Medical Knowledge Attention Enhanced Neural Model for Named Entity Recognition in Chinese EMR

Zhichang Zhang, Yu Zhang, Tong Zhou

College of Computer Science and Engineering, Northwest Normal University, China
zzc@nwnu.edu.cn zhangyu.wuyi@foxmail.com 1083993141@qq.com

Abstract. Named entity recognition (NER) in Chinese electronic medical records (EMRs) has become an important task of clinical natural language processing (NLP). However, limited studies have been performed on the clinical NER study in Chinese EMRs. Furthermore, when end-to-end neural network models have improved clinical NER performance, medical knowledge dictionaries such as various disease association dictionaries, which provide rich information of medical entities and relations among them, are rarely utilized in NER model. In this study, we investigate the problem of NER in Chinese EMRs and propose a clinical neural network NER model enhanced with medical knowledge attention by combining the entity mention information contained in external medical knowledge bases with EMR context together. Experimental results on the manually labeled dataset demonstrated that the proposed method can achieve better performance than the previous methods in most cases.

Keywords: Chinese Electronic Medical Record, Named Entity Recognition, Deep Learning, Knowledge Attention.

1 Introduction

Named entity recognition (NER), which aims to identify boundaries and types of entities in text, has been one of the well-established and extensively investigated tasks in natural language processing (NLP), and is an essential component for a large number of NLP applications such as entity linking (Chabchoub *et al*., 2016), relation extraction (Liu *et al*., 2014), question answering (Cao *et al*., 2011) and knowledge base population(Carlson *et al*., 2010).

The electronic medical record (EMR), sometimes called electronic health record (EHR) or electronic patient record (EPR), is one of the most important types of clinical data and often contain valuable and detailed patient information for many clinical applications. Clinical NER in EMR text is therefore a fundamental task in medical NLP and has been extensively studied (Ye *et al*., 2011). Nevertheless, most previous studies on clinical NER have primarily focused on EMR in English text. With the rapid growth of clinical NLP applicationsin China, NER from Chinese clinical text has also become a hot research topic for biomedical informatics or NLP researchers.

Most previous research on NER have been dominated by applying traditional machine learning (ML) based models, which require a set of informative features that are well engineered and carefully selected. However, feature engineering is very time-consuming and costly, and resulting feature sets are both domain and model-specific.

In the past few years, the advent of deep neural networks with the capability of automatic feature engineering has leveraged the development of NER models, and this kind of models has also been studied on clinical NER (Le *et al*., 2018). While deep neural network approaches for clinical NER have achieved better performance compared to traditional models, many existing domain knowledge bases have rarely been utilized or combined in these deep models. Medical knowledge bases (MKBs), however, contain a large amount of clinical entity name and definition or description of them. These entity names can be applied as reference lexicon for clinical NER, meanwhile their definition and description provide rich context and entity relationship information which is still helpful for entity recognition.

In view of the above, this paper proposes a novel model for clinical NER in Chinese EMRs. The model trains character-level embedding representation of words using Convolutional Neural Network (CNN), and combines them with pretrained character embedding vectors obtained from large-scale background training corpus, then sends the combined vectors to a deep neural network called BILSTM-CRF to train entity recognition model. To enhance the representation and distinguish ability of words and their contexts, we integrate the medical knowledge attention (MKA) learned from entity names and their definition or descriptions in MKBs. The experimental results on the labeled Chinese EMR evaluation corpus show that the model achieved the best performance without any artificial features, and the F-values is 92.03%.

The remainder of this paper is composed as follows. In section 2 we summarize the related work about clinical or medical NER. In section 3 we present our medical knowledge attention enhanced neural network NER model in Chinese EMRs. In section

4 we show the experimental results on the test data and give some analysis. Finally, we summarize our work and outline some ideas for future research.

2 Related Work

NER is typically treated as a sequence labeling problem and many researchers applied ML-based methods to learn named entity tagging decisions from annotated texts. Those techniques utilized for clinical NER are Support Vector Machines (SVM), structural SVM (SSVM), Conditional Random Fields (CRF), Maximum Entropy (ME). (Wang *et al.*, 2009) applied CRF, SVM and ME to recognize symptoms and pathogenesis in ancient Chinese medical records and showed that CRF achieved a better performance. (Wang *et al.*, 2012) conducted a preliminary study on symptom name recognition in clinical notes of traditional Chinese medicine. (Xu *et al*., 2014) proposed a joint model that integrates segmentation and NER simultaneously to improve the performance of both tasks in Chinese discharge summaries. (Lei *et al*., 2014) systematically evaluated the effects of different features and ML algorithms on NER in Chinese clinical text.

In recent years, unsupervised learned word embeddings have been seen tremendous success in numerous NLP tasks, including clinical or medical NER. (Tang *et al*., 2014) used CRF model and supplement artificial features with word embeddings to identify biological entity, achieved good performance on BioCreative II GM and JNLPBA corpus. (Chang *et al*., 2015) also utilized word embeddings in their CRF-based medical NER model and obtained performance improvement on JNLPBA corpora.

Deep neural network (DNN) architecture is also widely used in NER task. (Yao *et al*., 2015) trained a multilayer neural network model for biological entity recognition with word embeddings generated on unlabeled biological texts. (Li *et al*., 2016) used the Bi-directional Long Short Term Memory Network (BLSTM) method to achieve an 88.6% and 72.76% F-value on the Biocreative II GM and JNLPBA corpus respectively. (Dong *et al*., 2017) used the BLSTM model to identify named entiy in Chinese electronic medical records. (Liu *et al*., 2017) used the BLSTM-CRF with features to identify clinical notes, achieve an 89.98% F-value on the 2016 N-GRID.

Although clinical or medical NER in English text has been extensively studied and many kinds of traditional ML-based and DNN-based models have been proposed, there is limited work on clinical NER in Chinese EMRs using word embeddings or deep learning methods. Furthermore, the application of MKBs and their effectiveness on clinical NER should also be carefully studied and analyzed.

3 The Proposed Method

In this paper, we propose a neural network architecture combining BI-LSTM-CNN-CRF network with Medical Knowledge-Attention that will learn the shared semantics between medical record texts and the mentioned entities in the MKBs. The architecture of our proposed model is shown in Figure 1. After querying pretrained character embedding tables, the input sentence *s* will be transformed respectively to the corresponding sequences of pretrained character embeddings and random generated character embedding matrixes for every word. Then a CNN is used to form the character level representation and a bidirectional LSTM is used to encode the sentence representation after concatenating the pretained character embeddings and char-level representation of the sentence. Afterwards, we treat the entity information from MKBs as a query guidance and integrate them with the original sentence representation using a multimodal fusion gate and a filtering gate. At last, a CRF layer is used to decode.

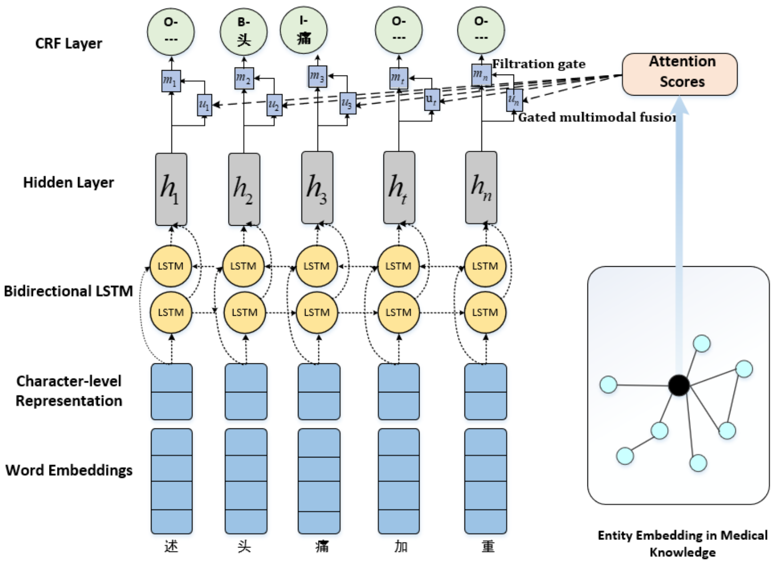

Fig. 1. The framework of NER in Chinese EMRs enhanced with medical knowledge attention

3.1 Feature Extractor

Character-Level Representation with CNN. As described in Figure 2, we firstly train character embeddings from a large unlabeled Chinese EMR corpus, then CNN is used to generate sentence character-level representation from the character embedding matrix sequence to alleviate rare character problems and capture helpful morphological information like special characters in EMRs. Since the length of sentences is not consistent, a placeholder (padding) is added to the left and right side of character embeddings matrix to make the length of every sentence character-level representation vector matrix sequence equal.

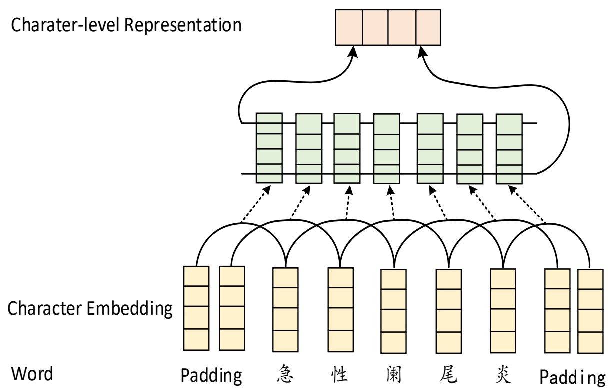

Figure 2. Character-level representation of a sentence by CNN.

Sentence Encoding with Bidirectional LSTM. Bidirectional LSTM encodes the sentence twice from the start to the end and from the end to the start. Thus, at each time state *t*, we can obtain two representations $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ of sentence. Two representations are concatenated to form the final encoding representation of the sentence:

$$h_t = \left[\overrightarrow{h_t},\ \overleftarrow{h_t}\right] \tag{1}$$

Medical Knowledge Attention. Concerning rich entity relation and definition information containing in MKBs, the medical knowledge attention is applied to integrate entity representations learned from external knowledge bases as query vector for encoding. We use a medical dictionary to encode entity information and entity relation information into attention scores as entity embeddins.

$$a_t = f\left(eW_A h_t\right) \tag{2}$$

Where e is the embedding for entity, and W_A is a bi-linear parameter matrix. We simply choose the quadratic function $f(x) = x^2$, which is positive definite and easily differentiate.

Gated Multimodal Fusion. Based on the output of LSTM and attention scoring, we design a gated multimodal fusion (GMF) method to fuse the features from EMR text context and external knowledge dictionary. When predicting the entity tag of a word, the GMF trades off how much new information of the network is considering from the query vector with the EMR text containing the word. The GMF is defined as:

$$h_{a_t} = \tanh(W_{a_t} a_t + b_{a_t}) \tag{3}$$

$$h_{h_t} = \tanh(W_{h_t} h_t + b_{h_t}) \tag{4}$$

$$g_t = \sigma(W_{g_t}(h_{a_t} \oplus h_{h_t})) \tag{5}$$

$$m_t = g_t h_{a_t} + (1 - g_t) h_{h_t} \tag{6}$$

Where W_{a_t}, W_{h_t}, W_{g_t} are parameters, h_{h_t} and h_{a_t} are the new sentence vector and new query vector respectively, after transformation by single layer perceptron. $\oplus$ is the concatenating operation, σ is the logistic sigmoid activation, g_t is the gate applied to the new query vector h_{h_t}, and m_t is the multimodal fused feature from the new medical knowledge feature and the new textual feature.

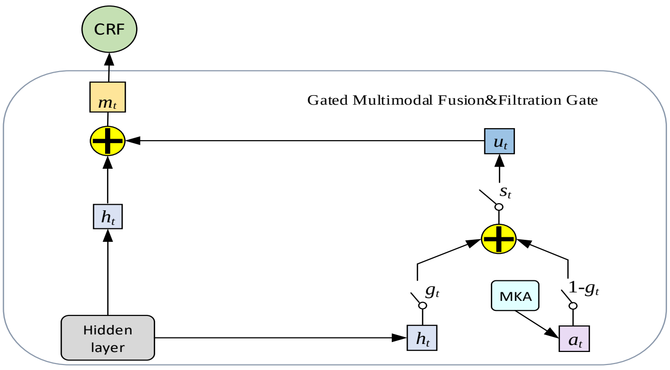

Fig.3. The architecture of gated multimodal fusion and filtering gate.

Filtering Gate. The architecture of gated multimodal fusion and filtering gate are shown in Figure 3. When decoding the combination of the word feature at position t and its corresponding multimodal fusion feature m_t, the impact and necessity of the MKB feature for different POS of word is different. We therefore use a filtering gate to combine different features. The filtering gate s_t is a scalar in the range of [0, 1] and its value depend on how much the multimodal fusion feature is helpful to label the entity tag of the word. s_t and the input feature to the decoder $\hat{m}_t$ are defined as follows:

$$s_t = \sigma(W_{s_t,h_t} h_t \oplus (W_{m_t,s_t} m_t + b_{m_t,s_t})) \tag{7}$$

$$u_t = s_t(\tanh(W_{m_t} m_t + b_{m_t})) \tag{8}$$

$$\hat{m}_t = W_{\hat{m}t}(h_t \oplus u_t) \tag{9}$$

where $W_{m_t,s_t}, W_{s_t,h_t}, W_{m_t}, W_{\hat{m}t}$ are parameters, h_t is the hidden state of bidirectional LSTM at time t, u_t is the reserved multimodal features after the filtering gate filter out noise, and $\oplus$ is the concatenating operation.

3.2 CRF Tagging Layer

Formally, we use $X = \{x_1,\cdots, x_n\}$ to represent a generic input sequence where x_i is the input vector of the ith word, and $y = \{y_1,\cdots, y_n\}$ denotes the set of possible entity tag sequences for X where y_i represent a generic tag for x_i. We use the maximum conditional likelihood estimation for CRF training. The logarithm of likelihood is given by:

$$L(p(y/X)) = \sum_i \log p(y/X) \tag{10}$$

In the decoding phrase, we predict the output sequence that obtains the maximum score given by:

$$y^* = \arg\max_{y' \in Y} p(y/X) \tag{11}$$

4 Experiments

In this section, we evaluate our method on a manually annotated dataset. Following (Nadeau *et al*., 2007), we use Precision, Recall, and F1 to evaluate the performance of the models.

4.1 Data Set

Table 1. Number statistics of different entity type in the evaluation dataset.

Statistics	Train	Test	Total
Symptom (症状)	823	412	1235
Disease (疾病)	1013	506	1519
Laboratory test (检查)	637	315	952
Body parts (身体部位)	761	380	1141
Medicine (药物)	547	273	820
Total	3781	1886	5667

We use our own manually annotated corpus as evaluation dataset, which consists of 800 de-identified EMR texts from different clinical departments of a grade-A hospital of second class in Gansu Province. Five entity types, including symptom, disease, laboratory test, body parts and medicine are labeled. The number of every entity type in dataset is shown in the Table 1.

4.2 Embedding

We use Google's Word2Vec to train character embeddings on our 30 thousand unlabeled Chinese EMR texts which is from a grade-A hospital of second class in Gansu Province. Random generated character embeddings are initialized with uniform samples from $[-\sqrt{\frac{3}{dim}},\sqrt{\frac{3}{dim}}]$, where we set dim = 30.

4.3 Parameter Setting

Table 2 gives the chosen hyper-parameters for all experiments. We tune the hyper-parameters on the development set by random search. We try to share as many hyper-parameters as possible in experiments.

Table 2. Parameter Setting.

Parameter	Value
Character embedding size	50
Word Embedding Size	100
Learning Size	0.014
Decay Rate	0.05
Dropout	0.5
Batch Size	10
LSTM State Size	200
LSTM Initial State	0
CNN Window Size	3
CNN Number of filters	50

4.4 Experimental results

We carry out the experiments to compare the performance of the following models.

CNN+BILSTM+CRF: It was proposed by (Ma and Hovy, 2016) and is a truly end-to-end system. The model was reported to have achieved the best 91.21% F1 score on the CoNLL 2003 test set. This model will be used as the baseline.

CNN+BILSTM+MKA+CRF: In this model, the fusion gate and the filtering gate are not used, and the medical knowledge attention score is simply added to the output value of LSTM as the query vector.

CBMFiC: This model is formed after leaving the fusion gate from the architecture described in Figure 1. It directly concatenates the features from different modalities. At each time step, we use a filtering gate to filter out the noise introduced by the attention score, then concatenate the external multimodal feature of the word with the text feature as CRF input.

CBMFuC: This model is formed after only leaving the filtering gate from the architecture described in Figure 1. After obtaining the fused features, both Medical Knowledge-Attention and text representation, we concatenate the fusion feature (query vector) with the text feature at this time step, then the CRF is applied for learning.

CBMFuFiC: This model is totally corresponding to the architecture in Figure1, and contains all components investigated in this study including.

Table 3. The performance of different models on the evaluation dataset.

Model	Precision (%)	Recall (%)	F1 (%)
CNN+BILSTM+CRF(Baseline)	88.88	88.31	88.59
CNN+BILSTM+MKA+CRF	90.21	88.72	89.46
CBMFiC	90.49	89.65	90.07
CBMFuC	91.27	**92.03**	91.65
CBMFuFiC	**92.19**	91.87	**92.03**

4.5 Effect of Dropout

Table 4 compares the experiment performance when the dropout layer is or not used. All other hyper parameters remain the same as in Table 2. It demonstrates the effectiveness of the dropout in reducing overfitting.

Table 4. Results with and without dropout

	Train	Dev	Test
No	99.63	93.74	90.21
Yes	**99.19**	94.87	**92.03**

4.6 Discussion

The experiment results of different models on our manually annotated datasets are shown in Table 3. Compared with the baseline model, all other models have improved performance and it shows that the updated neural network model is better than the traditional deep neural network model only composed of CNN, BILSTM and CRF on the clinical NER task.

The performance of the CBMC model is better than baseline. This result shows that, the rich information of entities and their relations from MKBs is useful for clinical NER in Chinese EMR. CBMFuC model is slightly better than CBMC model and shows that it is surely helpful for the clinical NER task in Chinese EMRs to fuse the features from EMR text context with external knowledge dictionary and utilize gated multimodal fusion (GMF) is help for our model. The supplement of external information in MKBs sometimes causes noise to the model. We therefore use a filtering gate to weight and combine different features. As shown by our experimental results, CBMFiC model is also slightly better than CBMC model. Therefore, the filtering gate is helpful to improve the performance of our model.

Due to their sublanguage characteristic of Chinese EMRs, the expression of clinical named entity is different from those in general text. Using the entity information in the MKB as the classification query vector can lead the decoder to focus on the entity itself. So we combine text features and MKB features together with a multimodal fusion gate as query vector. We then set up a filtering gate to filter out useless feature vectors. The experiment result shows that our model CBMFuFiC, which integrates CNN, BILSTM, medical knowledge attention, gated multimodal fusion, filtering gate, and CRF, achieves the best F1 score on our labeled evaluation corpus.

5 Conclusion

In this work, we proposed a medical knowledge-attention enhanced neural clinical entity recognition model, which makes use of the external MKBs in the way of attention mechanism. A gated multimodal fusion module is introduced to decide how much MKB features are fused into the query vector at each time step. We further introduced a filtering gate module to adaptively adjust how much multimodal information can be considered at each time step. We built a manually annotated Chinese EMR dataset for evaluation. The experiment results on the evaluation dataset show that our proposed approach improved the clinical NER performance obviously compared to the baseline models.

In the future, we will explore a fine-grained clinical entity recognition model for Chinese EMR and to extract entity relation in Chinese EMRs.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable comments. The research work is supported by the National Natural Science Foundation of China (No. 61762081, No. 61662067) and the Key Research and Development Project of Gansu Province (No. 17YF1GA016).

References

1. Cao Y-G, Liu F, Simpson P, Antieau L, Bennett A. AskHERMES: An online question answering system for complex clinical questions. Journal of Biomedical Informatics, 2011, 44 (2) :277-288.
2. Carlson A, Betteridge J, Wang R C, et al. Coupled semi-supervised learning for information extraction. DBLP, 2010:101-110.
3. Chabchoub M, Gagnon M, Zouaq A. Collective Disambiguation and Semantic Annotation for Entity Linking and Typing. Semantic Web Challenges, Springer International Publishing, 2016.
4. Chang F-X, Guo J, Xu W-R, Chung S-R. Application of Word Embeddings in Biomedical Named Entity Recognition Tasks. Journal of Digital Information Management, 2015,13(5):321-327.
5. Dong X, Chowdhury S, Qian L, et al. Transfer bi-directional LSTM RNN for named entity recognition in Chinese electronic medical records. In the proceedings of International Conference on E-Health Networking, Applications and Services. IEEE, 2017:1-4.
6. Le H-Q, Nguyen T, Vu S, and Dang T-H. D3NER: Biomedical named entity recognition using CRF-biLSTM improved with fine-tuned embeddings of various linguistic information. Bioinformatics, https://doi.org/10.1093/bioinformatics/bty356, 2018.
7. Lei J, Tang B, Lu X, Gao K, Jiang M, Xu H. A Comprehensive Study of Named Entity Recognition in Chinese Clinical Text. Journal of the American Medical Informatics Association, 21(5):808-814, 2014.

8. Li L, Jin L, Jiang Y, et al. Recognizing Biomedical Named Entities Based on the Sentence Vector/Twin Word Embeddings Conditioned Bidirectional LSTM[M]. In the proceedings of the Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Springer International Publishing, 2016:165-176.
9. Liu Y, Liu K, Xu L-H and Zhao J. Exploring Fine-grained Entity Type Constraints for Distantly Supervised Relation Extraction, in Proceedings of COLING 2014, Dublin, Ireland, August, 23-29.
10. Liu Z, Tang B, Wang X, et al. De-identification of Clinical Notes via Recurrent Neural Network and Conditional Random Field. Journal of Biomedical Informatics, 2017, 75S:S34.
11. Ma X, Hovy E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. https://arxiv.org/pdf/1603.01354, 2016.
12. Nadeau D, Sekine S. A survey of named entity recognition and classification. Lingvisticae Investigationes, 2007, 30(1):págs. 3-26.
13. Tang B-Z, Cao H, Wang X-L, Chen Q-C, Xu H. Evaluating Word Representation Features in Biomedical Named Entity Recognition Tasks. Biomed Research International, vol. 2014, Article ID 240403, 6 pages, doi:10.1155/2014/240403. 2014.
14. Wang S, Li S, Chen T. Recognition of Chinese Medicine Named Entity Based on Condition Random Field. J Xiamen University (Natural Science) 2009;48:349–64.
15. Wang Y, Liu Y, Yu Z, et al. A preliminary work on symptom name recognition from free-text clinical records of traditional chinese medicine using conditional random fields and reasonable features. Proceedings of the 2012 Workshop on Biomedical Natural Language Processing, Stroudsburg, PA, USA, 2012:223–30.
16. Xu Y, Wang Y, Liu T, et al. Joint segmentation and named entity recognition using dual decomposition in Chinese discharge summaries. Journal of the American Medical Informatics Association, 2014; 21:e84–92.
17. Yao L, Liu H, Liu Y, et al. Biomedical Named Entity Recognition based on Deep Neutral Network[J]. International Journal of Hybrid Information Technology, 2015, 8.
18. Ye F, Chen Y Y, Zhou G G, et al. Intelligent recognition of named entity in electronic medical records. Chinese Journal of Biomedical Engineering, 2011, 30(2):256-262.