International Journal of Computer Science and Information Technology
ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 2, Number 1, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v2n1.47
Journal homepage: https://wepub.org/index.php/IJCSIT/index

A Method for Ancient Book Named Entity Recognition Based on BERT-Global Pointer

Wen Jiang*

College of Software Engineering, Chengdu University of Information Technology, Chengdu, China

ABSTRACT

Correct identification of entities in ancient books and documents is the basic step of analyzing ancient Chinese texts, and provides an important prerequisite for in-depth mining of humanistic knowledge in ancient books and documents. In CCL2023 named entity recognition task of ancient books, according to the task definition and the re-quirements of the task organizer, this paper proposes the BERT Global Pointer named entity recognition model; Fine tune the field adaptation training based on the unlabeled 24 history ancient book text data; SWA, FGM, cross validation and post-processing are used to improve the recognition accuracy of the model. The experimental results show that the model and the strategy proposed in this paper have good recognition effect in the multi dynasties, cross domain ancient book entity recognition scene. F1 value on the final line reaches 95.083%.

KEYWORDS

Ancient Texts of Twenty-Four Histories, Named Entity Recognition, Domain-Adaptive Pretraining, Model Fusion, Adversarial Training

1. INTRODUCTION

Named Entity Recognition (NER) is an important task in natural language processing. Its purpose is to identify specific types of entities such as persons, locations, and organizations in text, providing crucial support for subsequent tasks like information extraction and sentiment analysis. Due to its irreplaceable role across various application scenarios, NER has become one of the focal points of research in both academia and industry.

Correctly identifying entities such as personal names, official titles, and book titles in ancient texts is a fundamental step in analyzing and processing classical Chinese texts. As illustrated in Figure 1, recognizing the entity "蒲家奴" in the text involves analyzing relationships between characters and changes in official titles, which are essential prerequisites for interpreting the underlying meanings and deeply exploring the humanities knowledge described in ancient texts.

In recent years, there has been considerable research interest in the recognition of named entities in ancient texts such as historical records, poetry, and traditional Chinese medicine. Zhou et al. [1]2022 conducted work on extracting entities from ancient poetry and constructing related knowledge graphs. Song et al. [2] 2020 enhanced entity recognition performance in traditional Chinese medicine texts using dictionary information. Yu et al.[3] 2020 proposed an end-to-end solution without domain knowledge to identify entities in the Twenty-Four Histories. Currently, there is a need for further exploration to improve the performance of entity recognition in the Twenty-Four Histories, which cover multiple dynasties and cross various domains.

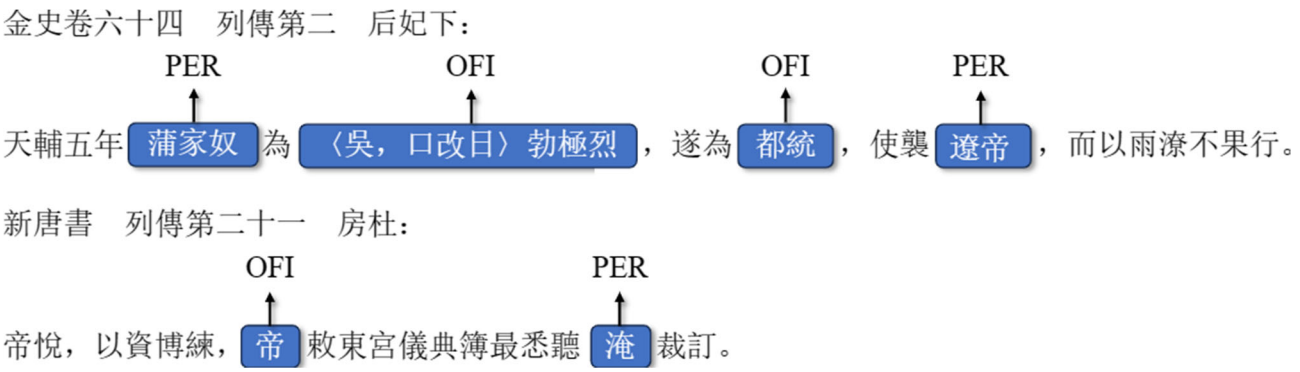

Figure 1. Descriptions of ancient texts from different dynasties

Compared to traditional entity recognition tasks in ancient texts, the task of entity recognition in the Twenty-Four Histories presents its own unique challenges. Firstly, the corpus for named entity recognition in the Twenty-Four Histories consists of texts from multiple dynasties and different fields. As illustrated in Figure 1, there are significant differences between the names of official positions during the Jin Dynasty and the Tang Dynasty, each possessing distinct characteristics specific to their respective dynasties.

Secondly, the corpus for named entity recognition in the Twenty-Four Histories contains characteristics such as full-text traditional Chinese characters, high language complexity, and ambiguity in entity references. As shown in Figure 1, the entity "帝" (emperor) exhibits different entity variations in different contextual settings, adding complexity and difficulty to the task of named entity recognition in the ancient texts of the Twenty-Four Histories.

Lastly, the limited amount of data provided for the evaluation task and the imbalance between entity types increase the difficulty of model recognition.

Therefore, in the evaluation task of ancient named entity recognition at CCL2023[4], this paper explores the task of named entity recognition based on the Twenty-Four Histories, according to the data provided by the task organizers and the requirements of the task definition. Addressing the aforementioned issues, this paper proposes a BERT-Global Pointer-based model for ancient entity recognition. Through domain adaptation fine-tuning based on unlabeled texts from the Twenty-Four Histories, and employing strategies such as Stochastic Weight Averaging (SWA)[5], Fast Gradient Method (FGM)[6], cross-validation, and post-processing, the overall recognition accuracy of the model is enhanced.

Experimental results demonstrate that the proposed BERT-Global Pointer model for ancient entity recognition performs well in the context of ancient texts spanning multiple dynasties and domains. Additionally, the domain adaptation pre-training strategy employed in this paper effectively improves the overall performance of entity recognition by approximately 0.31%. Techniques such as SWA and FGM contribute to enhancing model generalization and robustness. With cross-validation and rule-based post-processing correction, the final entity recognition performance achieves an F1 score of 95.083%.

2. METHODS

2.1. Problem Definition

Given a classical text $x = \{x_1, x_2, ..., x_n\}$ from the Twenty-Four Histories, where L is the length of the classical text x, and m is the number of entity types, Global Pointer can establish m L * L scoring matrices. These matrices assign scores to any two tokens corresponding to the input text. If the tokens from the x_i-th to the x_j-th token form an entity, then the corresponding entity prediction table predicts

true at the i-th row and j-th column, indicating the respective entity category, with the rest of the positions being false.

2.2. Model Structure

This paper builds upon the BERT-Global Pointer model for the task of Named Entity Recognition (NER) in the Twenty-Four Histories classical texts. The basic framework of this model is illustrated in Figure 2.

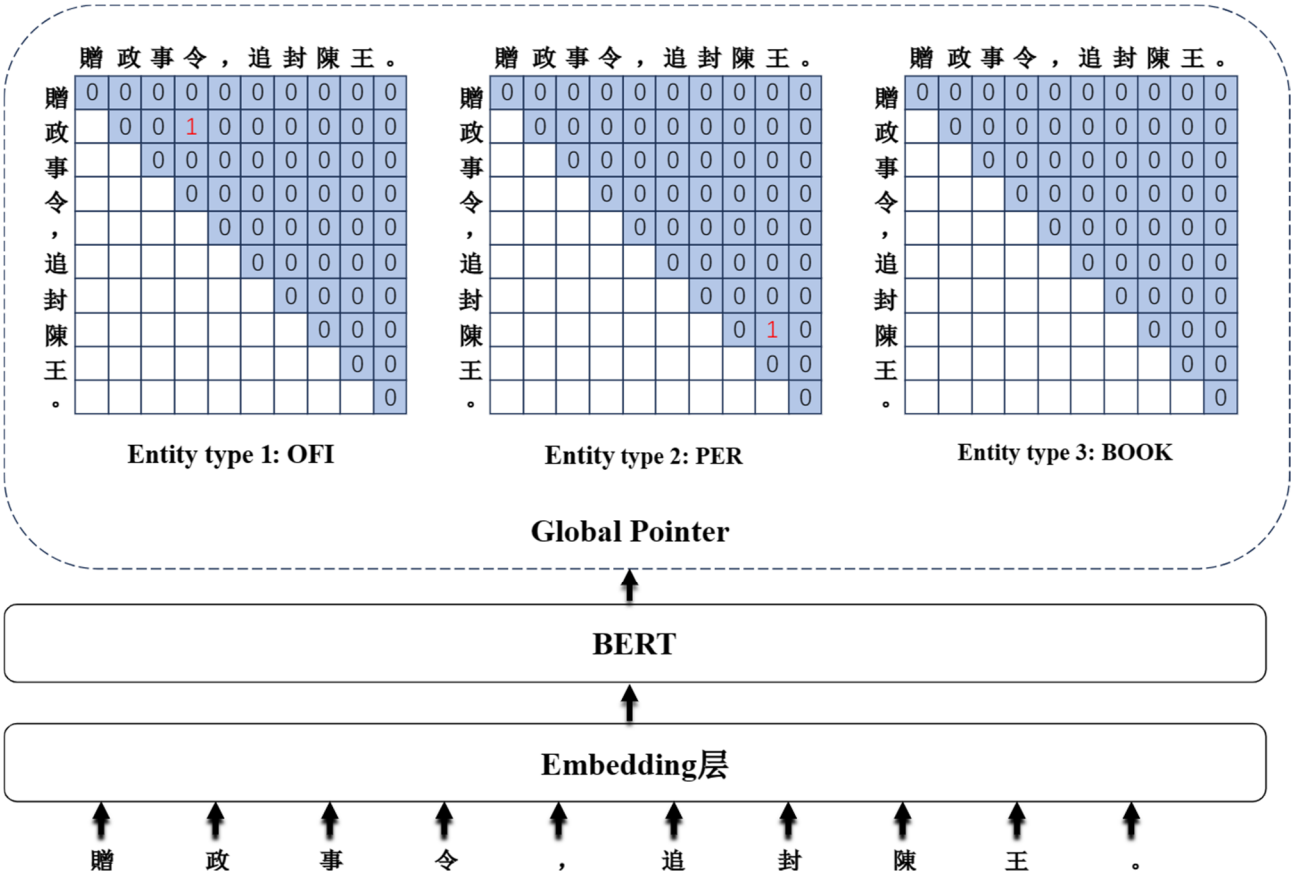

Figure 2. Model architecture

As shown in Figure 2, the classical texts from the Twenty-Four Histories are inputted at the character level. Firstly, character embeddings are obtained through an Embedding layer. Secondly, embeddings with rich contextual semantic information are acquired through the BERT [7] layer. Finally, decoding is performed using the Global Pointer [8].

2.3. Global Pointer

The main idea of Global Pointer[8] is as follows: Suppose we aim to identify entities in a text sequence of length n, assuming there is only one type of entity to be recognized, and each entity to be recognized is a contiguous segment of this sequence, of arbitrary length, and can be nested within each other (i.e., there may be intersections between two entities). Then, there are n(n + 1)/2 "candidate entities" in this sequence, meaning that a sequence of length n contains n(n + 1)/2 different contiguous subsequences, which encompass all possible entities. The task is to select the true entities from these n(n + 1)/2 "candidate entities", which is equivalent to a "n(n + 1)/2 choose k" multi-label classification problem. If there are m types of entities to be recognized, then m separate "n(n + 1)/2 choose k" multi-label classification problems are formulated.

2.4. Adversarial Training

Adversarial Training is a training method that introduces noise into the training process. By adding perturbations to the samples while trying to keep the distribution of the original samples unchanged, the model learns to ignore such perturbations, thereby enhancing the model's robustness. In this paper, we experimented with adversarial methods such as FGM (Fast Gradient Method), PGD (Projected Gradient Descent) [9], and a combination of FGM and PGD. Through online experiments, we found that FGM outperformed PGD and the combination of FGM and PGD in this task.

2.5. Model Fusion

The primary idea of Stochastic Weight Averaging (SWA) is to average the weights of the model during the training process to obtain a more stable and better-performing model in terms of generalization. Specifically, in SWA, multiple models generated during the training process are weighted averaged rather than just using the final model. This approach smooths out fluctuations in the model during training and reduces the risk of overfitting, thus improving the model's generalization ability. Moreover, SWA can improve the accuracy and efficiency of the model within a smaller number of iterations. The algorithmic process of SWA is as follows:

1. Train the model using standard stochastic gradient descent or other optimization algorithms.
2. During training, save a set of model parameters at regular intervals.
3. After training, perform weighted averaging on the saved sets of parameters to obtain the SWA model.
4. Use the SWA model for inference or further fine-tuning.

2.6. Cross-validation and post-processing

K-fold cross-validation is a common method for evaluating machine learning models. Its main purpose is to comprehensively assess the model using limited data and select the optimal model parameters. The original dataset is divided into K subsets, with one subset used as the validation set and the rest as the training set. Then, the model is trained on the training set and evaluated on the validation set. This process is repeated K times, with different subsets chosen as the validation set each time until all subsets have been used for validation. Finally, the evaluation results are averaged to obtain the final evaluation result of the model. In this paper, we adopt five-fold cross-validation and use the predictions from the five models generated by five-fold cross-validation for voting.

Voting among the predictions of models trained with five-fold cross-validation may lead to errors in the prediction results due to diversity in fusion. In this paper, we use rule-based post-processing methods to correct the prediction results to improve the precision of the predictions. The main cases to be addressed are as follows:

1. For entities with obvious errors in the prediction results, such as entities containing characters such as "、" and "。", the predicted entities are deleted and saved in non-entity form.
2. For entities that appear in the training set but are not correctly predicted in the test set, a method of building a corresponding entity list is used to match the string and replace it with the entity form.

3. EXPERIMENT

3.1. Dataset

The training dataset is based on the "Twenty-Four Histories" corpus, containing 22 volumes of text from 13 books. It is randomly truncated into segments of approximately 100 characters each. Each segment contains entities marked with curly braces "{}", followed by the entity type indicated by "|" such as person names (PER), book titles (BOOK), and official titles (OFI), totaling 154,000 characters (including punctuation). There are a total of 2347 segments in the training set. The distribution of entity counts is shown in Figure 3, and the distribution of entity lengths is shown in Figure 4. The test dataset consists of 224 segments.

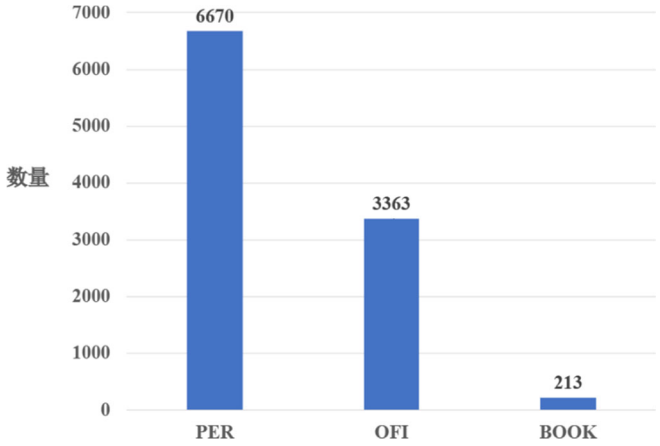

Figure 3. Entity counts

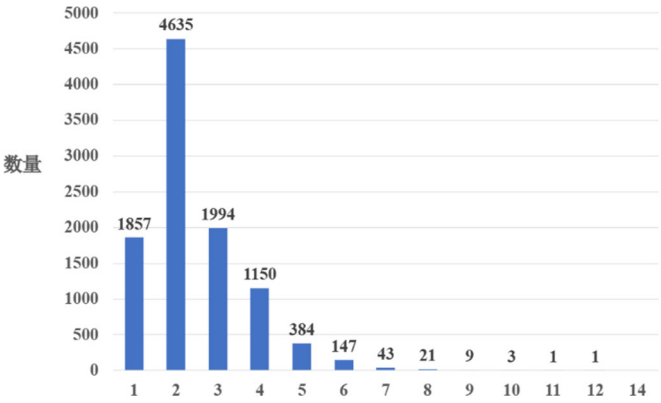

Figure 4. Entity lengths

3.2. Experimental Settings

When fine-tuning BERT using unlabeled data from the Twenty-Four Histories corpus for domain adaptation, a total of 137,042 unlabeled data samples were selected. The initial learning rate was set to 2e-5, batch size to 32, maximum sentence length to 128, and training epochs to 100.

For the proposed model in this paper, during training, the initial learning rate was set to 4e-5, AdamW optimizer was used, batch size was set to 32, maximum sentence length was set to 256. SWA was applied for model fusion, selecting epochs between 6 and 40. FGM conducted adversarial training for each epoch. SpatialDropout was set to 0.3, and the training epochs were set to 45.

3.3. Domain Adaptation Pretraining

In this paper, we utilized open-source unlabeled data from the Twenty-Four Histories, including 13 historical books such as "Northern History," "Later Tang History," and "History of Jin," totaling

137,042 unlabeled data samples. After removing duplicates from the test dataset, we employed domain adaptation pretraining to fine-tune BERT.

To achieve better pretraining results, we experimented with various pretrained models such as skiubert, bert-ancient-base[10], and the general domain model nezha-cn-base[11]. Through comparison, we found that bert-ancient-base performed the best. Bert-ancient-base offers the following advantages:

1. Many classical Chinese texts contain traditional Chinese characters and numerous rare characters, which are absent in part of the vocabulary of pretrained models. Bert-ancient-chinese, by learning from extensive corpora, further expands the vocabulary of pretrained models. Its final vocabulary size is 38,208, compared to 21,128 for bert-base-chinese and 29,791 for siku-bert. Bert-ancient-chinese possesses a larger vocabulary and includes more rare characters, which is conducive to enhancing the model's performance in downstream tasks.
2. Bert-ancient-chinese employs a larger training dataset. Compared to siku-bert, which only uses "Siku Quanshu" as the pretraining dataset, bert-ancient-chinese utilizes a larger dataset (approximately six times the size of "Siku Quanshu"), covering a wider range of content from various categories such as history, philosophy, literature, and medicine.
3. Based on the idea of domain-adaptive pretraining, bert-ancient-chinese continues training on top of bert-base-chinese by incorporating classical Chinese texts, aiming to obtain a pretrained model tailored for the automatic processing of classical texts.

Therefore, in this paper, we conducted domain adaptation pretraining based on bert-ancient-base. Since bert-ancient-chinese is derived from bert-base-chinese with additional training on classical Chinese texts, we continued to use the conventional MASK strategy and adopted a dynamic MASK policy during training, randomly masking 15% of tokens each iteration, generating new MASKed texts to enhance the robustness of model training.

3.4. Data preprocessingomain Adaptation Pretraining

The training data for this evaluation task consists of a total of 2347 samples. Through statistical analysis, it was found that there is some noise in the original training dataset, such as garbled characters shown in Figure 5, and issues like labeling errors and inconsistencies in the corpus. This paper preprocesses and cleans the text of both training and testing data, including removing noise characters and correcting entity positions. The preprocessing results are shown in Figure 5.

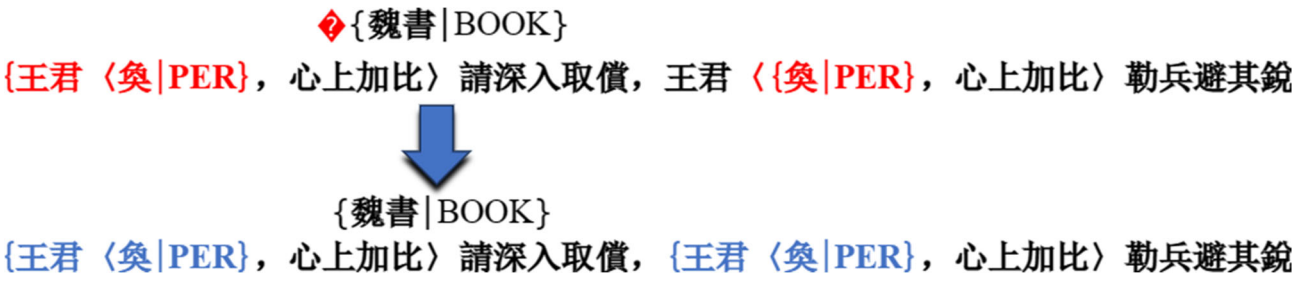

Figure 5. Preprocessed data results

3.5. Experimental Results

In the task of named entity recognition (NER) in the classical texts of the Twenty-Four Histories, this paper adopts character-level F1 score as the evaluation metric. To further validate the effectiveness of the proposed model in practical NER applications, this paper sequentially compares the online test results of the baseline model with enhancements including SWA, FGM, domain adaptation pretraining, cross-validation, and post-processing correction strategies. As shown in Figure 6, the

experimental results indicate that the BERT-Global Pointer model proposed in this paper exhibits good recognition capabilities in the multi-dynasty, cross-domain scenario of named entity recognition in the classical texts of the Twenty-Four Histories.

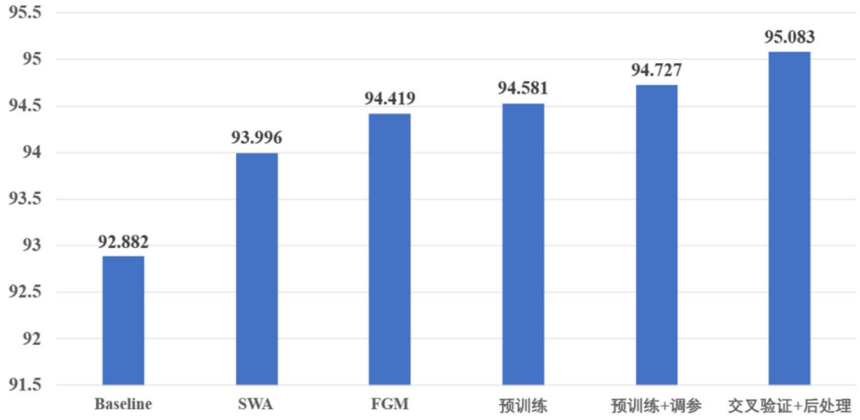

Figure 5. Experimental results

Firstly, as shown in Figure 6, after enhancing the model's generalization ability and robustness through SWA and FGM, the entity recognition performance of the model is improved by approximately 1.6% compared to the baseline. Further analysis in Table 1 reveals a significant improvement in the recognition accuracy of entities labeled as "PER" and "OFI", reducing erroneous entity predictions made by the baseline model.

Secondly, the proposed strategy of domain adaptation pretraining using unlabeled data introduces textual information from the domain, enhancing the model's learning of contextual semantic associations. Under appropriate parameters, the model's entity recognition capability is improved by approximately 0.31%. Observation from Table 1 indicates that the improvement mainly stems from the recognition of entities labeled as "OFI".

Table 1. Predicted label statistics

Method	Label		
	PER	OFI	BOOK
Baseline	681	332	11
SWA+FGM	674	328	10
Domain Pretraining	672	338	9
5-fold+post processing	672	340	9

Finally, by conducting cross-validation to obtain comprehensive entity recognition performance and applying post-processing rules for correction, the method proposed in this paper achieves a final online F1 score of 95.083%.

4. CONCLUSION

In the CCL2023 Named Entity Recognition (NER) evaluation task for classical texts, this paper, in accordance with the requirements of the task organizers and based on the task definition, addresses the issues encountered in the task by proposing the BERT-Global Pointer model for recognizing named entities in classical texts. The main techniques include:

1. Utilizing unlabeled data from the Twenty-Four Histories for domain adaptation fine-tuning of BERT to assess the impact of BERT's domain adaptation on the performance of named entity recognition in classical texts. Experimental results indicate that this method contributes to improving the overall annotation performance of named entity recognition in classical texts.

2. Introducing the BERT-Global Pointer model for recognizing named entities in classical texts. Compared to traditional sequence labeling models, this model demonstrates excellent entity recognition performance in multi-dynasty, cross-domain scenarios of classical text entity recognition. By employing Global Pointer for global recognition, it enhances the accuracy of entity recognition while also offering fast inference speed and effectively addressing features such as nested entities.
3. Employing an adversarial training strategy based on FGM for the Embedding layer in the model to improve its robustness. Utilizing random weighted averaging for model fusion with multiple saved checkpoints during training to enhance the model's generalization ability. Employing a rule-based post-processing correction method to modify erroneous cases in the prediction results, achieving optimal prediction results.

Experimental results indicate that the proposed methods contribute to improving the overall recognition performance of the Named Entity Recognition task for classical texts from the Twenty-Four Histories.

REFERENCES

[1] Zhou F, Wang C, Wang J. Named entity recognition of ancient poems based on Albert-BiLSTM-MHA-CRF model[J]. Wireless Communications and Mobile Computing, 2022, 2022. https://doi.org/10.1155/2022/6507719.

[2] Song B, Bao Z, Wang Y Z, et al. Incorporating lexicon for named entity recognition of traditional Chinese medicine books[C]//Natural Language Processing and Chinese Computing: 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14–18, 2020, Proceedings, Part II 9. Springer International Publishing, 2020: 481-489. https://doi.org/10.1007/978-3-030-60457-8_39.

[3] Yu P, Wang X. BERT-based named entity recognition in Chinese twenty-four histories[C]//International Conference on Web Information Systems and Applications. Cham: Springer International Publishing, 2020: 289-301. https://doi.org/10.1007/978-3-030-60029-7_27.

[4] Su Q, Wang Y, Deng Z, et al. CCL23-Eval (GuNER2023)(Overview of CCL23-Eval Task 1: Named Entity Recognition in Ancient Chinese Books)[C]//Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations). 2023: 34-40. https://aclanthology.org/2023.ccl-3.4.

[5] Izmailov P, Podoprikhin D, Garipov T, et al. Averaging weights leads to wider optima and better generalization[J]. arXiv preprint arXiv:1803.05407, 2018.https://doi.org/10.48550/arXiv.1803.05407.

[6] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[J]. arXiv preprint arXiv:1412.6572, 2014. https://doi.org/10.48550/arXiv.1412.6572.

[7] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018. https://doi.org/10.48550/arXiv.1810.04805.

[8] Su J, Murtadha A, Pan S, et al. Global pointer: Novel efficient span-based approach for named entity recognition[J]. arXiv preprint arXiv:2208.03054, 2022. https://doi.org/10.48550/arXiv.2208.03054.

[9] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv preprint arXiv:1706.06083, 2017. https://doi.org/10.48550/arXiv.1706.06083.

[10] Wang P, Ren Z. The uncertainty-based retrieval framework for Ancient Chinese CWS and POS[J]. arXiv preprint arXiv:2310.08496, 2023. https://doi.org/10.48550/arXiv.2310.08496.

[11] Wei J, Ren X, Li X, et al. Nezha: Neural contextualized representation for chinese language understanding[J]. arXiv preprint arXiv:1909.00204, 2019. https://doi.org/10.48550/arXiv.1909.00204.