Construction of NER Model in Ancient Chinese: Solution of EvaHan 2025 Challenge

Yi Lu
University of Toronto
tomlu@cs.toronto.edu

Minyi Lei
McMaster University
leim13@mcmaster.ca

Abstract

This paper introduces the system submitted for EvaHan 2025, focusing on the Named Entity Recognition (NER) task for ancient Chinese texts. Our solution is built upon two specified pre-trained BERT models, namely GujiRoBERTa_jian_fan and GujiRoBERTa_fan, and further enhanced by a deep BiLSTM network with a Conditional Random Field (CRF) decoding layer. Extensive experiments on three test dataset splits demonstrate that our system's performance, 84.58% F1 in the closed-modality track and 82.78% F1 in the open-modality track, significantly outperforms the official baseline, achieving notable improvements in F1 score.

1 Introduction

Named Entity Recognition (NER) is one of the most fundamental tasks in natural language processing (NLP), playing a crucial role in understanding ancient Chinese corpus. In ancient Chinese texts, identifying entities such as person names, geographical locations, official titles, book names, and time expressions is particularly challenging due to the archaic language, estoteric grammar, ambiguous boundaries, and diverse annotation schemas. In this work, we present a solution that leverages domain-specific pre-trained BERT models combined with a BiLSTM+CRF architecture. Our approach is designed to effectively capture both the semantic representations provided by the pre-trained language models and the sequential dependencies inherent in the text, which are critical for accurate and effective entity boundary detection.

2 Related Works

Named Entity Recognition (NER) refers to the task of tagging entities in text with their corresponding type. Early studies in NER mainly relied on using hand-crafted rules (Zhang and Elhadad, 2013) and dictionaries (Pomares-Quimbaya et al., 2016) to capture entity patterns, which obtained satisfiable performance on specific fields, while suffering from suboptimal generality and poor scalability on broader use cases. Statistical machine learning techniques, including Hidden Markov Models (HMM) (Baum and Petrie, 1966) and Conditional Random Fields (CRF) (Lafferty et al., 2001) were widely adopted for NER, incorporating contextual features and further improved the NER systems' performance. Recent advancements, including the application of BiLSTM-CRF (Huang et al., 2015) and pre-trained language models such as BERT (Devlin et al., 2019), GPT (Radford and Narasimhan, 2018), ELMo (Peters et al., 2018) and RoBERTa (Liu et al., 2019).

Specifically, BERT-based models utilize attention mechanisms (Vaswani et al., 2017), which allows models to dynamically focus on relevant parts of the input sequence, thereby mitigating the limitations of fixed-size hidden representations in RNN-based models, achieving human-comparable results on English NER benchmarks.

Prior work in modern Chinese NER, such as (Huang et al., 2015), has demonstrated the benefits of integrating contextualized embeddings with CRF for structured prediction. The CRF layer refines predictions by modeling label dependencies and enforcing valid output sequences, a feature that is particularly beneficial when dealing with the complex annotation schemes often encountered in ancient texts.

Ancient Chinese texts pose additional challenges due to significant linguistic differences from modern Chinese, sparse annotated data, and heterogeneous tag schemes. Some

Proceedings of the Second Ancient Language Processing Workshop associated with NAACL 2025, pages 112–116

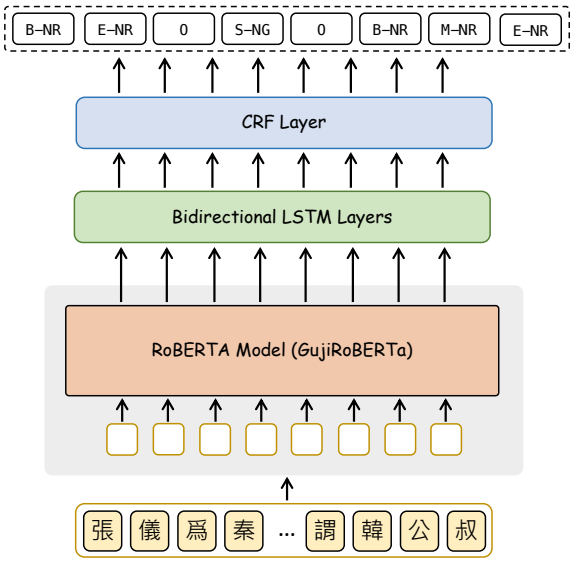

Figure 1: Visualization of our architecture.

works have addressed these issues by building domain-adapted pre-trained models and employing data augmentation or active learning strategies. Our work builds on these advances while specifically tailoring the model and training pipeline for ancient Chinese NER.

3 Model Architecture

We strictly follow the competition requirements by only using the provided pre-trained model: "GujiRoBERTa_jian_fan" (Wang et al., 2023) in the closed-modality track. Our solution is based on a two-branch setting:

Close-Modality. The pre-trained model: GujiRoBERTa_jian_fan (Wang et al., 2023) is applied as the backbone for extracting information from the corpus. Its output representations are fed into a 4-layer Bidirectional LSTM (BiLSTM) with a hidden dimension of 1024 to capture long-range dependencies. A fully connected layer maps the BiLSTM outputs to the label space, and finally, a CRF layer is employed to model the structural constraints among labels, as shown in Figure 1.

Open-Modality. For the open modality track, we adopted the same architecture, and the only difference is the use of the GujiRoBERTa_fan model. Key hyperparameters in our architecture are shown in table Table 1, which is also shared with the close-modality model training.

Specifically, we set the maximum sentence length to 256, as over 98% of sentences fall within this length, according to our analysis of both training and testing datasets. AdamW (Loshchilov and Hutter, 2019) is specifically selected to speed up model training and improve the model's generalization capability. Its base learning rate is set as $2e-5$, along with weight decay being set as 0.01. The gradient clipping is set to 5 to stabilize the model training.

Parameter	Value
Maximum Sentence Length	256
Model Training Batch Size	16
BiLSTM's Hidden Dimension	1024
Number of BiLSTM Layers	4
Optimizer	AdamW
Learning Rate Scheduler	CASwithW
Dropout (in BiLSTM layers)	0.1
Gradient Clipping	5

Table 1: Model and training configuration of our system. "CASwithW" refers to Cosine annealing schedule with warmup.

4 Feature Preprocessing

Our preprocessing pipeline involves the following components:

Sentence Splitting. During Explorative Data Analysis, we observed that multiple non-standard Unicode characters are present in the training corpus. Specifically, we observed that quotation marks contain multiple types. To accommodate the dataset, we utilize the Chinese period as the only splitting character. This ensures that the sentences fed to the model are coherent and that over 98% of sentences are within the maximum length.

Tokenization. We use the tokenizer provided by GujiRoBERTa, which has been pre-trained on an ancient Chinese corpus.

Label Mapping. The label vocabulary is constructed based on the training set with splits: A, B, and C. The training set splits are curated from three ancient documents: Shiji, Twenty-Four Histories, and Traditional Chinese Medicine Classics. Each dataset refers to a distinct NER task, whose tags of interest are shown in Table 2.

As the training data from datasets A, B, and C have heterogeneous tag schemes, we ensure that the mapping covers all tags (including prefixes such as B-, M-, E-) for non-"O" tokens. To ensure the existence of "O" label, we specifically set it as the first element in the label-to-id and id-to-label mappings, where it

corresponds to the id 0.

Dataset	Annotation	Meaning
Split A	NR	Person Name
	NS	Geographical Location
	NB	Book Title
	NO	Official Title
	NG	Country Name
	T	Time Expression
Split B	NR	Person Name
	NS	Geographical Location
	T	Time Expression
Split C	ZD	Disease
	ZZ	Syndrome
	ZF	Medicinal Formula
	ZP	Decoction Pieces
	ZS	Symptom
	ZA	Acupoint

Table 2: Combined Tagset for Named Entities in Datasets A, B, and C (without examples).

5 Experiments

In this section, we demonstrate the experiment results that we conducted while determining the model's architecture design.

5.1 Experimental Setup

We choose overall F1 as the metric for evaluating the model's performance, the metric is compared with the one produced by the baseline model, which is constructed by the committee. The baseline is a simple "SikuRoBERTa-BiLSTM-CRF" model, whose hyperparameters for constructing the BiLSTM and CRF module are not disclosed.

We split the training data with a 90%/10% ratio for the training and validation dataset split. Each model is trained for 40 epochs. During model training, we evaluate the current model at each epoch on the separate test sets of Split A, B, and C. For each split, we evaluate the model's performance on each split using the F1 score metric and update the best F1 that we obtained so far for this dataset split. The best-performed model will be saved and will be used for testing dataset inference after the model training.

The model is trained on a single NVIDIA A6000 40G card. Detailed model training hyperparameter settings are revealed as Table 1.

5.2 Model Architecture Ablation

We compared several configurations regarding the model's architecture. Apart from our final design, 'RoBERTa + BiLSTM + CRF' (where, for simplicity, we refer to the "GujiRoBERTa" model used in both close and open modality as "RoBERTa"), we also experimented multiple configurations. Using the same RoBERTa model, we experimented the use of a single CRF layer or SPAN layer. Due to the constraints of time and computational resources, experiments on model architecture are conducted after the submission deadline.

As Table 3 shows, directly applying a CRF layer on top of the RoBERTa output yielded an F1 score of 79.68%, which is slightly lower than the baseline. By replacing the CRF layer with a double-pointer SPAN layer, an F1 score of 80.62% is achieved, which is similar to the baseline. While such simple combinations do not yield substantial improvements, the combination of RoBERTa + BiLSTM + CRF achieved significant improvement, indicating that the incorporation of a deep BiLSTM layer is crucial for capturing sequential context and dependencies.

Model Arch.	Precision	Recall	F1 Score
Baseline	81.41%	79.82%	80.61%
R+C	78.62%	80.78%	79.68%
R+S	84.65%	76.96%	80.62%
R+B+C	**85.92%**	**83.28%**	**84.58%**

Table 3: Different model architecture's performance comparison (transposed). The highest metric values in each column are highlighted in bold. Acronyms: "R", "C", "B", "S" refers to RoBERTa Model, CRF, BiLSTM and SPAN, respectively.

5.3 Model Hyperparameter Ablation

To determine the optimal model configuration, we performed ablation experiments on two key hyperparameters: "hidden_dim", controlling the hidden dimension size of both the BiLSTM and CRF modules in our system, and "num_layer", setting the depth of the BiLSTM module. Experiment results are summarized in Table 4 and Table 5.

Hidden Dimension	512	1024	1536
Precision	80.67%	**85.92%**	81.87%
Recall	83.71%	83.28%	**84.45%**
F1 Score	82.16%	**84.58%**	83.14%

Table 4: Ablation on Model Width (with 4 BiLSTM layers), highest values are in **bold**.

As the results shown in both Table 4 and Table 5, an appropriate configuration of both

the size of hidden dimension and the number of layers is critical for achieving optimal performance. On NER task. In the ablation on model width (with the number of layers fixed at 4), increasing the hidden dimension from 512 to 1024 results in a substantial improvement in the F1 score (from 82.16% to 84.58%). However, further increasing the hidden dimension to 1536 causes a slight drop in performance (83.14%), suggesting that an excessively large hidden dimension may introduce redundancy or overfitting.

Depth	4 Layers	8 Layers	12 Layers
Precision	**85.92%**	81.79%	82.20%
Recall	83.28%	**83.37%**	81.65%
F1 Score	**84.58%**	82.57%	81.93%

Table 5: Ablation on BiLSTM Module's depth, each layer's hidden dimension are fixed as 1024.

Similarly, in the ablation on model depth (hidden dimension fixed at 1024), the best performance is achieved with 4 layers (84.58%). Increasing the number of layers to 8 and 12 leads to a decrease in the F1 score to 82.57% and 81.93%, indicating that although a deeper model might capture more complex patterns, it may also become prone to overfitting or suffer from optimization difficulties.

Overall, these experiments demonstrate that a balanced configuration, a hidden dimension of 1024, and 4 layers provide the most effective trade-off between model capacity and generalization performance.

6 Performance of Close and Open Modality Track

Following the finding of the above experiments, we utilized the most optimal model architecture design that we came up with: RoBERTa + BiLSTM + CRF, along with BiLSTM hidden dimension: 1024 and depth set as 4.

For the Closed Modality Track, we utilize "GujiRoBERTa_jian_fan" as requested by the competition, reporting an F1 of 84.58% as our final score in this track.

For the Open Modality Track, we reused the model configuration and hyperparameter settings while utilizing an external RoBERTa model: "GujiRoBERTa_fan", which is being pre-trained on ancient Chinese corpus only. We report the F1 as 82.78% as our final score.

7 Future Work

Despite our system's performance on both close and open modality significantly outperforms the baseline: 80.61% F1, there remain avenues for further improvement which were not fully discovered in this work due to time and resource constraints:

Data Augmentation. Employing augmentation strategies such as synonym replacement (with domain-specific ancient Chinese synonym dictionaries) or back-translation could increase data diversity and improve the model's tagging accuracy.

Adversarial Training. Integrating techniques such as the Fast Gradient Method (FGM) during finetuning BERT models could potentially improve the model's robustness.

Model Ensembling. Combining multiple models with diverse architectures (e.g., BERT+Attention+CRF) could further boost performance and improve the F1.

Open-Modality Exploration. For the open-modality track, leveraging Large Language Models (LLMs) and prompt-based approaches to transform the NER task into the generative task or utilizing ModernBERT to develop an ancient Chinese-specific BERT could lead to much stronger models that excel in ancient Chinese NER tasks.

Future work could explore these directions to further push the boundaries of ancient Chinese NER, especially under the challenges posed by heterogeneous tag schemes and limited annotated data.

8 Summary

To conclude, our system is built on strong foundations provided by domain-specific pretrained models and is enhanced by a BiLSTM+CRF architecture with optimal depth and width. Our solution achieves an F1 score of around 84.6% in the closed-modality track and around 82.8% in the open-modality track, significantly outperforms the baseline of 80.61% F1, and demonstrates our design's effectiveness.

9 Limitations

While our system demonstrates strong performance on the EvaHan 2025 NER task, there are several limitations. Firstly, our system

is built around the provided pre-trained "GujiRoBERTa" models, which may limit generalization to texts beyond the training domain or unseen linguistic variations in ancient Chinese corpuses. Secondly, due to time and computational resource constraints, we were unable to perform more comprehensive hyperparameter tuning or explore alternative architectures such as Transformer-CRF models in greater depth. Lastly, the diversity in annotation schemes across datasets A, B, and C poses challenges for unified modeling, which were only partially addressed in our current implementation.

References

Leonard E. Baum and Ted Petrie. 1966. Statistical inference for probabilistic functions of finite state markov chains. *Annals of Mathematical Statistics*, 37:1554–1563.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 282–289. Morgan Kaufmann.

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. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Alexandra Pomares-Quimbaya, Alejandro Sierra Múnera, Rafael Andrés González Rivera, Julián Camilo Daza Rodríguez, Oscar Mauricio Muñoz Velandia, Angel Alberto Garcia Peña, and Cyril Labbé. 2016. Named entity recognition over electronic health records through a combined dictionary-based approach. In *International Conference on ENTERprise Information Systems/International Conference on Project MANagement/International Conference on Health and Social Care Information Systems and Technologies, CENTERIS/ProjMAN/HCist 2016, Porto, Portugal, October 5-7, 2016*, volume 100 of *Procedia Computer Science*, pages 55–61. Elsevier.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

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

Shaodian Zhang and Noemie Elhadad. 2013. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. *J. Biomed. Informatics*, 46(6):1088–1098.