A Neural Framework for Chinese Medical Named Entity Recognition

Zhengyi Zhao, Ziya Zhou, Weichuan Xing, Junlin Wu, Yuan Chang, and Binyang Li(✉)

School of Information Science and Technology, University of International Relations, Beijing, China
byli@uir.edu.cn

Abstract. Named Entity Recognition (NER) in the medical field targets to extract names of disease, surgery, and the organ location from medical texts, which is considered as the fundamental work for medical robots and intelligent diagnosis systems. It is very challenging to recognize the named entities in Chinese medical texts, because (a) one single Chinese medical named entity is usually expressed with more characters/words than other languages, i.e. 3.2 words and 7.3 characters in average; (b) different types of medical named entities are usually nested together. To address the above issue, this paper presents a neural framework that is constructed by two modules: a pre-trained module to distinguish each individual entity from the nested expressions, while a modified Bi-LSTM module to effectively identify long entities. We conducted the experiments based on the CCKS2019 dataset, our proposed method can identify the medical entity in Chinese, especially for those nested entities embodied in long expressions, and 95.83% was achieved in terms of F1-score, and 18.64% improvement was achieved compared to the baseline models.

Keywords: Named Entity Recognition · Chinese electronic medical records

1 Introduction

Named Entity Recognition (NER) aims at identifying the named entities mentioned in the text, such as the name of persons, locations, and organizations, and classifying them into the predefined categories. To ensure the accuracy of extraction and classification, most of the current research work focuses on recognize the named entity in specific domain, including finance, medical, legal, and political field.

With the rapid development of electronic medical records and medical texts, NER in the medical field has received much attention from both academics and industry. Medical named entity recognition is to identify the names of disease, drug, surgery, afflicted organ, and classify them into predefined types [1]. e.g., CCKS has organized medical named entity recognition open challenge for Chinese electronic medical records (CEMRs) in 2019. These open challenges cannot only provide a batch of high-quality annotated datasets for subsequent research, but also boom many products, medical intelligent diagnosis robot, medical decision support system, and so on.

R. Xu et al. (Eds.): AIMS 2020, LNCS 12401, pp. 74–83, 2020.
https://doi.org/10.1007/978-3-030-59605-7_6

For this purpose, there are some study on the medical NER for English. Based on the compiled English medical text data set and the entity extraction model, the target entity can be extracted well from unstructured text. However, due to the expression style of medical texts and the morphological characters of Chinese, it is very challenging in Chinese medical named entity recognition. In Chinese, a medical named entity is usually constituted by several characters or words, and to express the same meaning, more Chinese characters/words will be required than that of English in most medical named entity expression. Figure 1 illustrates an example from CCKS 2019 dataset. There is only three English words "ulcerative rectal adenocarcinoma" to describe the type of cancer, but in the Chinese expression 浸润溃疡型直肠腺癌(ulcerative rectal adenocarcinoma), it consists of 3 words (9 characters) in total. According to our statistics on CCKS 2019 dataset, 3.2 words and 7.3 characters in average are required to form a Chinese medical entity.

Example: 患者术后病理示为浸润溃疡型直肠腺癌。
Translation: The ulcerative rectal adenocarcinoma was shown by postoperative pathology. showed.

Fig. 1. An example of medical NER from CCKS 2019 datasets.

More importantly, it is frequently occurring that one complex entity may cover multiple entities, and different types of medical named entities are nested together. In Fig. 1, for the entity 直肠腺癌(rectal adenocarcinoma), only part of its Chinese expressions, i.e. 直肠腺(rectal gland), 直肠(rectum), are also medical entities without any morphological changes. But the corresponding English expressions can be easily identified because different morphological structures are used in the entity.

To address the above issue, this paper presents a neural framework of BERT + Bi-LSTM + CRF, named BBC, that is constructed by two modules: a Bi-LSTM module that can capture the dependency with long distance within a sentence and extract the medical entities with multiple words or characters; a pre-trained model that can better express the rich semantics of medical texts and represent the relations between sentences [2]. The explements were conducted based on CCKS 2019 datasets, experiments show that BBC can identify the medical entity in Chinese, especially for those nested entities embodied in long expressions, and 95.83% was achieved in terms of F1-score, and 18.64% improvement was achieved compared to the baseline models.

2 Related Works

As a classic task of NLP, there are many studies on NER. The previous work can be generally divided into the following categories: rule-based, unsupervised learning, supervised learning, and deep learning based method. Since this paper is focusing on medical area, we also review some corpus as well as the techniques on medical NER.

(1) Traditional Methods

The rule-based NER methods relied on handcrafted rules, which performed well when the dictionary (such as WordNet [3]) was well constructed, but it could not be easily transferred to other domains, which would achieve the result with high precision and low recall. For those domains without resources, unsupervised learning is used to extract named entities from clustering groups based on contextual word similarity, calculate lexical resources, lexical patterns, and statistics on a large corpus and use them to predict the phrases of named entities. Huang et al. [4] applied unsupervised entity linking to link entity mentions to a domain-specific knowledge base. The framework learned the entity distribution through the global contexts, and a specific representation of entities from the knowledge base, and then identified the entity as well as the types.

In the supervised learning method, NER is generally regarded as a classification or sequence labeling task. Many existing machine learning algorithms were applied, which include Hidden Markov Models (HMM), maximum entropy Markov models (MEMMs), Support Vector Machines (SVM), and Conditional Random Fields (CRF). Li et al. describe an SVM model that used an uneven parameter to improve the performance for document classification [5].

(2) Deep Learning-Based Methods

In recent years, Huang et al. [6] first applied a BI-LSTM-CRF model to NLP benchmark sequence labeling datasets, that captured the features by Bi-LSTM effectively, and output the tag information through the CRF module Rei Maiek [7] uses the attention mechanism to dynamically use word vector and character vector information based on the RNN-CRF model structure. Akash Bharadwaj [8] later added phonological features to the original Bi-LSTM-CRF model and used the attention mechanism on character vectors to learn to focus on more efficient characters.

(3) Medical NER Corpus

Early NER tasks for medical texts were mostly rule-based and dictionary-based methods. Some corpus were constructed and widely used, such as MedLEE [9] 和GENIES [10]. Recently, Wang [11] annotated a corpus of medical progress notes with more than 15,000 medical named entities in 11 entity types and proposed a system combining CRF, SVM and maximum entropy models to reduce the misclassification. The corpus was used on the i2b2 challenge 2010 [12, 13] has also been applied to the medical text NER. In China, there are far fewer data sets based on Chinese medical texts than on English medical texts. More typical are the Chinese medical text evaluation dataset provided by the CCKS conference and the Q&A data of the medical online forum. The follow-up experiments in this paper are based on the CCKS 2019 dataset [14].

3 Methodology

3.1 Problem Definition

To better understand the task of medical named entity recognition, we firstly give the definition and the type of medical entity, and then introduce our model.

In this paper, we follow the definition of medical NER in CCKS 2019 evaluation [14], where the medical entities can be divided into 6 types: *disease, imaging test, laboratory testing, operation, medicine, and anatomic site.*

Without the loss of generality, we assume that we define that the input Chinese medical text is composed of a set of words, represented as $s = \{w_1, w_2 \cdots, w_n\}$. The objective of medical NER is to identify a successive sequence of m words $\{w_{\mathrm{i}-\mathrm{m}}, w_{i-m+1} \cdots, w_i\}$ as the medical named entity and classify it into one of the above types.

3.2 Model

Recall that the Chinese medical NER faces two challenges, the long entity recognition and nested structure. The former one will result in the incomplete identification of medical entity, while the latter one will classify the entity into an incorrect type, which will in turn affect the performance.

To solve the above problems, this paper presents a neural framework constructed by *BERT+Bi-LSTM+CRF* (BBC) modules. The overall structure of BBC is shown in Fig. 2. Unlike word2vec model [15], BERT is designed to pre-train the representation to further increase the generalization ability of the word2vec model and describe the rich features of character-level, word-level, sentence-level, and even inter-sentence relationships. So, the pre-trained module can well represent the inter-sentence contextual information to and accurately learn the characteristics of complex entities, thereby improving the accuracy of complex entity recognition [16].

Moreover, Bi-LSTM is able to solve the long-distance dependency of the medical text and make use of contextual features. Therefore, in this task, Bi-LSTM is trained to perform category prediction based on each word in the long entity, so as to express the actual meaning of the word more accurately in the long entity of the whole text.

In summary, our framework uses BERT as the model of the embedding layer and use the Bi-LSTM-CRF model to predict the label of each word in the medical text. This combination of model effectively increases the accuracy of the long entity recognition as well as complex entity recognition.

Pre-trained Module

In order to identify the nested entity, we should capture the relation between sentence, i.e. inter-sentence semantics. So, we design a pre-trained module in our model. On one hand, the pre-trained embedding can well represent the context semantics between sentences; on the other hand, the domain-dependent semantics can also be well represented. More detailly, in our pre-trained module, we employ the BERT-based Chinese model provided by Google [16]. The outputs generated by the last layer of BERT are the input of the subsequent neural network model.

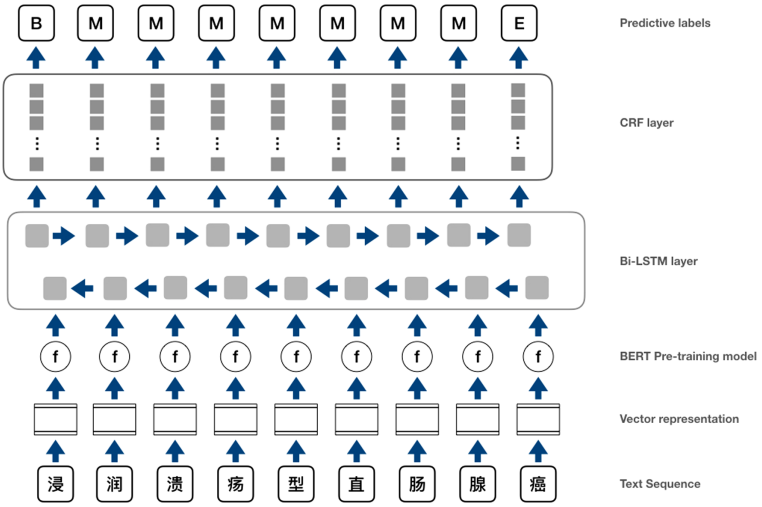

Fig. 2. Model architecture.

The BERT model follows the structure of GPU model and uses transformer encoder as the main model structure. Transformer abandons RNN circular network structure and models a text entirely based on attention mechanism [17].

Bi-LSTM Module

After the pre-trained module, the embedding layer is to convert the text into a set of corresponding word vectors, $x = \{x_1, x_2 \cdots, x_n\}$. We then attempt to capture the intra-sentence contextual information to identify the entity from the nested structure. Inspired by the excellent performance of [16], we also adopt Bi-LSTM. The structure of the Bi-LSTM neural network is shown in Fig. 3. Assume that there are input vector x_t, cell state C_t, temporary cell state $\tilde{C}_t$, hidden layer state h_t, forget gate f_t, memory gate i_t, output gate o_t at time t. We have access to calculate and pass on useful information for subsequent time by memorizing new information in the current unit state, while useless information is discarded, and the hidden state h_t is output at each time step.

Given the hidden state h_{t-1} at the last moment and input word vector x_t, we compute i_t, f_t, $\tilde{C}_t$, previous cell stateC_{t-1} as follows:

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right)$$

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right)$$

$$\tilde{C}_t = \tanh\left(W_C \cdot \left[h_{t-1}, x_t\right] + b_C\right)$$

The current cell state C_t and hidden layer state h_t can be calculated as follows:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

$$h_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right) * \tanh(C_t)$$

Therefore, we obtain the hidden vector $h = \{h_0, h_{01}, \cdots, h_{n-1}\}$.

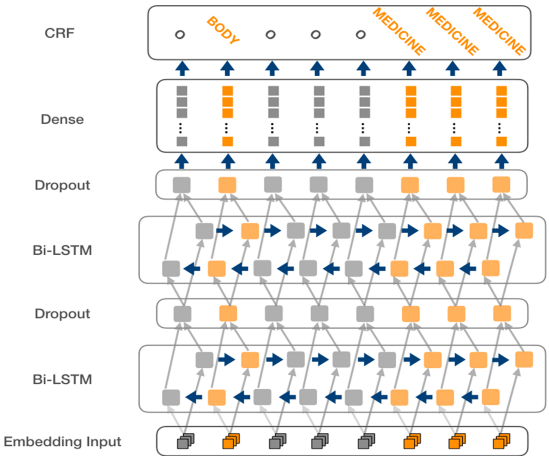

Fig. 3. Bi-LSTM-CRF structure.

Bi-LSTM-CRF

With the forward and backward passes over the network, the model has access to both past and future features. For the long entity 浸润溃疡型直肠腺癌, it can be maintained as a whole in the Bi-LSTM neural network. Due to the multi-dimensional vector expression of the pre-trained model, the entity features can be well represented, and the " 浸润溃疡型直肠腺癌" can be identified as an entire entity instead of the composition of other short entities. In addition, a dense layer is connected after the probability output to perform the possible splicing of all labels such that the layer can be sufficiently achieving local optimum.

CRF. In our model, we also adopt CRF for the labeling output, where the training is based on MLE and optimization algorithm [18]. For the training datasets, the parameters of the model can be adaptively obtained by maximizing the log-likelihood function of the data, where the log-likelihood function is:

$$L(w) = L_{\tilde{P}}(P_w) = \log \prod_{x,y} P_w(y|w)^{\tilde{P}(x,y)} = \sum_{x,y} \tilde{P}(x, y) \log P_w(y|x)$$

where $P(X, Y)$ is the empirical probability distribution.

After that, we utilize the improved iterative scaling method to continuously update the lower bound of the variable to maximize the log-likelihood function. For the case that the classification probability of each word in "invasive ulcerative rectal adenocarcinoma" is different, CRF can well identify the " 浸润溃疡型直肠腺癌" corresponding to the "*disease*" label by probabilistic labeling of long entities.

4 Experiment

4.1 Experiment Setup

Datasets
To investigate the performance of our BBC model, we conduct a set of experiments based on the CCKS 2019 dataset [14]. CCKS 2019 dataset is consisted of 1,000 real Chinese electronic medical records, and the entities were manually labeled and classified into six predefined categories: *disease, imaging test, laboratory testing, operation, medicine, and anatomic site.* Some statistics on the dataset were shown in Table 1, including the number and the proportion of long entity and complex entity in the six categories.

We argue that the performance of Chinese medical NER was usually affected by the nested entities or the long entities. For ease of comparison, we further classify them from two angles: the character amount and entity structure. The overall medical entity can be divided into simple entity and complex entity, where the structure of simple entity is relatively fixed without nested entities, such as " 左附件", " 奥沙利铂", etc. On the contrary, the complex entity is with more complicated structure, and some entities are nested together, such as " 直肠腺癌" includes " 肠", " 直肠", " 直肠腺" and " 腺癌".

We also classify the Chinese medical entities into long entity and short entity. The short entity is constituted by five or fewer Chinese characters, such as " 腹", " 胃癌", while the long entity contain more than five Chinese characters that may involve some complex entities, such as " 左侧盆腔淋巴结", " 转移性低分化腺癌".

Table 1. Statistics on the CCKS 2019 dataset.

Categories	#Long entities	Proportion (%)	#Complex entities	Proportion (%)
anatomic site	534	36.08	522	35.27
operation	683	89.40	269	35.21
disease	1,516	72.09	1,210	57.54
medicine	71	15.74	35	7.76
imaging test	81	36.82	70	31.82
lab test	89	56.04	42	14.53
total	2,974	54.49	2,148	40.48

It is obviously that the proportion of the long entity and complex entity is relatively big, which proved the motivation of our work.

Metrics
Similar with other NER tasks, we adopt *Precision*, *Recall*, and *F1-score* to evaluate the performance of our model. Let $S = \{s_1, s_2, \cdots s_m\}$ denote the output, while $G = \{g_1, g_2, \cdots g_m\}$ denote the correct result, the metrics can be computed as follows.

$$P = \frac{|S \bigcap_r G|}{|S|},\ R = \frac{|S \bigcap_r G|}{|R|},\ F-1 = \frac{2PR}{P+R}$$

where $\cap_r$ represents the intersection of .. and G.

Compared Methods
Since there was no method for Chinese medical NER, we redesigned a conventional CRF model [19] as our baseline model. We then compared our model with the baseline model and Huang's model [6]. To better demonstrate the effectiveness of BERT and Bi-LSTM, we also removed some modules of BBC and got the results.

4.2 Results

Table 2 showed the experimental results based on the CCKS 2019 dataset.

Table 2. Experimental results on CCKS 2019.

Types	Precision (%)	Recall (%)	F1-score (%)
Baseline model	79.94	74.63	77.19
Huang's model	76.92	85.07	79.13
Word2Vec+Bi-LSTM+CRF	85.62	84.66	83.90
BBC	**94.88**	**96.80**	**95.83**

In Table 2, BBC model achieved the best run on all the metrics, and 18.64% and 16.7% improvement were reached against the Baseline model and Huang's model in terms of F1-score. *Word2Vec+Bi-LSTM +CRF* performed better than *Huang's model*, it was because that Huang's model adopted a universal word embedding, while the *Word2Vec+Bi-LSTM +CRF* model used a medical domain-dependent word embedding. BBC incorporated pre-trained model outperformed the above two, it proved that domain-dependent information was very useful in medical NER. *Word2Vec+Bi-LSTM +CRF* outperformed Baseline model, and *Huang's model* beat the Baseline model, it proved that the Bi-LSTM module could well represent the intra-sentence semantics.

Table 3 and Table 4 provided the insights of each individual type of medical NER. We could find that the BBC could achieve 99.48% and 96.83% of F1-score in average on long entity and complex entity, respectively. As to the largest amount of long entity and complex entity, i.e. *disease* type, it could achieve 93.11% and 98.37% of F1-score, which was comparable with short entity and simple entity, respectively. That means our BBC model could effectively identify the long entity and the complex entity. Recall the example in Fig. 1, it could be successfully recognized by BBC model.

Table 3. Insights on the performance of long entity for each individual type.

Types	Long entity			Short entity		
	PRE	REC	F−1	PRE	REC	F−1
disease	94.69	91.58	93.11	94.99	98.96	96.93
operation	100.00	100.00	100.00	100.00	100.00	100.00
anatomic site	97.76	98.20	97.98	100.00	100.00	100.00
lab test	16.67	9.41	12.03	38.42	45.36	41.61
imaging test	94.44	97.14	95.77	100.00	100.00	100.00
medicine	100.00	94.44	97.14	95.21	95.47	95.34
total	99.17	99.79	99.48	93.36	95.71	95.42

Table 4. Insights on the performance of complex entity for each individual type

Types	Complex entity			Simple entity		
	PRE	REC	F−1	PRE	REC	F−1
disease	96.79	100.00	98.37	93.55	91.64	92.58
operation	100.00	100.00	100.00	100.00	100.00	100.00
anatomic site	99.05	99.52	99.29	99.55	99.55	99.55
lab test	9.38	100.00	17.14	36.73	32.60	34.54
imaging test	96.67	100.00	98.30	100.00	99.38	99.69
medicine	100.00	100.00	100.00	95.30	95.30	95.30
total	93.98	99.85	96.83	92.80	91.49	92.14

5 Conclusion

This paper targets on medical named entity recognition from Chinese medical records. To tackle with the challenges of long entity and nested structure in Chinese medical texts, this paper presents a neural framework by incorporating pre-trained module and Bi-LSTM module. Beneficial from the modules, both of intra-sentence semantics and inter-sentence semantics can be well captured so as to significantly improve the performance on Chinese medical NER. Based on the CCKS2019 evaluation dataset, those entities hidden in the nested structures and the entity with multiple words can be successfully identified. 95.83% was achieved in terms of F1-score, and 18.64% improvement was achieved compared to the baseline models.

Acknowledgements. This research is supported by the Natural Science Foundation of China (61976066, 61502115, U1636103), the Fundamental Research Fund for the Central Universities (3262019T29), the Joint funding (SKX182010023, 2019GA35) and Students' Academic Training Program of UIR (3262019SXK15).

References

1. Hopfield, J.: Neural networks and physical systems with emergent collective computational abilities. https://doi.org/10.1142/9789812799371_0043
2. de Benito-Gorron, D., Lozano-Diez, A., Toledano, D.T., Gonzalez Rodriguez, J.: Exploring convolutional, recurrent, and hybrid deep neural networks for speech and music detection in a large audio dataset. EURASIP J. Audio Speech Music Process. **2019**(1). https://doi.org/10.1186/s13636-019-0152-1
3. Fellbaum, C.: WordNet. In: Poli, R., Healy, M., Kameas, A. (eds.) Theory and Applications of Ontology: Computer Applications, pp. 231–243. Springer, Dordrecht (2010). https://doi.org/10.1007/978-90-481-8847-5_10
4. Huang, L., May, J., Pan, X., Ji, H.: Building a fine-grained entity typing system overnight for a new X(X = Language, Domain, Genre), 10 March 2016. arXiv:1603.03112v1
5. Li, Y., Bontcheva, K., Cunningham, H.: SVM based learning system for information extraction. In: Winkler, J., Niranjan, M., Lawrence, N. (eds.) DSMML 2004. LNCS (LNAI), vol. 3635, pp. 319–339. Springer, Heidelberg (2005). https://doi.org/10.1007/11559887_19
6. Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF models for sequence tagging, 9 August 2015. arXiv:1508.01991
7. Marek, R., Crichton, G.K.O., Pyysalo, S.: Attending to characters in neural sequence labeling models, 14 November 2016. arXiv:1611.04361
8. Bharadwaj, A., Mortensen, D., Dyer, C., Carbonell, J.: Phonologically aware neural model for named entity recognition in low resource transfer settings. https://doi.org/10.18653/v1/d16-1153
9. Friedman, C., Alderson, P., Austin, J., Cimino, J., Johnson, S.: A general natural-language text processor for clinical radiology. J. Am. Med. Inform. Assoc. **1**(2), 161–174 (1994)
10. Friedman, C., Kra, P., Yu, H., Krauthammer, M., Rzhetsky, A.: GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. Bioinformatics **17**(Suppl. 1), S74–S82 (2001)
11. Wang, Y.: Annotating and recognising named entities in clinical notes. https://doi.org/10.3115/1667884.1667888
12. Uzuner, O., South, B., Shen, S., Duvall, S.: 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J. Am. Med. Inform. Assoc. **18**(5), 552–556 (2011)
13. Kiritchenko, S., de Bruijn, B., Cherry, C.: NRC at i2b2: one challenge, three practical tasks, nine statistical systems, hundreds of clinical records, millions of useful features. In: Proceedings of the 2010 i2b2/VA workshop on challenges in natural language processing for clinical data (2010)
14. CCKS 2019 NER of CEMR. https://www.biendata.com/competition/ccks_2019_1/
15. Rong, X.: word2vec parameter learning explained, 11 November 2014. arXiv:1411.2738v4
16. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding, 11 October 2018. arXiv:1810.04805
17. Gong, C., Tang, J., Zhou, S., Hao, Z., Wang, J.: Chinese named entity recognition with Bert. ISBN: 978-1-60595-651-0 (2019)
18. Xishuang, D., Shanta, C., Lijun, Q.: Deep learning for named entity recognition on Chinese electronic medical records: Combining deep transfer learning with multitask bi-directional LSTM RNN. PLoS ONE (2019). https://doi.org/10.1371/journal.pone.0216046
19. Konkol, M., Konopík, M.: CRF-Based Czech named entity recognizer and consolidation of Czech NER research. In: Habernal, I., Matoušek, V. (eds.) TSD 2013. LNCS (LNAI), vol. 8082, pp. 153–160. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40585-3_20