Original Article

Model-based clinical note entity recognition for rheumatoid arthritis using bidirectional encoder representation from transformers

Meiting Li[1#^], Feifei Liu[2#], Jia'an Zhu[2], Ran Zhang[1], Yi Qin[1], Dongping Gao[1]

[1]Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing, China; [2]Department of Ultrasound, Peking University People's Hospital, Beijing, China

Contributions: (I) Conception and design: D Gao; (II) Administrative support: D Gao, J Zhu; (III) Provision of study materials or patients: J Zhu, F Liu; (IV) Collection and assembly of data: F Liu, M Li; (V) Data analysis and interpretation: M Li, F Liu; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Dongping Gao. Institute of Medical Information, Chinese Academy of Medical Sciences, No. 3 Yabao Road, Chaoyang District, Beijing, China. Email: gaodp_gaodp@126.com.

Background: Rheumatoid arthritis (RA) is a disease of the immune system with a high rate of disability and there are a large amount of valuable disease diagnosis and treatment information in the clinical note of the electronic medical record. Artificial intelligence methods can be used to mine useful information in clinical notes effectively. This study aimed to develop an effective method to identify and classify medical entities in the clinical notes relating to RA and use the entity identification results in subsequent studies.

Methods: In this paper, we introduced the bidirectional encoder representation from transformers (BERT) pre-training model to enhance the semantic representation of word vectors. The generated word vectors were then inputted into the model, which is composed of traditional bidirectional long short-term memory neural networks and conditional random field machine learning algorithms for the named entity recognition of clinical notes to improve the model's effectiveness. The BERT method takes the combination of token embeddings, segment embeddings, and position embeddings as the model input and fine-tunes the model during training.

Results: Compared with the traditional Word2vec word vector model, the performance of the BERT pre-training model to obtain a word vector as model input was significantly improved. The best F1-score of the named entity recognition task after training using many rheumatoid arthritis clinical notes was 0.936.

Conclusions: This paper confirms the effectiveness of using an advanced artificial intelligence method to carry out named entity recognition tasks on a corpus of a large number of clinical notes; this application is promising in the medical setting. Moreover, the extraction of results in this study provides a lot of basic data for subsequent tasks, including relation extraction, medical knowledge graph construction, and disease reasoning.

Keywords: Named entity recognition; rheumatoid arthritis (RA); artificial intelligence; bidirectional encoder representation from transformers (BERT); clinical notes

Submitted Jan 24, 2021. Accepted for publication Jun 24, 2021.
doi: 10.21037/qims-21-90
View this article at: https://dx.doi.org/10.21037/qims-21-90

^ ORCID: 0000-0003-1555-7728.

© Quantitative Imaging in Medicine and Surgery. All rights reserved.

Introduction

Rheumatoid arthritis (RA) is an immune system disease with joint erosion as the main clinical manifestation. The main site of injury is the synovial joint. If the disease is not controlled early, erosion of cartilage and bone will lead to joint destruction, causing joint deformities resulting in severe motor disability within 10 to 20 years. Also, studies have shown that patients with RA have a 1.5 to 2.0 times higher risk of cardiovascular disease than the general population (1). Rapid control of inflammation in the course of the disease by modifying anti-rheumatoid drugs (2) is the current treatment principle. Therefore, for RA disease, early diagnosis, and drug treatment are of great significance for the treatment and prognosis of the RA.

In many cases, artificial intelligence (AI) methods can be used to mine knowledge existing in clinical notes that clinicians cannot manage and process quickly. However, AI technology has its characteristic challenges in different fields. For example, in the medical field, real world data will greatly improve the accuracy of the model, and the results will be better used in interpreting medical data (3). Electronic medical records (EMRs) often contain the most valuable and clinically significant information about diseases. However, it is very difficult to extract valuable information from EMRs because of the large amount of unstructured data in the form of free text (4), such as course records, patients' chief complaints, doctor-patient communication records, doctor-patient agreements, and examination results. At present, no completely reliable entity extraction tool is available that does not rely on human beings for entity extraction from clinical texts. Named entity recognition in the medical domain refers to identifying entity information and classification of valuable entities in a large number of unstructured text data.

Bidirectional encoder representations from transformers (BERT) is a language pre-processing model for representing word vectors developed by Jacob Devlin *et al.* (5) in 2018. It has been proven to perform better than the most advanced algorithms in 11 natural language processing (NLP) tasks (6), including named encoder relation (NER), which is suitable for solving tasks with deep semantic characteristics. In the task of named entity recognition using a neural network model, the feature representation of a word vector has a significant impact on the effect of entity recognition. Bidirectional encoder representations from transformers use the deep bi-directional transformer as a feature extractor to dynamically generate word vectors with full consideration of context, which is superior to the traditional memory network in terms of learning effect and can realize the integrated bi-directional prediction with the help of the masked language model. The transformer uses the self-attention mechanism to solve the challenge of parallel processing and long-term dependency of a large amount of data in the corpus. BERT first builds masked-language modeling (MLM) on the general domain data during the pre-training period and performs the next sentence prediction (NSP) tasks. BERT is initialized with pre-trained parameters and fine-tuned with tagging data for other specific tasks such as entity recognition, text classification, and automatic questions and answers in the fine-tuning phase.

Based on the superior performance of a variety of natural language application scenarios and two-stage task features of BERT, our research used the BERT pre-training word vector embedding corpora text representation to characterize vectors. It then used the bidirectional network input from both short- and long-term memory studies, eventually using conditional random field (CRF) global optimal output sequence tags. Compared with the existing research, on the one hand, there is a large amount of high-quality and rich content of RA clinical notes entity data that are manually labeled; on the other hand, a BERT language model based on RA clinical text in Chinese was set with an F-score of 0.936. The model can be used for subsequent records automatically for entity tagging and knowledge map drawing, and disease reasoning in Chinese clinical NLP tasks. The experimental results show that the performance of the model using BERT is better than the traditional NER model (7).

The main contributions of this paper are as follows:

(I) Under the supervision of clinicians, the clinical notes included in 1,600 EMRs relating to RA were analyzed and processed, and the diagnosis, treatment, examination, and other basic information regarding RA were annotated in detail. This work provides authoritative and reliable data resources for further research in the RA field, such as constructing automatic questions and answers knowledge base.

(II) A large number of reliable clinical data were used to train a named entity recognition (NER) model in the RA field, which can be directly used for entity recognition in the RA field. The above entity extraction results of this research provide basic data for subsequent medical relationship extraction,

© Quantitative Imaging in Medicine and Surgery. All rights reserved.

medical knowledge graph construction, disease reasoning, and other tasks that have practical significance in scientific research and clinical practice.

(III) The most advanced models in the field of natural language processing have been adopted, which achieved good results in many tasks of natural language processing. In this research, artificial intelligence technology was applied in the field of RA, providing a preliminary exploration for the modernization of disease diagnosis and treatment. More technologies are expected to be applied in the field of RA in the future based on the data from this study.

Related work

In the entity recognition task of NLP, there are three main recognition methods: dictionary-and-rule-based, machine-learning-based, and neural-network-based methods. The dictionary-and-rule-based approach requires manual customization of many characteristic rules, which are usually specific to a particular document (8). The method based on machine learning needs to build a lot of feature engineering. The above methods have poor generalization ability, which is not conducive to the generalization of results and transfer learning of results.

In recent years, the method of deep learning without feature engineering has been paid more and more attention to and is being widely used in NER tasks due to its strong generalizability (9). Xia *et al.* (10) used the long short-term memory (LSTM) model to identify medical entities in EMR text and achieved 89.44% of the F1-score, which confirmed the effectiveness of deep learning algorithms in EMRs named entity recognition tasks. Cheng *et al.* (11) extracted clinical entities and attributes from various types of clinical narrative texts, such as operative records, discharge summaries, clinical data requests, etc., using a novel hybrid approach called clinical entity and attributes extractor (CEAER), which combines the rules and bidirectional LSTM networks with a conditional random field layer model. Finally, this bidirectional long short-term memory conditional random fields (BiLSTM-CRF) model achieved an F1-score of 87.00%. Yin *et al.* (12) used convolutional neural networks (CNNs) to encode the radical-level representation of characters, then utilized the self-attention mechanism to capture long-term dependencies between characters in a single sequence. Compared with the BiLSTM-CRF model, the performance of the AR-CCNER (Chinese Clinical Named Entity Recognition model based on Self-Attention mechanism and Radical-level features) model improved the F1-score by 0.55% on the China Conference on Knowledge Graph and Semantic Computing (CCKS) 2017 dataset and improved the F1-score by 1.1% on a CNER (Chinese Named Entity Recognition) dataset, which medical experts manually annotate.

The representation of word vectors has led to developing language models such as Word2Vec, embeddings from language models (ELMO), and generative pre-trained (GPT) models. Word2Vec is a static expression, which ignores the polysemy of words in the text. Embeddings from language models use LSTM instead of a transformer with better feature extraction and adopt bidirectional splicing fusion feature mode instead of integration fusion mode. Generative pre-trained models adopt the structure of the unidirectional language model, which can be intuitively considered because the representation ability of words is not as good as that of the bidirectional model. With the continuous enrichment of the experimental corpus and the development of artificial intelligence technology, the superiority of BERT in representational word vectors has been proven. Li *et al.* (13) added the BERT model pre-trained on the Chinese clinical data to the bidirectional LSTM layer and further improved the performance by adding dictionary and root features. They achieved 89.56% and 91.60% F1 scores on the CCKS 2018 and CCKS 2017 NER task datasets, respectively, showing superior performance compared with the existing models. Xie *et al.* (14) obtained contextualized word vectors through the BERT language pre-processing model and then constructed the BERT-BiLSTM-CRF model to conduct experiments on two typical corpus datasets, obtaining F1-scores of 94.65% and 95.67%. Zhang *et al.* (15) embedded the BERT pre-trained context to the BiLSTM-CRF model and achieved an F1-score of 93.52% in the NER task of breast cancer EMRs and an F1-score of 96.73% in relation extraction. Zhang *et al.* (16) proposed a pre-trained BiLSTM-CRF model, which employs BERT to enrich the semantic representation of a word vector and introduced it into the BiLSTM network as input for training. Experiments on the CNMER-2019 corpus showed that this model improved the performance of Chinese medical text named entity recognition to an F-score of 84.32%. In order to extract the entities in Chinese EMRs more efficiently and accurately, Li *et al.* (17) integrated the BERT word embedding model based on the traditional BiLSTM-CRF model and obtained

© Quantitative Imaging in Medicine and Surgery. All rights reserved.

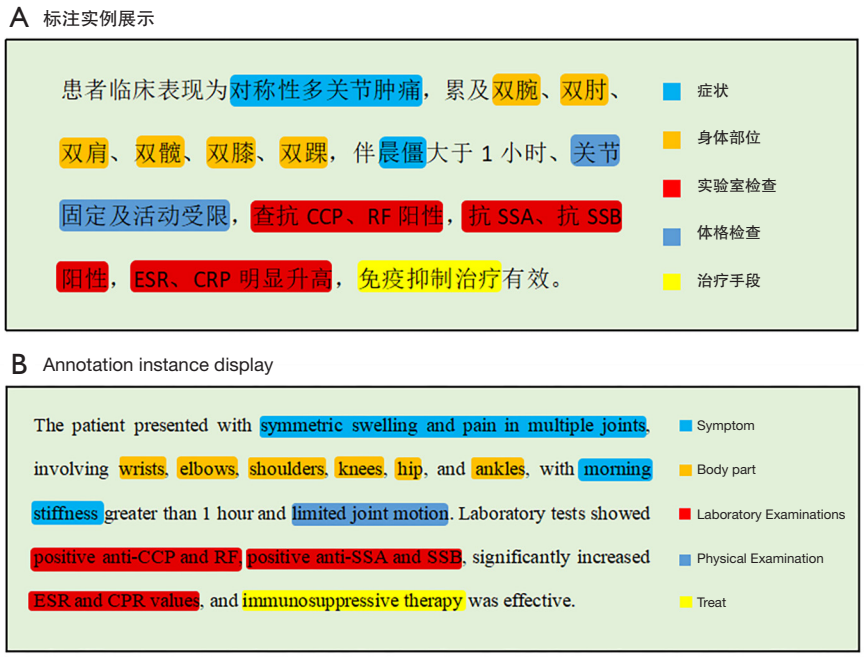

Figure 1 Example of annotation results.

an F-score of 88.45 on the named entity recognition task on the CCKS2019 dataset. Wang *et al.* (18) compared to the recognition effects of different levels of pre-training models on NER task in the dataset "labeled Chinese dataset for diabetes (LCDD)"; the results show that recognition gets better as the level gets deeper, and in addition, the specific effects of recognition are related to the corpus. Therefore, high-quality clinical notes data with manual annotation was adopted in this study to avoid the adverse impact of the recognition effect due to the quality of the corpus; moreover, a large number of data labels were obtained for further research.

Methods

Clinical notes data annotation

In this study, relevant research on named entity recognition of EMRs at home and abroad was investigated in advance. Combined with the actual characteristics of RA medical records, RA named entities with rich content were labeled in unstructured EMR data (19), forming the manual annotation corpus of 1,600 clinical notes data. The final annotation results (partial entities) are shown in *Figure 1*, highlighting RA's diagnostic information, symptoms, and treatment.

Embedding layer

In practical tasks, BERT's two-stage task strategy can save training time on the one hand and make up for the lack of training corpus, on the other hand, demonstrating its good performance in the field of natural language processing. The first stage is the pre-training of the language model, in which the prior semantic knowledge is acquired from a large amount of data of the unmarked corpus, and then the knowledge is transferred to the specific task by combining it with other models to improve the performance of the specific task model. The second stage is fine-tuning with specific tasks because there is no pre-training BERT language model for general Chinese clinical records. We used a task-specific fine-tuning method after training in the general generic domain, which transferred prior semantic knowledge of the generic domain in the original model from the source domain to the target task domain, in this case, the RA domain. The internal structure of BERT is derived from the encoder of the transformer model and stacked by the self-attention layer and normalization layer.

Bidirectional encoder representation from transformers is the first fine-tuning presentation model, which achieves the most advanced performance on a large number of tasks and is superior to many specific task architectures. The core architecture of this article is shown in *Figure 2*. In order to

© Quantitative Imaging in Medicine and Surgery. All rights reserved.

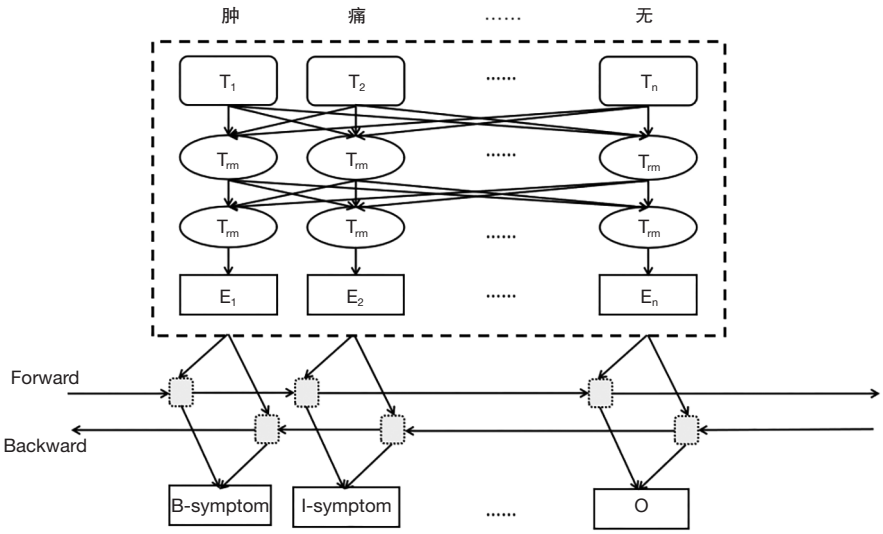

Figure 2 Bidirectional encoder representation from transformers-bidirectional long short-term memory-conditional random field (BERT-BiLSTM-CRF) model framework.

Table 1 Whole word mask style

Original text:
25 年前，患者因关节肿痛加重伴活动障碍 ……
Not whole word mask input:
25 年前，[mask] 者因关节肿 [mask] 加重伴活动障碍 ……
Whole word mask input:
25 年前，[mask][mask] 因关节 [mask][mask] 加重伴活动障碍 ……

improve the limitation of using unidirectional information only of the above methods and avoid the "see itself" effect produced in the training process of characters by the training method of two-way regulation, BERT adopts a "[mask]" mechanism. The specific operation was as follows: 15% of the words in the corpus were randomly selected for masking, and 80% of the masked characters were replaced with [mask] labels, 10% remained unchanged, and the remaining 10% were randomly replaced with other labels. This mechanism, similar to the cloze test, enables pre-training the deep bidirectional transformer model by fusing both left and right contexts to predict the current markup. At the same time, considering that the minimum semantic unit in Chinese is a phrase, the whole word masking strategy was adopted in this study, which fully considers the characteristics of Chinese clinical texts. In the whole word mask strategy, if part of a complete entity word is randomly masked, then the rest that belongs to the same word would also be masked. The specific representation is shown in *Table 1*: this strategy has been proven to have substantial performance gains in the field of natural language processing. The whole word masking strategy needs to perform Chinese word segmentation and then mask the tag belonging to the same word. The effect of this is to predict not a single masked tag but all the masked tags in the same entity word. Pre-training the "mask language model" (MLM) by using bidirectional transformer encoders, alleviating the unidirectional constraint of context by masking some tags randomly from the input context and greatly predicting the original masked words, which improves the performance of most NLP tasks. In addition, the training missions of BERT also include the function of "next sentence predicting", which aims to enhance the model's ability to understand the context of a long sequence. This mission is implemented as follows when sentence A and sentence B are both used as pre-training samples, there is a 50% chance that B will be the next sentence of A and a 50% chance that A has no semantically relevant relation to B. The core of BERT is a deep network structure based on the mechanism of "self-attention", which enhances the semantic ability of words by generating a weight matrix by comparing the degree of association between words in the same sentence, such that each word contains all information about the sentence in which it is located. The formalized

© Quantitative Imaging in Medicine and Surgery. All rights reserved.

representation of the weight matrix is shown in *Figure 3*. According to the "self-attention" mechanism, the degree of connection between words is related to the depth of color. For example, in the sentence "...multiple joint pain for 32 years..." ("…… 多关节肿痛 32 年 ……") for the prediction, the words "pain" (" 肿痛 ") and "joint" (" 关节 ") are more important than "32 years" ("32 年 "), that is to say, the semantic similarity of dark blue representation in the matrix is greater than that of light yellow representation. Q, K, and V are the same input as X multiplied by the initialization weight matrix, and then the transpose matrix of Q and K is used for matrix multiplication. In order to avoid the result being too large, the result is divided by the root mean square of the embedded dimension, normalized into a probability distribution, and finally multiplied by matrix V to get the final result.

$$\text{Attention}(Q,K,V)=\text{Soft max}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad [1]$$

In this paper, the multi-headed attention mechanism is projected through multiple different linear transformations, learning from multiple dimensions, and finally related to different attention results.

$$\text{Multihead}(Q,K,V)=\text{Concat}(\text{head}_1,\ldots,\text{head}_h)W^O \quad [2]$$

$$\text{headi}=\text{Attention}(QW_i^Q,QW_i^K,QW_i^V) \quad [3]$$

In the input layer of BERT, each character is composed of token embeddings, segment embeddings, and position embeddings, as shown in *Figure 4*. [SEP] is a marker, which separates two sentences. Each input sentence begins with [CLS], which indicates whether the next two sentences are semantically related.

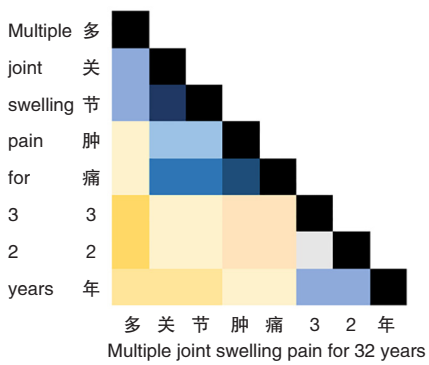

Figure 3 Bidirectional encoder representation from transformers (BERT) self-attention matrix.

BiLSTM-CRF layer

Long short-term memory is a kind of recurrent neural network (RNN) suitable for modeling text data. Compared with RNN, LSTM adds a memory unit consisting of three control gate structures: input gate, forget gate and output gate (20). Therefore, LSTM can capture contextual information of words and solve the problem of long-distance dependency of long text. Based on the above characteristics, LSTM can be used for the disease named entity recognition (21). The long and short-term memory network selectively preserves the input information at the current time t through the memory unit. For an input sentence sequence X = (x_1, x_2. . . x_t. . . x_n), the forget gate determines the information that should be discarded at the current moment, the input gate calculates the information needed to update the current cell state, and the output gate determines the hidden layer state of the final output at the current moment. Xt is the input vector corresponding to the current word in the sentence, and ht is the hidden state

Yunke was used for anti–inflammation and analgesia, and discharging after the swelling and pain of joints improved

Input	[CLS]	给	予	云	克	抗	炎	镇	痛	[SEP]	关	节	肿	痛	好	转	后	出	院	[SEP]
Token embeddings	$E_{[CLS]}$	$E_{给}$	$E_{予}$	$E_{云}$	$E_{克}$	$E_{抗}$	$E_{炎}$	$E_{镇}$	$E_{痛}$	$E_{[SEP]}$	$E_{关}$	$E_{节}$	$E_{肿}$	$E_{痛}$	$E_{好}$	$E_{转}$	$E_{后}$	$E_{出}$	$E_{院}$	$E_{[SEP]}$
+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Segment embeddings	$E_{[CLS]}$	E_A	E_A	E_A	E_A	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B	E_B	E_B	E_B	E_B	E_B
+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Position embeddings	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}	E_{11}	E_{12}	E_{13}	E_{14}	E_{15}	E_{16}	E_{17}	E_{18}	E_{19}

Figure 4 Bidirectional encoder representation from transformers (BERT) input vector representation.

© Quantitative Imaging in Medicine and Surgery. All rights reserved.

of the layer at the last moment. C_t is the current cell state and $\tilde{C}$ is a temporary cell state.

$$i_t = \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}C_{t-1} + b_i\right) \quad [4]$$

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}C_{t-1} + b_f\right) \quad [5]$$

$$o_t = \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}C_{t-1} + b_o\right) \quad [6]$$

$$\tilde{C} = \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right) \quad [7]$$

$$c_t = f_t c_{t-1} + i_t \tilde{C} \quad [8]$$

$$h_t = o_t \tanh\left(c_t\right) \quad [9]$$

σ is the activation function, i, f and o represent input gate, forget gate and output gate respectively, W is the weight matrix, while b is the bias vector parameter.

In fact, in the text of clinical notes, both past and future information are equally important in understanding vocabulary. Therefore, bidirectional contextual information needs to be considered in the named entity recognition model, while the unidirectional memory unit can only consider previous information (22,23). Considering the strong dependence on the context when predicting the classification of named entities, for example, when the text describes the following: "Patient-reported multiple joint swelling and pain for 9 years and aggravation for 1 month..." ("患者自述多关节肿痛9年，加重1月……"), the description of time in the end can also assist in the identification of the symptom entity of "joint swelling and pain" ("关节肿痛"). Yao *et al.* (22) first applied the bidirectional LSTM network to the word segmentation task, thus improving the performance of Chinese word segmentation. The bidirectional memory network divides each training sequence's forward and backward parts into two LSTMs, which are connected to an output layer later. For the input sequence $X=\{x_1,x_2,x_3,...,x_n\}$, the forward LSTM receives the input sequence from x_1 to x_n and calculates the hidden forward state $\vec{h}$, and the backward LSTM receives the input sequence from x_n to x_1 and calculates the backward hidden state $\overleftarrow{h}$. Finally, by connecting the hidden forward state with the backward hidden state, the complete past and future context information of each point in the input sequence of the output layer ht can be obtained. Bidirectional LSTM combines the forward and reverse prediction probability of the current label, combining the results of two different vectors to represent the ultimate hidden layers, which can yield more comprehensive information.

Among many machine learning methods applied to NER tasks, such as logistic regression (LR) (24), hidden Markov model (HMM) (25), support vector machine (SVM) (26), and conditional random field (CRF) (27), the conditional random field has been proven to have the best effect (6). Simply adding a linear layer after the neural network layer might ignore the strong dependence of output labels in NER. Therefore, we use CRF as the final output layer of BiLSTM to solve this problem. On the one hand, the characteristics of the context are well integrated; on the other hand, the dependency between the output labels is effectively considered (28).

The CRF score calculation formula consists of two parts. One part is the tag score matrix of the current position. For example, $P_{i,yj}$ is the confidence score of the jth tag of the ith position. The other part is the transfer score matrix between the current position and the previous position. For example, $T_{yi,yj}$ represents the transfer score of adjacent labels i and j. $T_{y0,yj}$ is the initialization score of tag j, the CRF layer can consider overall the current word and the previous word of the current word to get the most likely output.

$$S\left(X,y\right) = \sum_{i=0}^{n} T_{yi},y_{i+1} + \sum_{i=1}^{n} P_{i,yi} \quad [10]$$

All possible tag sequences were normalized to obtain the tag probability of sequence y. In the training process, we maximized the logarithmic probability of correct tag sequences y*.

$$\mathrm{Log}\left(P\left(y^*\middle|X\right)\right) = s\left(X,y^*\right) - \log\left(\Sigma_{\tilde{y}\in Y_X} e^{s(X,\tilde{y})}\right) \quad [11]$$

y* represents the result of the true tag, and Y_X represents all possible tags. After decoding, the output sequence with the maximum score is obtained:

$$y^* = \arg\max_{\tilde{y}\in Y_X} S\left(X,\tilde{y}\right) \quad [12]$$

Experiment

In the phase of entity data annotation in this study, the "BIO" marker was used to represent the type of entity and the position of characters in the entity for the word entity blocks appearing in medical record data. The "B" tag indicates the first position of the named entity; the "I" tag indicates any position in the entity other than the beginning position, and the "O" tag indicates that this

© Quantitative Imaging in Medicine and Surgery. All rights reserved.

Joint	关 B- 症状	B-Symptom
	节 I- 症状	I-Symptom
swelling	肿 I- 症状	I-Symptom
	痛 I- 症状	I-Symptom
pain	加 O	
	重 O	
aggravated,	……	
	查 O	
with	类 B- 实验室检验	B-Laboratory Examinations
	风 I- 实验室检验	I-Laboratory Examinations
rheumatoid	湿 I- 实验室检验	I-Laboratory Examinations
	因 I- 实验室检验	I-Laboratory Examinations
factor	子 I- 实验室检验	I-Laboratory Examinations
	阳 O	
positive	性 O	

Figure 5 "BIO" annotation sample.

Table 2 Experimental environment settings

Project	Environment
Operating System	Windows
Central processing unit (CPU)	Intel(R) Xeon(R) CPU E7-4850 v3 @ 2.20GHz
Computer memory	32.0GB
Python version	3.6
TensorFlow version	1.12.0

position is not part of the entity. Each clinical note is added with many medical entity names, location of entity words, and predefined entity categories in detail. Specific labeling samples are shown in *Figure 5*. Considering the characteristics of Chinese medical records writing, we treated each sentence ending with a period as a whole. In addition, the named entity recognition model of this experiment was based on the TensorFlow framework, and the specific experimental environment settings are shown in *Table 2*.

The model parameters of BiLSTM-CRF were set as follows: the dimension of the pre-training embedding vector was 300, the learning rate was 1E-3, the batch_size was 32, the dimension of the LSTM hiding layer was 300, the clip was 5, the dropout was 0.5, and the optimization algorithm was Adam.

The model parameters of BiLSTM-ATT-CRF were set as follows: the pre-training embedding vector dimensions was 200, the maximum sentence length was 300, the batch_size was 32, the dropout was 0.5, the learning rate was 1E-3, the attention-dim was 400, and the optimization algorithm was Adam.

The model parameters of BERT-BiLSTM-CRF were set as follows: the maximum sentence length was 128, the clip was 5, the dropout was 0.5, the learning rate was 1E-3, and the optimization algorithm was Adam.

Results

In the result evaluation stage, precision, recall, and F1-score were used to evaluate the effect of entity recognition comprehensively. Accurate represents the correct proportion of the predicted results of entities; recall rate represents the proportion of the correct number of predicted entities in all entities; true positive (TP) is the number of correctly identified medical entities, false-positive (FP) is the number of wrongly identified unrelated medical entities, and false-negative (FN) is the number of unidentified medical entities; F1-score is the average harmonic value of accuracy and recall rate, which can comprehensively reflect the performance of classification model in the NER task. The criteria for accurate medical entity prediction are that the entity's boundary and category are both predicted correctly.

$$\text{Precision} = \frac{TP}{TP + FP} \times 100\% \quad [13]$$

$$\text{Recall} = \frac{TP}{TP + FN} \times 100\% \quad [14]$$

$$\text{F1} = \frac{\text{precision} \times \text{recall} \times 2}{\text{precision} + \text{recall}} \times 100\% \quad [15]$$

In this study, 1,600 RA medical records were manually annotated at first, and then the data were divided into a training set and a test set according to a ratio of 4:1 for model training and test. The results of data set allocation are shown in *Table 3*.

(I) In the process of corpus labeling, "disease" refers to the accompanying diseases of RA patients that appear in medical records. For example, "high blood pressure" ("高血压"), "diabetes" ("糖尿病"), "osteoporosis" ("骨质疏松症") and so on.

(II) Symptoms refer to the disease in the body in the process of a series of function, metabolism, and morphology changes caused by the patient's subjective feeling regarding the symptoms or some objective pathological change. In the medical records relating to RA, joint symptoms may include

© Quantitative Imaging in Medicine and Surgery. All rights reserved.

Table 3 Number of training and test sets for the entities

Number	Named entity	Train set	Test set	Total
1	Disease	8,956	2,241	11,197
2	Symptom	26,166	6,541	32,707
3	Body parts	30,307	7,577	37,884
4	Laboratory examinations	24,912	6,228	31,140
5	Imaging examination	20,828	5,207	26,035
6	Physical examination	13,392	3,348	16,740
7	Treatment	24,413	6,103	30,516
8	Other examinations	301	75	376
9	Relieving factor	264	66	330
10	Aggravating factor	382	96	478

"pain in the joints" ("关节肿痛"), "morning stiffness" ("晨僵"), and "joint swelling" ("关节肿胀") and other symptoms may be "dry mouth" ("口干"), "dry eye" ("眼干"), etc.

(III) Body part refer to the joint part often involved in joint symptoms. Different parts are often affected in different patients due to different work and living habits.

(IV) Laboratory examination means that physical or chemical examinations have been conducted in the laboratory to determine the content, nature, concentration, quantity, and other characteristics of the substance to be tested. In the diagnosis of RA patients, laboratory examinations mostly refer to "rheumatoid factor" ("类风湿因子"), "anti-CCP antibody" ("抗 CCP 抗体"), "RF" ("RF"), "AFP" ("AFP"), "DIC total" ("DIC 全项"), and other hematological indicators.

(V) Imaging examinations include "joint color ultrasound" ("关节彩超"), "hand anteroposterior radiography" ("双手正位片"), "anteroposterior and lateral position of knee loading" ("膝关节负重正侧位"), "double patella position" ("双髌骨位"), etc.

(VI) Physical examinations refer to the detection and measurement of the human body's morphological structure and functional development level, which is particularly important in bone and joint diseases and is also an important disease indication. During RA medical record annotation, features such as "bone friction" ("骨摩擦感"), "floating patellar test" ("浮髌试验"), "4 sign test" ("4字征试验") and so on may be noted.

(VII) Other examinations carried out as necessary may refer to "bone marrow biopsy" ("骨髓活检") and "arthroscopy" ("关节镜检查"), etc.

(VIII) Treatment entities are labeled with terms including "intraarticular injection of sodium hyaluronate" ("关节腔内注射玻璃酸钠"), "artificial joint replacement" ("人工关节置换术") and drugs, such as "methotrexate" ("甲氨蝶呤"), "tripterygium wilfordii" ("雷公藤"), "leflunomide" ("来氟米特").

(IX) Relieving factors refers to mean factors that can temporarily relieve joint symptoms, mostly referring to "rest" ("休息").

(X) Aggravating factors are those that will aggravate the joint symptoms of RA patients, including environmental factors such as "cloudy and rainy days" ("阴雨天") and self-activity factors such as "going up and down" ("上下楼梯").

In order to prove the effectiveness of using BERT to generate word vectors in the named entity extraction of RA medical record data, the traditional BiLSTM-CRF model and the BiLSTM-CRF model with added attention mechanism were taken as the contrast and compared with the BERT-BiLSTM-CRF model. The experimental results are shown in *Table 4*.

As seen in *Table 4*, applying the self-attention mechanism of the BERT model to pre-process word vectors is greatly improved compared with the shallow language model, and adding an attention layer to the model is also effective. Therefore, the advantage of BERT's mask strategy in the learning context is fully implemented in the generation of

© Quantitative Imaging in Medicine and Surgery. All rights reserved.

Table 4 Model comparison results of a corpus test for rheumatoid arthritis

Model	Precision	Recall	F1-score
BiLSTM-CRF	0.896	0.865	0.880
BiLSTM-ATT-CRF	0.929	0.934	0.931
BERT-BiLSTM-CRF	0.922	0.951	0.936

Table 5 Entities training results

Number	Named entity	Precision	Recall	F1-score
1	Disease	0.905	0.824	0.863
2	Symptom	0.906	0.939	0.922
3	Body part	0.975	0.992	0.983
4	Laboratory Examinations	0.892	0.962	0.925
5	Physical examination	0.822	0.918	0.867
6	Treatment	0.887	0.921	0.904

word vectors. The whole word mask strategy adopted in this study enables the model to predict all masked words in many training samples and has a good recognition effect for various entities, as shown in *Table 5*. For the entity types with poor learning effects, the reason may be that the number of labels in the medical records is too small, or that the labeling standards are not uniform enough, or the labeling results are nested, etc., which will affect the effect of recognition.

Discussion

This research has achieved a model construction work in the field of named entity recognition in clinical notes of Chinese electronic medical records. We used the traditional BiLSTM-CRF model as the basic model structure, and found that compared with the shallow word vector generation model, the performance of the word vector model fine-tuned using BERT has been significantly improved. The F1-score of our model in the test data set reached 0.936. The superiority of BERT for word vector generation has been proven. The model in this research is suitable for the clinical notes data of rheumatoid arthritis, we have annotated a large number of medical entities in 1,600 clinical notes under the supervision of clinicians and used these annotation information to construct our model. The marked medical entity types include body part, symptoms, laboratory examinations, imaging examinations and medications that are of concern to clinicians who treat rheumatoid arthritis. For entities with characteristics of diagnosis and treatment of rheumatoid arthritis disease, such as imaging examination and medications, the annotation results have a certain contribution to the display of disease characteristics. Future research can use the research results of this work for tasks such as relationship extraction research, medical knowledge graph construction research, and related disease reasoning research, which has scientific research significance and clinical practice significance.

Acknowledgments

Funding: This research was supported by the National Key Research and Development Program of China, grant ID: 2020AAA0104905, the Major National Social Science Project, grant ID: 19ZDA041, and the National Natural Science Foundation of China, grant ID: 82071930.

Footnote

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://dx.doi.org/10.21037/qims-21-90). All authors reported that this research was supported by the National Key Research and Development Program of China, under Grant 2020AAA0104905, the Major National Social Science Project, under Grant 19ZDA041, and the National Natural

© Quantitative Imaging in Medicine and Surgery. All rights reserved.

Science Foundation of China, under Grant 82071930.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Open Access Statement: This is an Open Access article distributed in accordance with the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License (CC BY-NC-ND 4.0), which permits the non-commercial replication and distribution of the article with the strict proviso that no changes or edits are made and the original work is properly cited (including links to both the formal publication through the relevant DOI and the license). See: https://creativecommons.org/licenses/by-nc-nd/4.0/.

References

1. Wang J, Fang L, Zhang C, Liu X, Cui L, Nie T, Li R. Comparative analysis of the clinical features of rheumatoid arthritis complicated with coronary atherosclerotic herart disease between young and elderly patients. Chinese Medicine 2020;15:1505-8.
2. Lv N, Zhang Z, Zhu J. Influence of systemic nursing intervention on satisfaction and return visit rate of patients with rheumatoid arthritis. Clinical Research 2019;27:175-7.
3. Ng D, Du H, Yao MM, Kosik RO, Chan WP, Feng M. Today's radiologists meet tomorrow's AI: the promises, pitfalls, and unbridled potential. Quant Imaging Med Surg 2021;11:2775-9.
4. Yang J, Yu Q, Guan Y, Jiang Z. An Overview of Research on Electronic Medical Record Oriented named entity recognition and entity relation extraction. Acta Automatica Sinica 2014;40:1537-62.
5. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Paper present at: Proceedings of NAACL-HLT 2019. June 2-7, 2019; Minneapolis, Minnesota. doi: 10.18653/v1/N19-1423.
6. Zhao Y, Zhang Z, Liu H, Ding L. Classification of Chinese Medical Literature with BERT Model. Data Analysis and Knowledge Discovery 2020;4:41-9.
7. Xue K, Zhou Y, Ma Z, Ruan T, Zhang H, He P. Fine-tuning BERT for Joint Entity and Relation Extraction in Chinese Medical Text. Paper presented at: IEEE International Conference on Bioinformatics and Biomedicine (BIBM). NOV 18-21, 2019. San Diego, CA.
8. Lopes F, Teixeira C, Gonçalo Oliveira H. Comparing Different Methods for Named Entity Recognition in Portuguese Neurology Text. J Med Syst 2020;44:77.
9. Xu K, Yang Z, Kang P, Wang Q, Liu W. Document-level attention-based BiLSTM-CRF incorporating disease dictionary for disease named entity recognition. Comput Biol Med 2019;108:122-32.
10. Xia Y, Zheng J, Zhao Y, Xu X. Deeplearning Based Named Entity Recognition of Electronic Medical Record. Electronic Sci Tech 2018; 31:31-34+37.
11. Cheng M, Li LM, Ren Y, Lou YX, Gao JB. A Hybrid Method to Extract Clinical Information from Chinese Electronic Medical Records. IEEE Access 2019;7:70624-33.
12. Yin M, Mou C, Xiong K, Ren J. Chinese clinical named entity recognition with radical-level feature and self-attention mechanism. J Biomed Inform 2019;98:103289.
13. Li X, Zhang H, Zhou XH. Chinese clinical named entity recognition with variant neural structures based on BERT methods. J Biomed Inform 2020;107:103422.
14. Xie T, Yang J, Liu H. Chinese Entity Recognition Based on BERT-BiLSTM-CRF Model. Computer Systems & Applications 2020;(7):48-55.
15. Zhang X, Zhang Y, Zhang Q, Ren Y, Qiu T, Ma J, Sun Q. Extracting comprehensive clinical information for breast cancer using deep learning methods. Int J Med Inform 2019;132:103985.
16. Zhang M, Wang J, Zhang X. Using a Pre-Trained Language Model for Medical Named Entity Extraction in Chinese Clinic Text. Paper presented at: 10th IEEE International Conference on Electronics Information and Emergency Communication. July 17-19, 2020; Beijing, China. doi: 10.1109/ICEIEC49280.2020.9152257.
17. Li L, Yang J, Li B, Du Y, Hu W. Named Entity Recognition of Chinese Electronic Medical Record Based on BERT. Journal of Inner Mongolia University of Science and Technology 2020;39:75-81.
18. Wang Y, Sun Y, Ma Z, Gao L, Xu Y. Named Entity Recognition in Chinese Medical Literature Using Pretraining Models. Sci Program 2020;2020:8812754.
19. Yan H, Chen X, Wang W, Wang H, Yin M. Recognition model for French named entities based on deep neural network. Journal of Computer Applications 2019;3:1288-92.
20. Palangi H, Li D, Shen Y, Gao J, He X, Chen J, Song X, Ward R. Deep sentence embedding using long short-term memory networks: analysis and application to information retrieval. IEEE/ACM Transactions on Audio, Speech, and

© Quantitative Imaging in Medicine and Surgery. All rights reserved.

Language Processing 2016;24:694-707.
21. Jia X, Song W, Li W, Wang Q, Lei Y, Chen Z, Chang Z. Research and Implementation of the Construction of Slow Obstructive Pulmonary Knowledge Map. Journal of Chinese Computer Systems 2020;41:1371-4.
22. Yao YS, Huang Z. Bi-directional LSTM Recurrent Neural Network for Chinese Word Segmentation. Paper presented at: 23rd International Conference on Neural Information Processing (ICONIP). Oct 16-21, 2016; Kyoto, JAPAN. Accessed on 2016. doi: 10.1007/978-3-319-46681-1_42.
23. Zhang H, Kang X, Li B, Wang Y, Liu H, Bai F. Medical name entity recognition based on Bi-LSTM-CRF and attention mechanism. Journal of Computer Applications 2020;40:98-102.
24. Wang Q. Research on Text Classification Based on Attention Bi-LSTM.[dis-sertation]. Guangzhou: South China University of Technology; 2018.
25. Bikel DM, Miller S, Schwartz R, Weischedel R. Nymble: A High-performance Learning Name-finder. Paper presented at: Proceedings of the fifth conference on Applied natural language processing. March 31 - April 3 1997; Washington, DC, USA. Accessed on March 31 1997. doi: 10.3115/974557.974586.
26. Lafferty J, Mc Callum A, Pereira F. Conditional random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Paper presented at: Proc. 18th International Conf. on Machine Learning. June 2001; San Francisco. Accessed in 2001.
27. Ye F, Chen YY, Zhou GG, Li H, Li Y. Intelligent Recognition of Named-entity in Electronic Medical Records. Chinese Journal of Biomedical Engineering 2011;30:256-62.
28. Zhang Z, Liu Y. Chinese Word Segmentation Based on Bi-directional LSTM-CRF Model. Journal of Changchun University of Science and Technology (Natural Science Edition) 2017;40:87-92.

Cite this article as: Li M, Liu F, Zhu J, Zhang R, Qin Y, Gao D. Model-based clinical note entity recognition for rheumatoid arthritis using bidirectional encoder representation from transformers. Quant Imaging Med Surg 2022;12(1):184-195. doi: 10.21037/qims-21-90

© Quantitative Imaging in Medicine and Surgery. All rights reserved.