Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.053630

ARTICLE

Chinese Clinical Named Entity Recognition Using Multi-Feature Fusion and Multi-Scale Local Context Enhancement

Meijing Li*, **Runqing Huang and Xianxian Qi**

College of Information Engineering, Shanghai Maritime University, Shanghai, 200306, China

*Corresponding Author: Meijing Li. Email: mjli@shmtu.edu.cn

Received: 06 May 2024 Accepted: 28 June 2024 Published: 15 August 2024

ABSTRACT

Chinese Clinical Named Entity Recognition (CNER) is a crucial step in extracting medical information and is of great significance in promoting medical informatization. However, CNER poses challenges due to the specificity of clinical terminology, the complexity of Chinese text semantics, and the uncertainty of Chinese entity boundaries. To address these issues, we propose an improved CNER model, which is based on multi-feature fusion and multi-scale local context enhancement. The model simultaneously fuses multi-feature representations of pinyin, radical, Part of Speech (POS), word boundary with BERT deep contextual representations to enhance the semantic representation of text for more effective entity recognition. Furthermore, to address the model's limitation of focusing just on global features, we incorporate Convolutional Neural Networks (CNNs) with various kernel sizes to capture multi-scale local features of the text and enhance the model's comprehension of the text. Finally, we integrate the obtained global and local features, and employ multi-head attention mechanism (MHA) extraction to enhance the model's focus on characters associated with medical entities, hence boosting the model's performance. We obtained 92.74%, and 87.80% F1 scores on the two CNER benchmark datasets, CCKS2017 and CCKS2019, respectively. The results demonstrate that our model outperforms the latest models in CNER, showcasing its outstanding overall performance. It can be seen that the CNER model proposed in this study has an important application value in constructing clinical medical knowledge graph and intelligent Q&A system.

KEYWORDS

CNER; multi-feature fusion; BiLSTM; CNN; MHA

1 Introduction

The medical field has witnessed a rapid growth in medical information technology, leading to a significant focus on the informatization of Electronic Medical Record (EMR) [1]. During hospital visits, EMR are commonly utilized to record the patient's physical health status and capture the entire process of medical diagnosis. It is an indispensable medical data resource in healthcare services, as it provides patients with reliable medical evidence, assists doctors in grasping patients' physical health status, and supports clinical experiments and research [2]. It is usually stored as various data types, including unstructured free text that computers cannot automatically extract and recognize [3]. In order to effectively utilize the unstructured free texts, it is essential to employ entity extraction methods, such as Named Entity Recognition (NER) [4]. NER aims to extract valuable information from the text. While NER has achieved considerable success in English, Chinese Named Entity Recognition (CHNER) [5] poses greater complexity and difficulties. The complexity of CHNER is largely attributed to the abundance of homophones and the absence of clear boundaries in the language. These factors pose significant challenges, distinguishing it from other languages when it comes to recognizing named entities in Chinese text.

Previous research in CHNER has explored various approaches, including dictionary-based [6], rule-based [7], and machine learning-based methods [8]. These approaches have shown a degree of achievement in CHNER tasks. Although very accurate, these methods rely significantly on manual annotation and feature engineering, which can be a laborious and resource-intensive job [9]. With the ongoing advancements in science, technology, and computing power, there is a growing trend toward utilizing deep learning techniques in CHNER. Deep learning techniques have demonstrated superior performance across various domains. Particularly, Long Short-Term Memory (LSTM) [10] based neural network models have gained significant popularity in CHNER tasks. Among these models, the BiLSTM-CRF (Bidirectional LSTM with Conditional Random Fields) [11] model has emerged as a prominent approach and achieved noteworthy results in CHNER tasks. Nevertheless, the majority of existing CHNER models rely on character-based [12] or word-based [13] vector models. On the one hand, character-based models alone may not capture sufficient semantic information compared to word vectors. On the other hand, relying solely on unique word vector-based models may result in inadequate representation due to inaccuracies in the word-splitting tool, leading to subpar performance [14]. Moreover, existing CHNER methods frequently focus on global context information and overlook the importance of local context information, which are also essential for accurate entity recognition. Because of these reasons, CHNER models cannot fully consider semantic information when extracting the named entities from Chinese text. To solve these problems, we propose a multi-feature fusion and multi-scale local context enhancement for CNER model. Our contribution can be summarized as follows:

1. We propose a new feature extraction method based on multi-feature fusion and multi-scale local context enhancement, which comprehensively considers the multi-feature semantic of Chinese characters and simultaneously extracts deep global and local semantic information from Chinese Electronic Medical Record (CEMR) text.
2. We propose a multi-scale local context enhancement method based on multiple Convolutional Neural Networks (CNNs) with different kernel sizes to capture local contextual features from various scales, ranging from fine-grained to coarse-grained. This enables the model to delve deeper into the semantic information of the text.
3. We conduct extensive experiments on the publicly available CEMR datasets CCKS2017 and CCKS2019. The experimental results prove the validity of the model and verify the importance of each component in the model.

2 Related Work

In the initial phases of NER, dictionary-based and rule-based approaches were mainly used. The dictionary-based and rule-based methods primarily rely on manual formulation, where entities are recognized by domain experts, formulating specific rules, and then combining them with a dictionary using pattern matching. Still, due to the specificity of rule formulation and the incompleteness of the dictionary, the limitations of this method are significant, not only consuming a lot of effort and time but also not making it easy to expand the dataset. Machine learning-based approaches use supervised learning to convert NER tasks into sequence labeling tasks or classification tasks, in which the process often involves building many feature projects, typically Hidden Markov Models (HMMs) [15] models and Conditional Random Fields (CRFs) [16] models. Although this approach substantially improves over previous methods, it still requires extensive labeling by experts in the specialized domain. In addition, it still has a high time cost for training.

Deep learning technology has quickly advanced in recent years, making it the dominant strategy in NER research. Deep learning utilizes neural networks to automatically extract features of objects and has demonstrated success [17]. Huang et al. [11] proposed the BiLSTM-CRF model for sequence annotation tasks, which has significantly enhanced the accuracy of NER. Zhang et al. [18] proposed a lattice LSTM model specifically designed for NER, which incorporates the words' meaning into the word vector model, significantly improving Chinese boundary segmentation's ability. Wu et al. [19] proposed a network structure for NER based on CNN-LSTM-CRF, which jointly trained the NER and word segmentation models, enhancing the ability to accurately recognize entity boundaries in Chinese text. Xue et al. [20] proposed a centralized attention model that integrated BERT pre-training and collaborative learning to enhance feature representation in the parameter sharing layer, resulting in improved accuracy in extracting medical text entities and relations. Zhao et al. [21] proposed an adversarial training-based lattice LSTM model that integrated character and word embeddings, which incorporated adversarial perturbations into the lattice LSTM structure, with the specific goal of enhancing recognition performance in the context of Chinese clinical texts. Li et al. [22] enhanced CNER by including dictionary data and radical properties of Chinese characters to improve the contextualized representation of words in their model. Kong et al. [23] introduced a CNN model that employs a multi-level CNN structure to capture contextual information, making use of GPU parallelism and enhancing model performance. An et al. [24] enhanced CNER by incorporating a multi-head attention mechanism, which integrates a multi-head attention mechanism with a medical dictionary, enabling more efficient capturing of the relationship between Chinese characters and multi-level semantic features. Guo et al. [25] used the Transformer layer of the soft-dictionary structure to replace the traditional LSTM, and the soft-dictionary system of the Transformer layer not only supports parallel computation to save a lot of time but also captures more contextual dependencies and correlations.

In summary, deep learning-based NER has shown promising results and is gaining traction for practical NER tasks. Therefore, we propose a multi-feature fusion and multi-scale local context enhancement method for CNER. By extracting multi-feature embedding and fusing multi-scale local contextual features, our approach enables the model to better comprehend Chinese clinical text information, thereby improving overall model performance.

3 Proposed Method

The model's architecture, shown in Fig. 1, consists of four neural network layers: a feature embedding layer, a feature extraction layer, a multi-head attention (MHA) mechanism layer, and a CRF layer. Below are the detailed details of each layer:

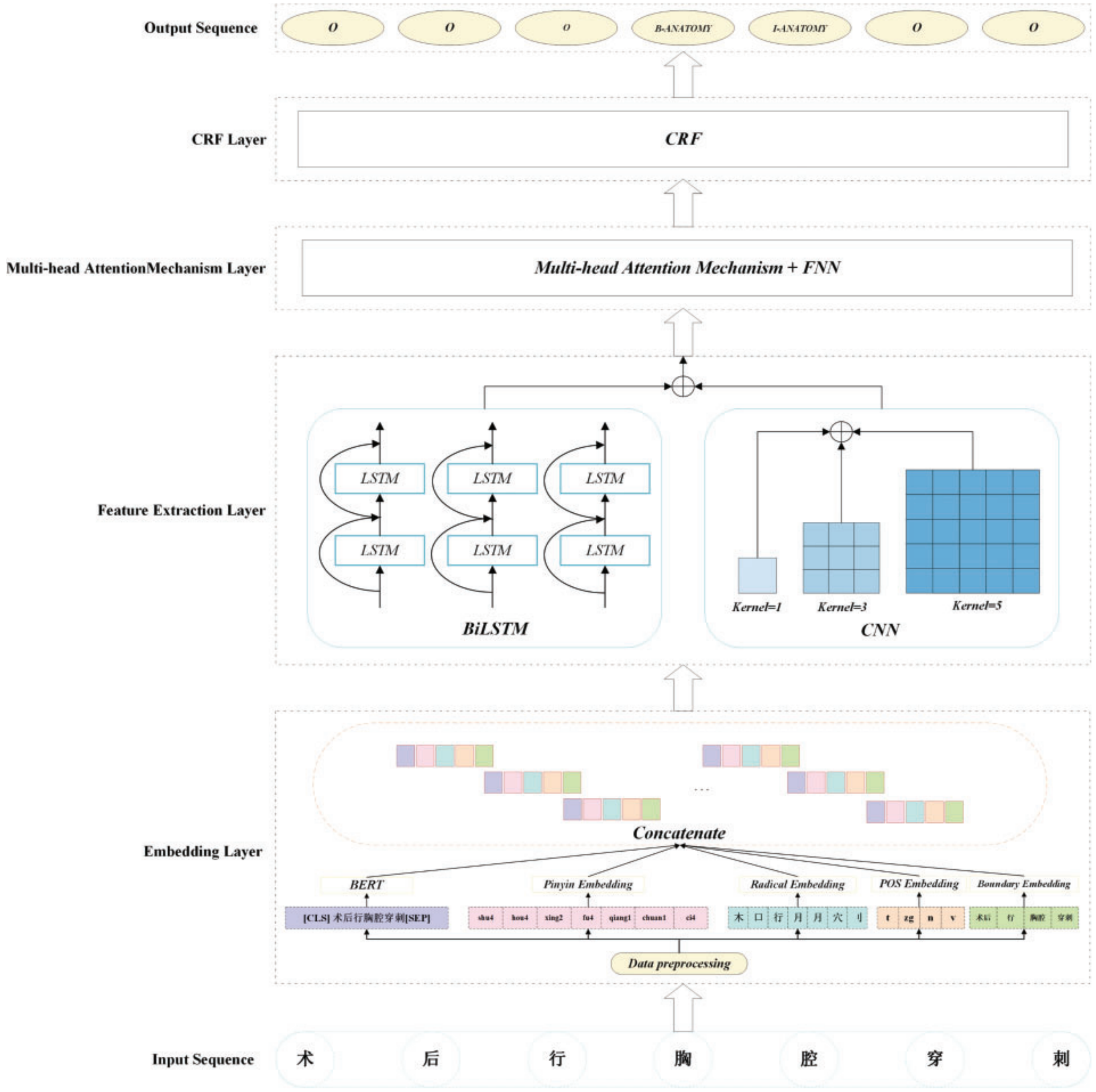

Figure 1: The overall model architecture

3.1 Feature Embedding Layer

To obtain as much rich semantic information as possible from the sequence, we extract sequence features from five different perspectives: pinyin feature, radical feature, Part of Speech (POS) feature, word boundary feature, and deep contextual word-level feature BERT.

3.1.1 Pinyin Feature

Pinyin, as the official standard for the pronunciation of Chinese characters, contains a wealth of semantic information. For dealing with low-frequency and unknown words, pinyin can also provide valuable pronunciation information. This feature is particularly important in medical texts that contain a large number of specialized clinical terms, because the same Chinese character may have completely different meanings under different pinyin. For example, although the Chinese character "中" in "中医" (ENG: Traditional Chinese Medicine) and "中风" (ENG: stroke) have the same character form, they have different pinyin, which leads to obvious differences in their semantics. The former refers to traditional Chinese medicine, while the latter represents an acute cerebrovascular disease. Therefore, simply converting Chinese characters into word-level vectors may lose this semantic information. In order to distinguish the different meanings expressed by polyphonic and homophonic characters in Chinese, we can assist the word-level semantic expression of Chinese characters by integrating pinyin features. Specifically, the pinyin vector consists of 27 dimensions: the first 26 dimensions correspond to the 26 letters in the pinyin system, while the last dimension is used to represent the tones of the Chinese character. To construct a pinyin vector, we first obtain the pinyin of each Chinese character in the corpus. Then, we count the number of occurrences of each pinyin letter in the first 26 dimensions of the vector, and combine it with the tone information in the last dimension to construct a comprehensive 27-dimensional pinyin vector. The construction process of the pinyin vector is as follows:

$$Q^{pinyin} = E^{pinyin}\left[F^{pinyin}\left(X\right)\right] \tag{1}$$

where X represents the input sequence, the F^{pinyin} function maps the input sequence to a sequence of pinyin sequence, E^{pinyin} represents the mapping table between pinyin and the input sequence, and Q^{pinyin} represents the pinyin vector.

3.1.2 Radical Feature

The radicals of Chinese characters are important in understanding the composition and meaning of Chinese characters. Radicals are often presented as specific symbols or shapes that help us decipher the pronunciation and meaning of a Chinese character. In the Chinese character system, characters sharing the same radicals are often semantically related. For example, "花" (ENG: flower) and "草" (ENG: grass) both contain the radical "艹", which is also commonly associated with herbs. In the medical field, many specialized clinical terminology naming entities also show consistent patterns of radicals. For example, many disease terms carry the radical "疒" such as "疹" (ENG: rash) and "痒" (ENG: itch). Traditional BERT models may not be able to capture subtle internal feature differences when dealing with low-frequency or unknown words, and thus extracting radical feature vectors can be helpful for vector bias of low-frequency or unknown words. To construct the radical vectors, we first create the radical of each character from the corpus and count the set of all occurrences of the radical. Then, we extract the radical of the current character and obtain the subscript of the position of the radical in the set as a one-dimensional radical vector for that character. The process of constructing its radical vector is as follows:

$$Q^{radical} = E^{radical}\left[F^{radical}\left(X\right)\right] \tag{2}$$

where X represents the input sequence, the $F^{radical}$ function maps the input sequence to a sequence of radical sequence, $E^{radical}$ represents the mapping table between radical and the input sequence, and $Q^{radical}$ represents the radical vector.

3.1.3 Pos Feature

POS is the grammatical property or lexical type that words have in a sentence. It describes the role and grammatical characteristics of a word in a sentence. In CNER, POS provides great help in extracting named entities. In addition, medical texts also contain a large number of terms with the same word form but different POS with very different meanings. For example, "感染" (ENG: infection), when used as a noun, denotes a concept or state that refers to the process of a pathogen spreading into an organism and causing an abnormal reaction. When used as a verb, it denotes an action or process that refers to the invasion of an organism by a pathogen that causes an infection. POS helps us to distinguish these terms and determine their roles and functions in the sentence. Therefore, extracting POS features helps the model to understand the text more deeply. To construct the POS vector, we first create the set of all POS, then extract the POS of each character and obtain the subscript of the position of that POS in the set as the one-dimensional POS vector of that character. The procedure of constructing the POS vector is as follows:

$$Q^{POS} = E^{POS}\left[F^{POS}(X)\right] \tag{3}$$

where X represents the input sequence, the F^{POS} function maps the input sequence to a sequence of POS sequence, E^{POS} represents the mapping table between POS and the input sequence, and Q^{POS} represents the POS vector.

3.1.4 Word Boundary Feature

In general domain datasets, names of places and organizations usually have distinct word boundaries, such as containing distinct boundary words like "省" (ENG: province) and "市" (ENG: city). However, in CNER, many entities do not have distinct boundaries, such as "横纹肌肉瘤" (ENG: rhabdomyosarcoma) and "肿瘤组织" (ENG: tumor tissue). Therefore, using the word boundary feature is crucial to addressing the issue of entity boundary ambiguity. To construct the word boundary vector, we first divide the sequence into words to obtain the word boundary sequence. We then encode the word boundary sequence using 3 dimensions one-hot encoding to obtain the final word boundary vector. Table 1 illustrates an example of constructing word boundary vector. The construction process of the word boundary vector is as follows:

$$Q^{boundary} = E^{boundary}\left[F^{boundary}(X)\right] \tag{4}$$

where X represents the input sequence, the $F^{boundary}$ function maps the input sequence to a sequence of word boundary sequence, $E^{boundary}$ represents the mapping table between word boundary and the input sequence, and $Q^{boundary}$ represents the word boundary vector.

Table 1: Example of word boundary vector construction

明	确	诊	断	脑	梗	死
0	1	0	1	0	0	1
0	0	0	0	0	1	0
1	0	1	0	1	0	0

3.1.5 Deep Context Word-Level Feature

To more accurately represent semantic information of Chinese clinical texts, we introduce BERT [26], an unsupervised deep bi-directional language model for obtaining the deep context representation of each word. BERT utilizes a deep bi-directional transformer encoder as its core architecture. The transformer architecture incorporates a self-attention mechanism and employs residual concatenation to mitigate network degradation, resulting in notable improvements in both training speed and model expressiveness. The sequence is BERT-encoded with a word embedding representation Q^{BERT}.

We splice the five obtained features to get the final fused representation vector:

$$Q^{fus} = Q^{BERT} \oplus Q^{pinyin} \oplus Q^{radical} \oplus Q^{POS} \oplus Q^{boundary} \tag{5}$$

3.2 Feature Extraction Layer

In order to obtain structural and semantic information of different levels of data, we simultaneously extract deep semantic features from both global and local perspectives using BiLSTM and multi-scale CNNs, respectively.

3.2.1 BiLSTM

To accurately represent the global semantic information of fusion vectors, we employ an LSTM network for feature extraction. The LSTM structure comprises three gates: the input gate, the output gate, and the forget gate. These gates allow the LSTM to select and utilize important information while handling input sequences. They enable selective storage and discarding of data, efficiently addressing the problem of gradient vanishing or exploding during the processing of lengthy text sequences.

To address the limitation of the hidden vector h_t in capturing contextual information in only one direction and learning semantic dependencies solely in a unidirectional sequence, we enhance the traditional LSTM by incorporating a BiLSTM. This modification enables better capture of semantic dependencies over longer distances. The BiLSTM utilizes contextual information from both forward and backward directions, generating two distinct semantic vectors: $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$. Finally, the hidden vectors from these two opposite directions are concatenated to obtain the complete context semantic vector $H_t = \left[\overrightarrow{h_t}; \overleftarrow{h_t}\right]$.

3.2.2 Multi-Scale CNNs

Due to the large number of clinical terms in CEMR, there may be strong correlations between neighboring characters, e.g., "胃癌" (ENG: stomach cancer) and "胃部CT" (ENG: stomach CT), the former is a disease and the latter is an examination. To capture local features between characters, multiple CNNs with varying kernel sizes are utilized to extract potential local contextual features within text sequences. This approach makes up for a deficiency in the limitations of BiLSTM, which primarily captures global features.

For multi-feature fusion sequence $Q = (Q_1, Q_2, \ldots, Q_n)$, we perform a convolution operation using multiple convolution kernels of different sizes. Each convolution kernel is of size k, which means that each convolution kernel captures the local context features between k neighboring characters. By applying multiple convolutional kernels of different sizes, we can obtain multiple sets of local context features of different sizes. The multi-scale convolutional method enhances the model's feature extraction in sequence data by capturing local semantic and structural features more effectively. The formula for the multi-scale CNNs is as follows:

$$O_t^{kn} = ReLU\left(w^T \cdot Q_{\lfloor t-\frac{k-1}{2}\rfloor : \lfloor t+\frac{k-1}{2}\rfloor} + b\right) \tag{6}$$

$$O_t = O_t^{k_1} + O_t^{k_2} + \cdots + O_t^{kn} \tag{7}$$

where $Q_{\lfloor t-\frac{k-1}{2}\rfloor : \lfloor t+\frac{k-1}{2}\rfloor}$ represents the embedding from $\left\lfloor t - \frac{k-1}{2} \right\rfloor$ to $\left\lfloor t - \frac{k+1}{2} \right\rfloor$, $ReLU$ is the activation function, O_t^{kn} represents the convolutional output with convolutional kernel size k, the "+" denotes the element summation operation, and O_t denotes the fusion feature embedding.

To improve CNER, we utilize a gate mechanism to effectively combine global and multi-scale local context semantic feature. This gate mechanism is capable of dynamically assigning weights and deciding how to utilize these features to label named entities. Its formula is as follows:

$$S_t = \sigma\left(W_{s_1} \cdot H_t + W_{s_2} \cdot O_t + b_t\right) \tag{8}$$

$$G_t = [S_t \circ H_t] \oplus [(1 - S_t) \circ O_t] \tag{9}$$

where S_t is used to evaluate the global and local contextual feature encoding, W_{s_1}, W_{s_2} are the trainable matrices, b_t is the bias term, O_t is the local context feature input, H_t is the global context feature input, and G_t is the output of the corresponding gate mechanism.

3.3 Multi-Head Attention Mechanism Layer

To better capture important features and correlations in a sequence, we employ a attention mechanism [27]. This mechanism automatically learns the distribution of attention weights at different locations and scales to enable feature selection and generate more expressive feature representations. The formula for the MHA is as follows:

$$Attention(Q,\ K,\ V) = softmax\left(\frac{QK^T}{d_k}\right)V \tag{10}$$

$$E_i = Concat(head_1, \ldots, head_n)\ W^o \tag{11}$$

$$where\ head_i = Attention\left(QW_i^Q,\ KW_i^K,\ VW_i^V\right) \tag{12}$$

where Q, K and V denote the query, key and value matrices, respectively. d_k denotes the scaling factor, which is used to adjust the range of values for the attention weights.

After obtaining specific contextual representations from multiple heads of attention, we use feed-forward neural networks (FNN) to better aggregate and encode features from different spaces.

The formula is as follows:

$$E_i = FNN(E_i) \tag{13}$$

3.4 CRF Layer

To assign labels to each character based on the final output vector, we employ CRFs for prediction. CRFs are commonly applied in tasks such as POS tagging and NER, taking advantage of their ability to model label dependencies. CRFs calculate the probability distribution of a certain random variable and utilize Viterbi's technique for decoding. This algorithm takes into account the relationships between adjacent labels to get the most effective overall label sequence.

Given an input sequence X and the corresponding hidden state sequence obtained from the model h, the conditional probability of the output label sequence Y can be computed using the definition of

CRF. The formula is as follows:

$$s(h,\ y) = \sum_{i=0}^{n} A_{y_i,\ y_{i+1}} + \sum_{i=1}^{n} P_{i,\ y_i} \tag{14}$$

$$P\left(\frac{y}{h}\right) = \frac{e^{score(h,\ y)}}{\sum_{y' \in Y(h)} e^{score\left(h,\ y'\right)}} \tag{15}$$

where A represents the transition score matrix between two labels. $A_{y_i,\ y_{i+1}}$ represents the probability of transitioning from label y_i at position i to label y_{i+1} in the sequence. $P_{i,\ y_i}$ denotes the probability of labeling position i as y_i. $P\left(\frac{y}{h}\right)$ corresponds to the normalized exponential function, and $Y(h)$ represents all possible label sequences.

4 Experiments

4.1 Datasets

We assessed our model's performance using two datasets: CCKS2017 and CCKS2019. The datasets provide an impartial evaluation of our model. Here are the dataset descriptions.

CCKS2017[1]: The dataset is a collection of CEMR released by the 2017 National Conference on Knowledge Graph and Semantic Computing and donated by Beijing Jimu Cloud Health Technology Co. (Beijing, China).The dataset comprises 1596 labeled samples, divided into 1198 samples for training and 398 samples for testing. The dataset has five categories of entities: Symptom, Disease, Check, Treatment, and Body. The statistics for each entity category are available in Table 2.

Table 2: Entity category of CCKS2017 dataset

CCKS2017	Symptom	Disease	Check	Treatment	Body
Train	7831	722	9546	1048	10,719
Test	2311	553	3143	465	3021

CCKS2019[2]: The dataset is a CEMR dataset that Yidu Cloud Technology C released as part of the 2019 National Conference on Knowledge Graph and Semantic Computing. The dataset consists of 1379 labeled samples, divided into a training set of 1000 samples and a testing set of 379 samples. The dataset contains six categories of entities: Anatomy, Disease, Exam, Medicine, Operation, and Check. The statistics for each entity category are contained in Table 3.

Table 3: Entity category of CCKS2019 dataset

CCKS2019	Anatomy	Disease	Exam	Medicine	Operation	Check
Train	1486	2116	318	456	765	222
Test	447	682	193	263	140	91

[1] CCKS2017: https://www.sigkg.cn/ccks2017 (accessed on 22/03/2024).
[2] CCKS2019: https://www.sigkg.cn/ccks2019 (accessed on 04/04/2024).

4.2 Evaluation Metrics

We employ common evaluation measures for CNER to evaluate the model's performance: precision rate (P), recall rate (R), and F1-score (F1). The formulas for each evaluation metric are as follows:

$$P = \frac{TP}{TP + FP} \tag{16}$$

$$R = \frac{TP}{TP + FN} \tag{17}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{18}$$

where *TP* denotes the count of entity types correctly predicted by the model, *FP* denotes the count of irrelevant entities predicted by the model, and *FN* denotes the count of entity types not successfully predicted by the model.

4.3 Experiment Setting

The parameter configurations utilized in this work are detailed in Table 4, encompassing a maximum word length of 128, a batch size of 16, 25 epochs per training session, and the AdamW optimization algorithm with a learning rate of 2e-5 and a dropout rate of 0.1.

Table 4: Experimental parameter settings

Parameter	Value
Maximum sequence length	256
Batch size	16
Number of epochs	25
optimizer	AdamW
Learning rate	2e-5
Dropout rate	0.1

4.4 Experiments and Analyses

4.4.1 Models Performance Comparison

This section presents a comparison between our model and other models. The results of the comparison between our model and benchmark models are presented in Table 5. Additionally, we compared our model with the latest models, as shown in Tables 6 and 7. The comparison models we selected include ELMo-lattice-LSTM-CRF [28], ACNN [23], RD-CNN-CRF [5], MKRGCN [29], MUSA-BiLSTM-CRF [24], AT-LatticeLSTM-CRF [21], FT-BERT-BiLSTM-CRF [30], ELMo-ET-CRF [31], RGT-CRF [32].

As shown in Table 5, our model exhibits excellent performance on both the CCKS2017 and CCKS2019 datasets compared to all benchmark models. For the CCKS2017 dataset, our model achieves 92.00% precision, 93.55% recall, and 92.74% F1 value. On the CCKS2019 dataset, our model achieves 89.02% precision, 86.78% recall, and 87.80% F1 value. Comparison with the benchmark model BERT-BiLSTM-MHA-CRF reveals a maximum F1 value difference of 3.31% and a minimum of 2.54%. This indicates that the BERT model, after fusing multi-feature embedding and multi-scale local contextual features, outperforms the BERT-only model in terms of feature representations, thereby validating the effectiveness of incorporating multi-feature embedding and extracting multi-scale local contextual features. Additionally, as illustrated in Tables 6 and 7, our model also outperforms the latest models. On the CCKS2017 dataset, our model increases the F1 value by 0.86% compared to the second-highest model and by 3.1% compared to the lowest model. On the CCKS2019 dataset, our model increases the F1 value by 1.11% compared to the second-highest model and by 2.78% compared to the lowest model. The excellent results shown by our model indicate that incorporating multi-feature embedding can greatly enhance the semantic representation of entities and make the model have better contextual representation, while using CNN to extract multi-scale local contextual features makes up for the shortcoming of using BiLSTM alone to extract global contextual features while ignoring local contextual features, and enhances the effectiveness of feature extraction.

Table 5: Result comparison with benchmark models

Models	CCKS2017			CCKS2019		
	P	R	F1	P	R	F1
BiLSTM-CRF	88.12	88.70	88.41	82.31	83.39	82.85
Lattice LSTM	89.01	89.65	89.32	83.20	84.86	84.02
BERT-BiLSTM-CRF	89.33	90.76	90.04	83.24	84.93	84.08
BERT-BiLSTM-MHA-CRF	89.81	90.69	90.20	84.74	84.45	84.49
Our model	**92.00**	**93.55**	**92.74**	**89.02**	**86.78**	**87.80**

Table 6: Result comparison of CCKS2017 with latest models

Models	CCKS2017		
	P	R	F1
ELMo-lattice-LSTM-CRF	90.20	90.06	90.13
ACNN	90.19	90.78	90.49
RD-CNN-CRF	90.63	92.02	91.32
AT-Lattice LSTM-CRF	88.98	90.28	89.64
FT-BERT-BiLSTM-CRF	92.06	91.15	91.60
MUSA-BiLSTM-CRF	**92.67**	90.97	91.81
MKRGCN	–	–	91.88
Our model	92.00	**93.55**	**92.74**

Table 7: Result comparison of CCKS2019 with latest models

Models	CCKS2019		
	P	R	F1
ELMo-lattice-LSTM-CRF	84.69	85.35	85.02

(Continued)

Table 7 (continued)

Models	CCKS2019		
	P	R	F1
ACNN	83.07	87.29	85.13
RGT-CRF	85.36	84.99	85.17
ELMo-ET-CRF	83.65	**87.61**	85.59
MSD-DT-NER	86.09	87.29	86.69
Our model	**89.02**	86.78	**87.80**

4.4.2 The Effect of Different Features on the Model

To investigate the impact of various features on the entity recognition performance of CEMR, we incorporated multi-feature into the BiLSTM-CRF model and carried out experiments. The impact of distinct characteristics on the model's entity recognition performance is displayed in Table 8.

Table 8: The effect of different features on the model

Models	CCKS2017			CCKS2019		
	P	R	F1	P	R	F1
Pinyin+ BiLSTM+CRF	70.22	71.05	70.63	63.11	**64.56**	63.83
Radical+BiLSTM+CRF	70.67	71.34	71.00	62.96	64.32	63.13
POS+BiLSTM+CRF	69.78	70.67	70.22	62.41	63.03	62.72
Word boundary+BiLSTM+CRF	68.33	69.76	69.04	63.24	62.93	63.08
Pinyin+Radical+POS+Word boundary+BiLSTM+CRF	**71.13**	**71.78**	**71.45**	**63.67**	64.14	**63.90**

Based on the experimental findings presented in Table 8, incorporating the pinyin feature, radical feature, POS feature, and word boundary feature into the BiLSTM-CRF model yielded comparable impacts on entity recognition, with some slight distinctions remaining. Thus, these four characteristics have a similar level of impact on the model. We combined the four feature vectors and fed them into the BiLSTM-CRF model. The model achieved the highest precision, recall and F1 values. Consequently, the values produced by combining many features in the model surpass the results achieved by using any single feature alone, demonstrating the superior usefulness of using multi-feature.

4.4.3 Impact of Different Number of Heads on the Model of the MHA

The MHA is commonly utilized in tasks like NER to capture inter-sequence relationships by employing many attention heads concurrently. The impact of the number of attention heads on model performance has not been thoroughly investigated. We conducted tests on two datasets, CCKS2017 and CCKS2019, to investigate how the number of heads impacts model efficiency.

Fig. 2 illustrates how increasing the number of attention heads at the beginning may improve the performance of the model. The model's performance peaks when the number of heads is increased to 8. Increasing the number of attention heads enhances the model's capacity to characterize complicated patterns and represent input sequences more effectively. Yet, as the number of heads increases further, the performance starts to decline. Increasing the number of attention heads results in more computational complexity, which impacts performance. We must balance performance improvement and computational complexity while selecting the number of attention heads.

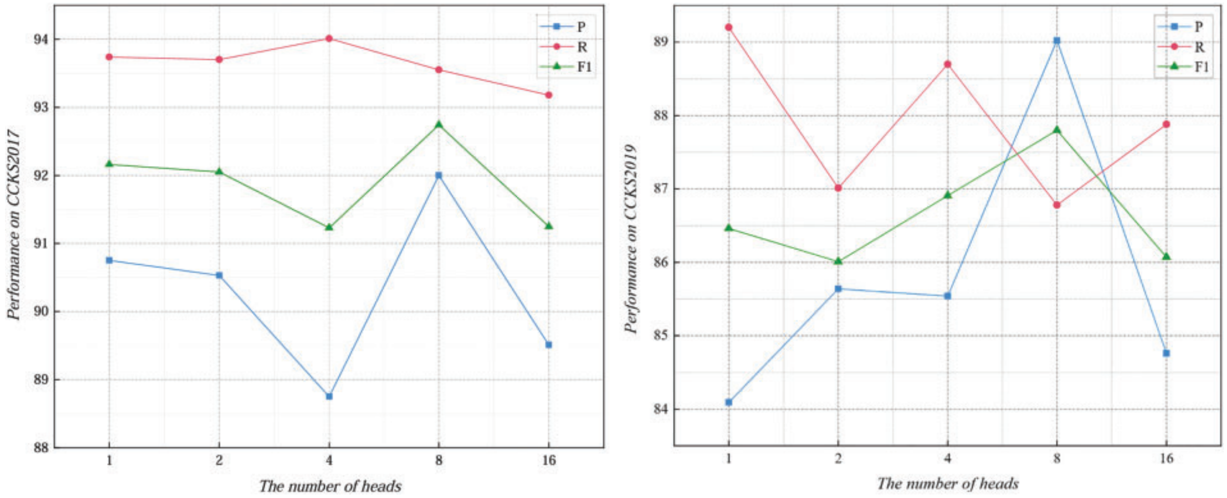

Figure 2: Impact of different head counts on CCKS2017, CCKS2019

4.4.4 Effectiveness of Multi-Scale CNNs

We performed ablation experiments on CCKS2017 and CCKS2019 datasets to assess the impact of the multi-scale CNNs module in the proposed model. The Table 9 displays the findings from the experiments conducted on CCKS2017 and CCKS2019 dataset. The results demonstrate that our entire model outperforms all others, and eliminating the multi-scale CNNs results in decreased performance. This highlights the necessity for the model to employ multi-scale local features and validates the efficiency of multi-scale CNNs.

Table 9: Impact of convolutional neural networks on the model

Models	CCKS2017			CCKS2019		
	P	R	F1	P	R	F1
Multi-feature+BiLSTM+MHA+CRF	90.18	**94.66**	92.30	85.66	**87.54**	86.59
Multi-feature+CNN+MHA+CRF	90.49	93.43	91.83	**89.20**	83.23	85.80
Our model	**92.00**	93.55	**92.74**	89.02	86.78	**87.80**

In addition, we used multiple combinations of convolutional kernels for comparison when extracting context-localized features using multi-scale CNNs. Fig. 3 displays the comparison results of several sets of convolutional kernels on CCKS2019. The precision, recall, and F1-score are highest when using 1, 3, and 5 convolutional kernels, and decrease as the window size increases. This decrease may be attributed to the loss of local contextual information when using larger convolutional kernels, leading to reduced performance. Therefore, in order to extract as many local contextual features as possible, we used convolutional kernels with three window sizes of 1, 3, and 5.

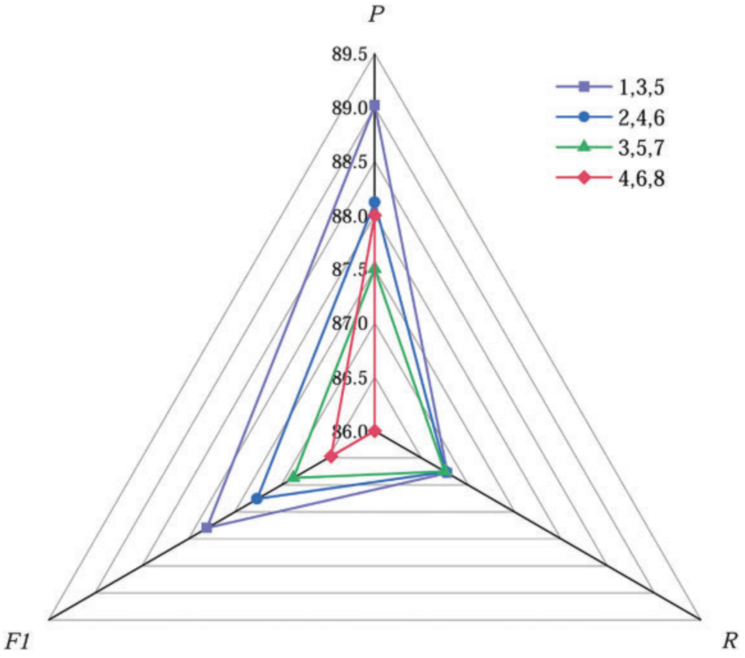

Figure 3: Comparison results of several sets of convolutional kernels on CCKS2019

4.4.5 *Separate P, R and F1 for Each Entity Category*

For a comprehensive evaluation of our model, Fig. 4 illustrates the Precision, Recall, and F1 for each entity category individually across the two datasets.

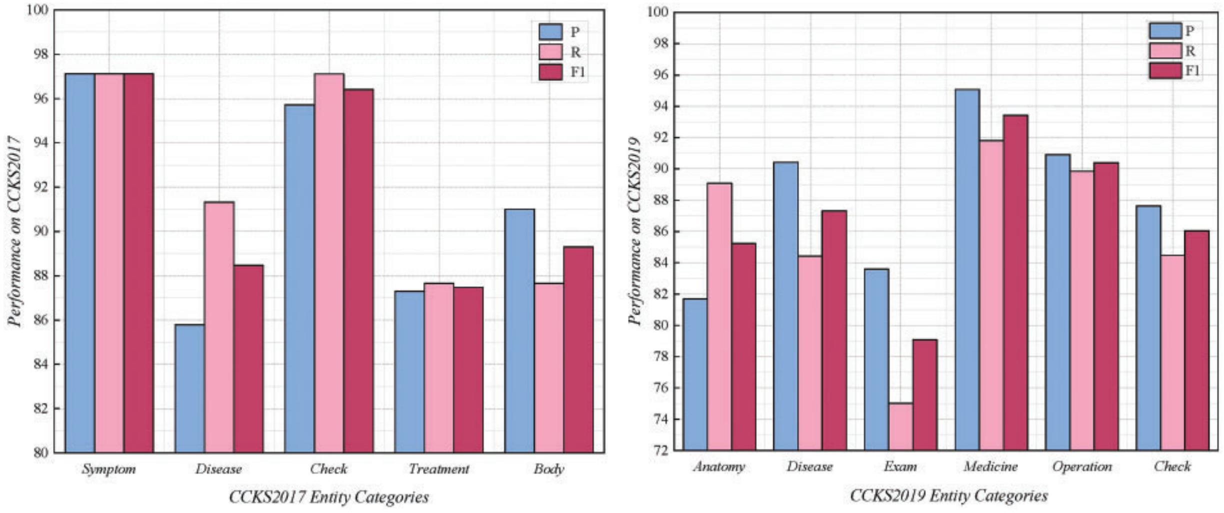

Figure 4: Precision, Recall, and F1 for each entity categories in the CCKS2017, CCKS2019

5 Conclusion and Future Work

This study introduces a CNER model that incorporates a multi-feature fusion and multi-scale local context enhancement approach. The model combines the features of pinyin, radical, POS, and word boundary while also leveraging the deep contextual representation of BERT. In addition, the fusion of multi-scale CNNs solves the limitation that the original BiLSTM can only extract global contextual features and enhances feature extraction, thus realizing a comprehensive understanding of the sentence information. Experimental assessments were carried out on two public datasets, showcasing the model's robust performance.

In future research, we will focus on exploring more effective fusion strategies and incorporating additional information from different dimensions to further improve recognition.

Acknowledgement: None.

Funding Statement: This study was supported by the National Natural Science Foundation of China (61911540482 and 61702324).

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Meijing Li, Runqing Huang; data collection: Runqing Huang; methodology: Meijing Li, Runqing Huang; analysis and interpretation of results: Xianxian Qi, Runqing Huang; writing—original draft: Runqing Huang; writing—review and editing: Meijing Li, Xianxian Qi. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Data contained in articles. Code no longer publicly available due to copyright restrictions.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] A. Dash, S. Darshana, D. K. Yadav, and V. Gupta, "A clinical named entity recognition model using pretrained word embedding and deep neural networks," *Decis. Anal. J.*, vol. 10, pp. 100426, Mar. 2024. doi: 10.1016/j.dajour.2024.100426.

[2] A. Boonstra and M. Broekhuis, "Barriers to the acceptance of electronic medical records by physicians from systematic review to taxonomy and interventions," *BMC Health Serv. Res.*, vol. 10, no. 1, pp. 231, Dec. 2010. doi: 10.1186/1472-6963-10-231.

[3] T. Wang, P. Xuan, Z. Liu, and T. Zhang, "Assistant diagnosis with Chinese electronic medical records based on CNN and BiLSTM with phrase-level and word-level attentions," *BMC Bioinform.*, vol. 21, no. 1, pp. 230, Dec. 2020. doi: 10.1186/s12859-020-03554-x.

[4] J. Lei, B. Tang, X. Lu, K. Gao, M. Jiang and H. Xu, "A comprehensive study of named entity recognition in Chinese clinical text," *J. Am. Med. Inform. Assoc.*, vol. 21, no. 5, pp. 808–814, Sep. 2014. doi: 10.1136/amiajnl-2013-002381.

[5] J. Qiu, Y. Zhou, Q. Wang, T. Ruan, and J. Gao, "Chinese clinical named entity recognition using residual dilated convolutional neural network with conditional random field," *IEEE Trans. Nanobiosci.*, vol. 18, no. 3, pp. 306–315, Jul. 2019. doi: 10.1109/TNB.2019.2908678.

[6] Q. Wang, Y. Zhou, T. Ruan, D. Gao, Y. Xia and P. He, "Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition," *J. Biomed. Inform.*, vol. 92, pp. 103133, Apr. 2019. doi: 10.1016/j.jbi.2019.103133.

[7] P. J. Gorinski *et al.*, "Named entity recognition for electronic health records: A comparison of rule-based and machine learning approaches," arXiv:1903.03985, 2019.

[8] M. Jiang *et al.*, "A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries," *J. Am. Med. Inform. Assoc.*, vol. 18, no. 5, pp. 601–606, Sep. 2011. doi: 10.1136/amiajnl-2011-000163.

[9] Y. Hu *et al.*, "Improving large language models for clinical named entity recognition via prompt engineering," *J. Am. Med. Inform. Assoc.*, vol. 13, pp. ocad259, Jan. 2024. doi: 10.1093/jamia/ocad259.

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. doi: 10.1162/neco.1997.9.8.1735.

[11] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," arXiv:1508.01991, 2015.

[12] J. Yin, S. Luo, Z. Wu, and L. Pan, "Chinese named entity recognition with character level BLSTM and soft attention model," *J. Beijing Instit. Technol.*, vol. 29, no. 1, pp. 1520–1532, 2020. doi: 10.1109/TASLP.2020.2994436.

[13] Z. Tang, B. Wan, and L. Yang, "Word-character graph convolution network for Chinese named entity recognition," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 1520–1532, 2020. doi: 10.1109/TASLP.2020.2994436.

[14] Z. Zhu, J. Li, Q. Zhao, and F. Akhtar, "A dictionary-guided attention network for biomedical named entity recognition in Chinese electronic medical records," *Expert. Syst. Appl.*, vol. 231, pp. 120709, Nov. 2023. doi: 10.1016/j.eswa.2023.120709.

[15] M. Awad and R. Khanna, *Hidden Markov Model, in Efficient Learning Machines*. Berkeley, CA, USA: Apress, pp. 81–104, 2015.

[16] J. D. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Int. Conf. Mach. Learn.*, San Francisco, CA, USA, 2001, pp. 282–289.

[17] T. Wang *et al.*, "A hybrid model based on deep convolutional network for medical named entity recognition," *J. Electr. Comput. Eng.*, vol. 2023, pp. 1–11, May 2023. doi: 10.1155/2023/8969144.

[18] Y. Zhang and J. Yang, "Chinese NER using lattice LSTM," in *Proc. Annu. Meet. Assoc. Comput. Linguist.*, Melbourne, Australia, 2018, pp. 1554–1564.

[19] F. Wu, J. Liu, C. Wu, Y. Huang, and X. Xie, "Neural Chinese named entity recognition via CNN-LSTM-CRF and joint training with word segmentation," in *Proc. World Wide Web Conf.*, San Francisco, CA, USA, 2019, pp. 3342–3348.

[20] K. Xue, Y. Zhou, Z. Ma, T. Ruan, H. Zhang and P. He, "Fine-tuning BERT for joint entity and relation extraction in Chinese medical text," in *2019 IEEE Int. Conf. Bioinform. Biomed.*, San Diego, CA, USA, 2019, pp. 892–897.

[21] S. Zhao, Z. Cai, H. Chen, Y. Wang, F. Liu and A. Liu, "Adversarial training based lattice LSTM for Chinese clinical named entity recognition," *J. Biomed. Inform.*, vol. 99, pp. 103290, Nov. 2019. doi: 10.1016/j.jbi.2019.103290.

[22] D. Li, J. Long, J. Qu, and X. Zhang, "Chinese clinical named entity recognition with ALBERT and MHA mechanism," *Evid. Based Complement. Alternat. Med.*, vol. 2022, pp. 1–9, May 2022. doi: 10.1155/2022/2056039.

[23] J. Kong, L. Zhang, M. Jiang, and T. Liu, "Incorporating multi-level CNN and attention mechanism for Chinese clinical named entity recognition," *J. Biomed. Inform.*, vol. 116, pp. 103737, Apr. 2021. doi: 10.1016/j.jbi.2021.103737.

[24] 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," *Artif. Intell. Med.*, vol. 127, no. C, pp. 102282, May 2022. doi: 10.1016/j.artmed.2022.102282.

[25] S. Guo, W. Yang, L. Han, X. Song, and G. Wang, "A multi-layer soft lattice based model for Chinese clinical named entity recognition," *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, pp. 201, Dec. 2022. doi: 10.1186/s12911-022-01924-4.

[26] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Conf. N. Am. Chapter Assoc. Comput. Linguist.: Hum. Lang. Technol.*, Minneapolis, MN, USA, 2019, pp. 4171–4186.

[27] A. Vaswani *et al.*, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 6000–6010.

[28] Y. Li *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.*, vol. 8, no. 9, pp. e19848, Sep. 2020. doi: 10.2196/19848.

[29] Y. Xiong *et al.*, "Leveraging multi-source knowledge for Chinese clinical named entity recognition via relational graph convolutional network," *J. Biomed. Inform.*, vol. 128, pp. 104035, Apr. 2022. doi: 10.1016/j.jbi.2022.104035.

[30] X. Li, H. Zhang, and X. H. Zhou, "Chinese clinical named entity recognition with variant neural structures based on BERT methods," *J. Biomed. Inform.*, vol. 107, pp. 103422, Jul. 2020. doi: 10.1016/j.jbi.2020.103422.

[31] Q. Wan *et al.*, "A self-attention based neural architecture for Chinese medical named entity recognition," *MBE*, vol. 17, no. 4, pp. 3498–3511, 2020. doi: 10.3934/mbe.2020197.

[32] J. Li, R. Liu, C. Chen, S. Zhou, X. Shang and Y. Wang, "An RG-FLAT-CRF model for named entity recognition of Chinese electronic clinical records," *Electronics*, vol. 11, no. 8, pp. 1282, Apr. 2022. doi: 10.3390/electronics11081282.