Medical Knowledge Extraction and Analysis from Electronic Medical Records Using Deep Learning

Peilin Li, Zhenming Yuan*, Wenbo Tu, Kai Yu, Dongxin Lu

Engineering Research Center of Mobile Health Management,
Ministry of Education, Hangzhou Normal University,
Hangzhou 311121, China

Key words: medical knowledge extraction; electronic medical record; named entity recognition; medical relation extraction; deep learning; bidirectional long short-term memory; conditional random field

OSID code

Objectives Medical knowledge extraction (MKE) plays a key role in natural language processing (NLP) research in electronic medical records (EMR), which are the important digital carriers for recording medical activities of patients. Named entity recognition (NER) and medical relation extraction (MRE) are two basic tasks of MKE. This study aims to improve the recognition accuracy of these two tasks by exploring deep learning methods.

Methods This study discussed and built two application scenes of bidirectional long short-term memory combined conditional random field (BiLSTM-CRF) model for NER and MRE tasks. In the data preprocessing of both tasks, a GloVe word embedding model was used to vectorize words. In the NER task, a sequence labeling strategy was used to classify each word tag by the joint probability distribution through the CRF layer. In the MRE task, the medical entity relation category was predicted by transforming the classification problem of a single entity into a sequence classification problem and linking the feature combinations between entities also through the CRF layer.

Results Through the validation on the I2B2 2010 public dataset, the BiLSTM-CRF models built in this study got much better results than the baseline methods in the two tasks, where the F1-measure was up to 0.88 in NER task and 0.78 in MRE task. Moreover, the model converged faster and avoided problems such as overfitting.

Conclusion This study proved the good performance of deep learning on medical knowledge extraction. It also verified the feasibility of the BiLSTM-CRF model in different application scenarios, laying the foundation for the subsequent work in the EMR field.

WITH the application of Internet technology in the field of medicine, great amount of electronic medical records (EMR) with rich medical knowledge have emerged. The electronic medical records are generated by medical staff for the purpose of recording the process of developing medical treatment with patients, including presentation of the disease, laboratory and imaging findings, therapeutic regimens and diseases progress, and so on. It implies the potential relation between the disease characteristics, drug usage and outcome of treatment.¹ The analysis and mining of these implicit knowledge can provide constructive help for medical decision-making and also provide data foundation that subsequently support the establishment of medical knowledge graph and visual display.

Electronic medical records are usually full of unstructured text, which can be post-structured by natural language processing to obtain the required medical knowledge.² Named entity recognition (NER) and medical relation extraction are two basic steps of the medical knowledge extraction. The logical relation between these two steps is series connection.

The main purpose of NER is to identify entities with specific meaning from the text. The original NER task mainly adopted a rule-based approach, which complete entity recognition and extraction tasks by formulating entity dictionaries and setting grammar rules. However, rules and grammars require supports of domain experts. The labor costs are considerable, and the flexibility is poor. The set grammar rules do not necessarily adapt to the new types of named entities. Later, some researchers have defined a NER task as the tagging task of the word sequence³ and applied the machine learning method to the NER task. Character sequence tagging specifies Chinese characters in a sentence as a label, such as the "BIESN" annotation method.⁴ The typical machine learning methods to deal with sequence tagging tasks are hidden Markov model (HMM) and conditional random field (CRF). Ye F⁵ used CRF and combined with multiple feature templates for NER tasks of electronic medical records. Li W⁶ combined CRF with rule method, which first completed sequence annotation with CRF, and then optimized the results of CRF with rules based on decision tree and clinical knowledge so as to achieve entity recognition of electronic medical records. However, CRF method is strongly dependent on manually extracted grammatical features. As the result, the setting and selection of

feature templates by researchers will directly affect the accuracy of NER.

Since the concept of relational extraction was initially introduced at the MUC-7 conference in 1998, it has been widely studied and developed in a variety of fields, such as finance, education, medical care and so on.⁷⁻⁹ In the field of medicine, medical relation extraction (MRE) is used to extract medical relations among items of entities from electronic medical record. The extraction of medical relations in the single sentence of electronic medical records was the preliminary step in a study.¹⁰ Due to the lack of classification criteria for relation types, six entity relation types are defined in advance, and manually selected features are used to extract medical relations through support vector machine (SVM) classifier. In the task published in 2010, I2B2 dataset proposed eight different relation type schemes,^{11,12} which established a scientific classification standard for the field of medical entity relation extraction. Considering the similarity of text types between contexts surrounding medical entities, Rink B¹³ used SVM classifier for classification, and the best F1 classification results at that time were obtained in I2B2 2010 dataset. However, the context information related to entities is diverse, which leads to the sparsity of feature vector representation. In order to solve this problem, the entity semantic relation in UMLS is used as the classification feature to improve the performance of medical entity relation recognition task.¹⁴ In addition, Lv X¹⁵ proposed a deep learning method for feature engineering, which uses automatic encoder to solve the problem of data sparsity and uses CRF as the classification model to classify the relation between entity pairs.

In recent years, deep neural network (DNN) has been widely used in various fields. The GloVe word embedding model¹⁶ brings great convenience to the NLP task based on DNN. Meanwhile, the recurrent neural network (RNN) with timing characteristics has been increasingly used in the field of NLP, and its effect has been proved generally better than other methods.¹⁷ Therefore, based on the recent deep learning algorithm, we built two application scenes of BiLSTM-CRF model in NER and MRE tasks, and validated it on the I2B2 2010 public dataset, where it showed good results.

MATERIALS AND METHODS

Long short-term memory

As a time-series recurrent neural network, long short-

term memory (LSTM) is suitable for processing and predicting important events with relatively long intervals and delays in the time series. Additionally, it can solve the gradient explosion problem when recurrent neural networks (RNNs) are used for long-term sequence predictions.

There are three control gates in each neuron of LSTM, namely, the input gate, forgetting gate and output gate. The output of the previous unit enters the LSTM unit and it is judged whether it is useful according to the cell. Only useful information is left, and the nonconforming information is forgotten by the forgetting gate. Equations (1) through (5) represent the parameter update process, where σ represents the sigmoid function, h_{t-1} represents the output of the previous unit of the LSTM, and h_t represents the current output. The structure of the LSTM unit is shown in **Figure 1**.

$$i_t = \sigma(W_{xi} + W_{hi}h_{t-1} + b_i) \quad (1)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \quad (2)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \quad (3)$$

$$C_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (4)$$

$$h_t = o_t \tanh(c_t) \quad (5)$$

BiLSTM uses two LSTM networks to propagate the sequence forward and backward, respectively, so that the context information of the sequence can be extracted more comprehensively.

Figure 1. LSTM neuron structure. The tanh is a classical neural network nonlinear activation function. The input gate, the forgetting gate, and the output gate in the LSTM unit are defined as i , f , and o , respectively. C_t is the state of the storage unit at the current time, and equation (4) represents the process of the state transition of the memory unit. The current state is calculated by the previous time state C_{t-1} , the result of the forgotten gate f_t , and the input gate i_t of the current time LSTM unit.

Data preprocessing

The data used in this study is the unstructured English EMR text in the I2B2 2010 evaluation task¹² with the officially labeled medical entity category and medical entity relation category. The dataset divides medical entities into three categories: medical problems, treatment, and tests. It classifies the relations contained between medical entities into eight categories, as shown in **Table 1**.

This study took a sentence in the EMR as a sequence and tried to identify and classifies all entities and medical relations contained in the sentence. As shown in the example in **Figure 2**, the sentence contains three entities, tests, medical problems and treatment, in which "CNIS" and "Carotid stenosis" constitute the medical relation of "TeRP".

In the input section, we used Stanford's GloVe word embedding model to convert EMR text data into 100-dimensional word vector data. We used the 822MB corpus from Wikipedia to combine the English electronic medical record corpus of I2B2 2010 as the

Table 1. Categories of medical relations and descriptions

Category	Description
TrIP	Treatment improves medical problems.
TrWP	Treatment worsens medical problems.
TrCP	Treatment causes medical problems.
TrAP	Treatment is applied to medical problems.
TrNAP	Treatment is not applied to medical problems.
TeRP	Tests reveal medical problems.
TeCP	In order to prove medical problems, need to be checked.
PIP	The relation between medical problems.

CNIS revealed **carotid stenosis** and she ultimately underwent **left carotid stenting** by vascular surgery on 10-30.

c="cnis" 83:0 83:0||t="test"
c="carotid stenosis" 83:2 83:3||t="problem"
c="left carotid stenting" 83:8 83:10||t="treatment"

c="cnis" 83:0 83:0||r="TeRP"||c="carotid stenosis" 83:2 83:3

Figure 2. Text and labeling example of electronic medical record (EMR). The example has three parts. The first part is the sentence containing the entity in the medical record, where "CNIS" is the abbreviation of calcineurin inhibitors (CnIs) and steroids treatment for vascular disease. The middle part shows the line:column number of the entities in the record and its category. The last part is the relation, TeRP, between the entities (CNIS and carotid stenosis) in the sentence.

training corpus of the GloVe word vector.

BiLSTM-CRF in NER

In the NER task, we classified the entities by the strategy of sequence labeling. We used a sentence as a sequence of model input and used the joint labeling strategy of "BIESO" and entity categories for the entities in the sentence. The model first used GloVe to transform the word into word vector. Then it got the hidden layer state of the node through BiLSTM, and finally carried out the joint probability distribution through the CRF layer. The model structure is shown in **Figure 3**.

BiLSTM-CRF in MRE

The result of NER is one of the input parts of the MRE task. So, the MRE task is in tandem with the NER task. The smallest input unit of the MRE model is the entity. Each medical entity is a set of word vectors containing the entity itself and its context information. In order to understand the model more clearly, **Figure 4** shows the neural network structure of a single entity(carotid stenosis) in the model. For each medical entity included in the sentence, we set a context window for it, which is demonstrated using a window size with a fixed context of 2 words.

The word vector sequence is the input of the BiLSTM layer for learning. Considering that the length of the entity is not fixed, we input the start and stop position information of the entity after the BiLSTM

layer, and only extract the hidden layer state of the corresponding entity position in the previous layer. The purpose of this step is to avoid the similarity of the feature information of adjacent entities and to affect the training effect of the model.

Figure 5 shows the complete BiLSTM-CRF medical relation extraction model. After passing through the neural network process shown in Figure 4, each entity performs probability calculation through the CRF layer to obtain a corresponding medical relation classification. The purpose of this layer is to transform the classification problem of a single entity into a sequence classification problem, and to link the feature combinations between entities. In fact, the model can be seen as a CRF model optimized for feature engineering by word vector and BiLSTM.

RESULTS

As mentioned earlier, this study uses the English electronic medical record data published in the I2B2 2010 evaluation meeting to conduct experiments. Of the 871 marked medical records, 397 were used for training and 477 were used for testing.¹⁸ The purpose of the experiment was to identify the correct physical and medical relations from the unstructured electronic medical record text. This study uses precision (P), recall (R) and F1-measure as indicators of evaluation. F1-measure is a weighted harmonic averaging of the precision and the recall. It is a commonly used evalu-

Figure 3. BiLSTM-CRF structure of named entity recognition(NER). "I1", "r1", "e1", etc., represents the different network layers in the model. The CRF layer is a combination label, where "B", "I", "E", "S", "O" are word segmentation labels and "Te", "P", "Tr" are entity category labels. BiLSTM, bidirectional long short-term memory; CRF, conditional random field.

Figure 4. Single entity neural network structure of medical relation extraction(MRE). “TeRP” in the CRF layer is a relation category label.

Figure 5. BiLSTM-CRF structure of MRE. FC is the fully connected layer.

ation standard in the field of Information Retrieval and is often used to evaluate the quality of a classification model. Let the result set of the model output be $\hat{y}$, and the result set of the manual label be y . The formula for the evaluation criteria is as follows:

$$\begin{aligned} P &= \frac{|\hat{Y} \cap Y|}{|\hat{Y}|} \\ R &= \frac{|\hat{Y} \cap Y|}{|Y|} \\ F1 &= \frac{2PR}{P+R} \end{aligned} \quad (6)$$

The result of NER

The experimental results of NER are shown in **Table 2**. We compared the model in this study with other

well-performing models on the dataset. The SVM and CRF model were proposed by Jiang M,¹⁹ which used the Systematized Nomenclature of Medicine (SNOMED) as an auxiliary resource. The LSTM model was chosen for a more intuitive comparison. The results show that the BiLSTM-CRF model has better expressiveness in the dataset. Moreover, the model converged faster and avoided problems such as overfitting.

The result of MRE

The experimental results of MRE are shown in **Table 3**. We compared the BiLSTM-CRF model with other well-performing models. Among them, SVM¹³ and maximum entropy (ME)²⁰ are both obtained by semi-supervised training. It is clear that the perfor-

Table 2. The F1-measure of models in NER

Models	Medical Problems	Treatment	Tests	Total
SVM	0.861	0.829	0.785	0.832
CRF	0.878	0.845	0.792	0.847
HMM ²⁰	0.875	0.851	0.804	0.852
LSTM	0.892	0.863	0.816	0.861
BiLSTM-CRF	0.902	0.896	0.832	0.879

NER, named entity recognition; SVM, support vector machine; CRF, conditional random field; HMM, hidden Markov model; LSTM, long short-term memory; BiLSTM-CRF, bidirectional long short-term memory combined conditional random field; Total means all categories of entities.

mance of our model is still better than others. From the implementation details, the classification of our model can effectively avoid the misclassification of non-entity texts. There will also be good performance in real scenes outside the dataset.

DISCUSSION

From the experimental results, the BiLSTM-CRF model constructed in this study has made a significant improvement over the previous research. The F1-measure of NER is close to 0.88, and of MRE is close to 0.78. However, there still remain some problems. The model has categories with a low recognition rate in both tasks. Especially in MRE, there are three categories where the F1-measures were less than 0.4. The general reason for this problem is that these categories are too small in the dataset and was difficult for the model to understand their semantics. For example, in the sentence "He was started on p.o. steroids and to CMED for **management** of **COPD exacerbation** but he appeared in more respiratory distress overnight", the "management" and "COPD exacerbation" should constitute a "TrWP" relation, but the model identified it as a "TrAP". The reason lied in that the patient's deterioration appears in the next short sentence, and the model didn't perfectly learn the feature.

In addition, the effects of the model on several

categories were higher than 0.9, of which "PIP" was all correctly classified. This is due to the large number of its examples and less interference.

This study constructed different structure of BiLSTM-CRF model and verified the good performance of deep learning model in medical knowledge extraction. This shows that the same model can be applied to different scenarios by reasonable modeling of the characteristics of the task. In view of the fact that clinical electronic medical records contain more domain experience knowledge, how to make corresponding adjustments and changes from the general method, and how to construct a more specific model will be questions that NLP needs to think about in the research field of clinical EMRs in the future. Besides, as two subtasks of MKE, the clinical accuracy of MRE is seriously affected by the results of NER. Although NER has a F1-measure close to 0.9, the error will be magnified during transmission. Therefore, when it comes to deeper research, it is imperative to further improve the effect of basic tasks in MKE.

Conflict of interest statement

The authors have no conflict of interests to disclose.

REFERENCES

1. Wu JW, Guan Y, Lv XB. Entity relation extraction from electronic medical records based on deep learning. *Intell comput appl* 2014; 4(3):35-8. doi: 10.3969/j.issn.2095-2163.2014.03.009.
2. Chen L, Li Y, Chen W, et al. Utilizing Soft Constraints to Enhance Medical Relation Extraction from the History of Present Illness in Electronic Medical Records. *J Biomed Inform* 2018; 87:108-17. doi: 10.1016/j.jbi.2018.09.013.
3. Xue NW, Shen LB. Chinese word segmentation as LMR tagging. *Proceedings of the second SIGHAN workshop on Chinese language processing*. 2003, Jul 11-12; Sapporo, Japan. Stroudsburg, PA, USA: Association for computation Linguistics; 2003. 17:176-9. doi:

Table 3. The F1-measure of models in MRE

Models	TrIP	TrWP	TrCP	TrAP	TrNAP	TeRP	TeCP	PIP	Total
SVM	0.23	0.05	0.496	0.806	0.17	0.872	0.45	0.87	0.737
ME	0.216	0.02	0.502	0.814	0.193	0.859	0.393	0.91	0.731
DNN+CRF ¹	0.225	0.03	0.534	0.86	0.225	0.916	0.451	0.96	0.752
BiLSTM-CRF	0.251	0.11	0.572	0.903	0.35	0.931	0.503	0.98	0.775

MRE, medical relation extraction; ME, maximum entropy; DNN, deep neural network.

- 10.3115/1119250.1119278.
4. Finkel JR, Grenager T, Manning C. Incorporating non-local information into information extraction systems by Gibbs sampling. *Proceedings of the 43rd annual meeting on association for computational linguistics* 2005; 363-70. doi: 10.3115/1219840.1219885
 5. Ye F, Chen W, Zhou GG, et al. Intelligent recognition of named entities in electronic medical records. *Chin J of Biomed Eng* 2011; 30(2):256-62. Chinese. doi: 10.3969/j.issn.0258-8021.2011.02.014.
 6. Li W, Zhao DZ, Li B, et al. Entity recognition of medical records combined with CRF and rules. *App Res Comput* 2015; 32(4):1082-6. doi: 10.3969/j.issn.1001-3695.2015.04.029.
 7. Bollegala D, Matsuo Y, Ishizuka M. Relation adaptation: learning to extract novel relations with minimum supervision. *Proceedings of the 22nd International Joint Conference of Artificial Intelligence (IJCAI)*. 2011 Jul 16-22; Barcelona, Spain. Menlo Park, California, USA: AAAI Press; 2011. p.2205-10. doi: 10.5591/978-1-57735-516-8/IJCAI11-368.
 8. Suchanek FM, Ifrim G, Weikum G. Combining linguistic and statistical analysis to extract relations from web documents. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2006 Aug 20-23; Philadelphia, USA. New York: ACM; 2006. p.712-7. doi: 10.1145/1150402.1150492.
 9. Qin B, Liu AA, Liu T. Unsupervised Chinese Open Entity Relation Extraction. *J Comput Res Dev* 2015; 52(5):1029-35. Chinese. doi:10.7544/issn1000-1239.2015.20131550.
 10. Uzuner O, Mailoa J, Ryan R, et al. Semantic relations for problem-oriented medical records. *Artif Intell Med* 2010; 50(2):63-73. doi: 10.1016/j.artmed.2010.05.006.
 11. Zhou G, Su J, Zhang J, et al. Exploring various knowledge in relation extraction. *ACL '05 Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. 2005 Jun 25-30; Michigan, USA. Stroudsburg, PA, USA: Association for computational Linguistics; 2005. p.427-34. doi: 10.3115/1219840.1219893.
 12. Uzuner O, South BR, Shen S, et al. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assn* 2011; 18(5): 552-6. doi: 10.1136/amiajnl-2011-000203.
 13. Rink B, Harabagiu S, Roberts K. Automatic extraction of relations between medical concepts in clinical texts. *J Am Med Inform Assn* 2011; 18(5):594-600. doi: 10.1136/amiajnl-2011-000153.
 14. Demner-Fushman D, Mork JG, Shooshan SE, et al. UMLS content views appropriate for NLP processing of the biomedical literature vs. clinical text. *J Biomed Inform* 2010; 43(4):587-94. doi: 10.1016/j.jbi.2010.02.005.
 15. Lv X, Yi G, Yang J, et al. Clinical relation extraction with deep learning. *Int J Inf Tech Decis* 2016; 9(7): 237-48. doi: 10.14257/ijhit.2016.9.7.22.
 16. Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014 Oct 25-29; Doha, Qatar. p.1532-43. doi: 10.3115/v1/D14-1162.
 17. Yin W, Kann K, Yu M, et al. Comparative study of CNN and RNN for natural language Processing 2017; arXiv:1702.01923.
 18. Yang JF, Yu QB, Guan Y, et al. A Review of research on name recognition and entity relation extraction of electronic medical records. *Acta Automatica Sinica* 2014; 40(8):1537-62. doi: 10.3724/SP.J.1004.2014.01537.
 19. Jiang M, Chen Y, Liu M, et al. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *J Am Med Inform Assn* 2011; 18(5):601-6. doi: 10.1136/amiajnl-2011-000163.
 20. De Bruijn B, Cherry C, Kiritchenko S, et al. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *J Am Med Inform Assn* 2011; 18(5):557-62. doi: 10.1136/amiajnl-2011-000150.