# GRoWE: A GujiRoBERTa-Enhanced Approach to Ancient Chinese NER via Word-Word Relation Classification and Model Ensembling

**Tian Xia, Yilin Wang, Xinkai Wang, Yahe Yang, Qun Zhao, Menghui Yang***

[1]School of Information Resource Management,Renmin University of China, Beijing, China,
[2]Midu Technology Co., Ltd., Shanghai, China

{xiat,wang_yilin}@ruc.edu.cn, xinkaiw18@gmail.com, owyangyahe@126.com,{zhao_qun,yangmenghui}@ruc.edu.cn

## Abstract

Named entity recognition is a fundamental task in ancient Chinese text analysis.Based on the pre-trained language model of ancient Chinese texts, this paper proposes a new named entity recognition method GRoWE. It uses the ancient Chinese texts pre-trained language model GujiRoBERTa as the base model, and the word-word relation prediction model is superposed upon the base model to construct a superposition model. Then ensemble strategies are used to multiple superposition models. On the EvaHan 2025 public test set, the F1 value of the proposed method reaches 86.79%, which is 6.18% higher than that of the mainstream BERT_LSTM_CRF baseline model, indicating that the model architecture and ensemble strategy play an important role in improving the recognition effect of naming entities in ancient Chinese texts.

## 1 Introduction

As an important carrier of Chinese history and culture, ancient Chinese texts preserve thousands of years of civilization and wisdom. It is essential to obtain the information contained in them. Named entity recognition (NER), a crucial natural language processing technique, plays an indispensable role in information extraction from ancient Chinese texts(Long et al., 2016). NER aims to extract entities such as person name, book title, official title and so on, providing a foundation for understanding ancient Chinese texts and constructing ancient Chinese knowledge graphs.

However, NER in ancient Chinese texts faces many challenges. First, Old Chinese (classical Chinese) is highly concise, lacks clear boundary markings between words, and the subject or object is often omitted in sentence structure, which greatly increases the difficulty of identifying named entities. Second, due to historical changes and the diversity of textual contexts, the same word or phrase may refer to completely different entities in different contexts. In addition, in order to promote the task of NER in ancient Chinese texts, the organizers of EvaHan 2025 release a unified dataset and pre-trained pedestal model, hoping to promote the progress of the named entity recognition task through a unified standard.

In this study, we propose a named entity recognition method GRoWE (GujiRoBERTa + Word-Word Relation Prediction + Ensemble) suitable for the characteristics of ancient Chinese texts: based on the pre-trained basic model of ancient Chinese texts GujiRoBERTa, the word-word relation processing of W2NER is reused, and a multi-model ensemble strategy is introduced. On the dataset published in EvaHan 2025, the method achieves significantly better results than the public baseline model.

## 2 Related work

Early NER research primarily relied on rule-based approaches and statistical models; however, the landscape transformed significantly with the advent of deep learning techniques. The BiLSTM-CRF architecture emerged as a pivotal innovation (Huang et al., 2015). This framework established a fundamental paradigm in the NER domain. The application of this method to person name recognition in ancient Chinese literature has shown promising results (Zhang et al., 2021).

In 2018,BERT was proposed(Devlin et al., 2019), characterized by its large-scale unsupervised pre-training and bidirectional Transformer architecture, which delivers robust contextual representations for NER tasks. Extending this technological foundation, the HistoryNER dataset came into being through the implementation of a BERT-BiLSTM-CRF architecture(Liu et al., 2021), specifically engineered for the identification of entity types within historical Chinese texts.

In terms of architectural advancements in NER,



*Proceedings of the Second Ancient Language Processing Workshop associated with NAACL 2025*, pages 141–145
May 4, 2025. ©2025 Association for Computational Linguistics

a unified MRC framework(Li et al., 2022b) materialized, reconceptualizing NER as a machine reading comprehension challenge. This innovative approach extracts entities via natural language queries, harnessing prior knowledge embedded within these queries to enhance the model's comprehension of entity categories. Concurrently, the field benefited from the development of the Global Pointer model(Su et al., 2022), which incorporates relative position encoding and multi-head attention mechanisms, substantially improving the detection of nested and lengthy entities. The evolution of NER methodologies further progressed with the conceptualization of the W2NER framework(Li et al., 2022a), which elegantly models adjacency relationships between entities through word-word relation classification. This novel perspective addresses critical limitations in conventional approaches when handling overlapping and discontinuous entities, contributing significantly to the field with its exceptional capability in processing complex entity structures.

# 3 Method

## 3.1 Model Selection

Mainstream named entity recognition methods use an encoder pre-trained language model to obtain the semantic information of the text, and integrate the features obtained from the encoder into the personalization module for further feature transformation, ultimately outputting probability distribution on different labels. The common personalization modules are as follows: LSTM+CRF combination method, GlobalPointer method which supports multi-head recognition of nested entities, and W2NER method for predicting relations between word pairs. In addition, named entity recognition can be regarded as a reading comprehension task, entities as problems, and text to be labeled as documents, and named entity recognition can be realized by marking the location of the problem in the document.

We cut the training dataset into consistent five-fold divisions, train the four methods mentioned above, and test the model performance. The specific experimental data are shown in Table 1. Based on the model performance, we select the W2NER as the personalized module for the ancient Chinese text named entity recognition task.

## 3.2 Architecture

To better adapt to the characteristics of ancient Chinese texts, we designed the network architecture GRoWE as shown in Figure 1. In order to reflect the characteristics of ancient Chinese texts, the encoder is GujiRoBERTa-jian-fan, a pre-trained model for ancient Chinese texts in the EvaHan2025 part, and the output results of the encoder are further input into the bidirectional LSTM layer, which encodes the features in both directions, and then is sent to the W2NER module, after encoding through Convolution Layer and Co-Predictor Layer, the relation between word pair is classified and the logits are calculated.

In order to explore the potential of model combination, we adopted a multi-model ensemble strategy: the training set was divided into five folds, and the model was trained using the data from four of the folds in turn, resulting in a total of five models. Then, the logits of these five models were directly summed up to form the final ensemble result vector, which was then decoded to obtain the final labeled result.

## 3.3 Main Process

The main process can be formally described as follows:

**Input Representation**: Given an input sentence $X = \{x_1, x_2, \ldots, x_N\}$, the BERT + BiLSTM model generates contextual word embeddings $H = \{h_1, h_2, \ldots, h_N\}$, where $h_i \in \mathbb{R}^{d_h}$ and $d_h$ represents the embedding dimension.

**Word Pair Embedding Computation**: Subsequently, the word-pair embedding $V_{i,j}$ is computed as:

$$V_{i,j} = \gamma_{i,j} \odot \left( \frac{h_j - \mu}{\sigma} \right) + \lambda_{i,j}$$

where:
- $\gamma_{i,j} = W_\alpha h_i + b_\alpha$
- $\lambda_{i,j} = W_\beta h_i + b_\beta$
- $\mu = \frac{1}{d_h} \sum_{k=1}^{d_h} h_{j,k}$
- $\sigma = \sqrt{\frac{1}{d_h} \sum_{k=1}^{d_h} (h_{j,k} - \mu)^2}$
- The symbol $\odot$ denotes element - wise multiplication.

**Multi - Layer Dilated Convolution Application**: Then, multi-layer dilated convolutions (DConv) are applied to $V$:

$$Q^l = \sigma\left(DConv_l(V)\right), \quad l \in \{1, 2, 3\}$$

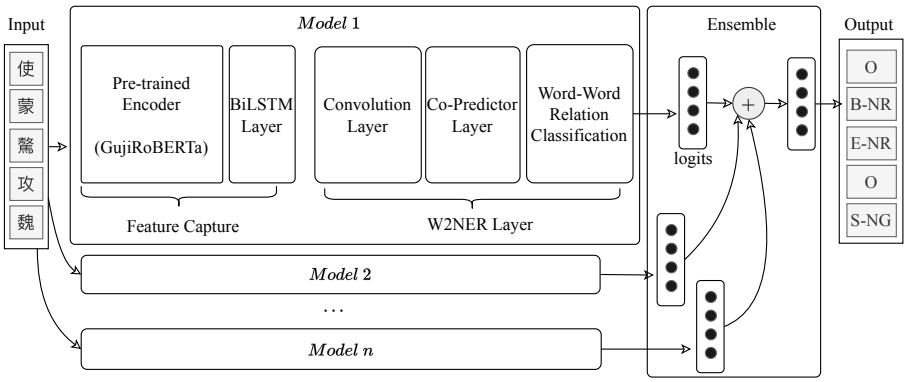


Figure 1: Architecture of GRoWE

The final representation is $Q = [Q^1; Q^2; Q^3] \in \mathbb{R}^{N \times N \times 3d_c}$, where $d_c$ is the hidden dimension of the convolution operation.

**Word-Pair Relation Probability Prediction**: After that, word-pair relation probabilities are predicted by combining Biaffine and MLP:

$$y_{i,j} = Softmax(y'_{i,j} + y''_{i,j})$$

where:

- Biaffine prediction: $y'_{i,j} = s_i^\top U o_j + W[s_i; o_j] + b$
- MLP prediction: $y''_{i,j} = MLP(Q_{i,j})$
- $s_i = MLP_2(h_i)$ and $o_j = MLP_3(h_j)$

**Inference and Aggregation**: During inference, the logits from each model's Co-Predictor Layer are aggregated. For $M$ models, the final logits are the sum of individual model outputs:

$$logits_{final} = \sum_{m=1}^{M} \left( y_{i,j}^{\prime(m)} + y_{i,j}^{\prime\prime(m)} \right)$$

The predicted relation is determined by selecting the label with the highest score after applying Softmax:

$$y_{i,j} = Softmax(logits_{final})$$

where $y_{i,j}^{\prime(m)}$ and $y_{i,j}^{\prime\prime(m)}$ denote the Biaffine and MLP outputs of the $m$-th model.

**Relation Decoding**: For relation decoding, a directed graph is constructed based on the predicted relations, and entities are extracted via depth-first search.

## 4 Data and experiments

### 4.1 Data preprocessing

The datasets used in this paper are from Shiji (dataset A), Twenty-Four Histories (dataset B), and Traditional Chinese Medicine Classics (dataset C) provided by evahan2025, and are divided into two parts, the training set and the test set. In this study, the preprocessing work is carried out on datasets A, B, and C, and data enhancement process is further carried out on the data in the training set. The preprocessing mainly includes sentence segmentation and de-duplication, while the data enhancement is based on sliding window operations to enrich the training data.

For the data preprocessing, in the initial stage, we cut all the datasets into sentences based on the single character break strategy, using "." , "!" ".", "!", "?" ", "!", "?", ";" as the sentence termination identifiers. For long sentences longer than 120 characters, we implemented a secondary cutting strategy, i.e., we selected the nearest comma or the stop sign to the length of 120 characters as the cutting point to ensure that the cut sentence remained within the appropriate length range. Subsequently, de-duplication is performed on the cut sentences, and the final number of sentences in the training set is obtained as follows: dataset A (8,218 sentences), dataset B (3,507 sentences), and dataset C (6,545 sentences), respectively. After carrying out the same operations, the number of sentences are obtained for the test set: dataset A (994 sentences), dataset B (884 sentences) and dataset C (1,115 sentences).

To enhance the data diversity, we use the sliding window technique to expand the samples in the training set. Specifically, for the preprocessed sentence sequence, $[s_1, s_2, s_3, s_4, \ldots, s_n]$, we implement a sliding window operation with an incremental step of 1 for each sentence $s_i$, generate candidate sets $[s_i, s_i + s_{i+1}, ..., \ldots, s_i + ... + s_n]$, and screen the combined sentences with a length of less than 120 characters as the expanded train-

ing samples.Through this processing, the number of training samples in datasets A, B, and C was expanded to 48,797, 14,629, and 65,227, respectively,totaling 128,653 training samples.

For the expanded sample, we implemented stratified sampling according to the sentence entity type, and divided the data into 5 equal parts for cross-validation. It's worth noting that this segmentation strategy can lead to high evaluation metrics due to the potential risk of data breaches. Specifically, when the original sentence that $s_i$ with its derivative samples $s_i + s_{i+1}$. Based on the fact that stratified sampling is randomly assigned to the training and test sets, the models may obtain some information of the test samples from the training data. However, considering that this division only serves the model selection session, and all the comparison models are evaluated under the same dataset and evaluation system, this division is acceptable in the context of this study.

After completing the model selection, we used the sequential cutting method to re-divide the data into five equal parts. In this way, we retrained and reevaluated the optimal models to ensure the reliability and fairness of the final results.

### 4.2 Parameter Settings

When comparing different methods, we use SikuBERT as the base model and the default parameters from each method's public code for training.

For the GRoWE method, the base model is GujiRoBERTa-jian-fan. The training settings are: batch size = 32, learning rate = 5e-6, seed=1234, and 14 epochs in total. We use the model from the last epoch to predict on the test set.

### 4.3 Comparative experiments with mainstream models

In this study, we systematically evaluate the performance of multiple architectures using BERT as encoder in the task of named entity recognizing in ancient Chinese texts, including BERT-BiLSTM-CRF, GlobalPointer, MRC, and W2NER. While keeping the encoder consistent, the experimental results are shown in Table 1, which show that the model based on W2NER architecture performs optimally among all the evaluated scenarios, with F1-score of 88.48%, which is significantly better than other architecture combinations.

Based on this finding, we further use the GRoWE architecture depicted in Figure 1 to train the model and validate it on the public test set.

### 4.4 Comparison of experiments on the Test set

The method of comparison is as follows:

Method1: RoBERTa-BiLSTM-CRF, the official baseline model announced by EvaHan 2025, the pre-trained language model corresponding to the encoder is SikuRoBERTa, and the personalized insertion module is BiLSTM+CRF.

Method2: RoBERTa-W2NER,following the official requirements of EvaHan 2025, the pretrained language model corresponding to the encoder is GujiRoBERTa-jian-fan, and the output of the W2NER Layer is directly used to decode the output for named entity recognition.

Method3: GRoWE,the method proposed in this paper, utilizes five models obtained from five-fold cross-training of the training set, and logits cumulative ensemble is used to obtain the final results.

As shown in Table 2, the combination of the proprietary pre-trained model and W2NER is significantly better than the baseline model, indicating that the W2NER layer embodies a stronger entity recognition ability through convolution and word-word relation prediction processing. GRoWE further adopted the ensemble learning strategy, and the performances were further improved, with the F1-score increasing by 6.18 percentage points compared with the baseline model, and outperforming other methods in all test subsets. It can be seen the word-word relation prediction model can improve the recognition effect of named entities, and the ensemble framework reduces the prediction bias of a single model by integrating the prediction advantages of multiple models.

## 5 Conclusion

This paper proposes a new named entity recognition method GRoWE. It uses the ancient Chinese texts pre-trained language model as the base model, and the W2NER word-word relation prediction model is superposed upon the base model to construct a superposition model. Then ensemble strategies are used to multiple superposition models.The public test set results show that the GRoWE method significantly outperforms the baseline model and improves the overall recognition effect of ancient Chinese texts NER.

## 6 Acknowledgements

This research is supported by the the National Social Science Fund of China (22BTQ068).

| Model Name | Indicator | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Average |
|---|---|---|---|---|---|---|---|
| BERT-BiLSTM-CRF | Precision | 85.46% | 85.88% | 85.85% | 84.38% | 84.92% | 85.30% |
| | Recall | 86.73% | 87.47% | 88.25% | 87.68% | 87.91% | 87.61% |
| | F1 Score | 86.09% | 86.67% | 86.52% | 86.00% | 86.39% | 86.33% |
| GlobalPointer | Precision | 86.21% | 86.25% | 86.08% | 85.63% | 85.07% | 85.85% |
| | Recall | 89.10% | 89.12% | 88.65% | 88.20% | 89.48% | 88.91% |
| | F1 Score | 87.61% | 87.64% | 87.32% | 86.88% | 87.19% | 87.33% |
| MRC | Precision | 84.58% | 86.39% | 86.29% | 82.84% | 85.03% | 85.03% |
| | Recall | 81.12% | 82.90% | 82.22% | 82.79% | 80.27% | 81.86% |
| | F1 Score | 82.81% | 84.61% | 84.20% | 82.81% | 82.58% | 83.40% |
| W2NER | Precision | 89.60% | 89.88% | 89.32% | 89.05% | 90.19% | 89.61% |
| | Recall | 87.99% | 87.68% | 86.64% | 87.06% | 87.55% | 87.38% |
| | F1 Score | 88.79% | 88.77% | 87.96% | 88.05% | 88.85% | 88.48% |

Table 1: Performance Comparison of Different Models with BERT Encoder across 5-fold Cross-validation

| Model Name | Indicator | Test A | Test B | Test C | Average |
|---|---|---|---|---|---|
| RoBERTa-BiLSTM-CRF (Baseline) | Precision | 85.90% | 87.09% | 71.84% | 81.41% |
| | Recall | 77.50% | 87.92% | 72.95% | 79.82% |
| | F1 Score | 81.48% | 87.50% | 72.40% | 80.61% |
| RoBERTa-W2NER | Precision | 88.17% | 89.55% | 79.53% | 85.47% |
| | Recall | **82.10%** | 89.96% | **88.64%** | **87.24%** |
| | F1 Score | 85.03% | 89.76% | 83.83% | 86.34% |
| GRoWE | Precision | **88.97%** | **90.22%** | **81.33%** | **86.64%** |
| | Recall | 81.45% | **90.34%** | 87.91% | 86.94% |
| | F1 Score | **85.04%** | **90.28%** | **84.49%** | **86.79%** |

Table 2: The Results of the Experiment on the Test set

## 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv preprint*. ArXiv:1508.01991 [cs].

Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2022a. Unified Named Entity Recognition as Word-Word Relation Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10965–10973. Number: 10.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2022b. A Unified MRC Framework for Named Entity Recognition. *arXiv preprint*. ArXiv:1910.11476 [cs].

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):163.

Yunfei Long, Dan Xiong, Qin Lu, Minglei Li, and Chu-Ren Huang. 2016. Named Entity Recognition for Chinese Novels in the Ming-Qing Dynasties. In *Chinese Lexical Semantics*, pages 362–375. Springer, Cham. ISSN: 1611-3349.

Jianlin Su, Ahmed Murtadha, Shengfeng Pan, Jing Hou, Jun Sun, Wanwei Huang, Bo Wen, and Yunfeng Liu. 2022. Global Pointer: Novel Efficient Span-based Approach for Named Entity Recognition. *arXiv preprint*. ArXiv:2208.03054 [cs].

Hailin Zhang, Hai Zhu, Junsong Ruan, and Ruoyao Ding. 2021. People name recognition from ancient Chinese literature using distant supervision and deep learning. In *2021 2nd International Conference on Artificial Intelligence and Information Systems*, ICAIIS 2021, pages 1–6, New York, NY, USA. Association for Computing Machinery.