A Novel Method of Chinese Electronic Medical Records Entity Labeling Based on BIC model

Yifan Wang[1], Guowei Teng[1*], Xuehai Ding[2], Guoqing Zhang[3], Yunchao Ling[3], Guozhong Wang[1]

[1] Shanghai Institute for Advanced Communication and Data Science, School of Communication and Information Engineering, Shanghai University, Shanghai, China.
[2] School of Computer Engineering and Science, Shanghai University, Shanghai, China.
[3] Bio-Med Big Data Center, Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences.

* Corresponding Guowei Teng. E-mail: tenggw@shu.edu.cn
Manuscript submitted May 4, 2020; accepted July 23, 2020.
doi: 10.17706/jsw.16.1.24-38

Abstract: In the field of bio-medicine, mass data are generated every day, such as Chinese electronic medical record (EMR), containing massive medical terminology and specific categories of entities. The way to analyze and obtain effective information from these sparse data is a difficulty in research. As the foundation of analyzing huge amount of biomedical text data, Named Entity Recognition (NER) is essential in Natural Language Processing (NLP) complementing with effective labeling data. One of the two basic sequence labeling methods is rule-based bulk corpus tagging, requiring domain experts to establish targeted recognition rule base. However, in the application field, this method is single, and the portability does not make the expectation, bringing great limitations; The other is complete manual labeling, but it is time-consuming and laborious. Based on Bidirectional Long Short-Term Memory network (BiLSTM), Iterated Dilated Convolution Neural Network (IDCNN) and Conditional Random Field (CRF), we proposed the BIC model. This paper proposes a method for EMR entity labeling based on BIC model, realizing automatic annotation of Chinese EMR data. Machine labeling data can be used after manual review, which will reduce the workload of manual labeling bestially. Compared with other models, F1 value of BIC model reached 91.90% in CCKS2017 dataset, and 78% in PACS report data. Experiments show that our method is superior to the others.

Key words: Chinese electronic medical record, named entity recognition, sequence labeling, BIC model, neural network.

1. Introduction

Intelligent precision medicine is now attracting more public's attention with a considerable speed, as a result of the big data explosion. The establishment of the knowledge graph [1] in the biomedical field can provide a knowledge base for intelligent precision medicine. The triple is the most basic unit of the knowledge graph [2], formed in "entity1-relationship-entity2". To extract effective triples, identifying the entities and relationship between the entities is needed. Medical named entity recognition (NER) commits to identify the entity that expresses patients' medical or health information in texts such as electronic medical record (EMR).

Electronic medical records refer to digital information generated by medical institution information

systems, such as text, symbol, chart, graph, data and image, which can realize the storage, management, transmission and reproduction of medical records [3]. Electronic medical records contain many entities, most types of them mainly composed of four categories of entities. (1) Private health information (PHI) entities, such as the name and number of patients, doctors, and medical institutions. (2) Entities closely related to the treatment of diseases, such as diseases, symptoms, inspections and treatments. (3) The therapeutic entities containing information about the name of drug, the dose, the mode of administration, the frequency of treatment, the duration of treatment, etc. (4) The time entities associated with patients' treatment and progression.

A lot of research has been carried out on named entity recognition in all kinds of fields around the world, especially in bio-medicine [4]. At present, most electronic medical record information extraction research is aimed at EMRs in English. The research on EMRs in Chinese has not yet been clear and systematic. The lack of training corpus restricts the research of Chinese EMR information extraction greatly in the past. With the extensive implementation of Chinese EMR systems, the number of EMRs has increased dramatically in recent years. However, it is a popular but difficult point to extract effective information from a huge number of EMRs. The construction of Chinese EMR tagging corpus is the basis of research. Open corpus is heavily limited in the field of bio-medicine since that manually labeling data requires a lot of human resources. It usually takes more than 10 times to annotate data manually than it does to annotate data by machine. Therefore, the way to reduce the workload of manual labeling and ensure the correctness of biomedical entity recognition has always been a difficulty in research.

The main contributions of our work are as follows:

1) We proposed the BIC model, an integrated model solving Named Entity Recognition (NER) task. This model can extract the global features and local detail features of the sentence, ensuring the integrity and accuracy of information extraction.
2) We transformed NER task into entity labeling task and implement it with algorithm, with which we can take less time to obtain more labeled data than before.

The first part of this paper is introduction, common methods of entity recognition task are outlined in the second part. Then, according to the treatment part in the EMRs, the tagging rule will be proposed according to the actual needs. In the third part, a new model "BIC" will be proposed, which is a sequence labeling model composed of BiLSTM, IDCNN and CRF. After that, the sequence labeling model will be used to annotate unlabeled data, and the data processing algorithms will be introduced to obtain more label data. The fourth part will prove the performance of the model and the correctness of the method. The fifth part summarizes the work of this paper and looks forward to the development trend of EMR information extraction.

2. Related work

Entity is the basic information element of text. NER is a basic task of natural language processing (NLP), to find the entity from a piece of text and mark the location and category of the entity. At present, researches on NER of medical texts mainly focus on EMRs and medical documents, with three main methods for identifying named entities: (1) dictionary-based and rule-based methods, (2) machine learning methods [5], (3) deep learning methods [6].

It is necessary to establish a named entity with a broad range of coverage dictionary in the method based on dictionary, and add the corresponding class, such as abbreviations, synonyms, deformation, etc. The matching algorithm is used to recognize named entities in the text, but lacks the compatibility of new named entities. In the medical field the dictionary updates frequently with a huge number of term quantity. Although the rule-based method compensates partly for the dictionary-based method failing to identify

words not included, it takes too much time for experts in specific fields to establish targeted recognition rule base.

Models based on machine learning include support vector machine (SVM) [7], hidden Markov model (HMM) [8], maximum entropy Markov model (MEMM) [9], and conditional random field (CRF) [10].

Approach based on deep learning begins from the initial multi-layer perceptron (MLP) to the long short-term memory network (LSTM) in recurrent neural network (RNN) [11]. Then, Huang et al. [12] proposed a seminal work of the BiLSTM-CRF structure, LSTM-CRF [13]-[16] model was formed by combining conditional random field (CRF). After that, bidirectional LSTM network appeared, forming BiLSTM-CRF model. BiLSTM-CRF model recognizes entity more effectively than extensive baseline method [17]. Habibi *et al*. [18] is the first to apply it in biomedical NER method. Wang *et al*. [19] made a great effort to clinical NER. Ma *et al*. [20] proposed that NER commonly uses CNN network as embedding layer, bidirectional LSTM network as the encoding end and CRF as the decoding end to form the BiLSTM-CNN-CRF model.

3. Preliminary

3.1. BILSTM

BiLSTM-CRF can be used for sequence labeling tasks with forward and backward input features, and it uses sentence level labeling information to make the algorithm more robust. BiLSTM contains forward LSTM layer and backward LSTM layer, and the input of sentences is the input of LSTM network in the form of one-hot coding through the word embedding layer. Let $W=\{w_1,\ldots,w_{t-1},w_t,w_{t+1},\ldots,w_n\}$, $w_t \in \mathbb{R}^d$, express a sentence of length n, and the t^{th} word of this sentence is a d-dimensional vector. The structure of the LSTM, shown in Fig.1, consists of a sub-net of circular connections called memory blocks. Each time step is an LSTM block that computes the current hidden layer vector h_t and the current cell state vector c_t based on the previous hidden layer vector h_{t-1}, the previous cell state vector c_{t-1}, and the current input word vector w_t.

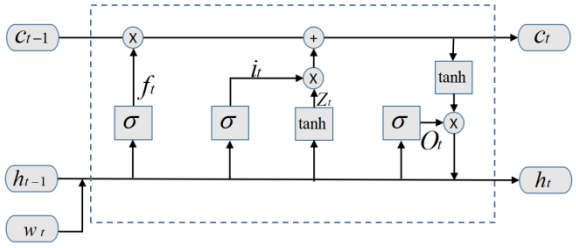

Fig. 1. LSTM block.

where i_t is the input gate, f_t is the forgotten gate, o_t is the output gate, LSTM memory cell can be expressed as follows:

$$i_t = \delta(W_{wi} w_t + W_{hi} h_{t-1} + b_i) \quad (1)$$

$$f_t = \delta(W_{wf} w_t + W_{hf} h_{t-1} + b_f) \quad (2)$$

$$z_t = \tanh(W_{wc} w_t + W_{hc} h_{t-1} + b_c) \quad (3)$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes z_t \quad (4)$$

$$o_t = \delta(W_{wo} w_t + W_{ho} h_{t-1} + b_o) \quad (5)$$

$$h_t = o_t \otimes \tanh(c_t) \quad (6)$$

In formulas (1)(2)(3)(5), $W_{(.)}$ represents the weight value, b represents bias value, and in formulas (1)(2)(4)(5)(6), c represents cell state vector. For each word vector w_t, the context information is included, and the forward LSTM layer is coded from w_1 to w_t, recorded as $\vec{h}_t$. The backward LSTM layer is coded from w_n to w_t, recorded as $\overleftarrow{h}_t$. The final combination of the context information to form the vector representation of the t^{th} word is $h_t = [\vec{h}_t, \overleftarrow{h}_t]$.

3.2. CRF

Given a training data set $D = \{(x^1, y^2), \ldots, (x^N, y^N)\}$, x^i is observation sequence of N data, which y^i is its corresponding marking sequence. CRF uses logarithmic likelihood function to maximize conditional probability for parameter estimation [21], conditional probability is shown in formula (7):

$$L(\mathbf{W}, \mathbf{b}) = \sum_{i=1}^{N} \log(p(y^i \mid x^i; \mathbf{W}, \mathbf{b})) \tag{7}$$

$\mathbf{W}$ is model parameter-weight, $\mathbf{b}$ is model parameter-bias. When the conditional probability likelihood function is maximized, the best label obtained is $\mathbf{y}^*$,

$$y^* = \arg\max_{y \in Y(\mathbf{x})} p(\mathbf{y} \mid \mathbf{x}; \mathbf{W}, \mathbf{b}) \tag{8}$$

3.3. IDCNN

The classical convolution filter acts on a continuous area of the matrix to slide, while the dilation convolution [22] adds a dilated width to the filter and skips the data in the middle of it when sliding on the input matrix. The filter matrix size is the same, but it can obtain a wider range of input matrix data. Suppose W_c is a convolution kernel with a width of r, and c_t is the output of convolution W_c operation on vector sequence x_t, their relations can be represented as follows:

$$c_t = W_c \bigoplus_{k=0}^{r} x_{t \pm k} \tag{9}$$

The dilated convolution is sliding a dilated width with step size δ when the convolution operation of W_c is performed on x_t, and the output c_t through the dilated convolution is shown in formula (10):

$$c_t = W_c \bigoplus_{k=0}^{r} x_{t \pm k\delta} \tag{10}$$

In formula (10), while δ equals 1, the dilated convolution becomes a normal convolution operation. While $\delta > 1$, the dilated convolution will obtain a wider range of receptive fields than the normal convolution. The Iterated Dilated Convolution (IDCNN) model proposed by Strubell E et al. [23] repeatedly applies the same small stacked dilated convolution block, and it takes the result of the previous dilated convolution as the input for each iteration. The dilated width increases exponentially alone with the number of layers increases, but the number of parameters increases linearly, so, the receptive field can quickly cover all the input data. The model is composed of four dilated convolution blocks of the same size, and the dilated width in each dilated convolution block is a three-layer dilated convolution of 1, 1 and 2 respectively. The sentence is input into the IDCNN model, and the feature is extracted through the convolution layer. IDCNN can also be used for sequence labeling tasks, with higher efficiency when the model accuracy is comparable to the BiLSTM model.

3.4. Evaluation

The common evaluation methods used mostly in information retrieval are precision, recall and F1 value. Precision represents the percentage how many of the entities marked by the machine are correct. Recall represents the percentage how many entities are marked correctly by machine in all entities. F1 value is the number comprehensively consider the precision rate and recall rate, then calculate the harmonic mean. Therefore, this paper uses the evaluation methods commonly used in information retrieval. The precision P, the recall rate R and the F1 value are shown in formula (11)(12)(13).

$$\text{Precision(P)} = \frac{\text{Number of entities system correctly identifies}}{\text{Number of entities system identifies}} \tag{11}$$

$$\text{Recall(R)} = \frac{\text{Number of entities system correctly identifies}}{\text{Number of entities in the document}} \tag{12}$$

$$F1 = \frac{2PR}{P+R} \tag{13}$$

4. Method

The sequence labeling model "BIC" combines deep learning (BiLSTM, IDCNN) and machine learning (CRF). The difficulty of CRF lies in selecting and constructing features. The advantage of BiLSTM, IDCNN is that they select structural features automatically based on training corpus, without requiring artificially construct and selecting features. Therefore, BiLSTM and IDCNN are combinedly used as the encoding end, and CRF is used as the decoding end.

Firstly, the corresponding medical entity tagging rule [24] will be given according to actual needs, and a small amount of data will be manually labeled which uses the BIO annotation method [25]. Secondly, the manual labeled data will be processed into training data required by the model. Thirdly, train the parameters to obtain the trained sequence labeling model. The entity annotation method block diagram is shown in Fig.2:

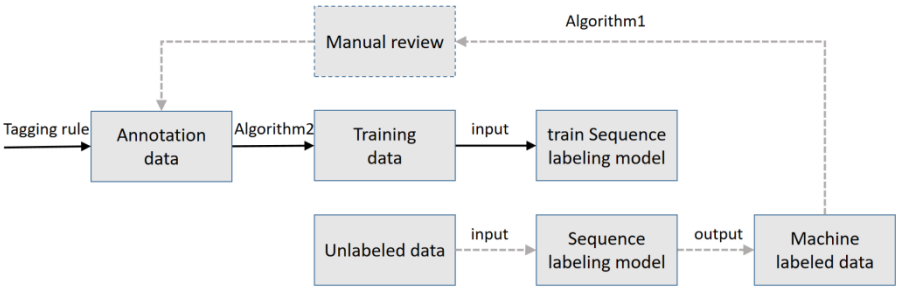

Fig. 2. Entity annotation method block diagram.

As shown in Fig.2, the unlabeled data will be input into the sequence labeling model, then we obtain the machine labeled data. After manually reviewed simply and quickly, the machine labeled data can be converted into the training data, which can be used to retrain a better sequence labeling model. Manual review obeys "Medical entity tagging rule" like manual label. But several entities has already been annotated by machine, and transformed by Algorithm 1.

The sequence labeling model will be more accurate with the amount of data increasing.

4.1. Medical Entity Tagging Rule

The open data set of EMR named entity identification in biomedical field we use is CCKS2017 evaluation

dataset[1], with entity types shown in Table 1.

Table 1. CCKS2017 Evaluation Data Entity Category Tag

Category	Tag	Description
Symptoms	Sas	Subjective feelings described by patients and Objective facts observed externally.
Inspection	Iat	The basis for clinical diagnosis and treatment provided by medical equipment.
Disease	Dad	Disease that patients have according to the symptoms.
Treatment	Tre	Medicine, surgery, etc.
Body parts	Oga	The part of the human anatomy where diseases, symptoms, and signs occur.

In medical imaging report in EMRs, we focus on the types of entities followed: diseases, drugs, symptoms, organs, states, periods, traits, indicators. The most important entities are organs and symptoms, which are the focus of clinical report in most condition. Based on the characteristics of the annotation data required by the model, we developed a data tagging rule. Tags facilitating the distinction between entities for each entity category are shown in the second column of Table 2. After defining the tag, select the appropriate feature to mark the entity. This paper uses the BIO annotation method containing three annotation types in Chinese NER to annotate the first batch of data. As shown in Table 3, 'B' represents the entity starting word, 'I' represents non-first word of the entity, 'O' represents the non-entity word. The entity tag is preceded by 'B', 'I' to distinguish different entities. Example of tagging is shown in Fig. 3.

Table 2. Entity Category Tag

Category	Tag	Example
Disease	Dis	chronic pancreatitis, gastric cancer
Medicine	Med	Static antibiotic, Ofloxacin eye drops
Symptom	Sym	Nodules, masses, calcifications
Organ	Oga	Pancreas, chest
State	Ste	Wavy, tortuous
Period	Ped	Portal phase, arterial phase
trait	Tra	Volume, shape, profile
Index	Idx	3.5mmX4.5mm

Table 3. BIO labeling Instructions

Type	Type Extension	Instruction
B	Begin	Starting word of the entity
I	Inside	Non-first word of the entity
O	Other	Not a word in any entity

Raw data：

胆囊腔内可见多发大小不一结节状致密影，胆囊壁毛糙. (Multiple nodular dense shadows of varying sizes can be seen in the gallbladder cavity, and the gallbladder walls are coarse.)

Marked data(data format 1)：

[1] https://github.com/info-wyf/NER_BI/blob/master/data/CCKS2017.zip

{{Oga-B:胆}} {{Oga-I:囊腔}} 内{{Act-B:可}}{{Act-I:见}} 多发大小不一{{Ste-B:结}}{{Ste-I:节状}}{{Dis-B:致}}{{Dis-I:密影}} {{Oga-B:胆}}{{Oga-I囊壁}} {{Ste-B:毛}}{{Ste-I:糙}} 。

Final Entity：

"胆囊腔(gallbladder cavity)"、"可见(seen)"、"结节状(nodular)"、"致密影(dense shadows)"、"胆囊壁(gallbladder walls)"、"毛糙(coarse)"

Label data description：

The original data is marked according to "data format 1'', which contains: "{{entity code-(BIO): original text}}". For example: "{{Oga-B:胆}} {{Oga-I:囊腔}}", the gallbladder cavity is an organ entity, and organ entity is marked as 'Oga'. "胆(gallbladder)"' is marked as "Oga-B'', "囊腔(cavity)" is marked as "Oga-I'', and 'B', 'I' can combine to form an organ entity.

4.2. BIC Model

The encoding end of the sequence labeling model is based on BiLSTM and IDCNN, the decoding end is based on CRF, as shown in Fig. 3. The input statement passes through the embedding layer, then it output word vector. The word vector is input to the BiLSTM model to extract the global features, and the output of the BiLSTM is input to the IDCNN. The model extracts the local detail features of the sentence, and then the predicted values of the labels obtained by IDCNN are input to the CRF layer in order to calculate the loss function. The model parameters are optimized according to formulas (7) and (8). After the forward and backward sentence features are extracted by BiLSTM, the high-dimensional features are extracted by IDCNN, and then the parameters are adjusted by CRF, which not only ensures the integrity of information extraction, but also ensures the accuracy of information. The feature selection structure is constructed by using BiLSTM, IDCNN, the obtained feature is decoded by CRF to obtain the result of the final sequence labeling. Combining deep learning and machine learning complements with each other, the theoretically obtained model works well.

The input of the model is Chinese text. According to different length sentences, it is divided into different training batches. Make each training batch have batch-size=20 sentences. A batch of training sentences is converted into vectors through the embedding layer, and each batch of training sentences is compensated. The space reaches the same length. For example, a sentence in a batch of training data has 21 characters and the 20 sentences of the same batch. The length of each sentence is 21 characters, converted into a vector through the embedding layer, and the dimension is [20, 21, 100] dimension vector.

The embedding layer contains "character embedding" and "phrase embedding". Considering that the word segmentation may be misinterpreted, we use word-based input to mark, without losing the word characteristics. Character embedding aims to convert 21 characters of each sentence in each training batch, using word2vec, into [20, 21, 100] dimensional vector; word embedding aims to process 21 characters of each sentence through word segmentation, corresponding to [20, 21, 20] dimensional vector. The output of the final embedding layer is a combination of characters and words, producing a [20, 21, 120] dimensional vector.

The embedding layer obtains the tensor of the input data, processed by the encoding end. The encoding end is formed by the combination of BiLSTM and IDCNN. The number of neurons in the BiLSTM hidden layer is 2*100=200, so the output of the BiLSTM layer is [20, 21, 200]. The IDCNN layer is formed by a combination of four iteration dilated convolutional neural networks. Set the size of each dilated CNN convolution kernel as [1, 3, 200, 100], and the dilated convolution has three layers. The dilation size of each layer is [1, 1, 2], filter width is 3, as shown in formula (10). The convolution operation is performed, and the output is [20, 21, 100]. The final four iterations of the dilated convolution result are spliced to form the output data [20, 21, 400] dimension tensor of the encoding end. The three layers of the dilated convolution,

the output of each layer is the input of the next layer, and the four iterations of the dilated convolution share the parameters between the same dilated convolutional layers, therefore greatly reducing the amount of calculation.

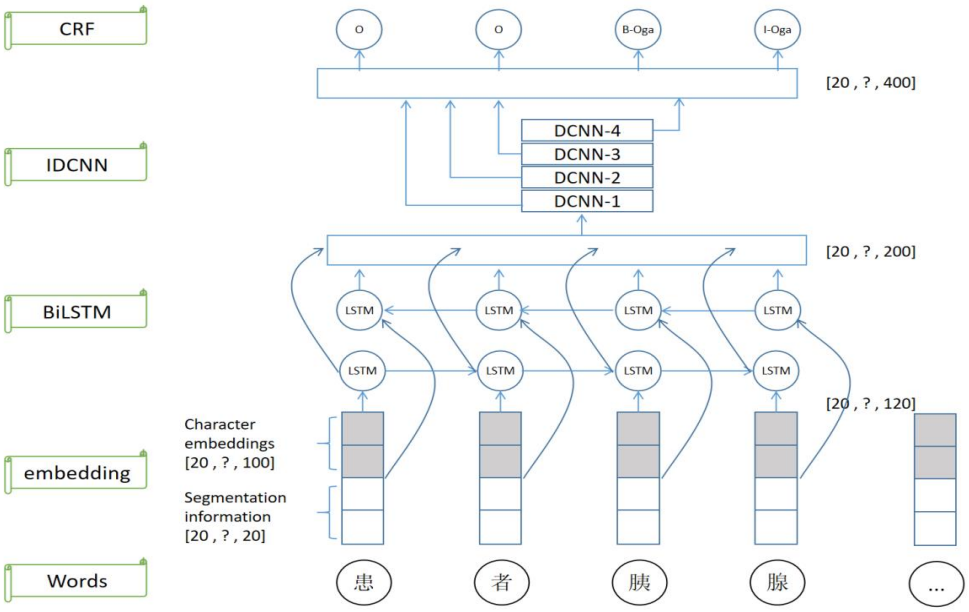

Fig. 3. BIC model diagram.

The decoding end predicts the tag corresponding to the input data by CRF. First, the output of the encoding end passes through a neural network. The weight of the network is [400, 33]. (Table 2 shows that there are 8 types of entities. The label of the table is BIO. The BIO format is converted into BIESO [26] format during the running of the model, so the final label type is 4x8+1=33.) After the neural network, we obtain a [20, 21, 33] dimension tensor, which is the logits for tags. Then we can use logits to calculate the loss of the model shown in formula (8), we get the best label when the conditional probability likelihood function is maximized.

4.3. Data Processing Algorithm

After training the sequence labeling model, the unlabeled data is input into the model, and the result of the machine labeling is output. According to the result of the machine labeling, the processing is formed as "{{tag : original}}", and the specific processing algorithm is shown in Algorithm 1.

Algorithm 1 Machine label data converts to data to be reviewed.

Input: source , entity , begin , end ; /* "begin" and "end" represents the location of "entity" in "source" */
Output: result; /*such as "*{{Oga-B:*}}*{{Oga-I:*}}*"*/

```
list(entity);
for j=len(entity); j > 1; j-- do
    if entity('start')==entity('end') then
        sour_str = source[entity('start')];
    else
        sour_str= source[start : end];
    end if
    tag_str = "{{"+entity('type')+"-B:"+sour_str+"}}";
    back =source[entity[j]['end'] : ];
    front = source[ : entity[j]['start']];
    result = front + tag_str + back;
end for
```

Algorithm 1. Machine annotation data converts to data to be reviewed.

When machine labeled data have been manually reviewed, the correct labeled data is formed. Then the data format required to convert the annotation data to the format model needs, which like "word by word, composed of original text, Tabs, entity mark, Line break." The specific processing algorithm process is shown in Algorithm 2.

Algorithm 2 Label data converts to training data.

```
Input: Tag_data /* such as "{{tag : original}}" */
Output: Train_data /* source,"\t",tag,"\n" */
for i = 1; j < len(entity); j++ do
    if str[i:i+1] == "{{" then
        if str[i+6] == 'B' then
            B_function();
        end if
        if str[i+6] == 'I' then
            I_function();
        end if
    end if
    if str[i:i+1] == "}}" then
        end_function();
    end if
end for
```

Algorithm 2. Label data converts to training data.

In Algorithm 1, the program inputs begin and end position and entity's tag information in source text, then outputs result of "data format 1" in Fig.3. Entities are processed in the form of "{{tag : original}}". The source text is processed by entity in reverse order, we process these entities from back to front and replace the source entity with tag, which ensures the correctness of data processing considering it does not affect the begin and end position of the previous entity.

In Algorithm 2, There are three different situations while processing the labeled data. (1) Contents starting from this character is the part of entity tag when "{{" encountered, (a) Only one character will be labeled as beginning of an entity in the form of "the original text, Tab, entity tag, Line break" when tag 'B' occurs. (b) Many characters will be labeled as the middle or end of an entity when tag 'I' occurs, which needs to be further processed. (2) It marks the end of the entity when "}}" encountered, meaning the following content is the source text or the next entity tag.

5. Experiment

5.1. Environment and Datasets

Our work was developed on Ubuntu 16.04 platform using Python 3.6 language in tensorflow framework, jieba version 0.39. The hardware equipment is as follows: one Intel i7 CPU, 16 GB memory and one GTX-1080Ti graphics card.

We performed the experiments on three datasets: Common dataset[2], CCKS2017 evaluation dataset (CCKS2017), Picture Archiving and Communication Systems report (PACS report).

•Common: common dataset, containing three categories of Person name, Location name, and Institution name.

•CCKS2017: a biomedical field EMR data set, containing five categories of Symptoms and signs, Inspection and testing, Disease and diagnosis, Treatment, and Body parts, which is labeled by tagging rule in Table 1.

[2] https://github.com/zjy-ucas/ChineseNER

•PACS report: provided by Shanghai Institutes for Biological Sciences, containing eight categories of Disease, Medicine, Symptom, Organ, States, Act, Trait, and Index, which is manually labeled by tagging rule in Table 2.

The experimental data is labeled by experts roughly divided into training set, development set and test set. The specific quantity distribution is shown in Table 4:

Table 4. Experimental Data Details

Data sets	Category(kinds)	Train(entry)	Development(entry)	Test(entry)
Common	3	20864	2318	4636
CCKS2017	5	360	40	20
PACS report	8	376	180	56

5.2. Experiment Settings

For the three datasets, we converted the labeled data in data format 1 to trained data in data format 2 shown in 4.1 with Algorithm 2. And took most of each dataset as the training set, and the remaining 10% of each as the development set and the test set. All the experiments take F1 as a metric.

The setting of BIC model[3] training parameters is as follows: Batch size of input data is 20 or 16. The characters are 100 dimensions in the character embedding layer and the words are 20 dimensions in the word embedding layer. The hidden layer of BiLSTM has 200 nodes and the number of IDCNN filters is 200. The dropout value is set to 0.5 or 0.6, the gradient cropping is set to 5 and then use Adam as its optimization algorithm. The epoch size is 100 and the learning rate is set from 0.0001 to 0.02.

5.3. Experiment Results and Analysis

We compared our model with the methods widely used for NER on each dataset. IDCNN-CRF model outperforms better than IDCNN and BiLSTM etc. So in our experiments IDCNN-CRF and BIC models were compared in four datasets. Comparison of F1 value can be seen in Table 5, details are shown in Fig.5 and Fig.6. A detailed comparison of the optimal results of each model in the CCKS2017 dataset in Table 6 and PACS report dataset in Table 7.

Table 5. Comparative Results between BIC Model with State-of-the-Art Deep Models

Model	Datasets		
	Common	CCKS2017	PACS report
IDCNN-CRF (Strubell et al.2017)	90.86	89.90	76.37
Ouyang et al.2017	-	88.85	-
Hu et al.2017	-	91.08	-
Wang et al.2019	-	91.24	-
BIC	**92.03**	**91.90**	**78.00**

It can be seen from Table 5 that BIC model achieves the best results in three datasets. In Common dataset our model achieved 92.03% F1, which is 1.17% higher than the IDCNN model, In CCKS2017 dataset our model achieves 91.90% F1, which is 2% higher than IDCNN model. Ouyang *et al*. [14] achieves 88.85% F1, Hu *et al*. [16] achieves 91.08% F1, and Wang *et al*. [19] achieves 91.24% F1. The best F1 of our model achieves 91.90%, which performes better than other state-of-the-art deep models. In PACS report dataset our model achieves 76.39% average F1, which is 0.86% higher than the IDCNN model. The best F1 of our model achieves 78.00%, 1.63% higher than the IDCNN model.

[3] https://github.com/info-wyf/NER_BI

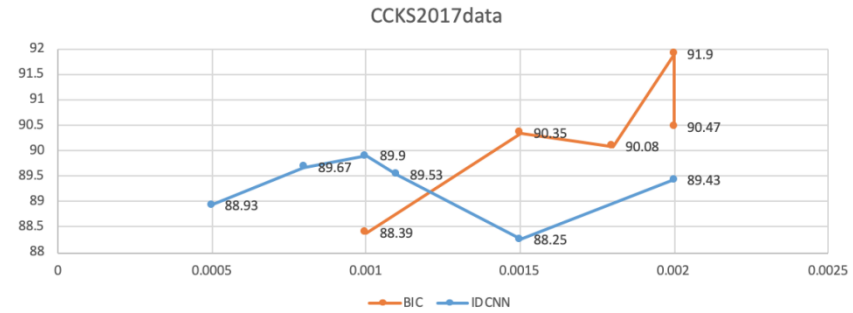

Fig. 4. Results in CCKS2017 evaluation data set.

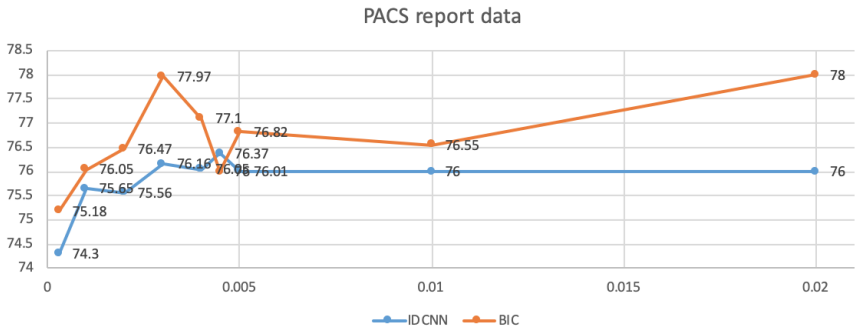

Fig. 5. Results in PACS report data.

From Fig. 4 and Fig. 5 we can see that F1 value of BIC model is 91.90% in CCKS2017 dataset and 78.00% in PACS report data. But F1 vlaue of our model in PACS report data does not achieve 90% because the dataset is relatively small. The result would be better if we have enough dataset.

From Table 6, we clearly observe that our model performs better than the IDCNN model, the higher precision and the higher recall rate. From the test data, IDCNN model recognizes 218 phrases and our model recognizes 227 phrases. Two models have recognized the disease and inspection entities and performed equally well. Our model recognizes more body parts and symptoms entities than IDCNN model, since the precision of symptoms is little lower. Our model performs better in both precision and recall rate in recognizing treatment entities.

Table 6. Results in CCKS2017 Dataset

Category	IDCNN			BIC		
	Precision	Recall	Phrases	Precision	Recall	Phrases
Disease	60.00	100.00	5	60.00	100.00	5
Inspection	95.00	100.00	40	97.44	100.00	39
Body parts	87.96	87.16	108	89.19	90.83	111
Symptoms	92.31	91.14	78	89.41	96.20	85
Treatment	76.92	83.33	13	84.62	91.67	13
Total	89.34	90.46	218	89.72	94.19	227

Table 7. Results in PACS Report Data

Category	IDCNN			BIC		
	Precision	Recall	Phrases	Precision	Recall	Phrases
Act	93.57	91.50	575	93.21	95.75	604
Index	51.75	66.07	143	47.37	64.29	152
Organ	77.79	75.00	1454	79.26	71.89	1408
State	75.02	84.85	1321	73.03	89.04	1424
Symptom	44.31	57.64	562	48.95	64.81	572
Trait	78.78	84.11	410	79.82	94.79	456
Total	74.03	78.86	3306	74.42	81.94	3435

As shown in Table 7, 6 entities are involved rather than all eight entities types due to the labeled PACS report dataset is small and inadequate. The recall rate of the results obtained by the BIC model in PACS report is higher than IDCNN model, but the precision is lower when some entities are identified. There might be some problems of incomplete identification of entities because of the large number of dense medical entities in PACS report dataset.

The experiments on two datasets show that the BIC model can recognize EMR entities more effectively. Since we have small dataset in biomedical field, we can apply our method to identity and label more unlabeled dataset, which would obtain at least 78.00% entities in one sentence. After being 39 manually reviewed quickly and simply, the machine labeled data can be converted into useful data with Algorithm 1.

6. Conclusion

In view of the wide variety of entities, professional descriptions and sparse data in Chinese EMRs, this paper proposes a novel method based on BIC model for sequence labeling. Comparing among the BiLSTM-CRF, IDCNN-CRF and BIC models, BIC model has a good performance whether in common field or in biomedical field. Our method utilizes a combination of machine labeling and manual review, F1 value of machine labeling can reach 78.00%. After manual reviewing, the data can be put into use. The entity categories can be divided more precisely according to the specific situation of actual problems, which helps to analyze the medical entity. In the future, attention mechanism can be added according to the characteristics of EMR data, and intensive learning can be carried out to improve the performance of the method further.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Yifan Wang and Guowei Teng conducted the research; Xuehai Ding and Yunchao Ling analyzed the data; Yifan Wang and Guowei Teng wrote the paper; Guoqing Zhang and Guozhong Wang reviewed and polished the paper; all authors had approved the final version.

Acknowledgment

This work is supported in part by National Key R&D Program of China, Grant NO. 2016YFC0901904, 2017YFC0907505, 2016YFC0901604 and Science and Technology Commission of Shanghai Municipality under grant (No. 16010500400).

References

[1] Rotmensch, M., Halpern, Y., Tlimat, A., Horng, S., & Sontag, D. (2017). Learning a health knowledge graph from electronic medical records. *Scientific Reports*, *7(1)*, 5994.

[2] Qi, G., Gao, H., & Wu, T. (2017). The research advances of knowledge graph. *Technology Intelligence Engineering*, *1*.

[3] Jin-Feng, Y., Qiu-Bin, Y., Yi, G., *et al*. (2014). An overview of research on electronic medical record oriented named entity recognition and entity relation extraction. *Acta Automatica Sinica*, *40(8)*,1537-1562.

[4] Leaman, R., Miller, C., & Gonzalez, G. (2009). November. Enabling recognition of diseases in biomedical text with machine learning: corpus and benchmark. *Proceedings of the 2009 Symposium on Languages in Biology and Medicine*.

[5] Zhang, Y., Wang, X., Hou, Z., & Li, J. (2018). Clinical named entity recognition from Chinese electronic health records via machine learning methods. *JMIR Medical Informatics*, *6(4)*.

[6] Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Ferrero, E., Agapow, P. M., Zietz, M., Hoffman, M. M., & Xie, W. (2018). Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, *15(141)*.

[7] Lee, K. J., Hwang, Y. S., & Rim, H. C. (2003). Two-phase biomedical NE recognition based on SVMs. *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*. (pp. 33-40).

[8] Rabiner, L. R., & Biing-Hwang, J. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine 3.1*, 4-16.

[9] McCallum, A., Freitag, D., & Pereira, F. C. (2000), Maximum entropy markov models for information extraction and segmentation.

[10] Lafferty, J, Mccallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the ICML*, *3(2)*, 282-289

[11] Wu, Y, Jiang, M, & Xu, J., *et al*. (2017). Clinical named entity recognition using deep learning models. Proceedings *of the AMIA Annual Symposium American Medical Informatics Association*.

[12] Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *Computer Science*.

[13] Lample, G., Ballesteros, M., Subramanian, S., *et al.* (2016). Neural architectures for named entity recognition.

[14] Ouyang, E., Li, Y., Jin, L., *et al*. (2017). Exploring n-gram character presentation in bidirectional RNN-CRF for Chinese clinical named entity recognition.

[15] Li, Z., Zhang, Q., Liu, Y., *et al*. (1976). Recurrent neural networks with specialized word embedding for chinese clinical named entity recognition.

[16] Hu, J., Shi, X., Liu, Z., *et al*. (1976). HITSZ_CNER a hybrid system for entity recognition from Chinese clinical text.

[17] Xu, K., Zhou, Z., Hao, T., *et al.* (2017). A bidirectional LSTM and conditional random fields approach to medical named entity recognition.

[18] Habibi, M., Weber, L., Neves, M., *et al*. (2017). Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, *33(14)*, i37-i48.

[19] Wang, Q., Zhou, Y., Ruan, T., *et al*. (2019). Incorporating dictionaries into deep neural networks for the chinese clinical named entity recognition. *Journal of Biomedical Informatics*.

[20] Ma, X., & Hovy, E. (2016). End-to-end sequence labeling via bi-directional Lstmcnns-crf.

[21] Dang, T. H., Le, H. Q., Nguyen, T. M., & Vu, S. T. (2018). D3NER: Biomedical named entity recognition using CRF-biLSTM improved with fine-tuned embeddings of various linguistic information. *Bioinformatics*, *34(20)*, 3539-3546.

[22] Yu, F., & Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions.

[23] Strubell, E, Verga, P, Belanger D., *et al*. (2017). Fast and accurate entity recognition with iterated dilated convolutions.

[24] Yang, J., Yu, Q., Guan, Y., & Jiang, Z. (2014). An overview of research on electronic medical record oriented named entity recognition and entity relation extraction. *Acta Automatica Sinica*, *40(8)*, 1537-1562.

[25] Ramshaw, L. A., & Marcus, M. P. (1999). Text chunking using transformation-based learning. *Natural Language Processing Using Very Large Corpora*, 157-176.

[26] Ratinov, L., & Roth, D. (2009). Design challenges and misconceptions in named entity recognition. *Proceedings of the Thirteenth Conference on Computational Natural Language Learning* (pp. 147-155).

Yifan Wang is with Shanghai Institute for Advanced Communication and Data Science, School of Communication and Information Engineering, Shanghai University, Shanghai, China. Yifan Wang's research interests are Chinese semantic analysis, entity relationship extraction and knowledge mapping applications.

Guowei Teng is with Shanghai Institute for Advanced Communication and Data Science, School of Communication and Information Engineering, Shanghai University, Shanghai, 200444, China and Key Laboratory for Advanced Display and System Application, Ministry of Education, Shanghai University, Shanghai, 200072, China.

Xuehai Ding is with School of Computer Engineering and Science, Shanghai University, Shanghai, China. Xuehai Ding is a lecturer of the High Performance Computering Center of Shanghai University. His research interests include artificial intelligence, image processing and analysis, data visualization analysis and mining, and high-performance computing. It mainly carries out the work of visual analysis and mining and processing of big data, especially the establishment and processing of image data in the database, and has carried out certain research work in medical image neural network identification, and has participated in many research projects.t certain research work in medical image neural network identification, and has participated in many research projects.

Guoqing Zhang is with BIO-Med BIG Data Center, Key Laboratory of Computational Biology, CASMPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences Zhang Guoqing: vice director and principal investigator of Bio-Med Big Data Center of Shanghai Institutes for Biological Sciences of Chinese Academy of Sciences. Zhang's main research interest is bioinformatics database and knowledge base, focusing on the integration and mining of omics data, literature data and clinical data in the fields, such as precision medicine, large population cohort, the development of personalized drug, microbiome and synthetic biology etc.

Yunchao Ling is with Bio-Med Big Data Center, Key Laboratory of Computational Biology, CASMPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences Yunchao Ling: Supervisor of database R&D department, Bio-Med Big Data Center of Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences. He received his degree in

bioinformatics from Beijing Institute of Genomics, Chinese Academy of Sciences. His research interests are bioinformatics, clinical genomics and knowledge base technologies. He is conducting research and development activities like multiple omics database construction, AI-based biomedical knowledge mining and integration, biological data analysis and visualization in the areas of human phenomics, precision medicine and clinical healthcare.

Guozhong Wang received the M.E. degree in computer application from Nanjing University of Science and Technology, Nanjing, China, in 1998. In 2006, he received the Ph.D. degree in system integration from East China Normal University, Shanghai, China. He is a professor at the School of Electronic and Electrical Engineering at Shanghai University of Engineering Science. His research interests include image processing, video coding, and video communication.