Mathematical Biosciences and Engineering

MBE, 19(3): 2206–2218.
DOI: 10.3934/mbe.2022103
Received: 17 October 2021
Revised: 05 December 2021
Accepted: 15 December 2021
Published: 04 January 2022

http://www.aimspress.com/journal/MBE

Research article

Entity recognition of Chinese medical text based on multi-head self-attention combined with BILSTM-CRF

Chaofan Li and Kai Ma*

School of Medical Information and Engineering, Xuzhou Medical University, Jiangsu 221004, China

* **Correspondence:** Email: cumtbmakai@126.com.

Abstract: Named entities are the main carriers of relevant medical knowledge in Electronic Medical Records (EMR). Clinical electronic medical records lead to problems such as word segmentation ambiguity and polysemy due to the specificity of Chinese language structure, so a Clinical Named Entity Recognition (CNER) model based on multi-head self-attention combined with BILSTM neural network and Conditional Random Fields is proposed. Firstly, the pre-trained language model organically combines char vectors and word vectors for the text sequences of the original dataset. The sequences are then fed into the parallel structure of the multi-head self-attention module and the BILSTM neural network module, respectively. By splicing the output of the neural network module to obtain multi-level information such as contextual information and feature association weights. Finally, entity annotation is performed by CRF. The results of the multiple comparison experiments show that the structure of the proposed model is very reasonable and robust, and it can effectively improve the Chinese CNER model. The model can extract multi-level and more comprehensive text features, compensate for the defect of long-distance dependency loss, with better applicability and recognition performance.

Keywords: clinical named entity recognition; electronic medical records; multi-head self-attention; bi-directional long-short term memory; conditional random fields

1. Introduction

The application of Hospital Information System (HIS) has accelerated the development of medical informatization, which carries a huge amount of Electronic Medical Records (EMR) containing patients' clinical treatment information. Facing the medical clinical information of

unstructured text storage type, how to perform effective data mining and knowledge discovery is the research focus of Natural Language Processing (NLP) in the medical field.

The main tasks of NLP include lexical analysis, sentence analysis, semantic analysis [1], information extraction [2] and some high-level tasks, such as human-robot language interaction [3,4], text summarization [5], dialogue system [6], etc. Named Entity Recognition, as one of the key fundamental tasks in NLP research, was first proposed in 1996 by R. Grishman at the MUC-6 conference [7]. It aims to recognize entities with specific meanings from unstructured text, such as proper names of people, places and organizations, and to make a basis for implementing tasks such as relationship extraction and knowledge mapping. NER early on was mainly based on rule and dictionary approaches, but it required a lot of effort for manual annotation and was less applicable [8,9]. Based on machine learning models including Maximum Entropy Models (MEM) [10], Hidden Markov Models (HMM) [11], Support Vector Machines (SVM) [12], Conditional Random Fields, CRF) [13], etc. Although they have good flexibility and robustness, they have the disadvantages of requiring a large number of annotated sets and over-dependence on the correctness of feature selection. With the emergence of deep learning neural network technologies, manual pre-processing of data is gradually abandoned for automated feature extraction [14]. Compared with the traditional approaches based on statistical rules or machine learning algorithms, it has the advantages of avoiding the high cost and complexity of manual feature extraction, reducing the dependence on word segmentation, and enhancing model generalization. Electronic Medical Records contain a series of important information closely related to patients' health status, such as their diseases, symptoms, examinations and treatments. Since unstructured electronic medical records contain a large number of special entities such as medical terminology and proper noun abbreviations, they need to rely highly on contextual information to extract the entities accurately. R. Collobert [15] was the first to use deep neural networks for named entity recognition. G. Lample [16] used a model combining BILSTM and CRF to achieve better results in NER tasks, and the BILSTM model has become the mainstream model architecture with its full consideration of contextual information over longer distances. Zhang [17] added a bi-directional GRU layers on the basis to form a deeply stacked neural network, which solves the problem of text representation where single-layer neural network encoding cannot capture depth features. Although the BILSTM-CRF based model can effectively establish contextual association information, it does not take into account the importance of different words and characters in the sentence, and there is still room for further improvement of recognition results [18]. Meanwhile, the study of named entity recognition of electronic medical record texts is different from the general field, and there are significant differences in the text structure of both. The sequence modeling of electronic medical record texts mainly includes serious drawbacks such as incomplete syntactic components of sentences, a large number of specialized terms, nested entities, and the existence of dependency relationships with entities far apart.

The introduction of the attention mechanism [19] effectively solves the problem of contextual information modeling by BILSTM. The attention mechanism combined with a deep neural network is used to assign weights to the output of the hidden layer [20,21], which effectively improves the effectiveness of the model. The self-attention [22] is a special kind of attention mechanism that calculates the association relationships between characters at different positions of a text sequence in order to obtain an interactive representation of the sequence. The multi-head self-attention obtains information about sequences in different subspaces by combining multiple parallel self-attention, thus allowing a more comprehensive text features to be obtained.

In addition to the above, unlike the English syntactic structure, there are two approaches for Chinese named entity recognition based on characters [23] and words [24]. The character-based approach reduces the influence of unfamiliar words, but individual characters contain insufficient semantic information and are generally represented in combination by adding pinyin, radicals, etc [25]. The word-based approach firstly faces the problem of accurate word separation, especially for special fields, such as EMR containing many intensive terms, and the accuracy of word separation directly affects the effectiveness of the model. Based on the combination of characters and words in the form of concatenation as input to improve the recognition effect of a single model, it still cannot effectively solve the ambiguity and diversity of Chinese word separation [26].

1.1. Motivation

A summary of the mentioned deep learning models for the CNER task reveals that the delineation of entity boundaries is closely related to individual characters and the effect of word segmentation. Whether individual characters or words are used as text feature inputs, the common features of characters and words in a text sequence are not sufficiently considered. How to perform an effective embedding method can directly affect the effectiveness of the CNER model. In the case of small sample size, the model cannot learn the corresponding linguistic knowledge well. Therefore, when using word vector tools for vectorized representation of text, the word vectors do not obtain a rich semantic representation. At the same time, the current BILSTM-CRF based models for feature extraction do not have the ability to highlight key features because it focuses on all features of the text, so it will cause feature redundancy in sequence modeling and reduce the performance of the model.

1.2. Proposed solution

In view of the above problems of the Chinese CNER model, we propose a method based on multi-head self-attention mechanism combined with BILSTM-CRF. The proposed method takes advantage of unsupervised learning to learn semantic vector representations that are more complete and closer to the true distribution on a large-scale corpus. Also, a combination of characters and words is utilized to obtain a more comprehensive representation of word embeddings. The model follows the parallel structure of multi-head self-attention mechanism module and BILSTM neural network module to learn the multi-level features of semantic association information of clinical text and capture the internal structural features and dependencies of sentences. The association weights of text features are obtained from multiple perspectives, and the highlighting of key text features is carried out to reduce the influence of redundant features on sequence modeling, so as to improve the effectiveness of Chinese clinical named entity recognition model.

2. Materials and methods

The CNER model based on multi-head self-attention combined with BILSTM-CRF (MHSA-BILSTM-CRF) is shown in Figure 1 below, which contains four main layers: Input Layer, Neural Network Layer, CRF Layer, and Output Layer. Firstly, a medical terminology dictionary is loaded, and then character vectors and word vectors are trained using word2vec on large-scale unlabeled medical record texts in the same field. The vector representation of the training data is performed using pre-

trained language model, which is then input to the BILSTM module and the multi-head self-attention module, respectively. Among them, the BILSTM module is used to learn the temporal features and contextual information of the text sequences, and the self-attention is used to obtain the global features of the text sequences and the association strength between words. Finally, the outputs of the two neural network modules are concatenated and fed to the CRF layer for sequence labeling, and the entity prediction labels are output.

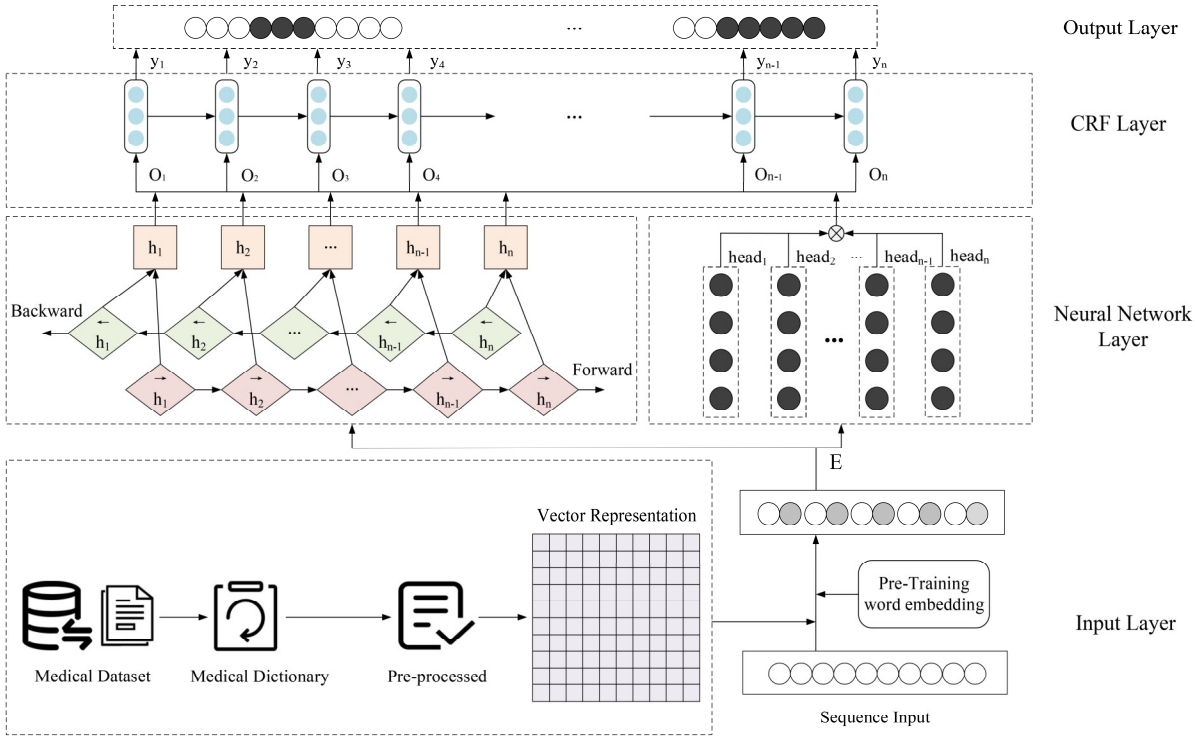

Figure 1. The overall model architecture of Chinese CNER.

2.1. Embedding layer

Data pre-processing operations such as word segmentation, removal of deactivated words and low frequency words are performed on the raw medical record text. The accuracy of word separation directly affects the effectiveness of the model, so a medical dictionary is cited to help improve the effectiveness of word separation. The word2vec can train the low-dimensional dense word vector, which reflects the relationship between words, but it cannot solve the problem of occurrence of unfamiliar words [3]. Although the character-based embedding approach is better, the semantic distance between the semantics of a single character and the semantics of a Chinese word is far from the same.

In order to effectively improve the problem of using a single form for representation, we used the word2vec to train character vectors and word vectors separately on a large-scale unlabeled corpus in the same field as the training set, defaulting the skip-gram model, making them more complete to learn the semantic vector representation. Set the input text sequence $S = (w_1, w_2, w_3, \ldots, w_n)$, then word vector lookup table for vectorized representation can be expressed as:

$$word\_emb_i = v_i \oplus ((p_{i1} + p_{i2} + \cdots p_{im}) / m), \forall i \in input \cap vec_{word} \cap vec_{char} \quad (1)$$

where, the randomly initialized word vector matrix of the training set is denoted as *word_emb*. w_i is the *i-th* word in *S*, w_i contains *m* characters, *input* indicates the word list of the training set, vec_{word} and vec_{char} represent words and characters of the large-scale pre-trained corpus, respectively. v_i is the word vector corresponding to w_i in vec_{word}, p_{ij} is the vector of each character contained in the word w_i in vec_{char}.

The text sequence $S = (w_1, w_2, w_3, ..., w_n)$ is transformed by *word_emb* to obtain the output matrix $E = (e_1, e_2, e_3, ..., e_n)$ of the embedding layer, e_i represents the word vector of w_i. The introduction of pre-trained language model can take advantage of unsupervised learning to learn a more complete and close semantic vector representation of the target characters and words to the true distribution on a large scale corpus. The CNER model uses *word_emb* for word vector mapping, which is based on the effective combination of characters and words, and then the word vector sequence is used as the input to the neural network for training.

2.2. Neural network layer

In order to better learn the contextual association information and the highlighting of key information, two parallel neural network modules are set up after the Embedding Layer: the BILSTM neural network module and the multi-head self-attention mechanism module. The output matrix $E = (e_1, e_2, e_3, ..., e_n)$ of the embedding layer will be input to the two modules separately for training, and then the respective output of the modules will be concatenated to obtain the final output of the neural network layer.

2.2.1. BILSTM module

LSTM effectively overcomes the gradient dispersion problem generated by RNN when processing long sequences through the gating mechanism, and is therefore suitable for text sequence annotation tasks. For the moment *t*, the input of the LSTM unit includes the current moment input vector e_t, the previous moment storage cell information c_{t-1} and the previous moment hidden layer output information h_{t-1}. The specific implementation of the LSTM unit is as follows:

$$i_t = \sigma(W_i e_t + U_i h_{t-1} + b_i) \quad (2)$$

$$f_t = \sigma(W_f e_t + U_f h_{t-1} + b_f) \quad (3)$$

$$\overline{c_t} = \tanh(W_c e_t + U_c h_{t-1} + b_c) \quad (4)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \overline{c_t} \quad (5)$$

$$o_t = \sigma(W_o e_t + U_o h_{t-1} + b_o) \quad (6)$$

$$h_t = o_t \odot \tanh(c_t) \tag{7}$$

where σ is the *sigmoid* function, W_i, W_f, W_c, $W_o \in R^{d_h \times d_e}$ are the weight matrix on the input vector e_t, U_i, U_f, U_c, $U_o \in R^{d_h \times d_e}$ are the weight matrix on the hidden layer state vector h_{t-1}, and b_i, b_f, b_c, $b_o \in R^{d_h}$ are the bias vector. i, o, $f \in R^{d_h}$ represent the input gate, output gate and forget gate, respectively.

Finally, the splicing of the output vectors of the forward and backward LSTM units is performed, and the feature vectors with bidirectional semantics are output as the output of the BILSTM neural network layer.

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

2.2.2. Multi-head self-attention module

BILSTM cannot display the importance of key information in the context during the calculation process, and it will lose important information for named entity recognition tasks when processing long sequence tasks. The attention mechanism was first applied to machine translation tasks for natural language processing [19], and subsequently self-attention was also applied to learn textual information representation [22]. The self-attention can be used to learn the dependencies between any two words in a sentence and to obtain internal structural information. Self-attention converts the input matrix E of embedding into three matrices with all dimensions d_k through different mapping operations: Q (Query), K (Key), and V (Value). Self-attention maps a set of Query and Key-Value to the output to obtain a weighted summation result for V, where the weight assigned to each V is calculated by the similarity function of the Q to the corresponding K:

$$Attention(Q, K, V) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{9}$$

where d_k denotes the vector dimension of the matrix Q, K, V, $soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)$ is the attention weighting score.

In the multi-head self-attention module, Q, K, and V respectively use h self-attention mechanisms for linear mapping to generate different weight matrices that represent the unique feature information of the word vector in different subspaces. Thus, the ability of the model to focus on different locations is extended. The input features are linearly mapped to different information subspaces through different weight matrices, and the same attention function calculation is done in each subspace to fully learn the structure and semantics of the sentence, where the $head_i$ calculation process is as follows:

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{10}$$

Multiple self-attention base units are stacked to form multi-head self-attention encoding module. Thus, the results of *h* parallel self-attention are combined and then linearly mapped once to obtain the final output *Att*.

$$Att = MultiHead(Q,K,V) = Concat(head_1, head_2, \cdots, head_h)W^O \tag{11}$$

where, W_i^Q, W_i^K, W_i^V, W^O represent the parameter matrix of the linear mapping.

The final output of the neural network layer O_t is calculated from h_t and Att_t as follows:

$$O_t = \tanh(h_t \oplus Att_t) \tag{12}$$

2.3. CRF layer

Compared with the judgment of the entity category by the *softmax* function, CRF can consider the dependency relationship between tags in the sentence-level sequence labeling, so as to obtain the global optimal labeling sequence [27]. Assuming a prediction sequence $y=\{y_1,y_2,\ldots,y_n\}$, the prediction probability of the model for sequence *y* is jointly determined by the output matrix *O* of the neural network layer and the CRF transition probability matrix *P*. The form of the CRF tagging process is as follows:

$$score(w,y) = \sum_{i=1}^{n} O_{i,y_i} + \sum_{i=1}^{n+1} P_{y_{i-1},y_i} \tag{13}$$

where, $O_{i,yi}$ indicates the probability that the *i-th* word is marked as y_i, P_{ij} indicates the probability of transfer from label *i* to label *j*.

The probability of generating a labeled sequence *y* conditional on the sentence *S* is:

$$p(y \mid S) = \frac{\exp(score(w,y))}{\sum_{y' \in Y_w} \exp(score(w,y'))} \tag{14}$$

The probability that the labeled result is correct is calculated during the training process using the maximum likelihood method.

$$\log(p(y \mid S)) = score(w,y) - \log(\sum_{y' \in Y_w} \exp(score(w,y'))) \tag{15}$$

The Viterbi algorithm is used in the prediction process to decode and select the best sequence of labels that maximizes the objective function:

$$y^* = \arg\max_{y'} score(w,y') \tag{16}$$

3. Results

3.1. Experimental data and labeling

The experimental dataset is a medical named entity recognition task for China Conference on Knowledge Graph and Semantic Computing (CCKS2019) that was conducted for public evaluation. The competition organizing committee provided 1000 high-quality data sets that were manually labeled, and defined six categories, including disease and diagnosis, imaging inspection, laboratory examination, surgery, medicine, and anatomy. The dataset for pre-training the language model was mainly derived from the unlabeled raw electronic medical records of the Ai'aiyi medical website and CCKS2017. The original dataset is annotated with characters using the BIO annotation method, "B-" for the first character of entity, "I-" for non-first character of entity, and "O" for non-entity. Based on this, entity categories are added in the form of "-type" for entity classification during entity annotation. The dataset was randomly divided into training, validation and test in the ratio of 8:1:1 for the experiments.

3.2. Evaluation indicators

The entity recognition model uses precision, recall and F-score as evaluation metrics. The precision rate can indicate the proportion of positive samples that are truly positive in the prediction results. Recall can indicate the proportion of positive samples from standard answers that are correctly predicted. The F-score is the summed average of precision and recall, and is a balanced measure of both precision and recall.

$$precision = \frac{correct}{recognized} \tag{17}$$

$$recall = \frac{correct}{entities} \tag{18}$$

$$F-score = \frac{2 \times precision \times recall}{precision + recall} \tag{19}$$

where, *correct* indicates the number of entities marked correctly, *recognized* indicates the total number of entities labeled, and *entities* is the total number of entities included in the standard answer.

3.3. Experimental results

The experimental environment was Windows 10, Python version 3.6, Keras 2.2.5, and TensorFlow 1.14.0. The parameters are optimized by the Adam algorithm during the training process, and the early-stop strategy is used to prevent the model from overfitting. After several experiments, the best hyperparameter of the model is that the character vertor and word vector are 100-dimensional, the initial learning is 0.01, the dropout rate is 0.3, the number of hidden units is 128, the number of heads for multi-head self-attention is 100, the batch size is 16, and the max epoch is 30.

In order to demonstrate the effectiveness of the proposed model, and to explore the effects of the model structure, embedding methods and attention mechanism on the NER model, 10 sets of comparison experiments are adequately set up for analysis. The models subscripted by "char" use character vectors, subscripted by "word" use word vectors, subscripted by "non-pretrained" is not using pre-trained language model, and the other embedding forms are referred to Section 2.1. The effects of each type of models are shown in Table 1, the effects of four different groups of model structures for entity category recognition are shown in Figure 2.

Table 1. Performance of NER for each type of model architecture.

No.			Precision	Recall	F-score
Ⅰ	Character Level	BILSTM-CRF$_{char}$	0.7664	0.7509	0.7586
Ⅱ		BILSTM-CRF-SA$_{char}$	0.7795	0.7716	0.7755
Ⅲ		MHSA-BILSTM-CRF$_{char}$	0.8211	0.8086	0.8148
Ⅳ	Word Level	BILSTM-CRF$_{word}$	0.7302	0.7462	0.7381
Ⅴ		BILSTM-CRF-SA$_{word}$	0.7679	0.7641	0.7660
Ⅵ		MHSA-BILSTM-CRF$_{word}$	0.8096	0.7998	0.8047
Ⅶ	Character-Word Fusion	BILSTM-CRF-MHSA$_{non-pretrained}$	0.7988	0.8003	0.7995
Ⅷ		BILSTM-CRF-MHSA	0.8163	0.8017	0.8089
Ⅸ		MHSA-BILSTM-CRF$_{non-pretrained}$	0.8104	0.8087	0.8094
Ⅹ		MHSA-BILSTM-CRF	0.8379	0.8212	0.8295

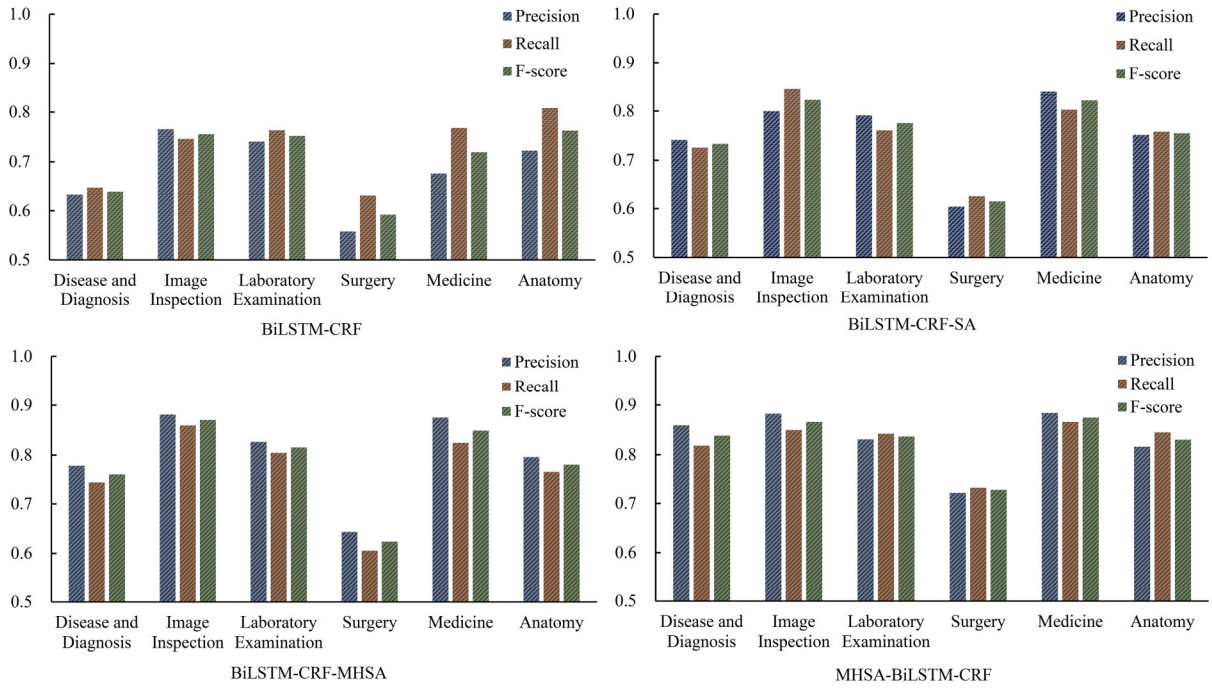

Figure 2. Recognition effect of entity categories.

4. Discussions

The MHSA-BILSTM-CRF proposed in this paper achieves the most excellent performance in the

results of several comparative experiments. The embedding layer uses a pre-trained language model for fusion of character vectors and word vectors, and the combined vectors are trained by the parallel structure of multi-head self-attention and BILSTM neural network, and then use CRF for sequence labeling. The overall F-score of the model reached 82.95%, which proved the effectiveness of the model for CNER tasks.

Comparing with embedding methods, including No.I-VI, it is obvious that the character-based embedding method has better performance than the word-based embedding method, mainly because the character-based overcomes the problems of deviations and unfamiliar words in the word separation [26]. Especially in medical texts, the entities are usually long and can cause ambiguity in word segmentation. The medical term dictionary that introduces prior knowledge can effectively improve the word segmentation effect, thereby improving the overall performance of the model.

Whether based on the form of character embedding or word embedding, it can be seen from No. I-II and IV-V that the introduction of the self-attention largely compensates for the lack of LSTM's ability to capture the association relationships between words when dealing with long sequences, and can capture a variety of semantic features and highlight key information from the character, word and sentence levels. Thus, the overall effect of the self-attention on the model is improved by about 1.69 and 2.79% at the character level and word level, respectively. The multi-head self-attention can capture the association relations between any position in the sentence, making it easier to learn the context-dependent information [28]. The attention mechanism uses the weight summation method to generate the output vector, making the propagation of the gradient in the network model easier than RNN and CNN. In addition, the multi-head self-attention is more capable of parallel execution and has a faster training speed. By selectively focusing on certain important information while ignoring other minor information, and assigning weights according to importance, more correlation weight features between clinical text words are obtained. Therefore, both No.VII-VIII of the series structure and No.III,VI,IX-X of the parallel structure have a greater improvement compared to the single self-attention, and the maximum improvement of F-score is about 4.29 and 6.35%, respectively.

As can be seen from No.VII-X, the model effect is effectively improved by introducing the pre-trained language model, with 0.94% improvement on the series form and 2.01% improvement on the parallel form. The introduction of the pre-training language model can perform unsupervised learning on the unlabeled medical field corpus to obtain the semantic vector representation of different characters or words [25], which can solve the problem of the lack of labeled data in the medical field to a certain extent. In addition, the pre-trained language model introduces external knowledge with a smaller probability of falling into a local optimum, which improves the generalization ability of the model. As shown in Figure 2, the recognition of imaging inspection and medicine is relatively good among the six types of entities. The application of the model proposed in this paper resulted in the largest recognition improvement in disease and diagnosis, nearly 19.8%, and the smallest in anatomy, only about 7.5%. For long entities such as surgery, the recognition effects of various models are relatively common. From the experimental results, the attention mechanism effectively improves the entity recognition for each category.

The proposed MHSA-BILSTM-CRF achieved 83.79, 82.12, and 82.95% in the precision, recall, and F-score, respectively. Compared with No. VIII, the overall performance has increased by about 2.05%, which clearly demonstrates the complementary effect of the parallel structure of the multi-head self-attention and the BILSTM neural network during data processing, and avoids the cumulative propagation of errors caused by the loss of information when BILSTM performs long sequence

processing. As a result, multi-head self-attention and BILSTM-CRF are combined very effectively and enhance the overall performance of CNER.

5. Conclusions

CNER is one of the most critical subtasks for NLP applications in the medical field. For the problem that the entity recognition model of Chinese EMR cannot effectively extract global features and association weight information within the text sequence. This paper proposes an embedding method that introduces a pre-trained language model and combines medical dictionary for the fusion of character vectors and word vectors, with a parallel structure of multi-head self-attention and BILSTM neural network for training the combined vector to obtain the feature representation. This method obtains the multi-level semantic feature information and the weight of the association relationship of the medical record text, thereby effectively improving the performance of the CNER. However, there are some limitations of the approach. The dataset used is high-quality labeled data from academic assessments, while the performance on other datasets still needs further validation. We will use other datasets in the field to test the scalability and generalization capabilities of the model in the future work.

Data availability statement

The datasets used to support the findings of this study are available from the biendata competitions website. (https://www.biendata.xyz/competition/ccks_2019_1/, https://www.biendata.xyz/competition/CCKS2017_2/). In addition, the datasets are available from the corresponding author upon request.

Acknowledgments

This research is supported by Key R & D Program for Xuzhou Science and Technology Plan Project, with granted number KC21308. In addition, it is also supported by Innovation and Entrepreneurship Project for University Students in Jiangsu Province, under granted No. 201810313047Y and 201910313004Z.

Conflict of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

1. W. Lu, Y. Zhang, S. Wang, H. Huang, Q. Liu, S. Luo, Concept representation by learning explicit and implicit concept couplings, *IEEE Intell. Syst.*, **36** (2021), 6–15. https://doi.org/10.1109/MIS.2020.3021188.
2. N. Zhang, H. Ye, S. Deng, C. Tan, M. Chen, S. Huang, et al., Contrastive information extraction with generative transformer, *IEEE/ACM Trans. Audio Speech Lang. Process.*, **29** (2021), 3077–3088. https://doi.org/10.1109/TASLP.2021.3110126.

3. R. Yu, W. Lu, H. Lu, S. Wang, F. Li, X. Zhang, et al., Sentence pair modeling based on semantic feature map for human interaction with IoT devices, *Int. J. Mach. Learn. Cybern.*, **12** (2021), 3081–3099. https://doi.org/10.1007/s13042-021-01349-x.
4. W. Lu, R. Yu, S. Wang, C. Wang, P. Jian, H. Huang, Sentence semantic matching based on 3D CNN for human–robot language interaction, *ACM Trans. Internet Technol.*, **21** (2021), 1–24. https://doi.org/10.1145/3450520.
5. M. Mohd, R. Jan, M. Shah, Text document summarization using word embedding, *Expert Syst. Appl.*, **143** (2020), 112958. https://doi.org/10.1016/j.eswa.2019.112958.
6. W. Li, W. Shao, S. Ji, E. Cambria, BiERU: Bidirectional emotional recurrent unit for conversational sentiment analysis, *Neurocomputing*, **467** (2022), 73–82. https://doi.org/10.1016/j.neucom.2021.09.057.
7. R. Grishman, B. Sundheim, Message understanding conference-6: a brief history, in *Proceedings of the 16th conference on Computational linguistics*, **1** (1996), 466−471. https://doi.org/10.3115/992628.992709.
8. M. Collins, Y. Singer, Unsupervised models for named entity classification, in *Proceedings of the Joint SIGDA T Conference on Empirical Methods in Natural Language Processing and V ery Large Corpora*, (1999), 100–110.
9. A. Abbas, E. Varoglu, N. Dimililer, ChemTok: A new rule based tokenizer for chemical named entity recognition, *BioMed Res. Int.*, (2016), 1–9. https://doi.org/10.1155/2016/4248026.
10. A. Ratnaparkhi, A Maximum entropy model for part-of-speech tagging, in *Conference on Empirical Methods in Natural Language Processing*, (1996), 133–142. https://aclanthology.org/W96-0213.
11. A. Borthwick, *A Maximum Entropy Approach to Named Entity Recognition*, US, New York University, 1999.
12. N. Cristianini, J. S. Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, UK, Cambridge University Press, 2000. https://doi.org/10.1017/CBO9780511801389.
13. A. Mccallum, W. Li, Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons, in *Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL*, **4** (2003), 188–191. https://doi.org/10.3115/1119176.1119206.
14. Q. Pan, C. Huang, D. Chen, A method based on multi-standard active learning to recognize entities in electronic medical record, *Math. Biosci. Eng.*, **18** (2021), 1000–1021. https://doi.org/10.3934/mbe.2021054.
15. R. Collobert, J. Weston, A unified architecture for natural language processing: deep neural networks with multitask learning, in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, (2008), 160–167. https://doi.org/10.1145/1390156.1390177.
16. G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (2016), 260–270. https://doi.org/ 10.18653/v1/N16-1030.
17. R. Y. Zhang, W. P. Lu, S. J. Wang, X. P. Peng, R. Yu, Y. Gao, Chinese clinical named entity recognition based on stacked neural network, *Concurrency Comput.: Pract. Exper.*, **33** (2021). https://doi.org/10.1002/cpe.5775.

18. N. Deng, H. Fu, X. Chen, Named entity recognition of traditional Chinese medicine patents based on BILSTM-CRF, *Wireless Commun. Mobile Comput.*, **2021** (2021). https://doi.org/10.1155/2021/6696205.
19. D. Bahdanau, K. Cho, Y. Bengio. Neural machine translation by jointly learning to align and translate, preprint, arXiv: 1409. 0473.
20. B. Z. Tang, X. L. Wang, J. Yan, Q. C. Chen, Entity recognition in Chinese clinical text using attention-based CNN-LSTM-CRF, *BMC Med. Inf. Decis. Making*, **19** (2019), 74. https://doi.org/10.1186/s12911-019-0787-y.
21. L. Luo, Z. H. Yang, P. Yang, Y. Zhang, L. Wang, H. F. Lin, et al., An attention-based BILSTM-CRF approach to document-level chemical named entity recognition, *Bioinformatics*, **34** (2018), 1381–1388. https://doi.org/10.1093/bioinformatics/btx761.
22. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., Attention is all you need, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, (2017), 6000–6010.
23. J. H. Qiu, Y. M. Zhou, Q. Wang, T. Ruan, J. Gao, Chinese clinical named entity recognition using residual dilated convolutional neural network with conditional random field, *IEEE Trans. NanoBioscience*, **18** (2019), 306–315. https://doi.org/10.1109/TNB.2019.2908678.
24. Z. Z. Li, Q. Zhang, Y. Liu, D. W. Feng, Z. Huang, Recurrent neural networks with specialized word embedding for Chinese clinical named entity recognition, in *Proceedings of the Evaluation Task at the China Conference on Knowledge Graph and Semantic Computing*, Berlin, German, Springer, (2017), 55–60.
25. X. M. Han, F. Zhou, Z. Y. Hao, Q. M. Liu, Y. Li, Q. Qin, MAF-CNER: A Chinese named entity recognition model based on multifeature adaptive fusion, *Complexity*, 2021. https://doi.org/10.1155/2021/6696064.
26. N. Ye, X. Qin, L. L. Dong, X. Zhang, K. K. Sun, Chinese named entity recognition based on character-word vector fusion, *Wireless Commun. Mobile Comput.*, 2020. https://doi.org/10.1155/2020/8866540.
27. C. Che, C. J. Zhou, H. Y. Zhao, B. Jin, Z. Gao, Fast and effective biomedical named entity recognition using temporal convolutional network with conditional random field, *Math. Biosci. Eng.*, **17** (2020), 3553–3566. https://doi.org/10.3934/mbe.2020200.
28. X. Y. Song, A. Feng, W. K. Wang, Z. J. Gao, Multidimensional self-attention for aspect term extraction and biomedical named entity recognition, *Math. Probl. Eng.*, 2020. https://doi.org/10.1155/2020/8604513.

AIMS Press