on these selected approaches, we obtained six combinations and compared their performance. Each combination was trained on training data A, B, and C and evaluated separately on validation data A, B, and C. In Table 1, we compare the performance of different combinations on the three data sets. For brevity, we only show the F1 Score. Based on the result of comparison, our model finally chosed CRF and LinearScheduleWithWarmup.

4.5 Testing

After the test datasets were released, We used trained models to test and got our NER results. After confirming that the number of characters is exactly the same as test datasets and that each line is completely aligned with the test datasets, we submitted our result documents before the deadline. The quantitative results of our model were informed by NER 2025, along with the baseline, which used SikuRoBERTa-BiLSTM-CRF. As shown in Table 2, our approach outperformed the baseline, especially on TestC. In Test Total, compared with baseline, Precision, Recall and F1 Score of our model increased by 1.51%, 4.74% and 3.13% respectively, demonstrating the effectiveness of our model.

However, our method has a slightly lower Recall on TestA and a slightly lower Precision on TestB. To explore the reasons, we carefully examined datasets and found that our model tends to get confused with annotations of certain official positions or time-related terms in TestA and TestB. However, many of the entities in dataset C are medical terms. The individual words that make up these terms appear less frequently in other entities, and the models are less easily confused facing with these terms. In future research, we will try to improve here.

5 Conclusion

In this report, we describe our named entity recognition system for EvaHan 2025 task, which proves the rationality of selecting CRF and LinearScheduleWithWarmup through experiments. Additionally, this report proves the effectiveness of the system by comparing to official baseline.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of

deep bidirectional transformers for language understanding. In *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.

Shuang Liu, Hui Yang, Jiayi Li, and Simon Kolmanič. 2021. Chinese named entity recognition method in history and culture field based on bert. *International Journal of Computational Intelligence Systems*, 14:1–10.

Weiming Liu, Bin Yu, Chen Zhang, Han Wang, and Ke Pan. 2018. Chinese named entity recognition based on rules and conditional random field. In *Proceedings of the 2018 2nd International conference on computer science and artificial intelligence*, pages 268–272.

Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. 2021. Named entity recognition and relation extraction: State-of-the-art. *ACM Computing Surveys (CSUR)*, 54(1):1–39.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Yuanhe Tian, Yan Song, Xiang Ao, Fei Xia, Xiaojun Quan, Tong Zhang, and Yonggang Wang. 2020. Joint chinese word segmentation and part-of-speech tagging via two-way attentions of auto-analyzed knowledge. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8286–8296.

Dongbo Wang, Chang Liu, Zhixiao Zhao, Si Shen, Liu Liu, Bin Li, Haotian Hu, Mengcheng Wu, Litao Lin, Xue Zhao, and 1 others. 2023. Gujibert and gujigpt: Construction of intelligent information processing foundation language models for ancient texts. *arXiv preprint arXiv:2307.05354*.

Dongbo Wang, Chang Liu, Zihe Zhu, Jiang, Feng, Haotian Hu, Si Shen, and Bin Li. 2021. Construction and application of pre-training model of siku quanshu oriented to digital humanities.

Miguel Won, Patricia Murrieta-Flores, and Bruno Martins. 2018. ensemble named entity recognition (ner): evaluating ner tools in the identification of place names in historical corpora. *Frontiers in Digital Humanities*, 5:2.

Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. 2017. Improving neural machine translation with conditional sequence generative adversarial nets. *arXiv preprint arXiv:1703.04887*.

Peng Yu and Xin Wang. 2020. Bert-based named entity recognition in chinese twenty-four histories. In *International Conference on Web Information Systems and Applications*, pages 289–301. Springer.