Journal of Computer Technology and Software

ISSN:2998-2383

Vol. 3, No. 5, 2024

Named Entity Recognition: A Comparative Study of Advanced Pre-trained Model

Zitao Zheng[1], Yiru Cang[2], Wangying Yang[3], Qiyuan Tian[4], Dan Sun[5]

[1] Independent Researcher, New Jersey, USA

[2] Northeastern University, Boston, USA

[3]University of Southern California, Los Angeles, USA

[4]George Washington University, Washington, USA

[5]Washington University in St. Louis, St. Louis, USA

Correspondence should be addressed to Wangying Yang; ywyazure@gmail.com

Abstract: The increasing demand for healthcare has positioned digital medical services as a pivotal trend in the medical industry's future. This paper examines the role of Medical Named Entity Recognition (NER) in enhancing the utilization of medical resources and promoting the intelligentization of clinical decision-making. Medical NER faces challenges due to the diversity and specialized nature of medical text data, but with the application of deep learning technologies, there has been a significant improvement in recognition accuracy and efficiency. A novel model, RoBERTa-FGM-MHA-CRF, is proposed in this study, which integrates adversarial training, multi-head attention mechanisms, bidirectional Long Short-Term Memory (LSTM) networks, and bidirectional Gated Recurrent Unit (GRU) networks to substantially enhance the model's generalization and accuracy in medical NER tasks. Experimental results using a dataset of medicine instruction manuals demonstrate the model's effectiveness in practical medical text processing.

Keywords: Digital Medical Services, Medical Named Entity Recognition, Pre-trained Models

1. Introduction

As people's demand for healthcare continues to grow, digital medical services are poised to become a trend in the future development of the medical industry. Medical institutions generate a large amount of medical text data every day. How to more intelligently and efficiently extract the information contained in these texts, and provide patients with more convenient and rapid medical services through digital platforms, effectively improving the utilization rate of medical resources, is of significant research importance.

Named Entity Recognition (NER) is an important research direction in the field of Natural Language Processing (NLP). The goal is to identify entities in the text that have specific meanings, such as names of people, places, organizations, institutions, etc. After decades of development, NER technology has made significant progress, evolving from early manual construction of rules or dictionaries to end-to-end solutions based on deep learning. NER technology can identify words in the text with specific meanings and determine the category to which the word belongs. In different application scenarios, the types of entities that NER technology needs to recognize are also different.

Due to the particularity of medical texts, compared with general domain NER, medical NER faces more difficulties and challenges. Firstly, medical texts are diverse, including semi-structured electronic medical record data and unstructured clinical literature data, and the description methods of different types of data are also inconsistent, which poses great difficulties for entity recognition. Secondly, medical texts are highly specialized, and the definition of entity categories is difficult to unify, and large-scale annotation will consume a lot of manpower, which to some extent restricts the training of large models. With the development of deep learning technology, technologies such as recurrent neural networks and pre-trained language models are further integrated with medical NER tasks, which can effectively improve the corpus recognition range and recognition accuracy of NER. This not only improves the efficiency of medical text mining but also helps to promote the intelligence of clinical decision-making and the optimization of the medical process, achieving precise use of medical information, and promoting the development of the medical industry. Therefore, in-depth research on medical NER technology has important application value and practical significance.

This paper proposes the RoBERTa-FGM-MHA-CRF model: to address the poor generalization ability of pre-trained models in some medical NER datasets, an adversarial training module is added to the pre-trained model to enhance the model's generalization ability. Then, after the output of the pre-trained model, modules such as multi-head attention mechanisms, bidirectional long short-term memory networks, and bidirectional gated recurrent unit networks are added to compare and analyze the performance of each module in the medical NER task. Finally, the conditional

random field is used as the output layer of the pre-trained model to improve the model's recognition accuracy. Experiments using medicine instruction dataset prove that adding adversarial training modules and multi-head attention mechanisms to the pre-trained model can effectively improve the model's accuracy and generalization ability in entity recognition tasks.

2. Background

Currently, deep learning-based Named Entity Recognition (NER) methods have become mainstream, with commonly used network models including Convolutional Neural Networks (CNN) [1], Recurrent Neural Networks (RNN) [2], Long Short-Term Memory (LSTM) networks [3][4], Gated Recurrent Unit (GRU) networks [5], Transformer models, self-attention mechanisms, and pre-trained models, etc.

Collobert et al. [6] proposed two solutions for the NER task: window-based and sentence-based. The window-based recognition method inputs the context window of the word being predicted and uses the neural network structure for NER; the sentence-based recognition method inputs the entire sentence, distinguishes the words in the sentence by adding relative positional information, and finally uses a convolutional neural network for recognition.

Yan et al. [7] combined BiLSTM with CRF in the field of NER, extracting long-distance dependencies in the input sequence with BiLSTM. They also compared LSTM, BiLSTM, CRF, and LSTM-CRF. The experimental results showed that BiLSTM-CRF has strong robustness compared to other models and can achieve high annotation accuracy without using word embeddings. Sahu et al. [8] proposed three network structures based on Long Short-Term Memory networks: B-LSTM, AB-LSTM, and joint AB-LSTM. Using word embeddings and distance embeddings as features, Bi-LSTM learns advanced feature representations, and SoftMax obtains the output.

In 2017, the Google team proposed the Transformer [9] model in the field of text translation. This structure fully relies on the self-attention mechanism to model the context and long-distance dependencies of the text, supporting parallel computing, which can greatly improve the efficiency and accuracy of model training. A series of pre-trained language models based on the Transformer structure emerged later, opening a new era in the field of natural language processing, and NER tasks in the medical field have also made significant progress.

In 2020, Lee et al. [10] proposed the BioBERT model, which is initialized with BERT weights and retrained in a large-scale biomedical corpus to specifically learn medical text knowledge. Compared with general language models, this model can better recognize medical information in the text and performs better in biomedical field tasks than general language models.

Su et al. [11], in response to the issue of NER models ignoring boundary information, proposed a segment-based NER framework called Global Pointer.

3. Method

3.1. Model Structure

Currently, medical Named Entity Recognition (NER) methods often combine pre-trained models with BiLSTM (Bidirectional Long Short-Term Memory) or BiGRU (Bidirectional Gated Recurrent Unit) modules to enhance the model's recognition accuracy. However, when the experimental dataset contains erroneous or anomalous data, it can lead to unstable model performance and a decrease in recognition accuracy. Generally, a good model should have strong robustness, maintaining relatively stable performance even under certain levels of interference and noise. Although BiLSTM or BiGRU modules can effectively model dependencies, they have poor modeling capabilities for long-distance dependencies. Through the analysis of the aforementioned issues, this paper introduces an adversarial training (Fast Gradient Method Adversarial Training, FGM) module and a Multi-Head Attention (MHA) module into the model. Adversarial training can add a certain degree of interference during model training, effectively enhancing the model's robustness against noisy data; incorporating the Multi-Head Attention module allows the model to focus on text features from different dimensions, obtaining richer and more diverse feature representations, and enhancing the model's expressive power. Additionally, this paper compares the Multi-Head Attention mechanism with BiLSTM and BiGRU modules, analyzing the performance of the three modules in the NER task. The RFM model consists of three parts: a feature extraction layer, an encoding layer, and a decoding layer. The model structure is shown in Figure 1.

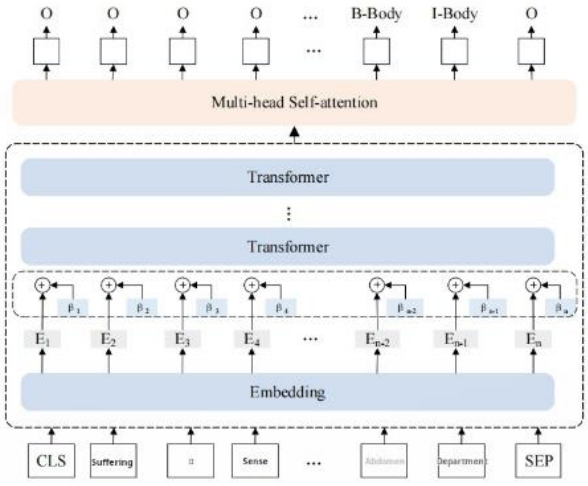

Figure 1. Model Structure

This paper utilizes the RoBERTa-wwm pre-trained model as the feature extraction layer of the entire model to obtain the semantic representation of the input sequence. It introduces a Multi-Head Attention (MHA) module to receive the output from the RoBERTa-wwm model, enabling dependency

modeling from multiple dimensions, followed by the use of Conditional Random Fields (CRF) as the decoding layer for sequence labeling. To enhance the robustness of the model, a perturbation mechanism is introduced in the embedding layer of the RoBERTa-wwm model, which adds noise during the training process to improve the model's ability to resist interference. Specifically, random perturbations are added to the word vectors in the embedding layer, allowing the model to adapt to varying degrees of input noise and preventing overfitting to the inputs. The semantic representation output by the RoBERTa-wwm model is then used as the input for the Multi-Head Attention module, which can learn long-distance dependencies within the sequence. By aggregating information from different positions, it achieves a more comprehensive and enriched feature representation, thereby improving the effect of sequence labeling. The main work of this paper is to enhance the robustness of the model by introducing perturbations in the feature extraction layer and to improve the feature representation by adopting the Multi-Head Attention mechanism, thereby further enhancing the performance of the RoBERTa-wwm model in the task of medical Named Entity Recognition (NER).

3.2. Feature Extraction Layer

To fully extract text features, this paper employs RoBERTa-wwm as the feature extraction module. RoBERTa-wwm is composed of multiple stacked Transformer Encoder blocks, which are divided into two parts: the Multi-Head Attention module and the fully connected feed-forward neural network.

The structure of the Encoder module is shown in Figure 2. This figure would typically illustrate the arrangement of the Multi-Head Attention and feed-forward neural network within the Encoder, along with the residual connections and layer normalization that are key components of the Transformer architecture, which contributes to the model's efficiency and effectiveness in processing sequence data.

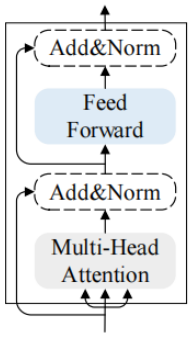

Figure 2. Multi-Head Attention Structure

Compared to BERT, RoBERTa-wwm has a larger number of parameters and uses a larger Batch_Size and more training data during the training process. RoBERTa-wwm has made the following improvements based on BERT:

Removal of the NSP (Next Sentence Prediction) task: NSP is a pre-training task used in BERT to train the model to understand the relationship between two sentences. In the training set, each piece of data consists of two sentences, A and B. If sentences A and B are adjacent, the output is true; otherwise, it is false. In the training samples, the ratio of positive to negative samples is 1:1.

Adoption of a dynamic masking strategy: In RoBERTa-wwm, the masking operation is not chosen to be performed during the preprocessing of the dataset but is postponed to when the model is input, making the masked positions different each time the data is input to the model, avoiding the model's mechanical result recording. Researchers have conducted experiments using the BERT-base model on the SQuAD2.0, MNLI-m, and SST-2 datasets. The F1 values for the dynamic masking strategy on the SQuAD2.0 and SST-2 datasets are higher than those for the static masking strategy, with only a slightly lower F1 value on the MNLI-m dataset. This indicates that the dynamic masking strategy is largely superior to the static masking strategy, and in addition, the dynamic masking strategy does not require copying training samples during the training process, which is also more computationally efficient than the static masking strategy.

3.3. Feature Encoding Layer

The feature encoding layer is responsible for receiving the output from the feature extraction layer, modeling long-distance dependencies and enhancing semantic features of the extracted vectors.

The traditional self-attention mechanism, during computation, uses every position in the entire input sequence as Query, Key, and Value to calculate attention weights. Based on the obtained weight values, it performs a linear combination of the input sequence to get the final output of the current module. In the Multi-Head Attention mechanism, there are multiple Attention Heads, each with its own Query, Key, and Value, which can be computed in parallel during model training. By incorporating the Multi-Head Attention mechanism, the model can learn different dimensional feature representations from the input text sequence, producing more comprehensive output features.

In the experiments of this paper, the feature representations output by the pre-trained model are first copied three times, serving as the input for the Multi-Head Attention module, specifically as Query (Q), Key (K), and Value (V). Then, the input is split according to the number of self-attention heads.

4. Experiment

4.1. Dataset

The Entity Recognition of the dataset originates from MedVidQA. The task is to identify and extract entity information from MedVidQA instruction texts. The MedVidQA dataset contains 3,010 manually created health-related questions paired with instructional videos and timestamps as visual answers.The training dataset contains 1,000 entries for model training and parameter tuning, while the test dataset contains 1,677 entries for evaluating the

model's actual performance. The dataset has annotated a total of 13 entity categories.

4.2. Setting

This paper conducts comparative experiments using multiple pre-trained models, and the weights of these models are all downloaded from Hugging Face. The model information is as follows:

BERT-base, BERT-wwm: Released by the iFLYTEK-HF Joint Laboratory. MacBERT-base, Ernie-1: Proposed by Baidu's team, the model structure is consistent with BERT. RoBERTa-wwm: Compared to BERT, RoBERTa-wwm uses dynamic masking and the whole word masking (wwm) strategy, with a larger dataset, Batch_Size, and model parameters.

4.3. Experiment Results Analysis

This section analyzes the experimental results of various pre-trained models on the MedVidQA instruction manual dataset. The selected pre-trained models include BERT, BERT-wwm, MacBERT-base, Ernie-1, and RoBERTa-wwm. In the experiments with the RoBERTa model, BiLSTM (Bidirectional Long Short-Term Memory), BiGRU (Bidirectional Gated Recurrent Unit), MHA (Multi-Head Attention), and FGM (Fast Gradient Method Adversarial Training) modules were added to compare their performance in the Named Entity Recognition (NER) task. Additionally, each experiment used Conditional Random Fields (CRF) as the decoding layer, with results shown in Table 1.

Table 1: Experiment Results

Model	Metrics		
	Precision	Recall	F1
BERT-base-crf	67.29	77.25	71.92
BERT-wwm-crf	66.51	77.63	71.64
MacBert-base	67.47	77.63	72.19
Ernie-1	67.91	78.16	72.67
RoBERTa-wwm	67.92	77.97	72.59
RoBERTa-wwm-BiLSTM	71.17	74.05	72.58
RoBERTa-wwm-BiGRU	69.42	76.9	72.96
RoBERTa-wwm-FGM	70.04	76.07	72.93
RoBERTa-wwm-FGM-MHA	70.31	76.67	73.35

From the experimental results on the instruction manual entity recognition dataset, it can be seen that the F1 score of the RoBERTa-wwm model is higher than that of BERT and other models, but slightly lower than that of Ernie-1. After adding the BiLSTM module to the RoBERTa-wwm model, the precision rate increased by 3.25%, but the recall rate decreased by 3.92%, and the F1 score decreased by 0.01%. This indicates that while the model cannot predict results comprehensively, it performs well in predicting some types of entities. Moreover, there was a certain degree of overfitting in the training process. Although under-sampling and over-sampling methods were used to preprocess the dataset's label distribution, the effects were not significant, and the recall rate dropped significantly. Other methods such as word replacement could be tried to expand the dataset.

The above experimental results prove that although the BiLSTM module can model the context of the input text and capture more comprehensive text features to enhance the representation ability, it requires a high distribution of entities in the dataset. Due to the uneven distribution of entities in the current training set, the model performs poorly in predicting some entities and exhibits overfitting. After adding adversarial training and Multi-Head Attention modules, the model achieved the best results, fully proving that the combination of Multi-Head Attention mechanisms and adversarial training modules is more suitable for handling sequence labeling tasks. The Multi-Head Attention mechanism can model from multiple angles of the input sequence, capturing features from different perspectives.

5. Conclusion

This paper focuses on the task of medical named entity recognition (NER), combining and improving upon the currently popular methods for NER. It compares the performance of different module combinations within medical NER datasets and proposes the RoBERTa-wwm-FGM-MHA-CRF model. The chapter begins by analyzing the robustness issues of current NER models, introduces the model used in this study, and provides a detailed introduction to the components of the model. It then introduces the datasets used in the experiments of this chapter, presents the text length distribution in the datasets, preprocesses long text data, and optimizes the issue of uneven entity distribution in the datasets. Finally, experiments are conducted under the same experimental conditions, analyzing the performance of each module in the NER task, and a fixed parameter experiment is conducted on the multi-head attention module to analyze the impact of the number of attention heads on the experimental results, thereby proving the effectiveness of the model used in this chapter in the task of medical named entity recognition.

References

[1] Lecun Y, Boser B, Denker J S, et al.Back propa gation applied to handwrit ten zip code recognition[J]. Neural Computation, 1989, 1(4):541-551.

[2] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors[J]. nature, 1986, 323(6088):533-536.

[3] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computati on, 1997, 9(8): 1735-1780.

[4] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representati ons using RNN encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv:1406.1078, 2014.

[5] Chung J, Gulcehre C, Cho K H, et al. Empirical evaluation of gated recurr ent neural networks on sequence modeling[J]. arXiv preprint arXiv:1412.3555, 2014.

[6] Collobert R, Weston J, Bottou L, et al. Natural language processing (almos t) from scratch[J]. Journal of machine learning research, 2011, 12(ARTICLE): 2493- 2537.

[7] Yan, X., Wang, W., Xiao, M., Li, Y., & Gao, M. (2024, March). Survival prediction across diverse cancer types using neural networks. In Proceedings of the 2024 7th International Conference on Machine Vision and Applications (pp. 134-138).

[8] Sahu S K, Anand A. Drug-drug interaction extraction from biomedical texts using long short-term memory network[J]. Journal of biomedical informatics, 2018, 86:15-24.

[9] Vaswani A,Brain G,Shazeer N,et al.Attention Is All You Need[J]. Advances in neural information processing systems,2017(Nips):5998–6008.

[10] Lee J, Yoon W, Kim S, et al.BioBERT: a pre-trained biomedical language r epresentation model for biomedical text mining[J].Bioinformatics. 2020;36(4):1234-12 40.

[11] 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.