Mathematical Biosciences
and Engineering

http://www.aimspress.com/journal/MBE

MBE, 17(4): 3498–3511.
DOI: 10.3934/mbe.2020197
Received: 29 December 2019
Accepted: 23 April 2020
Published: 09 May 2020

Research article

A self-attention based neural architecture for Chinese medical named entity recognition

Qian Wan[1], Jie Liu[1,2,*], Luona Wei[3] and Bin Ji[3]

[1] Science and Technology on Parallel and Distributed Processing Laboratory, National University of Defense Technology, Changsha 410073, China
[2] Laboratory of Software Engineering for Complex Systems, National University of Defense Technology, Changsha 410073, China
[3] College of Computer, National University of Defense Technology, Changsha 410073, China

* **Correspondence:** Email: liujie@nudt.edu.cn.

Abstract: The combination of medical field and big data has led to an explosive growth in the volume of electronic medical records (EMRs), in which the information contained has guiding significance for diagnosis. And how to extract these information from EMRs has become a hot research topic. In this paper, we propose an ELMo-ET-CRF model based approach to extract medical named entity from Chinese electronic medical records (CEMRs). Firstly, a domain-specific ELMo model is fine-tuned on a common ELMo model with 4679 raw CEMRs. Then we use the encoder from Transformer (ET) as our model's encoder to alleviate the long context dependency problem, and the CRF is utilized as the decoder. At last, we compare the BiLSTM-CRF and ET-CRF model with word2vec and ELMo embeddings to CEMRs respectively to validate the effectiveness of ELMo-ET-CRF model. With the same training data and test data, the ELMo-ET-CRF outperforms all the other mentioned model architectures in this paper with 85.59% F1-score, which indicates the effectiveness of the proposed model architecture, and the performance is also competitive on the CCKS2019 leaderboard.

Keywords: self-attention; ELMo; named entity recognition; Chinese electronic medical records; natural language processing

1. Introduction

With the combination of the medical field and big data, more and more consultation data and disease information are recorded in the form of electronic medical records (EMRs), and gradually become an important basis for assisting doctors in therapeutic diagnosis. EMRs record a large number of diagnostic information of patients: hospital records, course records, doctor's orders, case data and so on, including key entity information such as disease, surgery, drugs, etc. This information is a decisive factor for doctors to make treatment plans for patients [1]. It is of great significance to study how to extract key entity information from massive EMRs efficiently and accurately through intelligent methods.

Named entity recognition (NER) is a vital part of natural language processing (NLP) that meets the aforementioned requirements [2]. Its purpose is to recognize various named entities, e.g. names, place, organizations, etc., from raw text. Extracted entities can be taken as information for people, and can also pave the way for other NLP tasks, such as relationship extraction and knowledge graph construction. Recently, with the rise of deep learning technology, deep neural networks are utilized to achieve medical NER and have attracted much research attention.

So far, NER still faces huge problems in the field of Chinese electronic medical records (CEMRs). The main reasons are as follows: first of all, an entity may have multiple names due to the undefined text labeling standards [3]; secondly, the meaning of the same word or character may be completely different in different contexts, which causes confusions towards Chinese semantics; last but not least, Chinese has no natural vocabulary boundaries (spaces) as English does, so Chinese have no strict and correct vocabulary boundaries. In previous NER researches, the BiLSTM-CRF, which is the abbreviation of bi-directional Long-Short Term Memory (LSTM) joining with a conditional random field (CRF) layer, shows advanced performance and has become a prevalent architecture for various NER tasks [4,5]. This architecture outperforms traditional methods in that it eliminates the inefficient and complex method of manually designing feature templates and utilizes recurrent neural network (RNN) to automatically capture text features. However, lengths of CEMR texts are generally longer than traditional text, for CEMR texts contain at least several hundred Chinese tokens. Although it can capture long-term contextual dependencies [6], Long-Short Term Memory (LSTM) present inferior performance when text lengths exceed a certain step size [7]. In addition, at the beginning of the NLP task, each token in the text is represented by a low-dimensional dense vector [8]. Due to the highly domain-specialized medical field, universal pre-trained models are hardly adopted in practical tasks. The lack of Chinese medical corpus even makes it more difficult to pre-training medical domain-specific language models. In most cases, vector representations are generally randomly initialized, leading to a context-independent representation for each token [9], as a result, it can't tackle polysemy problems, and the limitations are very obvious.

Our work focus on Chinese medical NER in CEMRs, which has been a subtask of numerous influential academic conferences in NLP domain, e.g. China Conference on Knowledge Graph and Semantic Computing (CCKS), China Health Information Processing Conference (CHIP) etc.. These tasks not only accelerate Chinese medical NER research, but also provide several precious corpora for Chinese medical NER. In this paper, we first collect 4679 CEMRs, which are utilized to fine-tune the Chinese Embeddings from Language Models (ELMo) [9] trained with common field corpora, and then get a pre-trained model that can be used to dynamically generate context-dependent character embeddings for Chinese characters. Secondly, encoder from Transformer (ET) [10] is utilized as the model encoder instead of the traditional bi-directional Long-Short Term Memory (BiLSTM) in order to provide the proposed model with the ability to capture the long-term dependence of ultra-long CEMR texts efficiently. In ET, the distance between token and token is one, so there is no problem that the dependence is lost due to the lengthy distance between tokens. Our contributions are summarized as follows.

1) We fine-tuned a Chinese medical domain-specific ELMo model, which provides an authentic pre-trained language model for further research. A Chinese medical corpus with 4679 real-world CEMRs is constructed, containing about 1.8 million Chinese characters. Then a medical domain-specific model ELMo is fine-tuned by the efficiently application of Chinese medical corpus and the public available Chinese ELMo model.

2) We realize Chinese medical NER in CEMRs with ET-CRF model, which can tackle the long context dependencies better than the BiLSTM model with self-attention mechanism, and to the best of our knowledge this is the first time to apply ET-CRF model to Chinese medical NER.

3) Owning to the contributions above, the proposed ELMo-ET-CRF model achieves the best performance among all model architectures mentioned in this paper on the CCKS 2019 datasets, and the final F1-score is competitive to the current state-of-the-art performance.

2. Related word

NER has become one of the important tasks of information retrieval, data mining and NLP owing to of its extraordinary significance [11], and various solutions have been proposed in the existing literature.

2.1. Rule based approach

Matching entities through handwritten rules is the main method to deal with NER tasks in the early stage [12]. However, the construction of rules requires a certain level of expertise, and even a domain expert cannot enumerate rules that can model all entities. In addition, rules cannot be migrated because they rely on datasets. Thus, a same set of rules may not work under different datasets. This kind of handcrafted approach always leads to a relatively high system engineering cost.

2.2. Statistical machine learning based approach

The statistical machine learning method treats NER as a sequence labeling problem by inputting a set of sequences and outputing a set of predicted optimal tags sequences. Traditional methods include hidden Markov models [13,14], maximum entropy Markov models [15], conditional random fields [16,17], and support vector machines [18]. The most common implementation is feature template with CRF and different feature templates can be combined to form a new feature template. However, this statistical machine learning based method relies heavily on hand-crafted features, which cost a lot of overhead when find the most appropriate features.

2.3. Neural network based approach

In recent years, with the increase of computing power, training of deep neural network has become simple and feasible. The propose of word embeddings (e.g. word2vec, glove) make the useage of deep neural networks to deal with NLP tasks become a research focus [8,19]. Different from the traditional statistical machine learning training method, the training process of the neural network is regarded as an end-to-end process, which can automatically learn data features and avoid the extra overhead such as feature engineering.

More recently, the model structure which is using bidirectional LSTM to encode data with CRF as the decoder has got the most advanced results in the medical NER task [4,5], and the effect is significantly better than the traditional statistical model. The theory of model architecture was first proposed by Collobert et al. [20], Huang [21] and Lample [22] used LSTM-CRF for the first time to deal with sentence-level annotation problems. Ma et al. [23] used LSTM-CRF structure for the first time in English NER task, and achieved promising results, and then Dong et al. [24] first used LSTM-CRF to handle the Chinese NER task. LSTM has a cell structure and gate mechanism that allows the model to effectively capture long-term dependencies and has certain forgetting capabilities [6], and the advantage of splicing CRF after encoder layer is that it can use information that has already occurred during the sequence generation process to ensure that the output value is a set of optimal solution sequences. In addition, Zhang et al. [28] investigate a lattice-structured LSTM model for Chinese NER, which encodes a sequence of input characters as well as all potential words that match a lexicon. Zhang et al. [29] propose a convolutional attention layer to extract the implicit local context features from character sequence. Liu et al. [30] propose a Global Context enhanced Deep Transition architecture for sequence labeling. Qiu et al. [3] proposed the RD-CNN-CRF model, which effectively reduced the time required for training without losing the accuracy of the model. In terms of the inconsistency of the label, Ji et al. [1] proposed a hybrid model based on the attention mechanism, the attention mechanism effectively alleviates the problem of model accuracy decline caused by label inconsistency.

3. Methods

In our method, raw CEMRs are used as the input of ELMo to get the vector representation of sentences, sentence features are then extracted by ET before decoded through CRF to generate the annotation sequence. Figure 1 shows the structure of the model, which is detailed in this section.

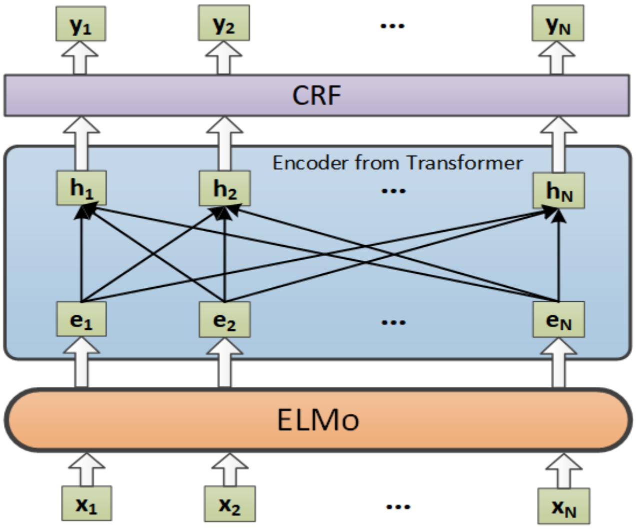


Figure 1. Architecture of ET-CRF model.

3.1. Embeddings from Language Models (ELMo)

In early NLP tasks, the input of RNN is a set of word embeddings generated by Word2Vec [8], Glove [19] etc. However, these embeddings are context-independent, i.e., each word corresponds to a unique static vector without changing with the context, so these methods are limited in the case of polysemy. The emergence of ELMo effectively solves this problem by applying stacked BiLSTM to model the entire sentence from two directions and mapping the sentences into a sequence of vectors. Since LSTM can capture context dependencies, it ensures that the output embedding sequence has a front-to-back correlation.

Given S sequence of N tokens, $\{w_1, w_2, \ldots, w_N\}$, these tokens first go through a token layer, which transfers the dimension of original character embedding to the input dimension of BiLSTM according to a weight matrix, then the output of token layer will be sent to a stacked BiLSTM to build a language model, as shown in Figure 2. Suppose the sentence length is N, and L represents the number of layers of BiLSTM, at each position t, each layer l of LSTM output a context-dependent hidden vector $h_{t,l}^{LM}\left(\left[\overrightarrow{h_{t,l}^{LM}}, \overleftarrow{h_{t,l}^{LM}}\right]\right)$ where $t = 1, 2, \ldots, N$ and $l = 1, 2, \ldots, L$. The hidden vector $h_{t,l}^{LM}$ of the last layer of LSTM output is used by the softmax layer to predict the token at the next moment.

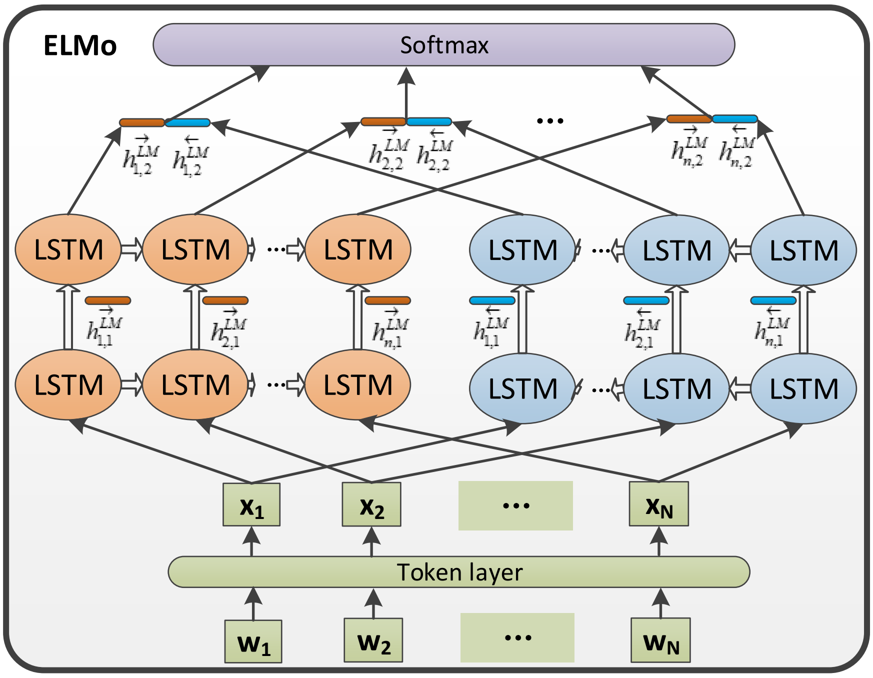

ELMo
Softmax
$\overrightarrow{h}_{1,2}^{LM}$ $\overleftarrow{h}_{1,2}^{LM}$
$\overrightarrow{h}_{2,2}^{LM}$ $\overleftarrow{h}_{2,2}^{LM}$
…
$\overrightarrow{h}_{n,2}^{LM}$ $\overleftarrow{h}_{n,2}^{LM}$
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
$\overrightarrow{h}_{1,1}^{LM}$
$\overrightarrow{h}_{2,1}^{LM}$
$\overrightarrow{h}_{n,1}^{LM}$
$\overleftarrow{h}_{1,1}^{LM}$
$\overleftarrow{h}_{2,1}^{LM}$
$\overleftarrow{h}_{n,1}^{LM}$
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
x_1
x_2
…
x_N
Token layer
w_1
w_2
…
w_N

Figure 2. Architecture of ELMo model.

When the training process finishes, each sentence will learn $2L + 1$ representations R_t, where L is the number of layers in the model:

$$R_t = \left\{x_t^{LM}, \overrightarrow{h_{t,l}^{LM}}, \overleftarrow{h_{t,l}^{LM}} | l = 1, \ldots, L\right\}$$
$$= \left\{h_{t,l}^{LM} | l = 0, \ldots, L\right\} \quad (1)$$

where $h_{t,0}^{LM}$ is the output of token layer and $h_{t,l}^{LM} = \left[\overrightarrow{h_{t,l}^{LM}}, \overleftarrow{h_{t,l}^{LM}}\right]$ for each BiLSTM layer.

3.2. Encoder from Transformer (ET)

Our model uses ET as the encoder layer. Compared with RNN and CNN, it takes a different approach that relies entirely on the self-attention mechanism to extract context features, which makes it highly parallel in the encoding process.

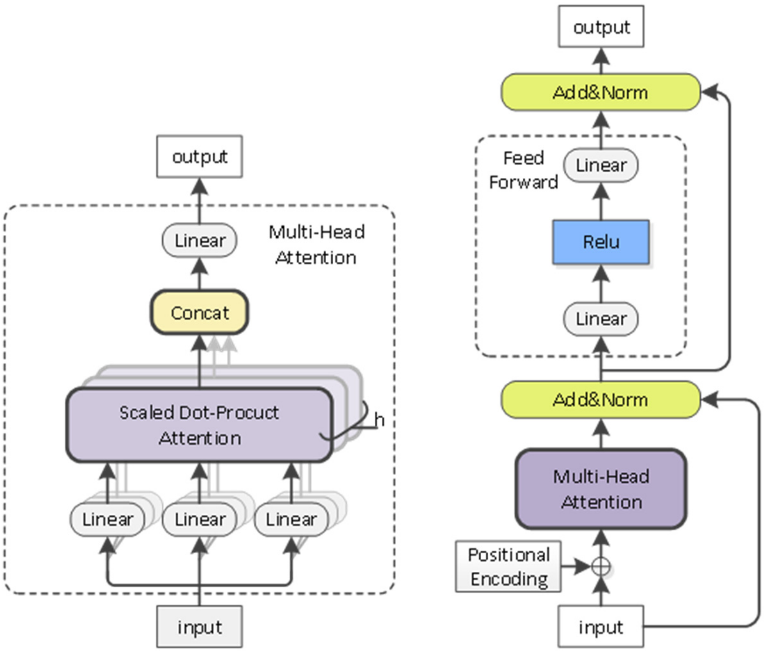


Figure 3. Architecture of ET.

Suppose the input S is a set of sequences $\{w_1, w_2, \dots, w_N\}$, $S \in \mathbb{R}^{N \times d_{model}}$, where N is the length of the sequence and d_{model} is the dimension of the input vector. We use multiple of the scaled dot-product attention components inside the multi-head attention layer to enhance the model's ability to encode sequences internally. We first add position information to the input sequence S, which follows the approach of Vaswani et al. [10], then perform a matrix transformation operation on S to obtain three weight matrices, namely key matrix K, query matrix Q and value matrix V. Finally, the output representation of a single self-attention is obtained by scaled dot-product attention:

$$Q, K, V = SW^Q, SW^K, SW^V \tag{2}$$

$$Attn(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

where $W^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W^K \in \mathbb{R}^{d_{model} \times d_k}$, $W^V \in \mathbb{R}^{d_{model} \times d_k}$, are learnable parameters and $softmax()$ is performed row-wise.

Single head attention may inhibit the information from different representation subspaces at different positions, so our model uses multi-head attention mechanisms [10]:

$$MultiHead(S) = [head_1, head_2, \ldots, head_n]W^O \tag{4}$$

where $head_i = Attn(Q_i,\ K_i,\ V_i)$ and W^O is learnable parameter.

Multi-head attention will be followed by a feedforward neural network, and the output of each sublayer in the encoder will have a residual link and layer normalization as shown in Figure 3.

Assuming that the output of the multi-attention is the $\bar{S}$, the final output of the encoder will be calculated as follows:

$$\bar{S} = layernorm\big(S + MultiHead(S)\big) \tag{5}$$

$$\tilde{S} = \ layernorm\big(\bar{S} + FFN(\bar{S})\big) \tag{6}$$

where $FFN(x) = \max(0, xW_1 + b_1)\,W_2 + b_2$ and $layernorm()$ represents the layer normalization [25].

3.3. Conditional random field (CRF)

In the NER task of CEMRs, the output tag sequence is strictly ordered. The CRF layer maintains a state transition matrix which stores the transition probability from the previous state to the current state, ensureing the tag prediction process has inner dependency.

Given the input sequence is $X = \{x_1, x_2, \ldots, x_N\}$, the score for defining the output sequence $Y = \{y_1, y_2, \ldots, y_N\}$ is expressed as follows :

$$s(X, Y) = \sum_{i=0}^{N} A_{y_i, y_{i+1}} + \sum_{i=1}^{N} P_{i, y_i} \tag{7}$$

where A is a matrix of transition scores, A_{ij} represents the score of a transition from the tag i to tag j. P is a matrix of tag scores which is obtained from the output of ET through a fully connected network, $P \in \mathbb{R}^{N \times k}$, where N is the length of the sequence, k is the number of labels, and p_{ij} represents the score of the j^{th} label of the i^{th} position in the sequence. y_0 and y_{N+1} are represented by <*bos*> and <*eos*>.

The model uses softmax to calculate the probability of all the tag sequences that may be generated by the input sequence *X*, and defines the logarithmic probability of maximizing the correct annotated sequence as the goal of the model optimization [22]:

$$p(Y|X) = \frac{e^{s(X,Y)}}{\sum_{Y\prime \in Y_{ALL}} e^{S(X,Y')}} \tag{8}$$

$$log(p(Y|X)) = s(X, Y) - log(\sum\nolimits_{Y\prime \in Y_{ALL}} e^{S(X,Y')}) \tag{9}$$

where Y_{ALL} represents all possible tag sequences for an input sentence X.

4. Experiments

4.1. Fine-tuning ELMo

The Chinese ELMo model used in this paper comes from the Research Center for Social Computing and Information Retrieval Harbin Institute of Technology. It was trained on Xinhua proportion of Chinese gigawords-v5, and takes roughly 3 days on an NVIDIA P100 GPU [26,27]. Since the datasets used in this paper belong to medical field, the corpus distribution has a strong specialized background, and there are also some uncommon characters in the text. Therefore, the pre-trained ELMo which is used in the general field cannot be directly applied to this task. We fine-tuned the ELMo with the medical corpus.

We collected 4679 CEMRs, which contains about 1.8 million characters, from CCKS2018, CCKS2019 and CHIP2018, and routinely preprocessed the text. In particular, it should be noted that in this paper the proposed neural network model is based on character embeddings. In the fine-tuning process, the hyperparameter setting is shown in the Table 1.

Table 1. Hyperparameter setting.

Hyperparameter name	Hyperparameter setting
Character Embedding	300
LSTM Layer	2
LSTM Cell Size	4096
LSTM Hidden Size	1024
Batch Size	1
Optimizer	Adam
Learning rate	0.001
Annealing rate	$0.9^{t/4679}$
Gradients clip	5
Dropout	0.1
Max epoch	100

After the model training is completed, we freeze the parameters of ELMo and use the weighted summation of each layer of ELMo as the output vector of ELMo, finally take the concatenation of the ELMo output and the original character embeddings as downstream model input:

$$V_t^{EMLo} = E(R_t; \Theta) = \lambda \sum_{l=0}^{L} \theta_l h_{t,l}^{LM} \tag{10}$$

$$c_t^{embed} = [c_t, V_t^{ELMo}] \tag{11}$$

where θ are softmax-normalized weights and the λ is a scalar which can scale the vector according to a certain ratio. These two parameters are learnable and is updated with downstream model training. c_t represents original character embedding which generated by Word2Vec, c_t^{embed} represents the final vector representation of the character at the *t* position in a sequence and it will be the input of ET.

4.2. Task and dataset

This paper focus on Chinese medical NER, which aims to detect entity boundary and categorize entity into pre-defined categories. The dataset used in this paper comes from CCKS2019, which is jointly provided by Yiducloud (Beijing) Technology Co., Ltd.

The training data contains 1000 manually annotated CMERs, while the test data contains 379 manually annotated CMERs. Each EMR contains two parts: raw CEMR and annotation information. The annotation information consists of several triples, which are formed of entity start index, entity end index and entity category. Through entity start and end indices, we can extract the entity from CMERs. All entities are categorized into six categories, disease and diagnosis, imaging examination, laboratory test, surgery, medicine and anatomy. Entities contained in this dataset is shown in Table 2.

Table 2. Entity category of dataset.

Category	Disease and diagnosis	Imaging examination	Laboratory test	Surgery	Medicine	Anatomy	Totally
Training data	4193	966	1194	1027	1814	8231	17425
test data	1310	344	586	162	483	2938	5823

4.3. Evaluation criteria

We use the standard evaluation criteria to validate the effectiveness of the model, namely precision, recall and micro F1-score. which can be calculated as follows:

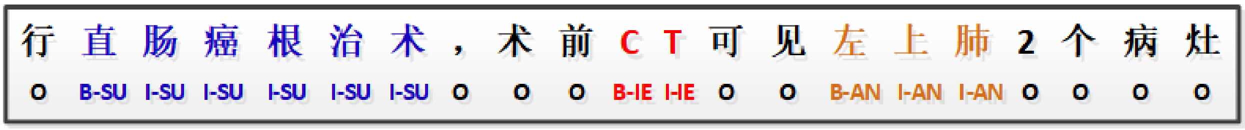
$$Precision = \frac{TP}{(TP + FP)} \tag{12}$$

$$Recal = \frac{TP}{(TP + FN)} \tag{13}$$

$$F1 = \frac{2 \times Precision \times Recall}{(Precision + Recall)} \tag{14}$$

where the TP, TN, FP and FN are true positive, true negative, false positive and false negative respectively.

In this paper, CMERs are encoded with BIO format (Begin, Inside and Outside). The token will label as B-label if the token is the beginning of a named entity, I-label means the token is inside a named entity, other case will label as O-label. It can be seen from Figure 4 that tag ‘O’ are negative samples, and other tags are positive samples.



行	直	肠	癌	根	治	术	，	术	前	C	T	可	见	左	上	肺	2	个	病	灶
O	B-SU	I-SU	I-SU	I-SU	I-SU	I-SU	O	O	O	B-IE	I-IE	O	O	B-AN	I-AN	I-AN	O	O	O	O

Figure 4. BIO tagging schema.

5. Results and discussions

We first use the training data to train the BiLSTM-CRF. In the experiment, the batch size was 10, Adam was selected as the model's optimizer, and the learning rate was set to 0.001. The initial character embedding is generated by word2vec method with a dimension of 300. The BiLSTM layer number is 1, and the hidden layer vector dimension is 300.

Table 3. F1-score with different number of ET layers and heads.

Layers	2	2	4	4	6	6
Heads	4	8	4	8	4	8
F1	83.95	84.05	83.68	84.01	83.55	83.98

For ET-CRF, since the layer numbers and self-attention heads have a critical impact on the model's performance, we trained six sets of models with different numbers of layers or heads, and compared the effects of the models under different parameters on the test set. The results are shown in Table 3. The hyperparameters for the best model are 2 layers and 8 heads. In addition, the model's input is 512 dimensions character embedding, which is also generated by the word2vec method with the same experimental configuration as BiLSTM-CRF.

Table 4. Results of the two models on test dataset.

Entity name	WV-BiLSTM-CRF			WV-ET-CRF		
	Strict index (%)			Strict index (%)		
	P	R	F1	P	R	F1
Disease and diagnosis	74.47	78.07	76.23	77.86	81.04	79.42
Imaging examination	82.22	89.31	85.62	77.78	96.55	86.15
Laboratory test	82.35	84.13	83.23	82.56	85.34	83.92
Surgery	79.04	81.27	80.14	80.00	87.63	83.64
Anatomy	81.34	84.02	82.55	82.83	85.78	84.28
Medicine	89.24	90.28	89.76	91.80	93.80	92.79
Total	80.39	83.39	81.86	82.08	86.12	84.05

The results of WV-BiLSTM-CRF and WV-ET-CRF are shown in Table 4 (WV represents word2vec). It is obvious that the F1-score of the medicine and the imaging examination part is ideal for the BiLSTM-CRF, but the recognition ability of disease and diagnosis is the short board of the model, F1-score only 76.23%. It is speculated that the entity length of the medicine and the anatomy part is generally short and there is no multiple standard, but the tag of the entity such as disease and diagnosis is subjective, context-dependent, and generally has a long length, so judging the boundary is quite difficultly for BiLSTM. The total F1-score of the ET-CRF is 84.05%, which is higher than that of BiLSTM-CRF, 2.19%, revealing that about 382 medical entities are rectified or extracted. And for entities such as disease and diagnosis, ET-CRF performs remarkable. This is an intuitive result, we speculated that due to the ET encoding process which offers effectively method of shortening the direct distance between characters, the context-dependent relationship become well captured and preserved, ensuring the good recognition rate of ET-CRF for longer entities.

Table 5. Test results of the two models with ELMo.

Entity name	ELMo-LSTM-CRF			ELMo-ET-CRF		
	Strict index (%)			Strict index (%)		
	P	R	F1	P	R	F1
Disease and diagnosis	76.93	80.07	78.47	79.57	82.53	81.02
Imaging examination	84.38	93.10	88.52	82.35	96.55	88.89
Laboratory test	84.94	86.78	85.85	83.72	86.54	85.11
Surgery	81.79	84.10	82.93	80.65	88.34	84.32
Medicine	84.74	86.36	85.54	84.75	87.92	86.31
Anatomy	91.06	92.13	91.59	90.32	93.80	92.03
Total	83.32	85.72	84.50	83.65	87.61	85.59

In order to verify the actual effect of ELMo on the NER task in CEMRs, we add ELMo components to the above two models, BiLSTM-CRF and ET-CRF. As a result, the input of the model becomes a dynamic context-dependent character embeddings with context information. The test results are shown in Table 5. For the ELMo-BiLSTM-CRF, the total F1-score is 2.64% higher than BiLSTM-CRF, which means that about 460 medical entities are rectified or extracted. The addition of ELMo has also improved the recognition ability of disease and diagnosis, which can be seen from the table that 2.24% has been improved. For the ELMo-ET-CRF, the total F1-score is 1.54% higher than that of ET-CRF, which shows that 268 medical entities are rectified or extracted. In summary, the addition of pre-trained ELMo enriches the information contained in the character embedding, which effectively improves the accuracy of the model.

Table 6. Results compared with CCKS2019.

Model	Strict index (%)		
	P	R	F1
WV-LSTM-CRF	80.39	83.39	81.86
WV-ET-CRF	82.08	86.12	84.05
ELMo-LSTM-CRF	83.32	85.72	84.50
ELMo-ET-CRF	83.65	87.61	85.59
CCKS2019-No.1	—	—	85.62
CCKS2019-No.2	—	—	85.59
CCKS2019-No.3	—	—	85.16

Table 6 illustrates that ELMo outperforms word2vec due to the obtained dynamic context-dependent model input, and the performance of ET-CRF in Chinese medical named entity recognition is significantly better than that of BiLSTM-CRF. Due to the excellent long context dependency capture capability of the self-attention mechanism, the ET-CRF's ability to recognize long entity is significantly better than the BiLSTM-CRF. In addition, we also found that the convergence speed of ET-CRF is significantly faster than BiLSTM-CRF. It can also be seen in the table that the final F1-score of the best model ELMo-ET-CRF in the paper is 85.59%, which is still competitive compared to the top three in the CCKS2019 competition (https://www.biendata.com/competition/ccks_2019_1/final-leaderboard/).

6. Conclusions

In this paper, we firstly fine-tune a medical domain-specific ELMo model through a small medical corpus which is contained with 4679 CEMRs. Then we apply the ET-CRF model to Chinese medical NER on CEMRs. Finally, the proposed ELMo-ET-CRF model use dynamic context-dependent ELMo character embeddings to incorporate more lexical, syntantic and semantic information, and alleviates long context dependency problem. Under the strict evaluation index, the F1-score of ELMo-ET-CRF on the test set is 85.59%, which is competitive to the state-of-the-art on this dataset, and indicates the effectiveness of the proposed model architecture.

Acknowledgments

We thank the anonymous reviewers for their constructive comments. This work was supported by grants from the National Key Research and Development Program of China (2017YFB0202104). Publication costs are funded by a grant of the National Key Research and Development Program of China (2017YFB0202104).

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. B. Ji, R. Liu, S. Li, J. Yu, Q. Wu, Y. Tan, et al., A hybrid approach for named entity recognition in Chinese electronic medical record, *BMC Med. Inform. Decis. Mak.*, **19** (2019), 64.
2. C. Zong, *Statistical natural language process*, Tsinghua University Press, 2013.
3. J. Qiu, Y. Zhou, Q. Wang, T. Ruan, J. Gao, Chinese clinical named entity recognition using residual dilated convolutional neural network with conditional random field, *IEEE Trans. Nanobiosci.*, **18** (2019), 306–315.
4. L. Li, L. Jin, Z. Jiang, D. Song, D. Huang, *Biomedical named entity recognition based on extended recurrent neural networks*, 2015 IEEE International Conference on bioinformatics and biomedicine, 2015. Available from: https://ieeexplore.ieee.org/abstract/document/7359761/.
5. R. Leaman, C. H. Wei, Z. Lu, tmChem: A high performance approach for chemical named entity recognition and normalization, *J. Cheminf.*, **7** (2015), S3.
6. S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Comput.*, **9** (1997), 1735–1780.
7. S. Lai, L. Xu, K. Liu, J. Zhao, *Recurrent convolutional neural networks for text classification*, Twenty-ninth AAAI conference on artificial intelligence, 2015. Available from: https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/viewPaper/9745.
8. T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, *arXiv preprint arXiv*, **2013** (2013), 1301.3781.
9. M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, et al., Deep contextualized word representations, *arXiv preprint arXiv*, **2018** (2018), 1802.05365.
10. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., *Attention is all you need*, Advances in neural information processing systems, 2017. Available from: http://papers.nips.cc/paper/7181-attention-is-all-you-need.
11. C. Lyu, B. Chen, Y. Ren, D. Ji, Long short-term memory RNN for biomedical named entity recognition, *BMC Med. Inform. Decis. Mak.*, **18** (2017), 462.
12. G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, et al., Mayo clinical text analysis and knowledge extraction system (cTAKES): Architecture, component evaluation and applications, *J. Am. Med. Inf. Assoc.*, **17** (2010), 507–513.
13. G. Zhou, J. Su, *Named entity recognition using an HMM-based chunk tagger*, proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2002. Available from: https://dl.acm.org/doi/10.3115/1073083.1073163.
14. M. Song, H. Yu, W. S. Han, Developing a hybrid dictionary-based bio-entity recognition technique, *BMC Med. Inform. Decis. Mak.*, **15** (2015), S9.
15. A. McCallum, D. Freitag, F. C. Pereira, *Maximum Entropy Markov Models for Information Extraction and Segmentation*, LCML, 2000. Available from: http://cseweb.ucsd.edu/~elkan/254spring02/gidofalvi.pdf.
16. A. McCallum, W. Li, *Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons*, Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4, 2003. Available from: https://dl.acm.org/doi/10.3115/1119176.1119206.
17. M. Skeppstedt, M. Kvist, G. H. Nilsson, H. Dalianis, Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study,

J. Biomed. Inform., **49** (2014), 148–158.

18. Z. Ju, J. Wang, F. Zhu, *Named entity recognition from biomedical text using SVM*, 2011 5th international conference on bioinformatics and biomedical engineering, 2011. Available from: https://ieeexplore.ieee.org/abstract/document/5779984/.
19. J. Pennington, R. Socher, C. Manning, *Glove: Global vectors for word representation*, Proceedings of the 2014 conference on empirical methods in natural language processing, 2014. Available from: https://www.aclweb.org/anthology/D14-1162.pdf.
20. R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, Natural language processing (almost) from scratch, *J. Mach. Learn. Res.*, **12** (2011), 2493–2537.
21. Z. Huang, W. Xu, K. Yu, Bidirectional LSTM-CRF models for sequence tagging, *arXiv preprint arXiv*, **2015** (2015), 1508.01991.
22. G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, *arXiv preprint arXiv*, **2016** (2016), 1603.01360.
23. X. Ma, E. Hovy, End-to-end sequence labeling via bi-directional lstm-cnns-crf, *arXiv preprint arXiv*, **2016** (2016), 1603.01354.
24. C. Dong, J. Zhang, C. Zong, M. Hattori, H. Di, Character-based LSTM-CRF with radical-level features for Chinese named entity recognition, in *Natural Language Understanding and Intelligent Applications*, Springer. (2016), 239–250.
25. J. L. Ba, J. R. Kiros, G. E. Hinton, Layer normalization, *arXiv preprint arXiv*, **2016** (2016), 1607.06450.
26. W. Che, Y. Liu, Y. Wang, B. Zheng, T. Liu, Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation, *arXiv preprint arXiv*, **2018** (2018), 1807.03121.
27. A. Kutuzov, M. Fares, S. Oepen, E. Velldal, *Word vectors, reuse, and replicability: Towards a community repository of large-text resources*, Proceedings of the 58th Conference on Simulation and Modelling, 2017. Available from: https://www.duo.uio.no/handle/10852/65205.
28. Y. Zhang, J. Yang, Chinese ner using lattice lstm, *arXiv preprint arXiv*, **2018** (2018), 1805.02023.
29. Y. Zhu, G. Wang, *CAN-NER: Convolutional attention network for Chinese named entity recognition*, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019. Available from: https://www.aclweb.org/anthology/N19-1342.pdf.
30. Y. Liu, F. Meng, J. Zhang, J. Xu, Y. Chen, J. Zhou, Gcdt: A global context enhanced deep transition architecture for sequence labeling, *arXiv preprint arXiv*, **2019** (2019), 1906.02437.