The 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)
Taipei City, Taiwan, October 20-21, 2023. The Association for Computational Linguistics and Chinese Language Processing

YNU-HPCC at ROCLING 2023 MultiNER-Health Task: A transformer-based approach for Chinese healthcare NER

Chonglin Pang, You Zhang and Xiaobing Zhou
School of Information Science and Engineering
Yunnan University
Kunming, China
Contact: {yzhang0202, zhouxb}@ynu.edu.cn

Abstract

Chinese healthcare NER is an essential task in natural language processing to automatically identify healthcare entities such as symptoms, chemicals, diseases, and treatments for machine reading and understanding. Previous studies used Bi-directional Long Short-Term Memory (BiLSTM) and Conditional Random Fields (CRF) to solve NER tasks. This paper uses the RoBERTa-large pre-trained language model combined with BiLSTM-CRF to build a NER model suitable for Chinese healthcare tasks. Dropout is used to improve the performance and stability of the model, and gradient clipping is added to prevent gradient explosion. Comparative experiments were conducted on the dev set to select the model with the best performance for submission. The best model managed to achieve a macro-averaging F1 score of 68.40, which ranked second in the ROCLING 2023 shared task.

Keywords: Chinese Healthcare Named Entity Recognition, RoBERTa, Bi-directional Long Short-Term Memory, Conditional Random Fields.

1 Introduction

The shared task of ROCLING 2023 is MultiNER-Health Chinese Multi-genre Named Entity Recognition in the Healthcare Domain, which requires predicting the named entity boundaries and categories for each given sentence. The data sources for this task include Formal texts (FT), Social media (SM), and Wikipedia articles (WA), describing a total of ten types of entities related to Chinese healthcare. Table 1 provides the definitions and examples of entity types. The task uses the common BIO (Beginning, Inside, and Outside) format for NER. Here, the B-prefix before a tag indicates that the character is the beginning of a named entity, and the I-prefix before a tag indicates that the character is inside a named entity. An O tag indicates that a token belongs to no named entity (Lee and Lu, 2021). For example, the input is 早起也能預防老化，甚至降低阿茲海默症的風險, the intelligence model is expected to extract two entities, including 老化 as SYMP, and 阿茲海默症 as DISE. Output according to BIO format as O, O, O, O, O, O, B-SYMP, I-SYMP, O, O, O, O, O, B -DISE, I-DISE, I-DISE, I-DISE, I-DISE, O, O, O.

Early named entity recognition methods mainly used the Hidden Markov Model (HMM) (Zhou and Su, 2002) or Conditional Random Fields (Sutton et al., 2012) to train named entity recognition models on a large amount of manually labeled corpus. The model learns knowledge from a large amount of labeled corpus without the need for manually defined rules. However, building a large-scale labeled corpus is time-consuming and laborious. In recent years, deep learning algorithms have been applied in the field of natural language processing (NLP). Named entity recognition methods based on deep learning mainly include Recurrent neuralnetwork (RNN) and Long short-term memory (LSTM). Since LSTM solves the problem of gradient disappearance during long sequence training, BiLSTM-CRF became one of the mainstream models at that time.

Once Transformer (Vaswani et al., 2017) emerged, it achieved great success in many NLP tasks. The transformer is different from traditional RNN and LSTM architectures. By introducing a self-attention mechanism and position encoding, Transformer greatly im-

Entity Type	Description	Examples
Body (BODY)	The whole physical structure that forms a person or animal including biological cells, organizations, organs and systems.	"細胞核" (nucleus), "神經組織" (nerve tissue), "左心房" (left atrium), "脊髓" (spinal cord), "呼吸系統" (respiratory system)
Symptom (SYMP)	Any feeling of illness or physical or mental change that is caused by a particular disease.	"流鼻水" (rhinorrhea), "咳嗽" (cough), "貧血" (anemia), "失眠" (insomnia), "心悸" (palpitation), "耳鳴" (tinnitus)
Instrument (INST)	A tool or other device used for performing a particular medical task such as diagnosis and treatments.	"血壓計" (blood pressure meter), "達文西手臂" (DaVinci Robots), "體脂肪計" (body fat monitor), "雷射手術刀" (laser scalpel)
Examination (EXAM)	The act of looking at or checking something carefully in order to discover possible diseases.	"聽力檢查" (hearing test), "腦電波圖" (electroencephalography; EEG), "核磁共振造影" (magnetic resonance imaging; MRI)
Chemical (CHEM)	Any basic chemical element typically found in the human body.	"去氧核糖核酸" (deoxyribonucleic acid; DNA), "糖化血色素" (glycated hemoglobin), "膽固醇" (cholesterol), "尿酸" (uric acid)
Disease (DISE)	An illness of people or animals caused by infection or a failure of health rather than by an accident.	"小兒麻痺症" (poliomyelitis; polio), "帕金森氏症" (Parkinson's disease), "青光眼" (glaucoma), "肺結核" (tuberculosis)
Drug (DRUG)	Any natural or artificially made chemical used as a medicine.	"阿斯匹靈" (aspirin), "普拿疼" (acetaminophen), "青黴素" (penicillin), "流感疫苗" (influenza vaccination)
Supplement (SUPP)	Something added to something else to improve human health.	"維他命" (vitamin), "膠原蛋白" (collagen), "益生菌" (probiotics), "葡萄糖胺" (glucosamine), "葉黃素" (lutein)
Treatment (TREAT)	A method of behavior used to treat diseases.	"藥物治療" (pharmacotherapy), "胃切除術" (gastrectomy), "標靶治療" (targeted therapy), "外科手術" (surgery)
Time (TIME)	Element of existence measured in minutes, days, years.	"嬰兒期" (infancy), "幼兒時期" (early childhood), "青春期" (adolescence), "生理期" (on one's period), "孕期" (pregnancy)

Table 1: Definitions and examples of entity types.

proves the effect of the model in long-distance dependency modeling and parallel computing. Based on the transformer architecture, many powerful pre-training language models (PLM) have emerged, such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019) and DeBERTa (He et al., 2020). These pre-trained models can be fine-tuned on NER tasks to adapt to the specific requirements of the task. In the fine-tuning process, the model is trained on a small amount of labeled data, and the model parameters are updated through the back-propagation algorithm. The advantage of a pre-trained model is that it has been initially learned through large-scale data, so it only needs a small amount of training on a specific task to achieve good results.

Most named entity recognition is researched based on English (Liu et al., 2021), English-named entities have obvious formal signs, and the identification of entity boundaries is relatively easy. In English, there are separators between words to identify boundaries, and each word has a complete meaning. Compared with English, the task of Chinese-named entity recognition is more difficult (Zhu et al., 2022). The difficulty of Chinese-named entity recognition lies in:

(1) Word boundaries are blurred. Chinese do not use spaces or other separators to represent word boundaries like in English and other languages. This feature leads to the problem of boundary ambiguity and recognition difficulties in Chinese named entity recognition.

(2) Semantic diversification. There are a lot of polysemous words in Chinese, and a word may be used in different contexts to express different meanings. Therefore, the named entity recognition model needs to have a stronger context understanding ability to correctly classify it.

(3) The morphological features are vague. In English, the first letter of some designated types of entities is usually capitalized, such as the name of a designated person or place. This information is an unambiguous feature that identifies the location and boundaries of some named entities. The lack of explicit features of Chinese morphology in Chinese-named entity recognition increases the difficulty of recognition.

Therefore, using a character-based training method for the Chinese healthcare NER task can effectively avoid the problem of difficult word segmentation in Chinese sentences, thereby obtaining better training results. Considering the limitations of its hardware equipment, this paper uses the RoBERTa-large as

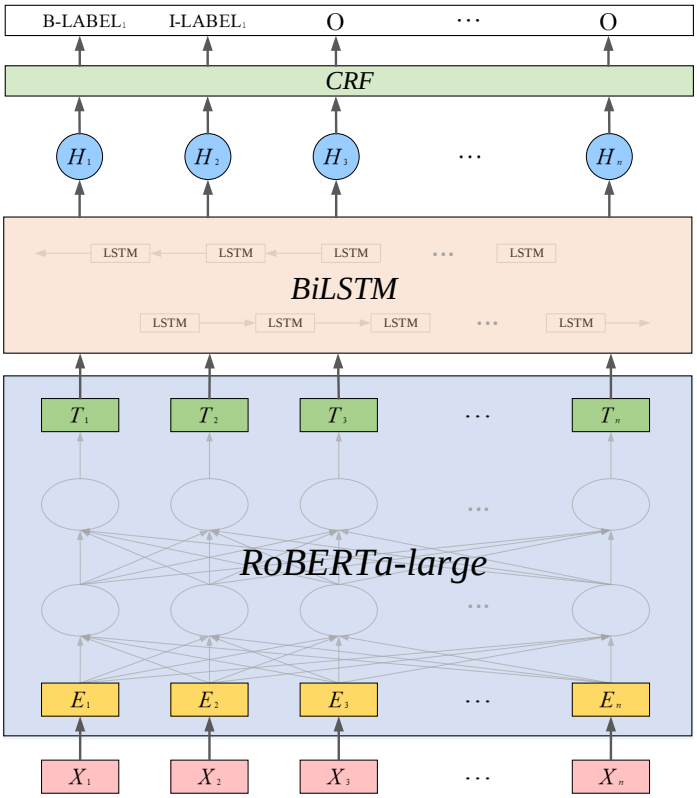

Figure 1: The overall architecture of the proposed method.

the input of BiLSTM-CRF, uses dropout (Srivastava et al., 2014) to improve the performance and stability of the model, and adds gradient clipping to prevent gradient explosion.

The rest of this paper is organized as follows. Section 2 describes the models used in this task. Experimental results are summarized in Section 3. The conclusion is finally drawn in Section 4.

2 Model Description

This section will describe the architecture of the proposed model in detail. This section has several components, including the pre-trained language model, BiLSTM, CRF, dropout, and gradient clipping. The model architecture is shown in Figure 1.

2.1 Pre-trained language model

BERT is a Transformer-based bidirectional language model. The main innovation of BERT lies in the pre-training method, which uses masked language model (MLM) and next sentence prediction (NSP) to capture the contextual semantics of words and sentences. MLM uses the [MASK] flag to randomly mask certain characters in the input and predict the masked word based on its context. Different from the unidirectional language model, MLM combines the text from the left and right directions at the same time, making full use of the semantics of the context. Compared with the traditional word vector model, it generates dynamic word vectors according to the context, which solves the problem of polysemy. In addition, BERT also uses NSP to capture sentence-level context. The model receives pairs of sentences as input and judges the order of the two sentences. Structurally speaking, BERT stacks multiple Transformer encoders together for feature extraction, and each Transformer encoder consists of a Self-attention layer and a feedforward neural network layer. The significance of using the Self-attention mechanism is that it not only encodes words based on the importance of the full text but also abandons the traditional cyclic neural network structure. It solves the long-term dependence problem of the traditional model and greatly improves the parallel computing capability of the model. The CKIP (Chinese Knowledge and Informa-

tion Processing) Group is a research team established in 1986 by the Institute of Information Science and the Institute of Linguistics, Academia Sinica. They released BERT Traditional Chinese pre-training language model ckiplab/bert-base-Chinese. RoBERTa is one of the optimized models after the emergence of the BERT model. RoBERTa uses larger-scale pre-training data than BERT, which increases the generalization ability and performance of the model. RoBERTa also uses dynamic masks instead of BERT's static masks, reducing the risk of model overfitting. Joint Laboratory of HIT and iFLYTEK Research (HFL) has released the traditional Chinese pre-training language model hfl/chinese-roberta-wwm-ext-large (Cui et al., 2021). The model can better capture semantic features at the Chinese word level, thus improving the overall performance.

2.2 BiLSTM

Traditional recurrent neural networks are mainly used to process sequence data. The data before and after the sequence data have a strong correlation. RNN can model the characteristics of the sequence data and store the data before and after. But over time, RNN often suffer from the problem of vanishing gradients. LSTM is a variant of RNN, which solves the problem of gradient disappearance generated during RNN training. LSTM cleverly uses the concept of gating to achieve long-term memory, and it can also capture sequence information. LSTM uses input gates, forget gates, and output gates to process information, which can discard some useless information and enhance the memory of neurons. However unidirectional LSTM cannot process context information at the same time, while BiLSTM is composed of forward LSTM and backward LSTM, which can obtain complete context information. Compared with LSTM, BILSTM can obtain more comprehensive feature information, thereby improving the performance of the model.

2.3 CRF

CRF is a classic discriminant probabilistic undirected graph model. This model calculates the optimal joint probability in a certain sequence. It optimizes the entire sequence instead of stitching together the optimal solutions at each moment. There is a dependency between the labels of the NER task, for example, the I-BODY label must appear after the B-BODY label or the I-BODY label. The prediction results in output by BiLSTM only consider the contextual information of Chinese healthcare data but do not learn the dependencies between labels. CRF can effectively constrain the dependencies between predicted tags, model the tag sequence, and obtain the global optimal sequence.

The calculation process of the CRF model is: take the output sequence $x = (x_1, x_2, x_3, \cdots\cdots, x_n)$ of BiLSTM as the input sequence of CRF, assuming that P is the output score matrix of BiLSTM, the size is $n \times k$, where n is the number of characters, k is the number of tags, and P_{ij} represents the score of the jth tag of the ith word. For the prediction sequence $y = (y_1, y_2, y_3, \cdots\cdots, y_n)$, its scoring function is shown as follows:

$$s\left(X, Y\right) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i} \qquad (1)$$

where A represents the transfer score matrix, A_{ij} represents the score transferred from tag i to tag j, and the size of A is $A + 2$. The probability generated by the predicted sequence Y is shown as follows:

$$p\left(Y|X\right) = \frac{e^{s(x,y)}}{\sum\limits_{\widetilde{Y} \in Y_X} s\left(X, \widetilde{Y}\right)} \qquad (2)$$

where $\tilde{Y}$ represents the real label sequence; Y_X represents all possible label sequences. Finally, the Viterbi algorithm is used to decode and find the highest-scoring Y^* among all Y, so that the global optimal sequence is obtained. The algorithm formula is as follows:

$$Y^* = \arg\max_{\widetilde{Y} \in Y_X} s\left(X, \widetilde{Y}\right) \qquad (3)$$

2.4 Dropout

Dropout is a regularization technique proposed for deep learning. When training the neural network, Dropout will discard the hidden layer nodes in the neural network structure according to a certain probability. Because the hidden layer nodes are randomly

ignored during the discarding process, this makes the network trained each time different. The hidden nodes appear randomly with a certain probability, so it cannot be guaranteed that each hidden node will appear at the same time every time so that the update of the weight value no longer depends on the joint action of the hidden nodes with a fixed relationship, which prevents certain features from A situation where it is only effective under certain other characteristics. Using dropout can effectively prevent the model from overfitting and improve the performance of the model.

2.5 Gradient clipping

During the training process of the deep learning model, the network parameters are updated through the gradient descent algorithm. After the model calculates the predicted value, it will calculate the loss between the target value and the predicted value according to the loss function. After getting the loss, enter the backpropagation stage to calculate the gradient value according to the loss. If the gradient value is too large, the parameter update amount will be too large, which will cause the model to fail to converge. Due to the large number of parameters of the RoBERTa-large model, gradient explosions tend to occur during the backpropagation phase. Therefore, this paper uses the gradient clipping algorithm to limit the size of the gradient by setting the maximum gradient threshold, which effectively avoids the occurrence of the gradient explosion problem.

3 Experimental Results

In this section, we conduct comparative experiments to select the best model for the final submission. The details of the experiments are as follows.

3.1 Dataset

The ROCLING 2023 shared task provides Chinese HealthcareNER Corpus and ROCLING-2022 CHNER Dataset (Lee et al., 2022a). The Chinese HealthNER Corpus provides train.json and test.json files. The data comes from Formal texts and Social media. The data formats include id, genre, sentence, word, word_label, character, and character_label. Because the task focuses on character-level labels, we choose character and character_label as input and output. Since the original dataset does not conform to the input format of the model, we performed data preprocessing on the dataset. The ROCLING-2022 CHNER Dataset provides the ROCLING22_CHNER_truth.txt file, and the data comes from Wikipedia articles. The original dataset conforms to the model's input format, so no additional adjustments are required.

3.2 Evaluation Metrics

Performance is evaluated by examining the difference between machine-predicted labels and human-annotated labels. We adopt standard precision, recall, and F1-score, which are the most typical evaluation metrics of NER systems at a character level. Precision is defined as the percentage of correctly named entities found by the NER system. Precision is defined as follows:

$$P = \frac{TP}{TP + FP} \tag{4}$$

Recall is the percentage of named entities present in the test set found by the NER system. The definition of Recall is as follows:

$$R = \frac{TP}{TP + FN} \tag{5}$$

F1-score is an indicator used in statistics to measure the precision of a binary (or multi-class) model, which takes into account the precision and recall of the classification model at the same time. The definition of F1-score is as follows:

$$F = \frac{2 \times P \times R}{P + R} \tag{6}$$

where TP is True Positive, FP is False Positive, FN is False Negative.

3.3 Implementation Details

The training data is divided into train data and dev data. First, we preprocess the train.json and test.json files of the Chinese HealthNER Corpus to obtain the train.txt and dev.txt files. The new file contains only two data, character and character_label, and is modified to meet the format required by the model input. We splice the ROCLING22_CHNER_truth.txt file and the

train.txt file to get a new train.txt file. Therefore, the train.txt file we use is composed of the data in the train.json file and the ROCLING22_CHNER_truth.txt file, and the dev.txt file is composed of the data in the test.json file.

This paper chooses multiple pre-trained language models for comparative experiments. While the DeBERTa model exhibits a more powerful performance compared to the BERT and RoBERTa models, its support for traditional Chinese is currently limited. Additionally, the Deberta model requires substantial GPU resources and incurs high running time costs. Therefore, we did not select the DeBERTa model for our experiments. The comparative experiments employed three pre-trained language models: ckiplab/bert-base-Chinese, hfl/chinese-roberta-wwm-ext, and hfl/chinese-roberta-wwm-ext-large.

3.4 Parameters Tuning

This paper uses the learning rate warm-up strategy, dropout, and gradient clipping to optimize the model. Warm up is a learning rate warm-up method mentioned in the ResNet (He et al., 2016) paper. A learning rate preheating method is mentioned in the paper. This method uses a small learning rate to train some epochs at the beginning of training and then modifies to the preset learning rate for training. Since the weight values of the model are initialized randomly at the beginning of training, the model may become unstable if a large learning rate is set. At the beginning of training, using the warm-up strategy can make the learning rate smaller for several epochs, and the model can gradually stabilize. When the model is relatively stable, modify it to the preset learning rate for training to make the model converge faster. In addition, we used dropout to prevent the model from overfitting. Since dropout will randomly ignore some hidden layer nodes when the dropout ratio is set too large, it will damage the learning ability of the model and affect the performance of the model; when the dropout ratio is set too small, dropout will not be effective, and the model will still There is a possibility of overfitting. The parameter tuning process is shown in the following Figure 2 and Figure 3.

During the training process, we found that

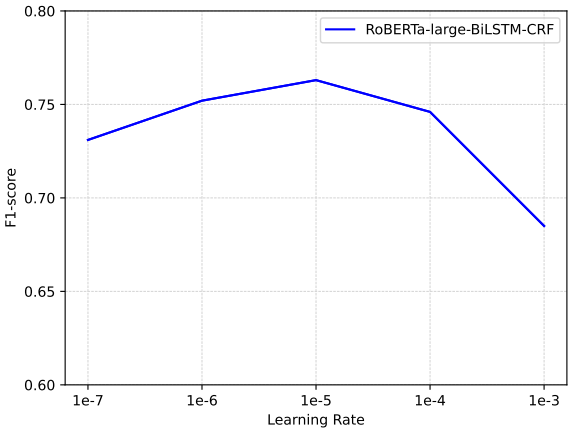

Figure 2: The performance of different learning rate on F1-score.

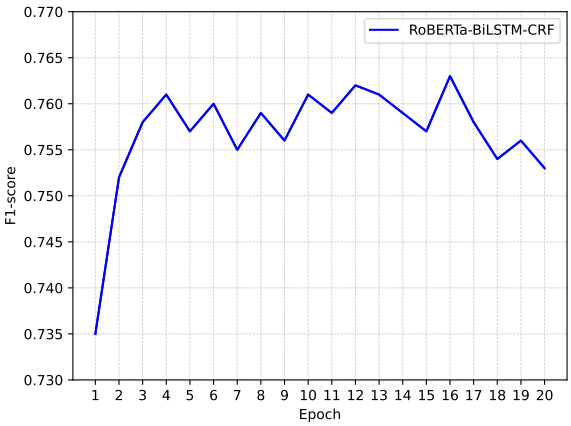

Figure 3: The performance of different epoch on F1-score.

the RoBERTa-large-BiLSTM-CRF model may experience a gradient explosion. Although BiLSTM can effectively prevent gradient disappearance, it cannot solve the problem of gradient explosion. So we added a gradient clipping algorithm to avoid the problem of gradient explosion. Set the maximum gradient threshold to 7.0, and the gradient explosion will no longer occur during the model training process. Moreover, the grid search is used to find the optimal parameters. Finally, the learning rate is set to 1e-5, the epoch is set to 20, the dropout is set to 0.4, the weight decay is set to 1e-7, and the warm-up ratio is set to 0.1.

3.5 Comparative Results

For experiments, we utilized the pre-trained language models as inputs for the BiLSTM instead of using word2vec, which was used in

Model	Precision	Recall	F1-score
BiLSTM-CRF(word2vec)	0.6631	0.7574	0.7072
BERT-BiLSTM-CRF	0.7349	0.7577	0.7461
RoBERTa-BiLSTM-CRF	0.7401	0.7796	0.7593
RoBERTa-large-BiLSTM-CRF	**0.7307**	**0.7974**	**0.7626**

Table 2: The score of each model in dev data.

Model	Precision	Recall	F1-score
BiLSTM-CRF(word2vec)	0.6631	0.7574	0.7072
RoBERTa-large-CRF	0.7335	0.7851	0.7584
RoBERTa-large-BiLSTM	0.7105	0.7964	0.7509
RoBERTa-large	0.7036	0.7954	0.7467
RoBERTa-large-BiLSTM-CRF	**0.7307**	**0.7974**	**0.7626**

Table 3: Ablation experiment comparison results in dev data.

the base model. The predictions from the BiLSTM were then processed through a CRF layer to obtain the final prediction results and calculate the F1 scores. The scores for each model, along with the scores of the baseline model, are presented in Table 2.

As indicated, the RoBERTa-large-BiLSTM-CRF model achieved the best results. Due to the existence of dropout, the results of each training of the model with the same parameters may also be different. Therefore, we used the RoBERTa-large-BiLSTM-CRF model training with the same parameter configuration to obtain three model files, used these three models to predict the test set, and finally submitted three results. Among them, the RUN2 test file achieved the best results in the test data set (Lee et al., 2023), with a macro-averaging F1 score of 68.40. The test results are better than the official baseline model: the BERT-BiLSTM-CRF model (Lee et al., 2022b).

3.6 Ablation Study

To examine the functions of each component of the module, we conduct ablation experiments by removing the RoBERTa-large pre-trained language model, BiLSTM model, and CRF layer individually. The results of the ablation experiments are presented in Table 3.

Based on the data in the table, it is evident that the F1 score drops significantly after removing the RoBERTa-large pre-trained language model, indicating that the pre-trained language model plays a critical role in this model. On the other hand, removing the BiLSTM model has a relatively minor impact on the F1 score. This is because the pre-trained language model exhibits powerful performance and can provide comprehensive context information. However, the results show that integrating the BiLSTM model slightly improves the F1 score, which leads us to retain the BiLSTM model in our final structure. Furthermore, removing the CRF layer also leads to a certain decrease in the F1 score.

3.7 Case Study

The models that did not use CRF in the ablation experiments all had varying degrees of dependency confusion between labels. For example, 德國麻疹疫苗 is a complete entity, its corresponding entity category is DRUG, and the expected prediction result should be B-DRUG, I-DRUG, I-DRUG, I-DRUG, I-DRUG, I-DRUG. However, the model without CRF yielded predictions of B-DRUG, I-DRUG, I-DISE, I-DISE, I-DRUG, I-DRUG. By comparing the train set, it can be found that the entity category corresponding to 麻疹 when it appears alone is indeed DISE. However, due to the existence of entity boundaries and the constraints of label dependencies, it is wrong to predict 麻疹 as DISE here. This demonstrates the importance of the CRF layer in resolving label dependencies. The NER task uses an exact matching rule, that is, each predicted entity category and entity boundary needs to be successfully matched to be considered a successful prediction. Therefore, the prediction score of the model using CRF under the exact matching rule is significantly higher than that of the model without CRF.

4 Conclusions

This paper provides a detailed description of the model structure utilized and the experimental process. The final test results yielded a macro-averaging F1 score of 68.40, securing a second-place ranking. Throughout the experiment, we compared various pre-trained language models and ultimately adopted the RoBERTa-large-BiLSTM-CRF architecture. In addition, we incorporated the learning rate warm-up strategy and dropout

techniques to enhance the model's performance. Future works will try to use a pretrained language model with a more powerful performance to explore whether the performance of the model can continue to improve.

References

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Lung-Hao Lee, Chao-Yi Chen, Liang-Chih Yu, and Yuen-Hsien Tseng. 2022a. Overview of the rocling 2022 shared task for chinese healthcare named entity recognition. In *Proceedings of the 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022)*, pages 363–368.

Lung-Hao Lee, Tzu-Mi Lin, and Chao-Yi Chen. 2023. Overview of the rocling 2023 shared task for chinese multi-genre named entity recognition in the healthcare domain. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*.

Lung-Hao Lee, Chien-Huan Lu, and Tzu-Mi Lin. 2022b. NCUEE-NLP at SemEval-2022 task 11: Chinese named entity recognition using the BERT-BiLSTM-CRF model. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1597–1602, Seattle, United States. Association for Computational Linguistics.

Lung-Hao Lee and Yi Lu. 2021. Multiple embeddings enhanced multi-graph neural networks for chinese healthcare named entity recognition. *IEEE Journal of Biomedical and Health Informatics*, 25(7):2801–2810.

Kang Liu, Qingsong Yu, and Shanhao Zhong. 2021. Chinese named entity recognition based on bi-directional quasi-recurrent neural networks improved with bert: new method to solve chinese ner. In *2021 the 5th International Conference on Innovation in Artificial Intelligence*, pages 15–19.

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. *arXiv preprint arXiv:1907.11692*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Charles Sutton, Andrew McCallum, et al. 2012. An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

GuoDong Zhou and Jian Su. 2002. Named entity recognition using an hmm-based chunk tagger. In *Proceedings of the 40th annual meeting of the association for computational linguistics*, pages 473–480.

Peng Zhu, Dawei Cheng, Fangzhou Yang, Yifeng Luo, Dingjiang Huang, Weining Qian, and Aoying Zhou. 2022. Improving chinese named entity recognition by large-scale syntactic dependency graph. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:979–991.