Named Entity Recognition in Electronic Medical Records Incorporating Pre-trained and Multi-Head Attention

Haotian Yang, Li Wang, Yanpeng Yang

***Abstract*—Chinese Named Entity Recognition (NER) for Electronic Medical Records (EMR) is a fundamental task in building a digital hospital and is widely considered to be a sequence annotation problem in the Natural Language Processing domain. (NLP). However, existing deep learning sequence annotation models cannot fully use the large amount of unannotated data for Chinese EMRs that contain a vast number of professional unregistered words, named entities, and inter-of-entity relationships carrying rich professional knowledge. Moreover, the syntactic structure of EMR sentences is complex, and the text is long; the features of the EMR documents often cannot be captured deeply. Aiming at these two problems, this paper proposes a deep learning method that combines Multi-Head Attention with a pre-trained language model. (BERT-BiGRU-Att-CRF). The method uses the BERT pre-trained model to obtain dynamic word vectors combined with contextual information, extracts global semantic features through a Bi-directional Gated Recurrent Unit (BiGRU), obtains augmented semantic features by using the Multi-Head Attention. Finally, using Conditional Random Field (CRF) decoding, outputs the globally optimal label sequence with greatest probability. The model is trained using the CCKS2019 Chinese EMR dataset containing six types of entities: anatomical sites, surgeries, diseases and diagnoses, medicines, laboratory tests, and imaging tests, and good results are achieved with an F_1 score of 86.97%.**

***Index Terms*—electronic medical records, Multi-Head Attention, pre-trained, fine-tuned.**

I. Introduction

ELECTRONIC Medical Record (EMR) generated by medical staff in the course of medical procedures carried out using a medical institution's information system [1]. Three primary forms of data in an EMR are tables, free text, and images. Among them, unstructured data in free text is a vital component of EMR. In order to better serve subsequent secondary development, such as medical decision trees or similar medical record research, the structured processing of EMR is necessary. Therefore, accurately identifying named entities in the EMR text is essential for structuring EMR. The structured processing of textual information for the biomedical and clinical healthcare domains has been widely carried out, and the NER in clinical EMR is one of these tasks.

Currently, Bi-directional Long-Short-Term Memory network with Conditional Random Field (BiLSTM-CRF) is the most efficient method for NER for clinical EMR [2]. Compared to machine learning methods [3], deep neural network techniques may lessen the composition of feature engineering by automatically learning high dimensional abstractions from the data. However, it is still unable to effectively utilize a large amount of existing unlabelled data, while the long-distance dependency between words makes the model unable to capture the EMR text's features in-depth due to the long text of the EMR.

In order to resolve the two problems mentioned above, this paper proposes a model incorporating a pre-trained language model and a Multi-Head Attention layer (BERT-BiGRU-Att-CRF). It has been demonstrated that the large unlabeled corpus can be efficiently used by the pre-trained. [4]. The Bi-directional Gated Recurrent Unit network followed by the CRF (BiGRU-CRF), which incorporates Multi-Head Attention, can extract distinct clinical features of EMR from multiple perspectives and levels. This structure significantly improves the accuracy of NER. On the CCKS2019 dataset, the model proposed this paper achieved the F_1 score of 86.97%.

II. Related Work

NER refers to recognizing entities in text with specific meanings. NER plays a crucial role in structuring textual information. NER for Clinical EMR is a branch of NER. There are three standard methods for NER: rule and dictionary, machine learning, and deep learning.

Establishing extensive knowledge bases and dictionaries is the primary foundation of rule and dictionary approaches, and by manually constructing rule templates and selecting feature information for matching, the system performance is heavily dependent on the rules and manual intervention, resulting in a more subjective design of its rules, and often does not list all the rules. The existing dictionary is challenging to include all the entities, leading to a lower recognition recall rate [5].

NER is handled as a sequence annotation problem in conventional machine learning techniques. Each sentence component is labeled using annotation models trained on a huge corpus. In the NER, models like the generative model are often utilized: HMM [6], CRF, etc. The CRF model has emerged as the standard NER model because it solves the problem of dependency between labels [7].

Manuscript received June 29, 2023; revised February 24, 2024. This work was supported by the Special Fund for Scientific Research Construction of University of Science and Technology Liaoning.

Haotian Yang is a postgraduate student at School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, CO 114051 China (e-mail: m18524337033@163.com).

Li Wang is a professor of the College of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, CO 114051 China (corresponding author to provide e-mail: wangli9966@ustl.edu.cn).

Yanpeng Yang is a senior engineer of Network Information Centre, University of Science and Technology Liaoning, Anshan, CO 114051 China (e-mail: yyp@ustl.edu.cn).

However, the method depends highly on pre-defined artificial features, which require high feature selection, feature engineering, and high labor costs.

Deep learning is being steadily used to NER problems because to its quick growth, and it has shown good results in the general area of NER. RNN and CNN are the primary models used in deep learning techniques. [8]. As the improved structure of RNN, LSTM network [9] and GRU [10] have become the mainstream methods of the current research on NER due to its solution to the defective of gradient vanishing and gradient exploding in the training process of lengthy sequences. Benefiting from the nonlinear transformation of deep learning, more intricate characteristics may be learned from data using deep learning models, avoid constructing many artificial features, and perform better than traditional NER methods.

The main task of NER for EMR in this paper is to identify pre-defined clinical terms from the unstructured EMR text, including six types of entities: anatomical sites (anatomy), surgeries (operation), diseases and diagnosis (disease), medicines(drug), laboratory tests (laboratory), and imaging tests (image). NER for EMR laid the foundation for electronic medical information data extraction, clinical diagnosis treatment information mining, clinical knowledge graph construction, etc.

Collobert [11] proposed a model grounded on convolution for resolving the sequence annotation issue. This approach automatically utilizes a CNN to extract features of the input sentence vector, which exceeds the effectiveness of previous machine learning requiring manual feature extraction. However, this method is limited because it can only utilize information within a fixed window size and thus cannot process long-range contextual information. Moreover, the use of LSTM models can improve the above problems. Hammerton [12] first proposed the application of LSTM for text entity recognition and experimentally verified the model's effectiveness. Yang [13] proposed a recognition method for named entities in EMR based on BiLSTM, and the experiments demonstrated the BiLSTM network model's efficacy for entity recognition. The LSTM-CRF model has gradually become a typical structure for entity recognition. Li [14] raised a NER model for EMR based on BiLSTM-CRF and carried out comparative experiments on the CCKS2018 Chinese EMR dataset, which demonstrated that the BiLSTM-CRF model performs noticeably better than the CRF model. However, The majority of these algorithms can only extract contextual information and need a substantial quantity of labeled data, which does not address long-distance dependencies between words. Therefore, this paper proposes a semi-supervised approach incorporated pre-trained language models and Multi-Head Attention, which is currently understudied in the CNER task of NER.

III. Named Entity Recognition Framework

This paper follows the same approach as mainstream research on NER, which treats NER for EMR as a sequence annotation task. The model in this paper comprises two parts: a pre-trained model and a sequence annotation model, as shown in Figure 1. This paper proposes a deep learning method incorporating the BERT as pre-trained model and the Multi-Head Attention called BERT-BiGRU-Att-CRF model.

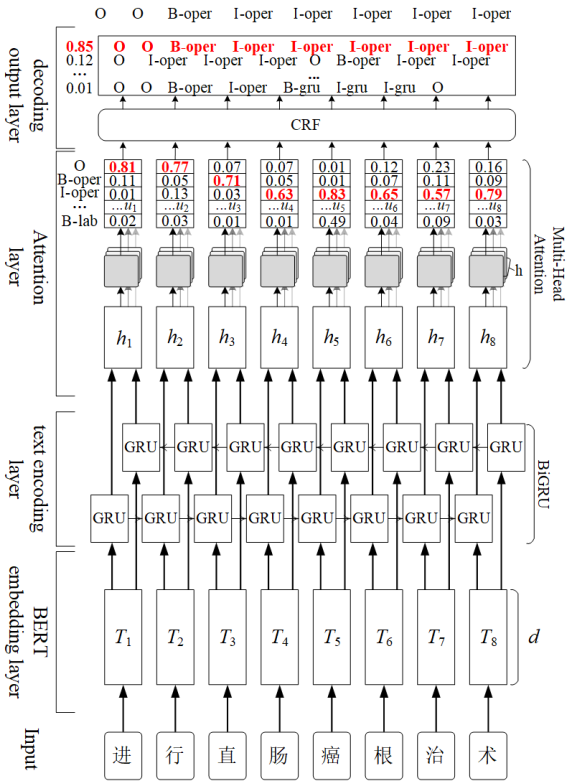

Fig. 1. Structure of BERT-BiGRU-Att-CRF Model

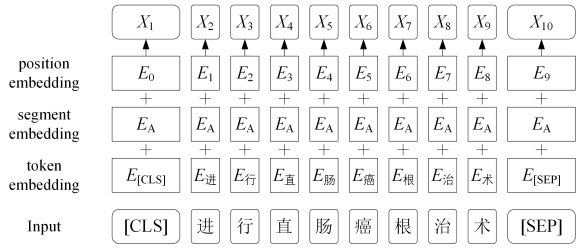

Fig. 2. Example Diagram of BERT Input

The pre-trained model is located in the embedding layer and the text encoding layer, the attention layer, and the decoding output layer make up the sequence annotation model.

A. Embedding Layer

BERT [15] is an unsupervised pre-trained language model proposed by Google in 2018, and it is a landmark model in NLP in the last several years. In NLP problems, the BERT model has been used extensively, but it is pre-trained from an extensive universal domain corpus, which does not perform at its level in vertical domains. Therefore, the bi-directional language model used as an embedding layer in this paper is fine-tuned [16] based on the BERT model to better perform downstream tasks in the medical domain.

Three components make up the input to the BERT model: token embedding, segment embedding, and position embedding. [CLS] is the sentence head vector, and [SEP] is the sentence center and end vector. In Figure 2, an example input is shown. The core network structure of BERT is composed of multi-layer stacked and bi-directional Transformer Encoder. Its structure is shown in Figure 3. $X_1, X_2, \ldots, X_n$ obtained by summing the three embeddings are represented as the final input sequence of BERT. $T = T_1, T_2, \ldots, T_n$ is the list of output word vectors from the BERT model. $T \in R^{n \times d}$,

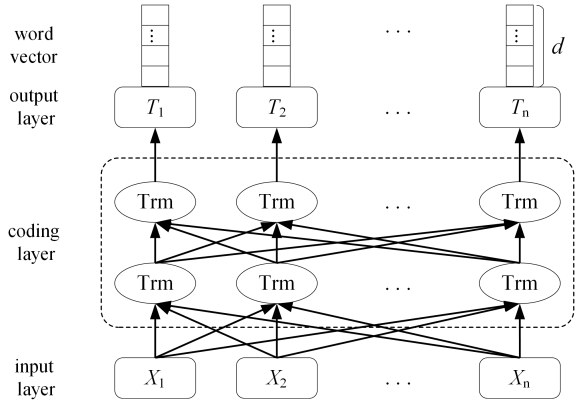

Fig. 3. Structure of BERT Model

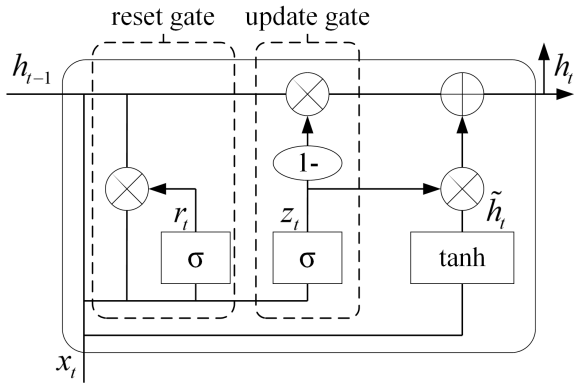

Fig. 4. Internal Structure of GRU Model

d denotes the vector dimension. The BERT model borrows the GPT idea of using a Transformer Encoder (including Multi-Head Attention) as a feature extractor, which enhances the ability of semantic feature extraction. BERT employs the CBOW training method used by Word2Vec concerning ELMO's bi-directional coding method. It gains a higher capacity for semantic extraction by making use of each word's contextual information.

The BERT model structure used in this paper has 24 Transformer block layers and 16 Self-Attention heads. Two unsupervised tasks are performed in the pre-trained phase: Masked Language Model (MLM) and Next Sentence Prediction (NSP). In order to train the semantic understanding of words, MLM randomly masks a certain percentage of words in each phrase and utilizes the contexts of those words to create predictions. The NSP task is a paragraph reordering task, which only considers two sentences and determines whether they are a text's preceding and following sentences. The goal of the task is to teach understanding between sentences using a sentence-level binary classification task. The combined application of the two tasks yields a thorough and exact vector representation of the text sequence that was entered. The problem of words have multiple meanings in different settings may be resolved by using the BERT model, which creates dynamic word vectors in response to contextual changes.

BERT is also fine-tuned using a large amount of unlabeled medical domain corpus so that the input text sequences generated by the BERT model are more accurate in the medical domain [17]. For example, "gastric cancer" and "leukaemia" have little connection in daily Chinese expressions, but in the medical domain "leukaemia" can also be referred to as "blood cancer", which belongs to the same category under the classification of cancer. Thus, the spatial distance between these two words should be closer to the medical domain than the generic one.

Therefore, the effective use of a large amount of unlabeled medical domain corpus can enable BERT to thoroughly learn a large number of unregistered professional vocabularies, named entities, and inter-of-entity relationships in the medical domain carrying rich professional knowledge so that BERT generates a vector representation of input text sequences that is more closely aligned with the knowledge system of the medical domain, which in turn improves the accuracy of the whole CNER for EMR task [18].

B. Sequence Annotation Model

Three layers make up the sequence annotation model: text encoding, attention, and decoding (output).

1) Encoding Layer: In order to extract global features from the word embedding layer's output vectors, the text encoding layer uses the BiGRU model. GRU is also a deformed structure of RNN. Like LSTM, it controls the passage of information through the 'gate' structure. It can learn the long-term dependency information, which improvements the trouble of long dependency and gradient disappearance in the backpropagation of traditional RNNs.

GRU is more straightforward and has only two gates. Its structure is shown in Figure 4. The input gate and forgets gate in an LSTM are combined by GRU to create a new gate, called update gate as z_t in Figure 4. Another gate of GRU is called reset gate, as r_t in Figure 4. $\oplus$ denotes matrix addition, $\otimes$ denotes the matrix multiplied by elements. GRU may significantly increase training efficiency since compared to an LSTM, it is simpler to train and has less parameters because of its simpler structure. The GRU entity recognition model has an impact similar to that of LSTM. The GRU forward propagation Equations are 1 to 5:

$$r_t = \sigma(W_r \otimes [h_{t-1}, x_t]) \tag{1}$$

$$z_t = \sigma(W_z \otimes [h_{t-1}, x_t]) \tag{2}$$

$$\tilde{h_t} = \tanh(W_{\tilde{h}}[r_t \otimes h_{t-1}, x_t]) \tag{3}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h_t} \tag{4}$$

$$y_t = \sigma(W_o \otimes h_t) \tag{5}$$

The information input at the current time is denoted by x_t, the hidden state from the prior time is represented by h_{t-1}, and the hidden state sent to the next moment is denoted by h_t.The candidate hidden state is denoted by $\tilde{h_t}$. W_r, W_z, $W_{\tilde{h_t}}$ are the weighting matrices, while the hyperbolic tangent function and sigmoid function are represented by $tanh$ and σ respectively, and y_t is the hidden node output. Where [ ] denotes two vectors connected, $\otimes$ denotes the matrix multiplied by elements.

The unidirectional GRU states are output from front to back, which cannot fully consider the following information. Therefore, this paper adopts a bi-directional GRU network, shown in Figure 5.

BiGRU uses forward and reverses GRU to extract contextual information features, weights and sums the outputs,

and then uses a linear layer to convert d dimensional vectors into m dimensional vectors. This produces a list of labeled vectors $H = h_1, h_2, \ldots, h_n$, $H \in R^{n \times m}$ as the BiGRU final output. Where n is the text sequence length and m is the number of entity type labels.

The calculation procedure is shown in Equation 6 to 8:

$$\overrightarrow{h_t} = GRU(x_t, \overrightarrow{h_{t-1}}) \tag{6}$$

$$\overleftarrow{h_t} = GRU(x_t, \overleftarrow{h_{t-1}}) \tag{7}$$

$$h_t = [\overrightarrow{h_t} \oplus \overleftarrow{h_t}] \tag{8}$$

The input x_t at the present time, the forward hidden layer output $\overrightarrow{h_{t-1}}$, and the reverse hidden layer output $\overleftarrow{h_{t-1}}$ at the prior time determine the present time hidden layer output state h_t of BiGRU.

2) Attention Layer: Longer statements occupy a large proportion of EMR text data, and BiGRU extracts text features that cannot be obtained over long distances by introducing an Attention that allocates more Attention (larger feature weights) to features that are related to entities of the EMR and allocates less Attention (smaller feature weights) to features that are not related. By including Attention, the model's capacity to extract local features is enhanced and contextual semantic characteristics linked to the available data may be obtained more effectively [19]. The BiGRU model's long-distance dependency problem is partially resolved as attention weights are only based on word vectors and are not impacted by the spacing between words.

The output of the hidden layer h_t is passed through the fully connected layer to obtain u_t. u_t is the vector of Attention weights for the relevance of the current information to the contextual information. where the bias term is b_t, and the weight matrix is W_t. As shown in Equation 9:

$$u_t = tanh(W_t h_t + b_t) \tag{9}$$

The attention score vector a_t is obtained by normalizing the weight vector u_t using the softmax function. As shown in Equation 10:

$$a_t = softmax(u_t) = \frac{esp(u_t)}{\sum_{t=1}^{n} esp(u_t)} \tag{10}$$

The BiGRU layer output h_t outputs a weighted global semantic feature vector s_t after weight allocation by the Attention, as shown in Equation 11:

$$s_t = \sum_{t=1}^{n} a_t h_t \tag{11}$$

However, the shortcomings of the Attention are also apparent: the model will focus too much on its position when encoding the information of the current location and cannot closely relate to the contextual information. Therefore, for long-text electronic medical records, this paper adopts the Multi-Head Attention to obtain the features from the EMR text at various points and angles.

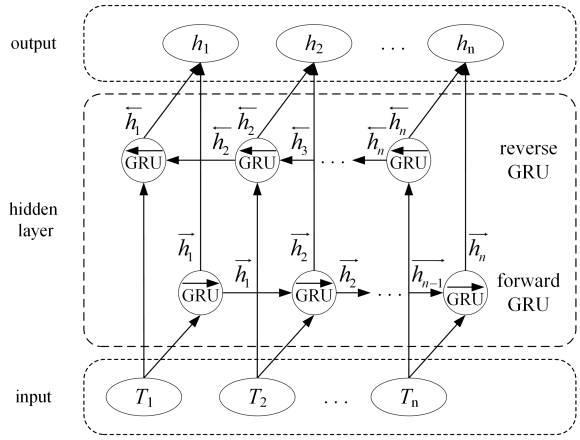

Fig. 5. Structure of BiGRU Model

3) Decoding Layer: BiGRU-Attention solves the problem of long-distance dependency in text encoding process and obtains the optimal output labels by calculating the specific score of each label obtained. However, it cannot solve problems such as dependency between labels (e.g., labels B-disease must exist before label I-disease). Therefore, the model cannot use its output labels as reasonable predictions. In order to provide a globally optimum sequence of acceptable labels, the CRF's primary function is to limit the entanglement between annotations by transferring the score matrix.

The output score matrix of BiGRU-Attention is S, $S \in R^{n \times m}$, s_{ij} denotes the jth label score of the ith character x_i in the text sequence.

First calculate the score of the text sequence $X = x_1, x_2, \ldots, x_n$ for the predicted label sequence $Y = y_1, y_2, \ldots, y_n$ as shown in Equation 12:

$$score(X, Y) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} S_{x_i, y_i} \tag{12}$$

Where A is the transfer score matrix, $A \in R^{(m+2) \times (m+2)}$, $A_{y_i, y_{i+1}}$, where $A_{y_i, y_{i+1}}$ is the score of labels yi transferred to labels y_{i+1}, S_{x_i, y_i} is the probability that the ith character of the input text sequence is predicted to be label y_i.

Secondly, the probability of outputting a sequence of labels Y is determined after using normalization, softmax, on it. As shown in Equation 13:

$$P(Y \mid X) = \frac{esp^{score(X,Y)}}{\sum_{\widetilde{Y} \in Y_X} esp^{score(X,\widetilde{Y})}} \tag{13}$$

Where Y_X is the set of all possible label sequences and $\widetilde{Y}$ is the proper label sequence.

Lastly, the globally optimum label sequence Y_{pre} of the text sequence X is obtained by using the Viterbi algorithm. The collection of labels with the greatest output probability is Y_{pre}. As shown in Equation 14:

$$Y_{pre} = arg\ maxscore(X, \widetilde{Y}) \tag{14}$$

IV. Experiment

A. Experimental Environment and Parameter Settings

The experimental environment is shown in Table I. The experimental parameters are set as shown in Table II.

TABLE I
EXPERIMENTAL ENVIRONMENT

Name of experimental environment	Configuration
Operating System	Windows 10
Programming Language	Python 3.7
Deep Learning Frameworks	Pytorch 1.7.0
Graphics Card	NVIDIA GTX 1070
Random Access Memory	64G

TABLE II
EXPERIMENTAL PARAMETER SETTINGS

Name of parameter	Value of a parameter
maximum length of sequence	256
hidden dimension of GRU	512
Dropout rate	0.25
Learning rate	0.0001
Optimizer	Adam
batch size	10

TABLE III
STATISTICAL RESULTS FOR DIFFERENT ENTITY TYPES

Type	Train set	Validation set	Test set
Diseases	2116	635	682
Image	222	67	91
Laboratory	318	95	193
Operation	765	230	140
Drug	456	137	263
Anatomy	1486	446	447

B. Dataset and Annotation Strategy

The dataset utilized in this paper originates from the CCKS2019 CNER tasks for EMR and is a modification of the CCKS2017 dataset, comprising a train set and a test set. The dataset was prepared by Yidu Cloud (Beijing) Technology Co.

Each row of data is a json, and the json key is ['originalText," entities'], the original text and the list of entities. json['entities'] for the list, each element represents an entity entity, which has the entity in the original text of the start position (start_pos), end position (end_pos), and entity type.

The dataset comprises a complete record of outpatient and inpatient medical cases, consisting of 1379 Chinese EMRs in plain text. The set includes 1000 train sets and 379 test sets. Including six types of EMR named entities: diseases and diagnoses, imaging tests, laboratory tests, surgeries, medicines, and anatomical sites. In this paper, the train set in the CCKS2019 dataset was randomly disrupted untimely.After that, the train and validation sets have been divided with a weight of 7:3. The statistical outcomes of the different entities in the dataset are shown in Table III.

This paper employs the BIO entity annotation approach to represent entity boundaries in the NER problem, therefore converting it into a sequence annotation problem for deep learning. 'B' indicates the start of an entity, 'I' indicates its middle, and 'O' indicates that it is not an entity. There are a total of 13 labels for six types of entities: B-Diseases, B-Image, B-Laboratory, B-Operation, B-Drug, B-Anatomy, I-Diseases, I-Image, I-Laboratory, I-Operation, I-Drug, I-Anatomy and O.

The unlabelled data is provided by another subtask of CCKS2019 Healthcare Entity and Attribute Extraction (Cross-Hospital Migration), with 1,000 pieces of unlabelled data for each scenario. This dataset is the input corpus for pre-trained language models and word embeddings along with the train set after the disrupted segmentation process.

TABLE IV
MEANING OF EQUATION PARAMETERS

Parameters	Meaning
T_P	Entities correctly recognised
F_P	Entities incorrectly recognised
F_N	Entities present in corpus but not recognised

C. Methods of Sentence Division

A conventional method for representing text is founded on word embedding. This technique correlates words to their context by transforming them into low-dimensional dense vectors using neural networks. This resolves the challenge of high dimensional sparse features in the Bag of Word (BOW) model [20]. Word vectors allocate semantic details of words in the vector space, thus resembling words nearer to each other in the vector space [21]. However, in Chinese, word separation tools frequently fail to ensure the precision of the outcomes because of the presence of unclear word boundaries. These inaccuracies could worsen and impede the accuracy of the resulting processing. Thus, this paper implements character vectors for text representation to preclude the adverse effects stemming from word separation errors.

D. Experimental Results and Analysis

1) Indicators for Model Evaluation: The evaluation criteria for named entity recognition include the Exact-Match (EM) and the Relaxed-Match (RM) criteria. In the EM criterion, it is required to both accurately identify the entity boundaries and correctly classify the entity types so that the recognition can be regarded as successful. In contrast, the RM criterion considers that in the case of correct entity identification, as long as the boundaries match, the recognition can be regarded as correct. Since the RM criterion cannot analyze the errors in the experiment, the EM criterion is used in this experiment to evaluate the model.

NER is usually evaluated by three metrics: recall (R), precision (P), and F_1. F_1 is the harmonic mean of R and P. The formulae for each indicator are shown in Equation 15 to 17, and the parameters in the equation and their meanings are shown in Table IV.

$$R = \frac{T_P}{T_P + F_N} \times 100\% \tag{15}$$

$$P = \frac{T_P}{T_P + F_P} \times 100\% \tag{16}$$

$$F_1 = \frac{2PR}{P + R} \times 100\% \tag{17}$$

Among them, the F_1 score can comprehensively evaluate the model's performance, so this paper adopts the F_1 score to assess the model performance.

TABLE V
EFFECT OF DIFFERENT ATTENTION LAYERS ON NER

Type of layers	Precision	Recall	F_1
None	85.72	80.29	83.01
Attention	86.24	82.57	84.41
Multi-Head Attention	87.84	86.10	86.97

TABLE VI
THE ROLE OF LANGUAGE MODELS IN STRUCTURE

Type of language model	Precision	Recall	F_1
None	75.28	84.06	79.43
BERT	87.89	83.51	85.57
our method	87.84	86.10	86.97

2) Experimental Results and Analysis: In order to verify the improvement of the introduction of the Multi-Head Attention for NER in long text sequences, a comparative validation is conducted experimentally, where one model removes the Multi-Head Attention layer and one uses the Attention layer to compare with the model this paper proposed. The results are presented in Table V.

As demonstrated in Table V, enhanced Attention or Multi-Head Attention is able to extract global information of long text in multi-level and multi-angle to deeply capture the own features of long text, and both attention layers have improved the F1 score, which is 1.40% and 3.96%, respectively.

In order to verify the effect of trained language models on the increase in recognition rate, this paper conducted experiments, one of which eliminated the pre-trained model and used only the annotated model. The other used the unfine-tuned BERT model. The experimental results are shown in Table VI.

The experimental findings show that utilizing the pre-trained language model with fine-tuned enhances NER, leading to a 7.54% increase in F_1 score relative to recognition without the use of the language model. This result is due to the pre-trained language model's features leveraging a more significant amount of information from the vast pools of unlabeled data within the general domain, compared to recognition solely on annotated data. The BERT model showed an F_1 score increase of only 1.40% in comparison to the model not fine-tuned. This can be attributed to the model's fine-tuned process, which provides it with additional knowledge of the medical domain, acting as a knowledge enhancement. However, the improvement was limited due to the smaller size of the dataset used for fine-tuned in comparison to the BERT pre-trained corpus.

In order to verify the effectiveness of the BERT-BiGRU-Att-CRF model proposed in this paper on NER, two groups of comparison experiments of different methods for feature extraction were conducted.

The first set of experiments will be compared with mainstream NER methods based on recurrent neural networks. The first group of comparison experiments includes this paper's model and other models: Lattice LSTM model [22] [23], BiLSTM-CRF model [24], BiGRU-CRF, BiLSTM-CRF+ELMO model [25], BERT-BiLSTM-CRF model [26] [27], BERT-BiGRU-CRF, BERT-BiLSTM-Att-CRF model [28].

TABLE VII
RESULTS OF DIFFERENT RECURRENT NEURAL NETWORK MODELS

Models	Precision	Recall	F_1
Lattice LSTM	79.41	79.24	79.32
BiLSTM-CRF	80.37	80.45	80.41
BiGRU-CRF	80.65	80.57	80.61
BiLSTM-CRF+ELMO	82.06	81.83	81.94
BERT-BiLSTM-CRF	83.45	81.34	82.40
BERT-BiGRU-CRF	83.81	82.59	83.20
BERT-BiLSTM-Att-CRF	86.15	85.75	85.95
BERT-BiGRU-Att-CRF	87.84	86.10	86.97

TABLE VIII
RESULTS WITH CONVOLUTIONAL NEURAL NETWORK MODELS

Models	Precision	Recall	F_1
CNN-CRF	75.66	77.12	76.39
IDCNN-CRF	79.41	79.24	79.32
BERT-IDCNN-CRF	82.29	81.76	82.02
BERT-BiGRU-Att-CRF	87.84	86.10	86.97

The overall recognition effect of each model in the Group 1 comparison experiment is shown in Table VII.

(1) In the experimental process, the training time of the BERT-BiGRU-Att-CRF model is 129.25 min, and the training time of the BERT-BiLSTM-Att-CRF model is 186.32 min, which shows that replacing the BiLSTM with BiGRU can significantly save the time cost. Moreover, the GRU model slightly outperforms the LSTM model in all aspects of entity recognition, indicating that the proposed model in this paper ensures the recognition effect based on saving the training cost.

(2) Comparing models 1, 2, 3, 4 and 5, 6, 7, 8 it can be seen that all the indexes are improved after the introduction of BERT pre-trained model, which proves the effectiveness of BERT model. Because the BERT pre-trained model generates word vectors with full consideration of the semantic information in different contexts, it obtains dynamic word vectors of the input text sequence. It solves the problem of multiple meanings of words.

(3) Comparing models 5, 6 and model 7, 8 it is evident that the model's capacity for entity recognition improves even more with the addition of the Attention mechanism. This suggests that the addition of Attention improves the model's capacity for extracting semantic features from lengthy texts, which contributes to the model's improved entity recognition performance.

(4) Comparing model 3 with model 8, the introduction of the pre-trained model and the Multi-Head Attention significantly improves the recognition of named entities, where it improves the precision by 7.19%, the recall by 5.53%, and the F_1 score by 6.36%.

The second set of experiments will be compared with a named entity recognition method containing a convolutional operation. The first group of comparison experiments includes this paper's model, CNN-CRF, IDCNN (Iterated Dilated Convolutions)-CRF model [29], and BERT-IDCNN-CRF model.

The overall recognition effect of each model in the second comparison experiment is shown in Table VIII.

Viewing Table VII and Table VIII simultaneously, it indicated that recurrent neural networks are better in the NER task as compared to convolutional neural networks; among them, the entity recognition effect based on the BERT-BiGRU-CRF model is significantly better than that based on the BERT-IDCNN-CRF model, with the precision, recall, and F_1 score higher than that of the BERT-IDCNN-CRF model by 5.55%, 4.34%, and 4.95%, respectively. Table VIII demonstrates that the model this paper proposed has a superior than the NER model using convolutional operations for entity recognition.

Combining the above analyses of the results of the two single-variable experiments and the two sets of comparison experiments, with the F_1 score increased to 86.97%, it was confirmed that the pre-trained model and the Multi-Head Attention were effective in enhancing the accuracy of the named entity identification task, and that the BERT-BiGRU-Att-CRF model was valid and feasible.

V. Conclusion

This paper proposes a BERT-BiGRU-Att-CRF CNER for the EMR model incorporating the Attention mechanism and a pre-trained model, BERT. Employing an unlabeled medical corpus, the BERT model was fine-tuned to increase the representational relevance of word vector embedding in the medical domain and to get character-level dynamic vectors of input text sequences to address the conundrum of various meanings of words. Using BiGRU to obtain global semantic features of the input text sequence saves the model's training time cost. Using the Attention to enhance the semantic information of feature vectors, solving the long-distance dependence problem in feature extraction, improving the accuracy of the model entity recognition, and improving it to Multi-Head Attention, avoiding excessive Attention to its location information, closely linking the contextual information, and enhancing the ability to extract features. The CRF is used to deal with the interdependence of labels and improve the effects of CNER for EMR. Experiments show that the proposed model can effectively identify anatomical sites, diseases, and operations entities, and comparative experiments on the CCKS2019 dataset obtain that the F_1 score of the proposed model reaches 86.97%, which is higher than that of other models. In the next stage, we will emphasize to improving the pre-trained model, further reducing the time cost, improving the model training effect, and applying the proposed model to other NER fields.

References

[1] D. CM, C. EG, R. SR, D. K, and F. T. et al., "Electronic Health Records in Ambulatory Care–a National Survey of Physicians," *2008 Massachusetts Medical Society*, vol. 359(1), pp. 50–60, 2008.

[2] L. Y, W. X, H. L, Z. L, L. H, and X. L. et al., "Chinese Clinical Named Entity Recognition in Electronic Medical Records: Development of a Lattice Long Short-Term Memory Model with Contextualized Character Representations," *JMIR Med Inform. 2020 Sep 4*, vol. 8(9), p. e19848, 2020.

[3] G. Zhou and J. Su, "Named Entity Recognition Using an HMM-Based Chunk Tagger," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL 02, 2002, p. 473–480.

[4] C. Ma and C. Zhang, "Joint Pre-Trained Chinese Named Entity Recognition Based on Bi-Directional Language Model," *Int. J. Pattern Recognit Artif Intell*, vol. 35, pp. 2 153 003:1–2 153 003:16, 2021.

[5] Y. Sari, M. F. Hassan, and N. Zamin, "Rule-based Pattern Extractor and Named Entity Recognition: A Hybrid Approach," in *2010 International Symposium on Information Technology*, vol. 2, 2010, pp. 563–568.

[6] A. Ekbal and S. Bandyopadhyay, "A Hidden Markov Model Based Named Entity Recognition System: Bengali and Hindi as Case Studies," in *Pattern Recognition and Machine Intelligence*, 2007, pp. 545–552.

[7] N. Patil, A. Patil, and B. Pawar, "Named Entity Recognition Using Conditional Random Fields," *Procedia Computer Science*, vol. 167, pp. 1181–1188, 2020.

[8] T. Gui, R. Ma, Q. Zhang, L. Zhao, Y.-G. Jiang, and X. Huang, "CNN-based Chinese NER with Lexicon Rethinking," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, vol. IJCAI-19, 2019, pp. 4982–4988.

[9] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF Models for Sequence Tagging," *ArXiv*, vol. abs/1508.01991, 2015.

[10] N. Banik and M. H. H. Rahman, "GRU Based Named Entity Recognition System for Bangla Online Newspapers," in *2018 International Conference on Innovation in Engineering and Technology (ICIET)*, 2018, pp. 1–6.

[11] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural Language Processing from Scratch," *arXiv*, vol. 1103.0398, 2011.

[12] J. A. Hammerton, "Named Entity Recognition with Long Short-Term Memory," *Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL 2003*, vol. 2003, pp. 172–175.

[13] Y. Hongmei, L. Lin, and Y. R. et al., "Recognition Model of Named Entities in Electronic Medical Records Based on Bidirectional LSTM Neural Network," *Chinese Tissue Engineering Research*, vol. 22(20), pp. 3237–3242, 2018.

[14] L. L and H. L., "Named Entity Recognition in Chinese Electronic Medical Records Based on the Model of Bidirectional Long Short-Term Memory with a Conditional Random Field Layer," *Stud Health Technol Inform. 2019 Aug 21*, vol. 264, pp. 1524–1525, 2019.

[15] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. Volume 1 (Long and Short Papers), Jun. 2019, pp. 4171–4186.

[16] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to Fine-Tune BERT for Text Classification?" *arXiv*, p. 1905.05583, 2020.

[17] G. B, L. H, and N. L, "Integration of Natural and Deep Artificial Cognitive Models in Medical Images: BERT-based NER and Relation Extraction for Electronic Medical Records," *Front Neurosci*, vol. 2023 Sep 4, p. 1266771, 2003.

[18] W. Liu, P. Zhou, Z. Zhao, Z. Wang, Q. Ju, H. Deng, and P. Wang, "K-BERT: Enabling Language Representation with Knowledge Graph," *arXiv*, p. 1909.07606, 2019.

[19] V. Ashish, S. Noam, P. Niki, U. Jakob, J. Llion, G. A. N, K. L. ukasz, and P. Illia, "Attention is All you Need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[20] L. Li, W. Xu, and H. Yu, "Character-level Neural Network Model Based on Nadam Optimization and Its Application in Clinical Concept Extraction," *Neurocomputing*, vol. 414, pp. 182–190, 2020.

[21] X. Rong, "Word2Vec Parameter Learning Explained," *ArXiv*, vol. abs/1411.2738, 2014.

[22] L. Yongbin, Z. Liping, L. Weihai, and W. Xiaohua, "Research on Chinese Clinical named entity recognition: Lattice LSTM with Contextualized Character Representations," *JMIR Medical Informatics*, vol. 8, 05 2020.

[23] Y. Zhang and J. Yang, "Chinese NER Using Lattice LSTM," *ArXiv*, vol. abs/1805.02023, 2018.

[24] B. Ji, S. Li, J. Yu, R. Liu, J. Tang, Q. Li, and W. Xu, "A BiLSTM-CRF Method to Chinese Electronic Medical Record Named Entity Recognition," *Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence*, 2018.

[25] L. Nan, L. Ling, D. Zeyuan, S. Yawen, Y. Zhihao, and H. Lin, "DUTIR at the CCKS-2019 Task1: Improving Chinese Clinical Named Entity Recognition using Stroke ELMO and Transfer Learning," *Proceedings of the Evaluation Tasks at the China Conference on Knowledge Graph and Semantic Computing (CCKS-Tasks 2019)*, 2019.

[26] X. Li, H. Zhang, and X. Zhou, "Chinese Clinical Named Entity Recognition with Variant Neural Structures Based on BERT Methods," *Journal of Biomedical Informatics*, vol. 107, p. 103422, 2020.

[27] R. Vunikili, N. SupriyaH., V. G. Marica, and O. Farri, "Clinical NER Using Spanish BERT Embeddings," *IberLEF@SEPLN*, 2020.

[28] Y. An, X. Xia, X. Chen, F.-X. Wu, and J. Wang, "Chinese Clinical Named Entity Recognition Via Multi-Head Self-Attention Based

BiLSTM-CRF," *Artificial Intelligence in Medicine*, vol. 127, p. 102282, 2022.
[29] H. Wenming, Z. Juan, X. Yannan, H. Zheng, and D. Zhenrong, "Named Entity Recognition in Chinese Judicial Domain Based on Self-attention mechanism and IDCNN," *2020 8th International Conference on Digital Home (ICDH)*, vol. 19-20 Sept. 2020, pp. 51–56, 2020.