Named Entity Recognition Model of Chinese Clinical Electronic Medical Record Based on XLNet-BiLSTM

Shen Zhou Feng (571674396@qq.com)
Shanghai University of Engineering Science
Su Qian Min
Shanghai University of Engineering Science
Guo Jing Lei
Shanghai University of Traditional Chinese Medicine

Research Article

Keywords: electronic medical record, named entity recognition, XLNet, Bidirectional Long Short-Term Memory(Bi-LSTM) network, Multi-headed attention, Conditional random field(CRF)

Posted Date: March 25th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-218833/v1

Named Entity Recognition Model of Chinese Clinical Electronic Medical Record Based on XLNet-BiLSTM

Shen Zhoufeng,Master candidate, School of Electronic and Electrical Engineering, Shanghai University of Engineering Science. E-mail: 571674396@qq.com;
Su Qianmin,PhD, associate professor, School of Electronic and Electrical Engineering, Shanghai University of Engineering Science.
Guo Jinglei,Ph.D., associate professor, Shanghai University of Traditional Chinese Medicine.

Abstract: The recognition of named entities in Chinese clinical electronic medical records is one of the basic tasks to realize smart medical care. Aiming at the insufficient text semantic representation of the traditional word vector model and the inability of the recurrent neural network (RNN) model to solve the problems of long-term dependence, a Chinese clinical electronic medical record named entity recognition model XLNet-BiLSTM-MHA-CRF based on XLNet is proposed. Use the XLNet pre-training language model as the embedding layer to vectorize the medical record text to solve the problem of ambiguity; use the bidirectional long and short-term memory network (BiLSTM) gate control unit to obtain the forward and backward semantic feature information of the sentence; Then input the feature sequence to the multi-head attention layer (multi-head attention, MHA), use MHA to obtain information represented by different subspaces of the feature sequence, enhance the relevance of context semantics and eliminate noise; finally, input the conditional random field CRF to identify the global maximum 优 sequence. The experimental results show that the XLNet-BiLSTM-Attention-CRF model has achieved good results on the CCKS-2017 named entity recognition data set.
Keywords: electronic medical record; named entity recognition; XLNet; Bidirectional Long Short-Term Memory(Bi-LSTM) network; Multi-headed attention; Conditional random field(CRF)

1. Research background and related work

Electronic Medical Record (EMR) is a Medical Record stored, managed and transmitted by a computer information system, including digital information about the patient's history, clinical manifestations and treatment methods recorded by Medical staff in the process of diagnosis and treatment of patients [1]. Because most electronic medical records are semi-structured and unstructured, their analysis and data mining are severely restricted. Named Entity Recognition (NER) is an important branch of Natural Language Processing (NLP) task, which is to discover and recognize proper nouns and meaningful words in natural texts and classify them into predefined categories [2]. The purpose of this paper is to identify and classify medical named entities in electronic medical records automatically by using named entity recognition technology.

The traditional research on named entity recognition of electronic medical records is mainly divided into rule-based and machine-learning-based methods. The rule-based method mainly relies on domain dictionaries constructed by domain experts for recognition, and for entities that do not appear in the dictionary, medical named entities can be identified by manually edited rules [3]. Due to the dependence of dictionary construction and rule making on domain experts, the method of named entity recognition based on machine learning in electronic medical records has been widely used. In recent years, deep learning has made significant progress in many fields such as speech

recognition, image recognition and video analysis [4]. A large number of researchers have applied deep learning to the field of entity recognition of electronic medical records. Through training and learning in large-scale annotated data, contextual semantic features can be better extracted for representation.

Named entity recognition method based on the depth of neural network, all need through the word embedded method to convert text to serialize vector, the current relatively popular in the word embedding method is founded in 2013 by Mikolov Word2Vec [5], etc, will be one of the traditional word - hot said into a low latitude, dense vector, each word by dozens or hundreds of dimension of real-valued vector said. However, the word vector trained by Word2vec is static, that is, the vector representation of the same word in different statements is unchanged, so it is impossible to obtain multiple meanings of the same word, and it is impossible to remove the ambiguity of word meaning with context in the training process [6]. There is often polysemy phenomenon in electronic medical records. For example, the word "disease" has different meanings in different words, which can be either a noun disease or an adjective intense. In recent years, many context-related word embedding methods, such as ELMO (Embeddings from Language Models) and OpenAI-GPT (Generative Pre-Training), have been proposed in academic circles to address these problems [7]. However, the language representation of the above two language models is unidirectional, so it is impossible to obtain the semantic information of the text of the electronic medical records in the two directions at the same time.

In order to solve the above problems, this study intends to introduce the bidirectional autoregressive pretraining language model XLNET into the NER task of electronic medical records, and proposes the XLNET - BILSTM-MHA-CRF named entity recognition model. And using the model of medical electronic medical records in the predefined diseases, symptoms, treatment, examination, 5 kinds of parts of the body entity named entity recognition, experiments show that using the training language model to build word embedded, and join in the BiLSTM - CRF long attention mechanism, multi-angle extracting text feature, effectively improve the named entity recognition effect. The algorithm described in this paper achieved F1 value of 91.74% in the named entity recognition task of CCKS2017.

2. XLNET-BILSTM-ATTENTION -CRF Named Entity Recognition Model

The overall structure of XLNET-BILSTM-MHA-CRF named entity recognition model is shown in Figure 1. The first layer of the model is the XLNET word embedding layer. Through the XLNET pre-training language model, the word vector of low dimension is used to represent every word in the medical record to get the serialized text input. The second layer is the BILSTM layer, which automatically extracts the forward and backward features of sentences by using the bidirectional short and long time memory neural network and then splits them into the next layer. The third layer is the MHA layer, which obtains the long distance dependence feature of the sentence by calculating the attention probability from multiple angles, and obtains the new feature vector. The fourth layer is the CRF layer. The input text features are sequentially labeled through calculation and the optimal label is output. The model is explained in detail below.

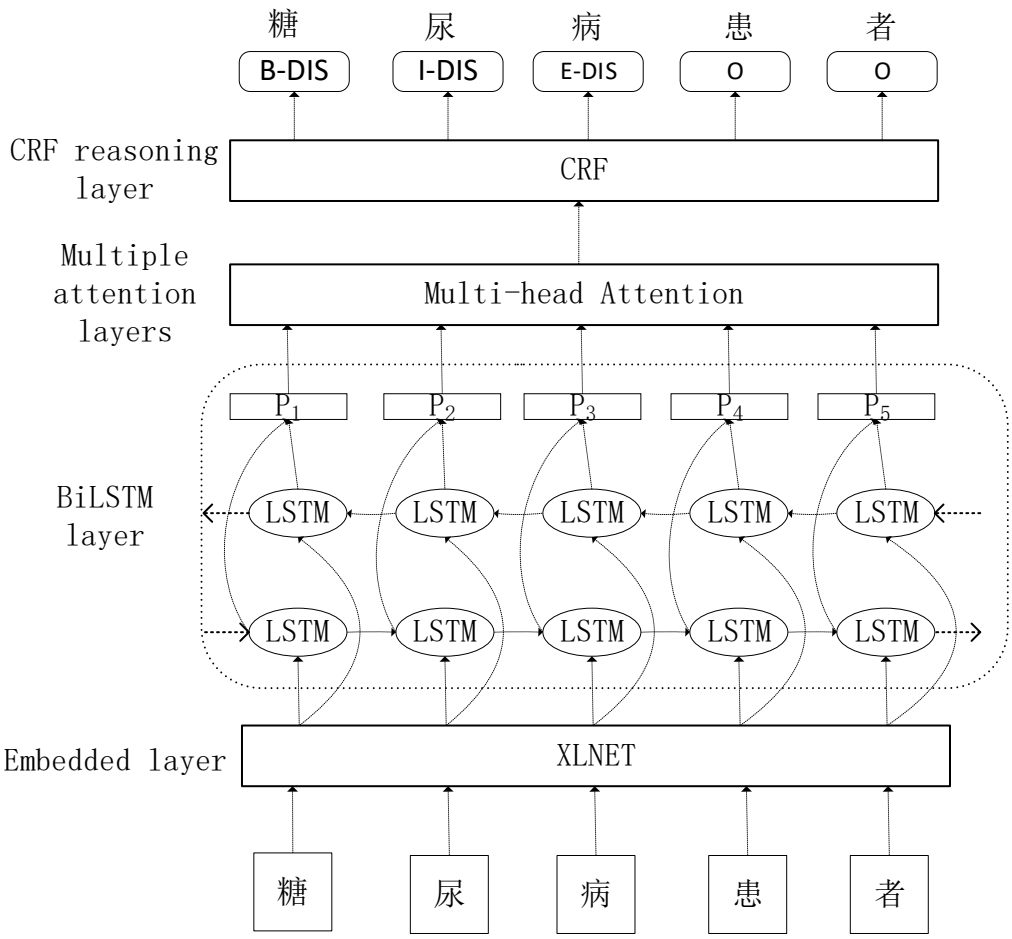

Fig.1 The structure diagram of ALBERT-BIGRU-MHA-CRF named entity recognition model

2.1 XLNET Pre-Training Language Model

XLNET model is a generalized autoregressive pre-training method proposed by CMU and Google team in 2019 based on the advantages and disadvantages of BERT, which realizes two-way prediction on the traditional autoregressive language model [8]. By using the Attention Mask method in Transformer module to obtain different arrangement and combination of input text, the model can fully extract context information for training, and overcome the effective information loss in the Bert model under the mask mechanism. XLNet mask mechanism of concrete as shown in figure 2, when the model input sentences for [sugar, urine, disease, cancer,], a group of randomly generated sequence for [disease, sugar, develop, and urine], so the calculation to the word "sugar" after relining can make use of the word "disease" of information, so in the first row only retained the third location information (expressed in solid), the location of the other information is covering (hollow). Another example is that the word "pee" is in the last position after rearrangement, and the information of the other four characters can be used, that is, the second line is all solid except for the second position.

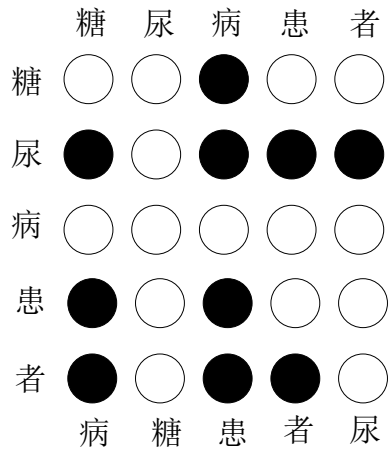

Fig.2 XLNet model mask mechanism example diagram

Most of the existing pre-trained language models use Transformer architecture [9], but there are still some deficiencies in capturing long-distance dependencies. To solve this problem, XLNET uses Transformer XL architecture with the introduction of loop mechanism (RNN) and relative location encoding. RNN is used to extract the long distance implicit dependency information of the previous segment, and then the memory units stored between the segments are used for the prediction of the next segment to fully capture the long distance text features. Figure 3 shows how information is passed between fragments. The dotted box represents the memory information extracted from the previous segment, which is transferred to the next segment through the memory unit to realize the information transfer.

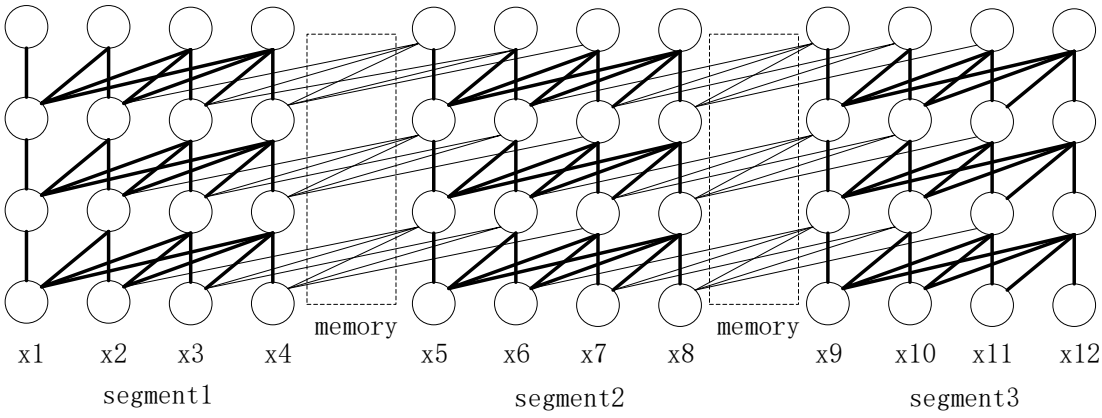

Fig.3 XLNet cycle mechanism fragment information transfer diagram

In terms of position coding, relative position coding is used to replace absolute position coding to solve the problem of word ambiguity and enhance the integrity of text feature extraction. The self-attention formula after adding relative position coding is as follows:

$$A_{i,j}^{rel} = E_{x_i}^T W_q^T W_{k,E} E_{xj} + E_{x_i}^T W_q^T W_{k,R} R_{i-j} + u^T W_{k,E} E_{xj} + v^T W_{k,R} R_{i-j} \quad (1)$$

Where, E_{xi}andE_{xj}respectively represent the text vectors of i and j, W represents the weight matrix, R_{i-j}represents the relative positions of i and j, $u^T$and $v^T$are the parameters to be learned, $W_{k,E}$ and $W_{k,R}$ respectively learn key vectors based on content and key vectors based on position.

The XLNET pre-training language model based on Transformer XL overcomes the insufficiency of single-item information transmission in the autoregressive language model through attention mask, circulation mechanism and encoding of relative position, and makes full use of semantic information of context to extract potential internal relations and train word vector representation with more

complete features.

2.2 Bidirectional short and long time memory network (BILSTM) model

For the cycle of traditional neural network (RNN) in addressing the problem of sequence annotation, the phenomenon of the gradient disappeared and gradient explosion [10], Hochreiter and Schmidhuber length is put forward in 1997 when the memory network (long short term memory, LSTM), the network is improved on the basis of RNN, its unit structure as shown in figure 4, by setting the left door, input and output door three threshold mechanism selectively forgotten and transfer of processing information, in order to capture the text sequence dependency information for long distances, It effectively solves the problem of gradient disappearance [11].

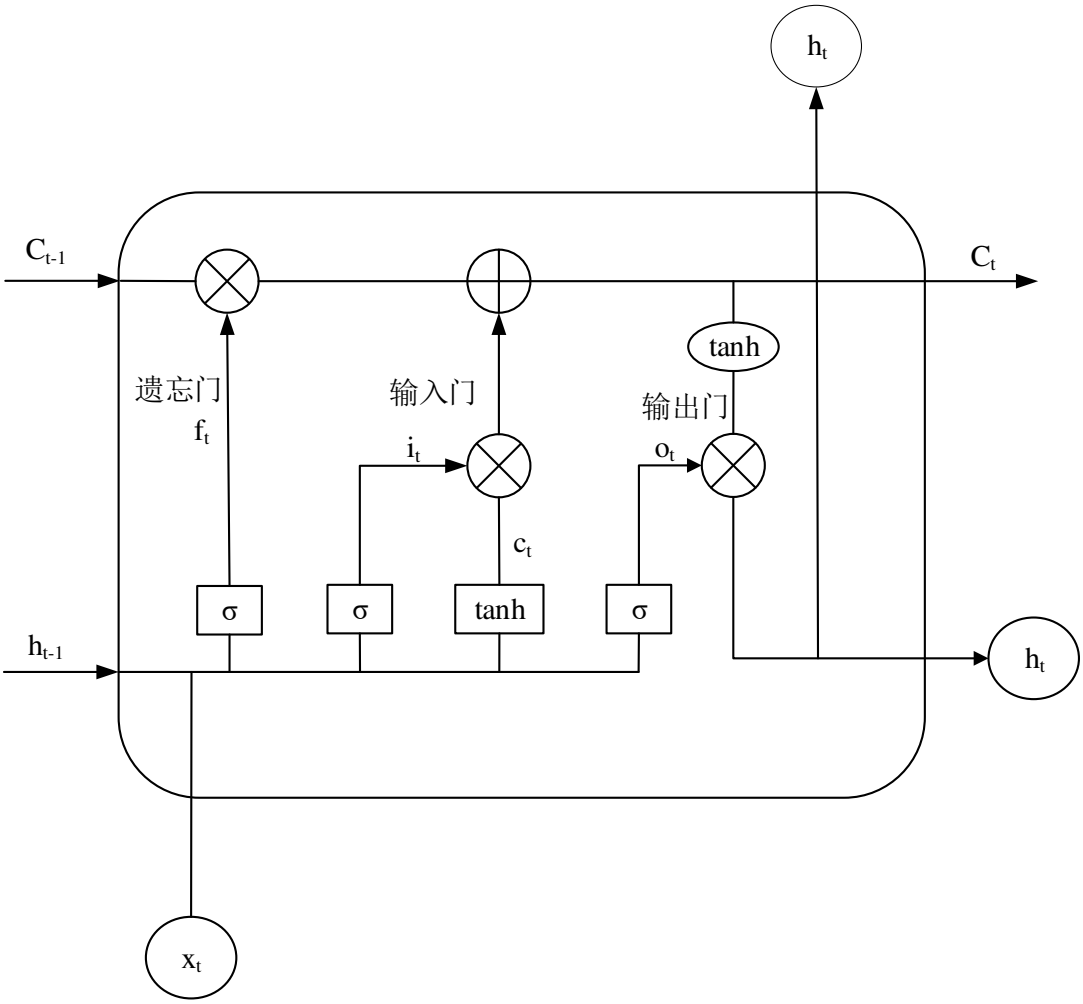

Fig.4 LSTM unit structure diagram

The calculation process of the hidden layer of a unit in LSTM network is as follows:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (2)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_f) \quad (3)$$

$$\tilde{C}_t = tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (4)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (5)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (6)$$

$$h_t = o_t * tanh(C_t) \quad (7)$$

f_t， i_t、i_t and C_t respectively t moment forgotten door, enter the door, output and memory cells, σ is sigmoid activation function that is hyperbolic tangent activation function tanh, W, b respectively connected to the two layers of weight matrices and bias vectors, x_t as the input vector, h_{t-1} for time t - 1 output, output of h_t is t time, $\tilde{C}_t$ said intermediate state.

Because LSTM can only deal with the current unit information before and after don't have access to information, so the length of the bidirectional memory network, is to use two layers of LSTM, respectively for text sequence forward information and to the information and then joining together after get the final characteristics of the hidden layer, said fully capture the context semantic information to improve the effect of named entity recognition.

2.3 Multi-head attention (MHA) model

In 2017, the Google machine translation team creatively proposed the multi-head attention model [12] by combining multiple self-attention. The specific structural model is shown in Figure 5. Input the text sequence $X = (X1，X2 \ldots，Xn)$ into the BiLSTM layer and output matrix $Y = (Y1，Y2, \ldots Yn)$ is the input of Q, K and V. There are H layers in Scaled Dot-Product Attention unit, and the attention calculation of each layer is shown in Equation (8). Then combine h single-headed attention outputs and make a linear transformation as shown in Equation (9), and the obtained MHA is the h-headed attention weight output of the t word. In NER task, the multi-head attention model can fully capture the long-term temporal dependence of sentences and obtain the global features.

$$head_i = soft\,max\left(\frac{QK^T}{\sqrt{d_k}}\right)V = soft\,max\left(\frac{(Y_tW_i^Q)(Y_tW_i^K)^T}{\sqrt{d_k}}\right)\left(Y_tW_i^V\right) \quad (8)$$

$$MHA_t = concat(head_1, head_2 \cdots head_h)W^O \quad (9)$$

Where, W_i^Q、$W_i^K$and W_i^V are the weight of parameters to be trained. $\sqrt{d_k}$ is the smooth term of the k dimension. ,softmax () as the normalized function. $concat()$ is a concatenation function. Wo is the weight parameter.

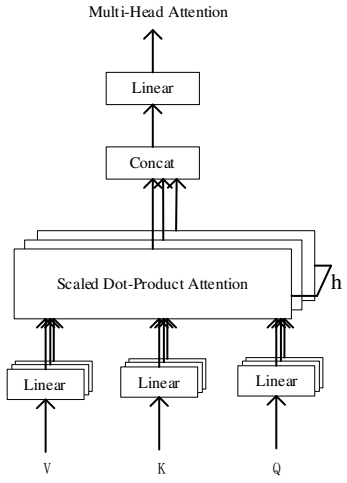

Fig.5 The model of Multi-head attention

2.4 Conditional random field (CRF) model

In 2001, Lafferty proposed a linear conditional random field (CRF) model to calculate a given sequence of random variables X=(X1, X2... , Xn), the random variable sequence Y=(Y1, Y2... Yn) the conditional probability distribution of P (Y | X) [13]. The model assumes that the sequence of random variables satisfies Markov property:

$$P(Y_i|X, Y_i, \dots, Y_N) = P(Y_i|X, Y_{i-1}, Y_{i+1}) \quad (10)$$

Where, X represents the input observation sequence; Y represents the corresponding sequence of states. In the named entity recognition task of electronic medical record, there is a constraint relationship between the label of each word and its adjacent label. For example, the O tag will not be followed by the I tag, and the I-dis will not be followed by the B-bod. CRF can obtain the optimal probability of the occurrence of the tag sequence corresponding to each word based on the output results of the previous layer network and the contextual semantic tag information.

Assume that the output sequence of the MHA model is X and one of the predicted sequence is Y, then the evaluation score S(X,Y) can be obtained:

$$S(X,Y) = \sum_{i=0}^{n} M_{y_i y_{i+1}} + \sum_{i}^{n} P_{i,\ y_i} \quad (11)$$

Where, $M_{yi,yi+1}$represents the transition probability from y_i tag to y_{i+1} tag; $P_{i,yi}$ represents the probability that the ith word is marked as y_i; N is the sequence length. Finally the maximum likelihood method is adopted to solve the maximum a posteriori probability $P(y|x)$, get loss function values of the model.

$$log\ P\ (y \mid x) = S(x,y) - \sum_{i=0}^{n} S(x, y_i) \quad (12)$$

3. Experiment and result analysis

3.1 Test data and labeling strategy

In this experiment, 400 medical annotated data from CCKS-2017 task 2 were selected as the data set, which were divided into training set, test set and prediction set according to the 7:2:1 method. The data set includes a total of 39,539 entities divided into five categories: symptoms, disease, treatment, examination, and body parts, with a total of 7,183 sentences. In this paper, the labeling method of BIOE is adopted, that is, B- represents the beginning of the entity, I- the middle part of the entity, E- represents the end of the entity, and O represents that the word does not belong to the specified entity category. The symbols and quantities of each type of entity are shown in Table 1.

Table 1 Medical entity notation

Number	Entity class	Start tag	Middle tag	End tag	Training set	Test set	Prediction set
1	疾病 Disease	B-Dis	I-Dis	E-Dis	893	255	127
2	症状 Symptom	B-Sym	I-Sym	E-Sym	7100	2028	1014
3	检查 Check	B-Che	I-Che	E-Che	8884	2537	1268
4	治疗 Cure	B-Cur	I-Cur	E-Cur	1059	303	151
5	身体部位 Body	B-Bod	I-Bod	E-Bod	9618	2748	1374

3.2 Evaluation Indicators

Entity identification and relation extraction experiments usually adopt the following indicators to evaluate the advantages and disadvantages of the model:

$$\text{Accuracy: } P = \frac{T_P}{T_P + F_P} \times 100\% \quad (13)$$

$$\text{Recall rate: } R = \frac{T_P}{T_P + F_N} \times 100\% \quad (14)$$

$$\text{F1 value: } F_1 = \frac{2 \times P \times R}{P + R} \times 100\% \quad (15)$$

Where: T_P represents the number of positive examples in the test set that are correctly predicted to be positive examples; F_P represents the number of positive examples in the test set that are misclassified as negative ones; $F_N$represents the number of negative examples in the test set that are misclassified as positive ones.

3.3 Experimental environment and parameter setting

The named entity recognition model of the experiment in this paper is based on the TensorFlow framework, and the specific experimental environment Settings are shown in Table 2.

Table 2 Experimental environment

Project	Environment
operating system	Windows10
CPU	I7-10750H@2.60GHz
GPU(Memory size)	RTX2060(8G)
Python version	3.6.5
Tensorflow version	1.14.0

The specific experimental parameters are set as follows: the size of the hidden layer of the BILSTM model is 128, and the number of network layers is 1. ReLU is selected as the activation function of the model, and the ratio of Dropout is set to 0.1 in the training stage. The batch size was set as 16, the maximum sequence length was set as 128, the learning rate was set as 1E-5, and the loss rate was set as 0.1. The ADAM optimizer was used for training.

3.4 Analysis of experimental results

In order to verify the performance of XLNET-BILSTM-MHA-CRF model proposed in this paper, it was compared with the following three groups of models: 1) BILSTM-CRF model [14]; 2) Bert-Bilstm-CRF model [15]; 3) XLNET-BILSTM-CRF model.

Table 3 shows the experimental results of different models. By comparing the experimental results of all models in Table 4, it can be seen that the accuracy rate, recall rate and F1 value of XLNET-BILSTM-MHA-CRF model are the highest in the five medical entities of symptom, disease, treatment, examination and body part, which are increased by 3.46%, 1.14% and 2.31% respectively compared with the baseline model of BILSTM.

Table 3 The results of each model experiment were compared

Model	Evaluation index	Entity type					As a whole
		Disease	Symptom	Check	Cure	Body	
BiLSTM-CRF	P	76.52	94.15	93.27	72.56	83.95	88.61
	R	75.84	95.76	91.80	75.77	85.35	90.27
	F1	76.18	94.95	92.53	74.13	84.64	89.43
	P	78.16	94.59	94.25	73.51	84.76	89.73

Bert-BiLSTM-CRF	R	76.43	95.89	93.81	76.43	86.71	90.14
	F1	77.29	95.24	94.03	74.94	85.72	89.93
XLNet-BiLSTM-CRF	P	80.73	95.02	94.96	74.98	85.61	91.43
	R	77.84	96.10	94.67	75.32	88.59	90.82
	F1	79.26	95.56	94.81	75.15	87.07	91.12
XLNet-BiLSTM-MHA-CRF	P	81.61	95.47	95.28	75.65	87.48	92.07
	R	78.85	96.41	95.72	76.23	89.79	91.41
	F1	80.21	95.94	95.50	75.94	88.62	91.74

In all models, the F1 value of the three medical entities of symptom, examination and body part is generally high, while the F1 value of the entity identification of disease and treatment is just the opposite. Through analysis, it can be found that the amount of training data of these two categories is obviously too small, leading to serious over-fitting phenomenon in the process of model training. In addition, most of the disease entities and treatment entities are long-word structures, such as "left orbital soft tissue cleft", "open reduction and internal fixation of left distal radius fracture", etc., while the solid structures of symptoms, examinations and body parts are simple and have a large amount of training data, so the model can fully learn the text characteristics of such entities. Therefore, in the later entity recognition of electronic medical records, the accuracy of the model can be improved by adding corpus. At the same time, the structure of the entity of long words can be further studied to dig deeper semantic information. For example, dictionary information can be introduced to increase semantic features and improve the generalization ability of the model.

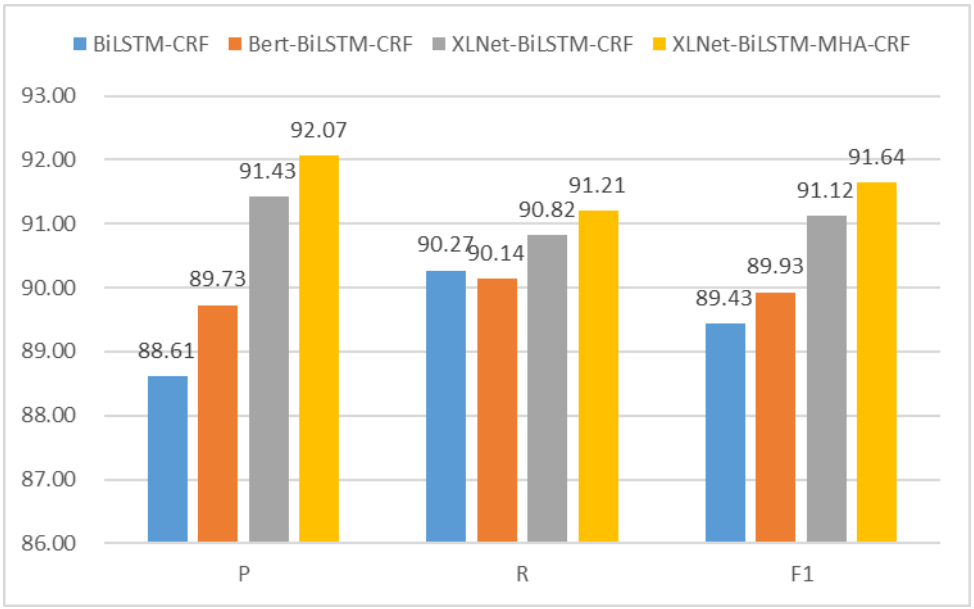

Fig.6 Comparison of experimental results of each model

It can be clearly seen from the results in Figure 6 that the performance of the models based on the pre-training language model XLNET and Bert is better than that of the BILSTM-CRF model, mainly because the latter uses the word vectors obtained by traditional Word2vec and cannot solve the problem of polysemy and the same word. At the same time, it is proved that the dynamic word vector constructed by pre-trained language model can improve the expression ability of text's intrinsic semantic information. By comparison, it is found that the performance of XLNET BILSTM-CRF model is 0.5%~2% higher than that of Bert based model, mainly because XLNET

makes up for the deficiency of Bert through Attention Mask and Transform-XL module, which leads to the improvement of recognition effect. Compared with XLNET-BILSTM-CRF model, the model proposed in this paper has a small improvement in accuracy rate, recall rate and F1, indicating that the addition of multi-head attention mechanism can make the text information representation more complete.

4. Summary and outlook

In this paper, XLNET-BILSTM-MHA-CRF named entity recognition model for medical electronic medical records is proposed. The dynamic word vector trained by the pre-trained language model in the large-scale corpus is used to replace the traditional static word vector to serialize the electronic medical records, which effectively solves the problem of polysemism and makes the semantic representation of the context more accurate. The generalized autoregressive prediction model XLNET can effectively make up for the deficiency of BERT model. The addition of MHA mechanism can capture long-distance dependency characteristics in electronic medical record text. Experimental results in task two data set of CCKS2017 show that the F1 value of XLNET-BILSTM-MHA-CRF model is 91.64%. Compared with other models, the model achieves better recognition effect and can better complete the named entity recognition task of medical electronic medical records. It has certain reference value for the research of entity recognition in medical field. Since there are only 400 pieces of electronic medical record data in this experiment, there are few types of entities and the number of entities is unbalanced. Therefore, more electronic medical record data should be obtained later to enrich the recognition types of the model and prepare for mining the medical information hidden in Chinese electronic medical records.

Compliance with Ethical Standards statements

1.Funding:This study was funded by the 2017 "Science and Technology Innovation Action Plan" of Shanghai

2.Conflict of Interest:The authors declare that they have no conflict of interest

3. Informed Consent: The participants agreed to contribute to the paper

Authorship contributions

Shen Zhoufeng was responsible for testing the model and writing the paper.

Su Qianmin and Guo Jinglei supervised the writing of the thesis.

References:

[1] YANG Jinfeng, YU Qiubin, GUAN Yi, et al. A review of named entity recognition and entity relationship extraction in electronic medical records [J]. Acta Automatica Sinica,2014,40(08):1537-1562.

[2] Xu Jing. Research on the Key Techniques of Open Text Information Extraction for Chinese Knowledge Graph [D]. Changsha: National University of Defense Technology,2018.

[3] ZHOU Kun. Research on Named Entity Recognition Based on Rules [D]. Hefei: Hefei University of Technology,2010.

[4] FU Wenbo, SUN Tao, LIANG Ji, et al. A review of the principles and applications of deep learning. Computer Science,2018,45(6A):11-15,40.

[5] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and

theirCompositionality [J]. Advances in Neural Information Processing Systems,2013, 26:3111-3119.

[6] Li Ying. Improvement and Application of Garbage Barrage Recognition Based on BERT-DPCNN [D]. Shanghai: Shanghai Normal University,2020.

[7] YU Tongrui, JIN Ran, HAN Xiaozhen, et al. A review of pre-training models for natural language processing [J/OL]. Computer Engineering and Applications: 1-11 [2020-11-12].

[8] Li Zhoujun, Fan Yu, Wu Xianjie. A review of pre-training techniques for natural language processing [J]. Computer Science,2020,47(03):162-173.

[9] Guo Xiaoran, Luo Ping, Wang Weilan. Chinese Named Entity Recognition Based on Transformer Encoder [J/OL]. Journal of jilin university (engineering science) : 1-8 [2021-02-06]. HTTP: / / https://doi.org/10.13229/j.cnki.jdxbgxb20200640.

[10] Yang Li, Wu Yuqian, Wang Junli, Liu Yili. A review of cyclic neural networks [J]. Journal of Computer Applications,2018,38(S2):1-6+26.

[11] Chu Deping, Wan Bo, Li Hong, Fang Fang, Wang Run. Geological entity identification based on ELMO-CNN-BILSTM-CRF model [J/OL]. Earth science: 1-22 [2021-02-06]. http://kns.cnki.net/kcms/detail/42.1874.P.20201109.1600.008.html.

[12] E Haihong, ZHANG Wenjing, XIAO Siqi, et al. A review of deep learning entity relationship extraction [J]. Journal of Software, 2019, 30 (0 6) :1793-1818.

[13] YING Yulong, LI Miao, WU Dabala, et al. Mongolian part of speech tagging based on conditional random field [J]. Computer Applications,2010,30(08):2038-2040.

[14] Wang Lijun, Zhou Yue, Gui Jie, Zhai Yun. Research on Chinese Traditional Chinese Literature Word Segmentation Model Based on BILSTM-CRF [J]. Computer Application Research,2020,37(11):3359-3362+3367.

[15] Xie Teng, Yang Junan, Liu Hui. Entity Recognition Based on Bert-Bilstm-CRF Model [J]. Computer Systems and Applications,2020,29(07):48-55.

Figures

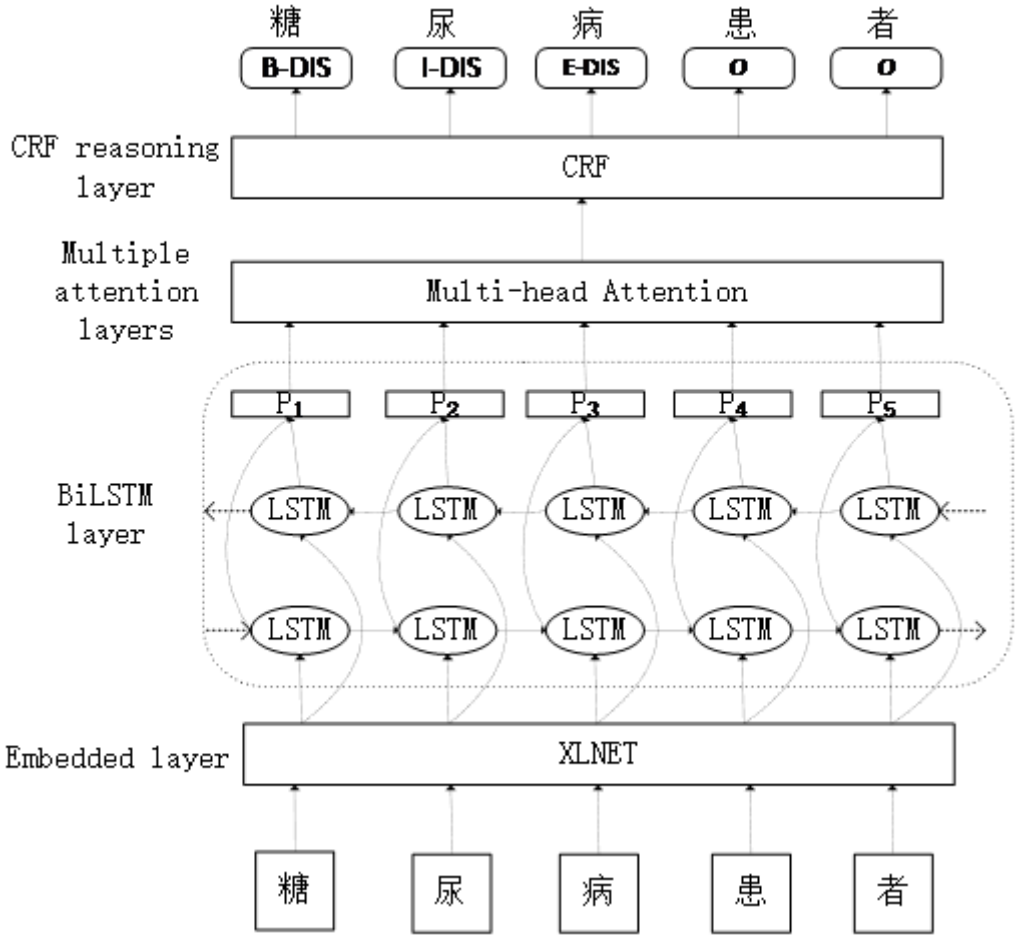

Figure 1

The structure diagram of ALBERT-BIGRU-MHA-CRF named entity recognition model

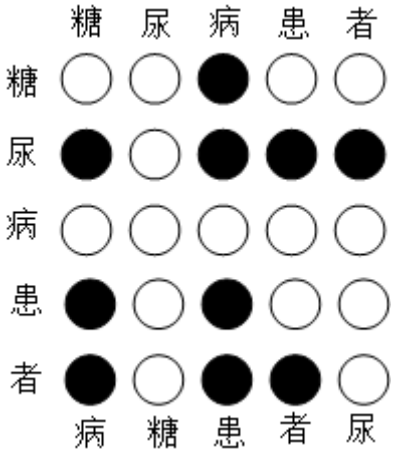

Figure 2

XLNet model mask mechanism example diagram

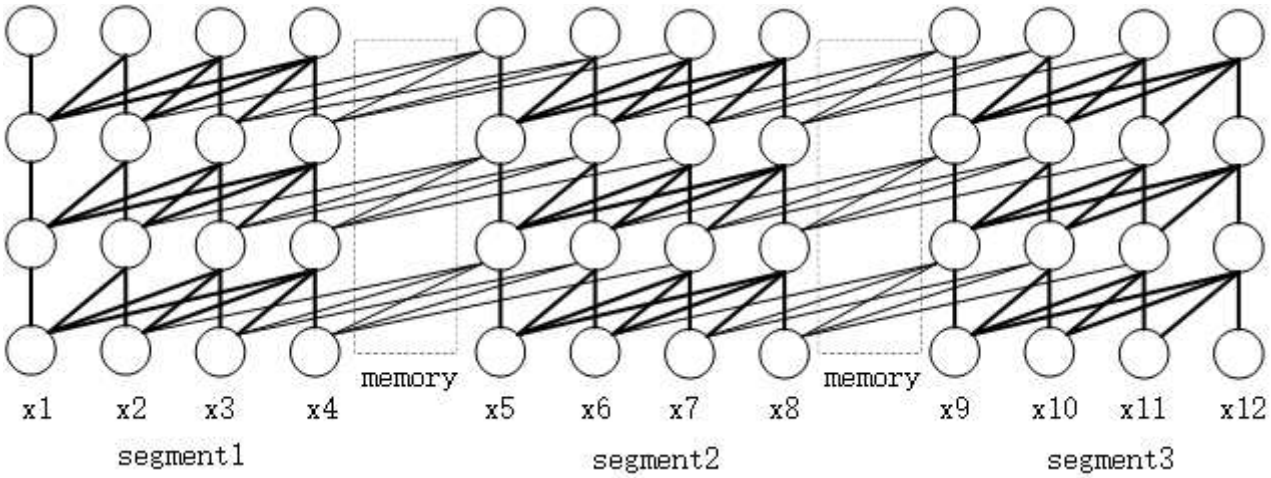

Figure 3

XLNet cycle mechanism fragment information transfer diagram

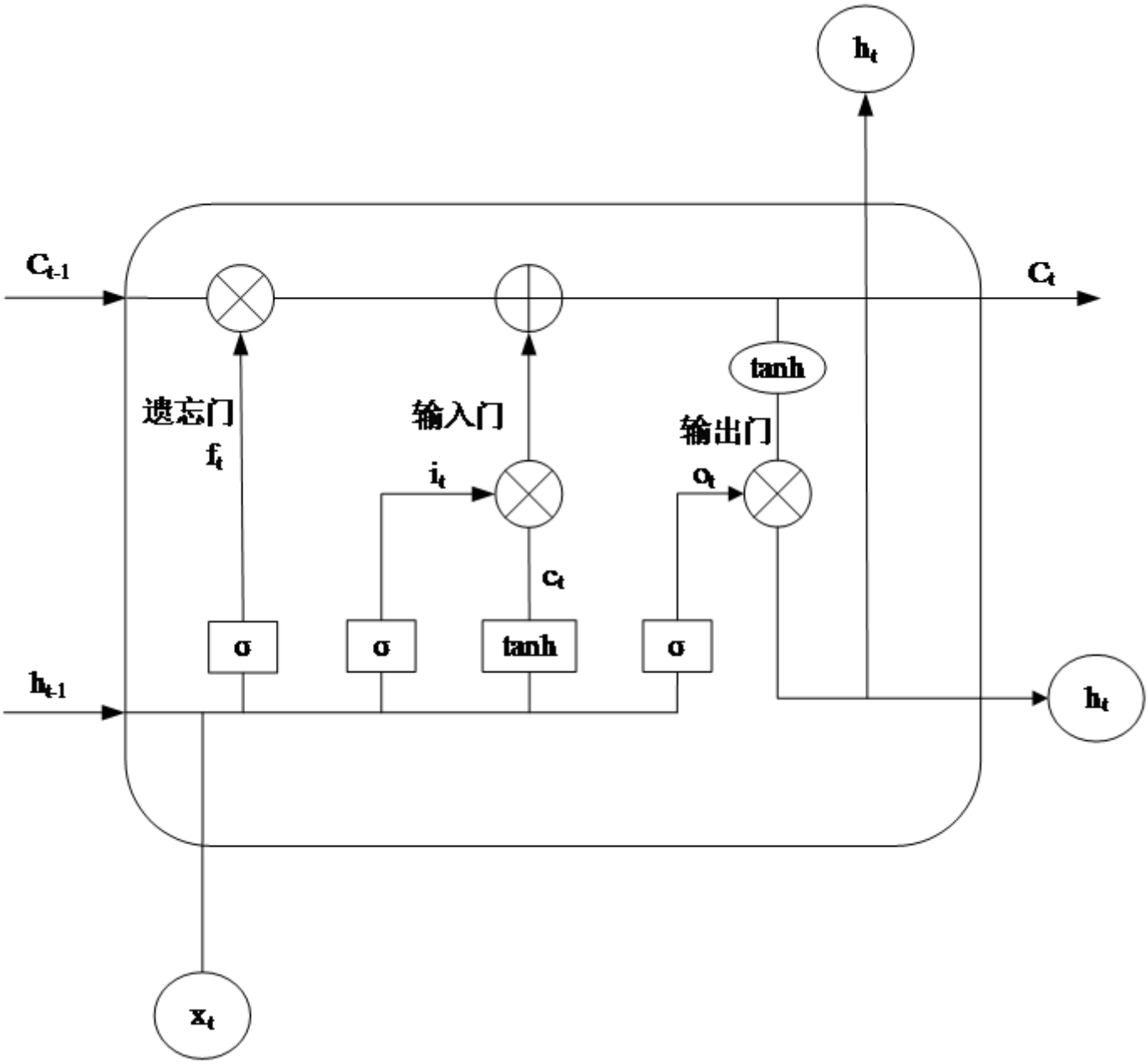

Figure 4

LSTM unit structure diagram

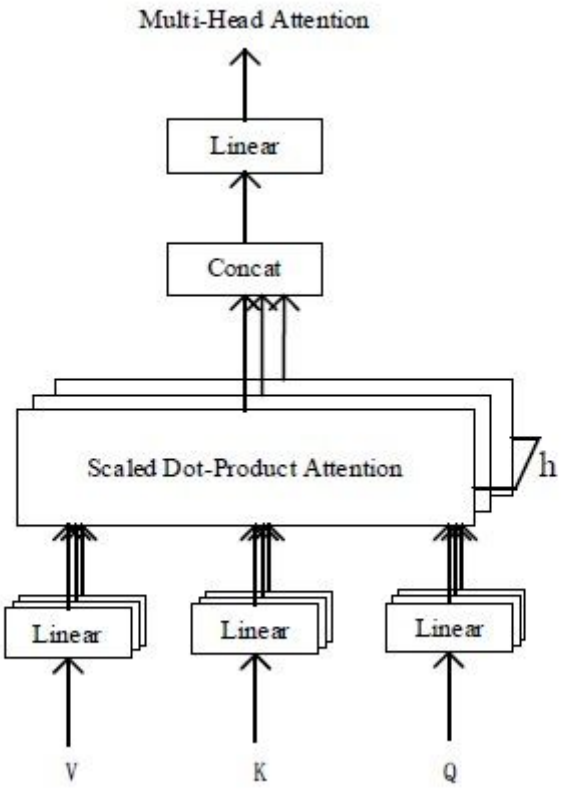

Figure 5

The model of Multi-head attention

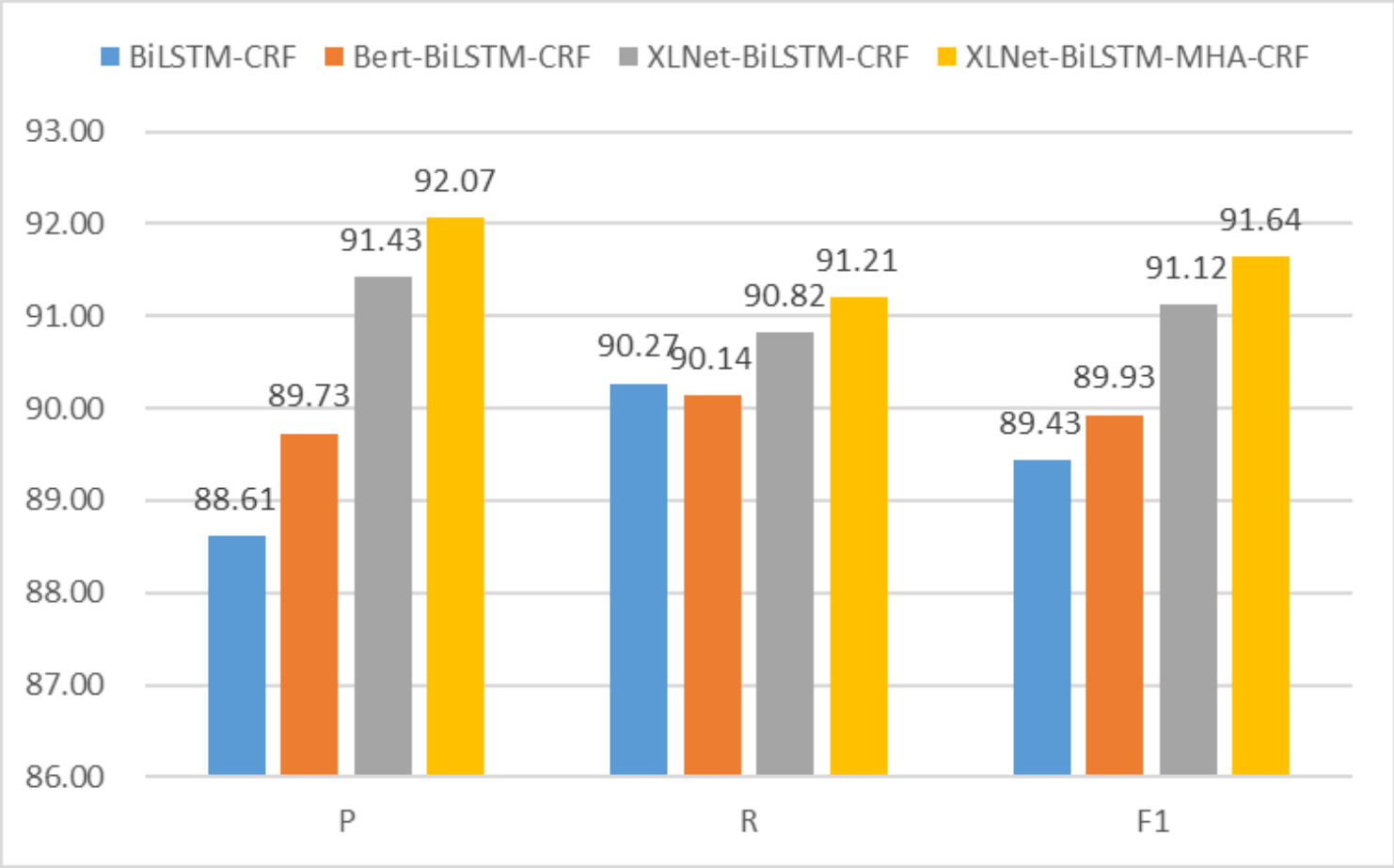

Figure 6

Comparison of experimental results of each model