Named Entity Recognition Using BERT BiLSTM CRF for Chinese Electronic Health Records

Zhenjin Dai*, Xutao Wang*, Pin Ni*†, Yuming Li*†, Gangmin Li*‡ and Xuming Bai§
*Research Lab for Knowledge and Wisdom, Xi'an Jiaotong-Liverpool University, China
†Department of Computer Science, University of Liverpool, UK
§Department of Interventional Radiology, Second Affiliated Hospital of Soochow University, China
Email: ‡Gangmin.Li@xjtlu.edu.cn, §2005baixuming@163.com

***Abstract*—As the generation and accumulation of massive electronic health records (EHR), how to effectively extract the valuable medical information from EHR has been a popular research topic. During the medical information extraction, named entity recognition (NER) is an essential natural language processing (NLP) task. This paper presents our efforts using neural network approaches for this task. Based on the Chinese EHR offered by CCKS 2019 and the Second Affiliated Hospital of Soochow University (SAHSU), several neural models for NER, including BiLSTM, have been compared, along with two pre-trained language models, word2vec and BERT. We have found that the BERT-BiLSTM-CRF model can achieve approximately 75% F1 score, which outperformed all other models during the tests.**

***Index Terms*—Named Entity Recognition, BiLSTM-CRF, Electronic Health Records**

I. INTRODUCTION

In views of electronic health records (EHR) in management, an increasing number of medical institutions have gradually begun to make use of EHR instead of the traditional paper records. These medical records contain a large amount of significant medical information, such as disease symptoms and diagnosis of patients, which can be extremely useful in the healthcare field, including disease prediction and the construction of medical knowledge graph. Thus, massive data mining and analysis on EHR have received extensive attention in recent years. During the medical information extraction, named entity recognition (NER) is one of the most essential natural language processing (NLP) tasks. In the context of EHR, this task aims to identify medical named entities from EHR and then classify them into the predefined categories, such as disease, drug and operations. These entities can be helpful for follow-up tasks. However, since there is no definite word boundary within the Chinese text, how to resolve the NER task in Chinese EHR has been a very hard issue. Additionally, the absence of the unified Chinese annotated standard in the medical field has also significantly increased its difficulty. In this paper, we have compared several neural models, including CNN-LSTM, BiLSTM, and BiLSTM-CRF, for Chinese medical named entity recognition. The pre-trained language models, word2vec and BERT, have been also tested by using them as word embedding. Based on CCKS 2019 and SAHSU dataset, it is found that the BERT-BiLSTM-CRF model can achieve about 75% F1 score that surpassed all other models during the tests.

II. LITERATURE REVIEW

Chinese named entity recognition (CNER) task is often modelled as the problem of serialization annotation [1]. Traditional NER methods like rule-based and template-based require significant cost and over-reliance on steps, such as rule building or feature engineering. The neural network-based approach as a data-driven approach can achieve an end-to-end whole procedure, not like a traditional pipeline and does not rely on feature engineering.

In the last few years, methods based on the neural network have become increasingly popular in CNER task [2]–[4]. These articles are grounded on the LSTM-CRF framework for CNER tasks, where the Long Short Term Memory (LSTM) network is used to learn the veiled representation of characters and Conditional Random Field (CRF) is directed at joint marker decoding. Wu et al. [5] proposed a unified framework grounded on CNN-LSTM-CRF for CNER. Meanwhile, a method of automatically generating pseudo-marker samples using existing marker data was also put forward. However, the weaknesses of their models lie in the lack of delimiters (spaces) and strong identifiers (uppercase), along with insufficient training data which lead to ambiguity of entities. Guillaume et al. [6] used Bidirectional LSTM (BiLSTM) for learning the veiled representation of the text and CRF for tag decoding.

Zhu et al. [7] proposed the CNER task with Convolutional Attention Network, which contains a character-level Convolutional Neural Network (CNN) locally self-attention layer, a Gated Recurrent Unit (GRU) and global attention layer for obtaining information about the context of the sentence from adjacent characters. Compared to other models, the model is more practical since it does not depend on any external resources. Luo et al. [8] proposed a neural model consisting of a BiLSTM layer based on the attention mechanism and a CRF layer (Att-BiLSTM-CRF), for NER in the chemical field. The method they proposed uses the literature-level information captured by the attention mechanism to enhance the labelling consistency among numerous instances which all obtain the same token in the text, achieving a correct rate of 91.14%.

978-1-7281-4852-6/19/$31.00 ©2019 IEEE

Figure 1: 'BIO' Tagging Format

The introduction of pre-trained language models is also one of the most important parts in neural network approaches. It has been shown that using the pre-trained language model as the word embedding layer can improve the models' performance of sequence tagging [9]. Mikolov et al. [10] proposed the word2vec model, which is a classical context-free language model that can generate a single vector representation for each word in the vocabulary. BERT (Bidirectional Encoder Representation from Transformers) [11] is a newly proposed contextual language model that can change the relationship between the pre-training generated word vectors and the downstream specific NLP tasks. The bidirectional attention-based mechanism makes the context in the text well captured in both directions, thus providing a strong guarantee for the subsequent tasks.

III. Methodology

As an essential NLP task during the medical information extraction, medical named entity recognition is the task of identifying and classifying medical named groupings, such as diseases, drugs and operations, within the health records. More specifically, this task aims to assign an entity label, or tag, to each word in the record. In addition, this paper has adopted BIO tagging scheme to label entities, where a token is labelled as B-label if the token is the beginning of a named entity, or I-label if the token is inside a named entity but not first, otherwise O-label. An example of BIO formatting can be seen in Figure 1, where 'DIS' representing disease.

Many neural models, including CNN-LSTM, BiLSTM, Bi-GRU, BiLSTM-CRF, have been proposed for English named entity recognition. Our work is to compare the performance of these models on CNER in the medical field. The BiLSTM-CRF model can not only efficiently make use of both past and future input features, but also take advantage of the sentence-level tag information offered by the CRF layer. During the NER task, the given fields, including the drugs and symptoms of a disease, should fall into the same label type, and the application of the CRF layer creates this dependency. Moreover, CRF can also avoid the independent situations that an B-drug is followed by an I-disease. By using BiLSTM layer alongside the CRF layer as the output layer, tagging decisions can be modelled jointly rather than independently. The main architecture of BiLSTM-CRF model is shown in Figure 2. We have also tried to combine BERT and word2vec with BiLSTM-CRF model. These pre-trained language models can be used as the word embedding layer of other neural models to perform the desired NLP tasks. In addition, the training and testing of

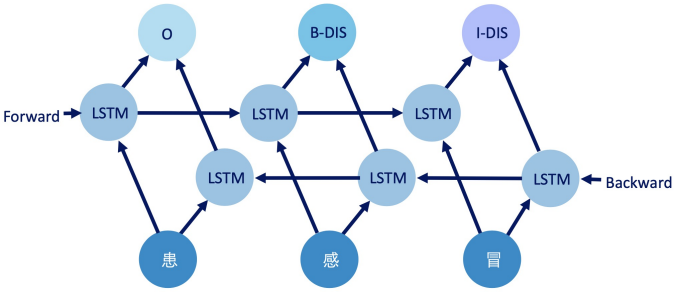

Figure 2: BiLSTM Structure

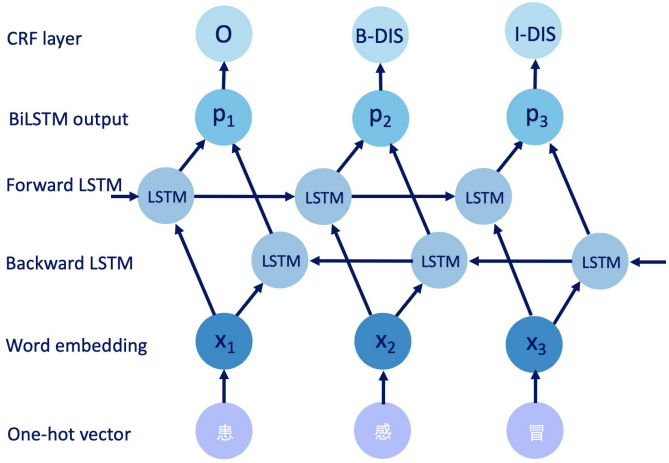

Figure 3: BiLSTM-CRF Structure

models are run on the academic evaluation task (CCKS 2019) and real-world business scenario of Chinese electronic health records datasets (SAHSU).

A. Method Descriptions

The process of named entity recognition using the BiLSTM-CRF model comprises the following steps:

1) Enter the first layer of the model, the look-up layer, and map each word in the sentence to a word vector based on the pre-trained word embedding;

2) Enter the second layer of the model, the Bidirectional Long Short-Term Memory (BiLSTM) layer, and automatically extract the sentence features;

3) Enter the third layer of the model, the Conditional Random Field (CRF) layer, and perform sequence labelling between sentences;

4) Repeat steps 1) through 3) above until all the data have been labelled. Figure 3 demonstrates the architecture of the BiLSTM-CRF model.

B. Method Details

More specifically, the pre-trained language model is firstly used as the word embedding layer of the BiLSTM-CRF model. After that, the BiLSTM layer with the CRF layer is used to tag the original text, and then the predicted word segmentation results are obtained. Finally, the supervised learning method is

used to iteratively learn the word segmentation results, thereby improving the model's performance to obtain the accurate results. More formally, the process can be described in the following six steps: 1) Define the BIO tagging scheme and generate a NER dataset based on the scheme; 2) For every sentence, a sentence containing m words (sequence of words) is represented as $x = (x_1, x_2, ..., x_m)$, where x_i indicates the index of the ith word of the sentence in the vocabulary, and so the corresponding one-hot vector of each word is obtained; 3) The first layer of the model is known as the look-up layer or word embedding layer that can map each word x_i in the sentence from one-hot vector to low-dimensional dense word vector (character embedding) $x_i \in R^d$ and d is the dimension of embedding, based on the pre-trained or randomly initialized word embedding matrix; 4) The next layer of the model is the BiLSTM layer that has the ability to extract the features of sentences. The word embedding sequence $(x_1, x_2, ..., x_m)$ of each sentence is taken as the input to BiLSTM in each step, then the output sequence of hidden states of the forward LSTM $(\overrightarrow{h_1}, \overrightarrow{h_2}, ..., \overrightarrow{h_m})$ and the corresponding output sequence of backward LSTM $(\overleftarrow{h_1}, \overleftarrow{h_2}, ..., \overleftarrow{h_m})$ are combined according to the position $h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}] \in R^n$ and get the complete sequence of hidden states:$(h_1, h_2, ..., h_m) \in R^{m \times n}$; 5) The function of the following linear layer is to map the hidden state vector from n-dimension to k-dimension, where k is the number of labels defining in the tagging scheme. As a result, the sentence features is extracted that are represented as a matrix $P = (p_1, p_2, \ldots, p_n) \in R^{m \times k}$; 6) The parameters of the CRF layer are represented by a matrix $A \in R^{(k+2) \times (k+2)}$, and A_{ij} denotes the score of the transition from the ith label to the jth label. Consider a sequence of labels $y = (y_1, y_2, ..., y_m)$, the following formula used to calculate the score of the labels sequence:

$$score(x, y) = \sum_{i=1}^{m} P_{i,y_i} + \sum_{j=1}^{m+1} A_{y_{j-1}, y_j} \tag{1}$$

The score of the whole sequence is equal to the sum of the score of all word within the sentence, which is determined by the output matrix P of BiLSTM layer and the transition matrix A of CRF layer. Moreover, the Softmax function can be used to obtain the normalized probability:

$$P(y|x) = \frac{e^{(score(x,y))}}{\sum_{y'=1}^{k} e^{(score(x,y'))}} \tag{2}$$

By maximizing the logarithmic likelihood function during model training, the following equation gives the logarithmic likelihood of a training sample (x, y^x):

$$\log P(y^x|x) = score(x, y^x) - log\sum_{y'} \exp(score(x, y')) \tag{3}$$

The model is trained by maximizing the log likelihood function and using the Viterbi algorithm

$$y^* = \arg\max_{y'} score(x, y') \tag{4}$$

to solve the optimal path in the prediction by using dynamic programming.

IV. Experiment

A. Experiment Environment

Our experimental environment is as follows: CPU Intel Xeon E5-2678 v3, RAM: Dual 2.50GHz, GPU: Dual Nvidia GeForce GTX 1080 Ti. The specific parameters of the training process of our models are as follows: batch size=16, epoch=50, LSTM units=256, GRU units=256.

B. Dataset Description

All the tests were run on the CCKS-2019 NER dataset and Second Affiliated Hospital of Soochow University (SAHSU) dataset, which are Chinese medical dataset for NER. These datasets are both consist of 1000 Chinese medical records. The CCKS-2019 NER dataset is an academic evaluation task of 2019 China Conference on Knowledge Graph and Semantic Computing (CCKS 2019), which is the largest and unique public named entity recognition task for Chinese EHR. The SAHSU dataset is 100 pieces of data randomly selected from electronic health records provided by the Second Affiliated Hospital of Soochow University, which is labeled by the KnoWis Lab of XJTLU[1] according to CCKS 2019 NER task rules. All medical entities were annotated using the same six predefined categories, including Diseases and Diagnosis, Image Inspection, Laboratory Inspection, Operation, Drug, Anatomic Site. And all the models were trained on the CCKS-2019 NER dataset.

V. Results and Analysis

We train CNN-LSTM, BiLSTM, BiGRU-CRF and BiLSTM-CRF models with different word embedding, including random embedding, word2vec and BERT, for two Chinese medical NER datasets, CCKS-2019 and SAHSU dataset. Table I has illustrated the models' performance on the testing dataset in terms of precision, recall and F1 score. According to the experiments on those models with random embedding, the performance of BiGRU-CRF model and BiLSTM-CRF model are similar for CCKS-2019 dataset (About 69% in F1 score), which are much better than other two models. However, BiGRU-CRF model can outperform BiLSTM-CRF model for the SAHSU dataset. Based on the observations on the results of each category, the main reason that causes the difference between them is that BiGRU-CRF model can perform much better (About 5% higher in F1 score) in identifying those entities belonging to 'operation' category, even if BiLSTM-CRF model can be slightly superior to it in recognizing medical entities in other categories. Figure 4 demonstrates the change of models'

[1]Research Lab for Knowledge and Wisdom, Xi'an Jiaotong-Liverpool University: https://www.knowis.org

Table I: The results for CCKS-2019 dataset and Second Affiliation Hospital of Soochow University (SAHSU) dataset

Model	CCKS-2019 Dataset			SAHSU Dataset		
	Precision	Recall	F1 Score	Precision	Recall	F1 Score
CNN-LSTM	0.4512	0.5723	0.5013	0.4002	0.5000	0.4424
BiLSTM	0.5820	0.6690	0.6217	0.5532	0.5811	0.5612
BiGRU-CRF	0.6667	0.7106	0.6877	0.6729	0.6520	0.6564
BiLSTM-CRF	0.6732	0.7055	0.6888	0.6266	0.6006	0.6089
W2V-BiLSTM-CRF	0.6448	0.7128	0.6765	0.6175	0.6558	0.6307
BERT-BiLSTM-CRF	0.7384	0.7531	0.7453	0.7529	0.7424	0.7457

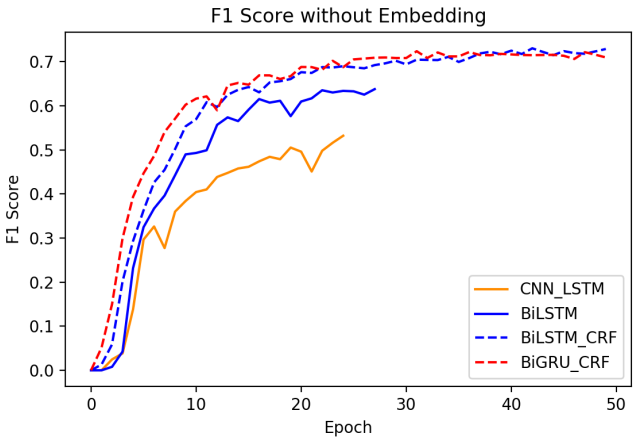

Figure 4: F1 score without Word Embedding

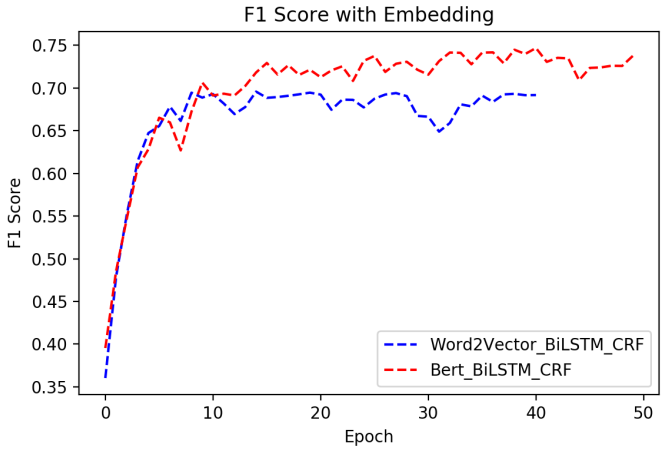

Figure 5: F1 score with Word Embedding

F1 score during the training process. Furthermore, we have also combined different word embedding with BiLSTM-CRF model, including word2vec and BERT. The results have shown that using word2vec as the embedding layer fails to increase the models' performance and even decline it, which probably results from the absence of large-scale medical corpus. In contrast, BERT can be capable of improving the performance of BiLSTM-CRF a lot and this model can achieve approximately 75% F1 score for both datasets. Figure 5 shows the training process of models with word embedding.

VI. CONCLUSIONS

In conclusion, this paper has compared a variety of approaches for Chinese medical named entity recognition and the experiment results have shown that BERT-BiLSTM-CRF model can outperform other models for this NER task and it can achieve about 75% in F1 score. While using random embedding as the word embedding layer, BiLSTM-CRF and BiGRU-CRF can have similar performance (About 69% in F1 score, respectively). In addition, we have found that BiLSTM-CRF model can surpass BiGRU-CRF model in most of predefined categories, except for the 'operation' category. The rationale behind this phenomenon will be explored in the following research. Furthermore, we have also explored the effects of language models for this task, including word2vec and BERT. The results show that BERT can improve the performance of BiLSTM-CRF (approximately 5% in F1 score), compared with the model with random embedding, while the use of word2vec can have an negative effect on the model's performance.

In the future work, more language models will be explored, such as BERT with Whole Work Masking [12], to improve our model's performance. In addition, we are also planning to produce a medical corpus to re-train the existing language models to make them more suitable for the medical field.

VII. ACKNOWLEDGEMENT

This work is partially supported by the AI University Research Centre (AI-URC) through XJTLU Key Programme Special Fund (KSF-P-02) and KSF-A-17. And it is also partially supported by Suzhou Science and Technology Programme Key Industrial Technology Innovation programme with project code SYG201840. We appreciate their support and guidance.

REFERENCES

[1] J. Gao, M. Li, C.-N. Huang, and A. Wu, "Chinese word segmentation and named entity recognition: A pragmatic approach," *Computational Linguistics*, vol. 31, no. 4, pp. 531–574, 2005.

[2] C. Dong, J. Zhang, C. Zong, M. Hattori, and H. Di, "Character-based lstm-crf with radical-level features for chinese named entity recognition," in *Natural Language Understanding and Intelligent Applications*. Springer, 2016, pp. 239–250.

[3] C. Dong, H. Wu, J. Zhang, and C. Zong, "Multichannel lstm-crf for named entity recognition in chinese social media," in *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*. Springer, 2017, pp. 197–208.

[4] J. Xu, H. He, X. Sun, X. Ren, and S. Li, "Cross-domain and semisupervised named entity recognition in chinese social media: A unified model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2142–2152, 2018.

[5] F. Wu, J. Liu, C. Wu, Y. Huang, and X. Xie, "Neural chinese named entity recognition via cnn-lstm-crf and joint training with word segmentation," *arXiv preprint arXiv:1905.01964*, 2019.

[6] L. Guillaume, B. Miguel, S. Sandeep, K. Kazuya, and D. Chris, "Neural architectures for named entity recognition in proceedings of naacl-hlt," 2016.

[7] Y. Zhu, G. Wang, and B. F. Karlsson, "Can-ner: Convolutional attention network forchinese named entity recognition," *arXiv preprint arXiv:1904.02141*, 2019.

[8] L. Luo, Z. Yang, P. Yang, Y. Zhang, L. Wang, H. Lin, and J. Wang, "An attention-based bilstm-crf approach to document-level chemical named entity recognition," *Bioinformatics*, vol. 34, no. 8, pp. 1381–1388, 2017.

[9] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of machine learning research*, vol. 12, no. Aug, pp. 2493–2537, 2011.

[10] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[12] Y. Cui, W. Che, T. Liu, B. Qin, Z. Yang, S. Wang, and G. Hu, "Pre-training with whole word masking for chinese bert," *arXiv preprint arXiv:1906.08101*, 2019.