This is the peer reviewed version of the following article: Zhang, R, Lu, W, Wang, S, Peng, X, Yu, R, Gao, Y. Chinese clinical named entity recognition based on stacked neural network. Concurrency Computat Pract Exper. 2020;e5775, which has been published in final form at https://doi.org/10.1002/cpe.5775. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Chinese Clinical Named Entity Recognition based on Stacked Neural Network

Ruoyu Zhang[1] | Wenpeng Lu[1] | Shoujin Wang[2] | Xueping Peng[3] | Rui Yu[1] | Yuan Gao[4]

[1]School of Computer, Qilu University of Technology (Shandong Academy of Sciences), Shandong, China
[2]Department of Computing, Macquarie University, New South Wales, Australia
[3]Centre of Artificial Intelligence, University of Technology Sydney, New South Wales, Australia
[4]96781 PLA Troops, Shaanxi, China

Correspondence
*Wenpeng Lu. Email: Wenpeng.Lu@qlu.edu.cn

Present Address
No.3501, Daxue Road, Changqing District, Jinan 250353 Shandong Province, PR China

Summary

The precise named entity recognition is a key component in Chinese clinical natural language processing. Though clinical NER systems have attracted widespread attention and been studied for decades, the latest NER research usually relies on a shallow text representation with one-layer neural encoding, which fails to capture deep features and limits its performance improvement. To capture more features and encode the clinical text efficiently, we propose a deep stacked neural network for Chinese clinical NER. The neural network stacks two bidirectional LSTM and GRU layers to encode the text twice, followed by a CRF layer to recognize named entities in Chinese clinical text. Extensive empirical results on three real-world data sets demonstrate that the proposed method significantly outperforms six state-of-the-art NER methods. Especially, compared with the conventional CRF model, our method has at least 3.75% F_1-score improvement on these public data sets.

KEYWORDS:
Named entity recognition, Electronic medical record, Chinese clinical text, LSTM, GRU

1 | INTRODUCTION

Clinical named entity recognition (NER) is the task of identifying entities like symptoms, diseases, exams, treatments, medications and body parts in clinical text[1,2]. A large amount of important medical information is contained in the narrative clinical text, which is not directly accessed by biomedical processing systems that rely on the structured data. Recognizing clinical named entities is usually implemented as the first step in clinical text mining, which is the key foundation to support the downstream biomedical processing systems and tasks. For the biomedical relation extraction task, NER is critical to find the target entities[3]. For the hospital mortality prediction task, NER is necessary to analyze clinical notes[4]. Clinical NER can benefit many applications in medical data mining, such as clinical surveillance, comorbidity analysis, pharmacovigilance and drug interactions[5,6].

With the development of medical informationization and hospital information system, more and more electronic health records (EHR) are generated. EHR systems store large amounts of data associated with patients, including diagnoses, laboratory test, prescriptions, clinical notes, etc., which have been used for such tasks as clinical decision support systems, disease inference, medical concept extraction[7]. For example, a piece of clinical text is *patients' color ultrasound results show moderate fatty liver*, where *fatty liver* is the name of the disease and *liver* is the name of body part. Another text is *Memory loss began 3 years ago*, where *memory loss* is the name of clinical symptom. It is obvious that to recognize clinical named entities accurately is critical for all the downstream tasks. However, build a NER system is not easy because of the richness and diversity of clinical text in

EHR systems. Moreover, NER in Chinese clinical text is more difficult because it lacks word boundaries and is more complex than Romance languages.

Early NER systems are rule-based systems that rely on the prepared clinical dictionaries, such as cTAKES [8] and MetaMap [9], which implement NER by looking up terminology dictionary. The performance of rule-based methods depends on the quality and coverage of the clinical vocabulary and labeling rules. As new terminologies emerge in endlessly, it is hard to maintain an exhaustive dictionary and update the labeling rules in time.

Traditional supervised learning methods view NER as a sequence labeling problem. In NER systems, a various of supervised machine learning algorithms are applied, such as conditional random fields (CRF) [10] and support vector machine (SVM) [11,12]. The performance of traditional supervised methods depends on the manually pre-defined features. As the diversity and complexity of clinical text, it is impossible to design a perfect feature set containing all features.

Recently, along with the development of deep learning methods [13,14,15], a series of neural network models have been applied in clinical NER task, which usually combine long-short term memory (LSTM) and CRF to find the best label sequences (i.e., BIO format labels [12]) for an input sequence (i.e., words in clinical text) [16,17]. Neural network models are popular because they do not require any prepared resources like rules and dictionaries, and manually defined features. Though various neural architectures have been proposed, most NER works usually rely on a shallow text representation with one-layer neural encoding, which are unable to capture more deep features and limit their performance improvement.

In order to achieve an outstanding performance in clinical NER task, the complex text features should be captured as more as possible. In this paper, we propose a deep stacked neural network for Chinese clinical NER. The neural network stacks two bidirectional LSTM and gated recurrent unit (GRU) layers to encode the clinical text twice, then employs a CRF layer to find optimal tag sequences so as to recognize named entities. The extensive experiments on three real-world data sets demonstrate the superiority of our proposed model.

The main contributions of this paper are as follows:

- We propose Chinese clinical named entity recognition method based on stacked neural network. Our model stacks bidirectional LSTM and GRU layers hierarchically to encode the clinical text twice to capture the deep features and generate a more ideal text representation, then employs a CRF layer to predict the right tag of each word to recognize named entities in clinical text.
- We implement a series of NER models and carry on an extensive comparative study for clinical NER task on three real-world data sets. This may provide an objective reference for related researchers to evaluate the abilities of different NER models.
- Extensive experiments on the real-world data sets demonstrate that our proposed NER method significantly outperforms the state-of-the-art methods.

The rest of the paper is structured as follows. We discuss the related work about clinical NER in Section 2. Section 3 describes our proposed model based on stacked neural network in detail, followed by the experiments in Section 4. Finally, we conclude the paper and outline the future work in Section 5.

2 | RELATED WORK

Briefly, we generally divide related work into the following categories: (1) rule-based methods, (2) feature-engineered supervised methods and (3) neural network methods.

Rule-based methods. Due to the complexity and specialization of medical domain, the early NER methods are implemented based on manually created rules and domain dictionaries, which work well when the rules and dictionaries are exhaustive. Gaizauskas et al. [18] design LaSIE system for information extraction, which identifies named entities by matching the input against the pre-stored lists of proper names and common nouns that act as named entities. Guergana et al. [8] realize cTAKES system for clinical text analysis, which recognizes named entities by looking up terminology dictionary. Aronson et al. [9] introduce MetaMap system, which provides the access to the concepts in the unified medical language system Metathesaurus from biomedical text, which can be viewed as a dictionary for clinical rule-based NER system. The performance of rule-based methods depends on the quality and coverage of the clinical dictionary and labeling rule. Because new terminologies and language phenomenons emerge in endlessly, rule-based methods are hard to maintain an exhaustive dictionary and labeling rule system, which lead to narrow and limited applications.

Feature-engineered supervised methods. Based on the labeled training data, feature-engineered supervised methods are trained to make expected predictions on example inputs, and predict right tags on unknown inputs, which rely on the manual defined features and training data[19,20,21]. Conditional random fields (CRF), maximum entropy models (MEM), support vector machines(SVM) and hidden markov models (HMM) are applied widely in NER task. Settles et al.[10] propose to recognize biomedical named entities with CRF and a variety of novel features, including orthographic features and semantic features. Skeppstedt et al.[12] apply CRF model on Swedish clinical text to recognize disorder, finding, pharmaceutical drug and body structure entites. Wang et al.[22] propose a NER method for electronic medical records, which establishes a hierarchical CRFs framework, and separates the complicated electronic medical records into relatively simple and interrelated sub-layers. Finkel et al.[23] develop biomedical NER system based on MEM model, which makes full use of local and syntactic features within texts, and external resources such as Web and gazetteers. Qu et al.[24] annotate electronic medical records, and implement NER systems on the records with the combination of MEM, SVM and CRF. These feature-engineered supervised methods demonstrate good performance in the data sets with enough training data. However, these supervised methods depend heavily on the manually annotated training data and manually designed feature engineering, which are huge burdens to prepare them and limit their performance on real-world data sets[25,26].

Neural network methods. With the development of deep learning[27,28,29], a variety of neural networks are proposed for NER task, which avoid the requirement for most feature engineering in the supervised methods[30]. Collobert et al.[31] propose the first single convolutional neural network architecture for NER task with manually defined orthographic features, which is trained jointly on multiple tasks with weight-sharing mutlitask learning. Gridach et al.[16] put forward a neural network that benefits word-level and character-level representations, with a combination of bidirectional LSTM and CRF models. Habibi et al.[17] present a generic NER method based on deep learning and statistical word embedding with LSTM and CRF models, which outperforms the existing entity-specific NER system by a large margin. As neural network methods do not require the manually pre-defined features and annotated training data, they are popular in the research community[32]. However, most existing NER models usually rely on a shallow text representation with one-layer neural encoding, which fail to capture enough deep features and encode the clinical text efficiently. This limits their performance improvement.

3 | OUR PROPOSED STACKED NEURAL NETWORK

3.1 | Framework of the proposed model

In NER task, we denote the input character sequence as $X = \{x_1, x_2, \dots, x_n\}$, the labeled tag sequence as $Y = \{y_1, y_2, \dots, y_n\}$. Given an input sequence X, the goal of NER system is to predict the right tag sequence Y. The framework of our proposed model is shown in Fig.1. This framework can be divided into three modules: input (top), encoder (middle) and output (bottom) modules. The input module is responsible to convert to input characters into their embeddings. The encoder module consists of the stacked neural network, which includes two sub-layers, i.e., bidirectional LSTM and GRU, which captures the deep features and encodes the input sequence. The output module predicts the tag for each input character.

3.2 | Input Module

In this module, the characters in input sequence $X = \{x_1, x_2, \dots, x_n\}$ are inputted with one-hot representations by look-up layer, which are further converted into embedding representation. All embedding vectors consist of a matrix, which is transferred to the next module, i.e., encoder module. In order to avoid the errors induced by Chinese word segmentation, we choose to use character embedding instead of word embedding, which is trained by Skipgram[33].

3.3 | Encoder Module

In this module, we encode the input character sequence with the stacked neural network, which includes bidirectional LSTM and GRU layers. Both LSTM and GRU are able to learn local and long-term dependencies among the sequence. We firstly handle the character embeddings with LSTM layer, where various features are captured from the character sequence. Next, the encoded features are further handled by GRU layer. Finally, the local and long-term dependencies learned by GRU are viewed as the final representation of the input sequence.

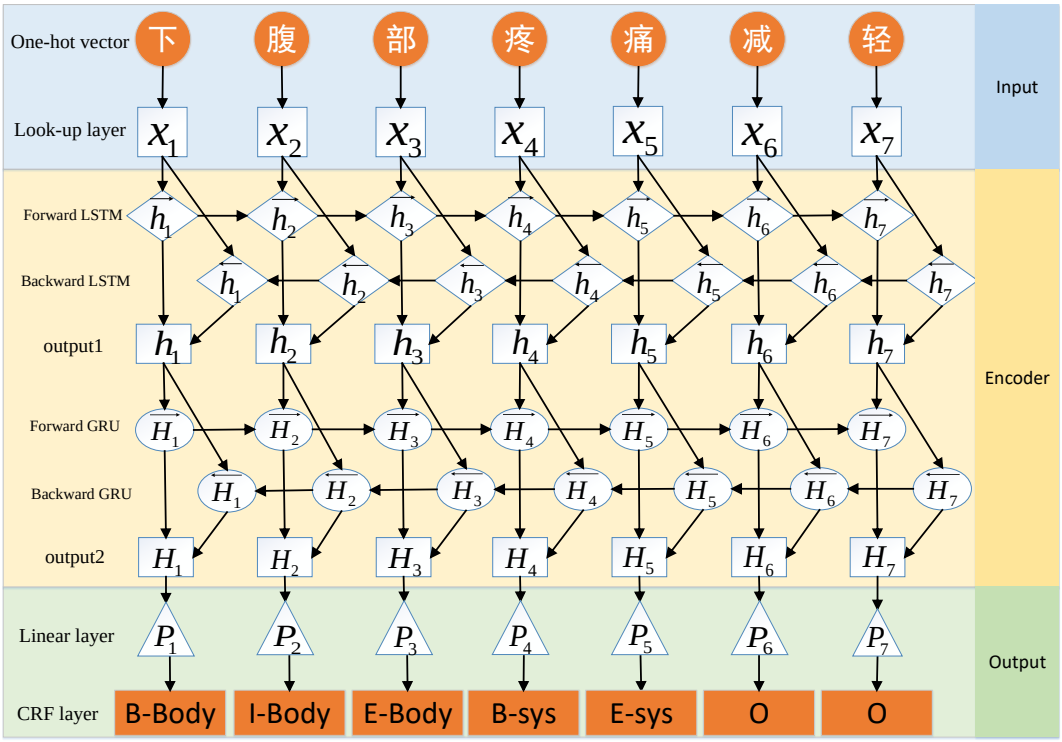

FIGURE 1 Structure of stacked neural network for Chinese named entity recognition

3.3.1 | LSTM

The bidirectional LSTM layer automatically extracts features of the input sequence. The character matrix representation from look-up layer is inputted into forward LSTM and backward LSTM as shown in Fig.1. Then, each hidden state in hidden sequence of the forward LSTM $(\overrightarrow{h_1}, \overrightarrow{h_2}, \dots, \overrightarrow{h_n})$ is concatenated with that of the backward LSTM $(\overleftarrow{h_1}, \overleftarrow{h_2}, \dots, \overleftarrow{h_n})$, as described in Equ.(1). A complete sequence of hidden states is obtained as $(h_1, h_2, \dots, h_n) \in R^{n*m}$.

$$h_t = \left[\overrightarrow{h_t}, \overleftarrow{h_t}\right] \in R^{n*m} \tag{1}$$

3.3.2 | GRU

Although bidirectional LSTM is able to capture the sequence features, it only encodes the input sequence one time with one-layer neural encoding, which is hard to capture enough deep features for the following output module. GRU is similar with LSTM and performs better in solving gradient dispersion. As the complementation of LSTM, we stack bidirectional GRU over LSTM. This layer takes the output of the previous LSTM as the input of GRU, each hidden state in hidden sequence of the forward GRU $(\overrightarrow{H_1}, \overrightarrow{H_2}, \dots, \overrightarrow{H_n})$ is concatenated with that of the backward GRU $(\overleftarrow{H_1}, \overleftarrow{H_2}, \dots, \overleftarrow{H_n})$, as described in Equ.(2). After the encoding by stacked neural network model, the hidden state sequence $(H_1, H_2, \dots, H_n) \in R^{n*m}$ generated by bidirectional GRU is returned as the final representation of the input sequence, which is transferred to the output module.

$$H_t = \left[\overrightarrow{H_t}, \overleftarrow{H_t}\right] \in R^{n*m} \tag{2}$$

3.4 | Output Module

This module consists of two sub-layers, i.e. linear layer and CRF layer. The former converts the dimension of representation embeddings generated by encoder module. The latter is responsible to output the predicted tag sequence.

In this module, we firstly handle the final representation generated by encoder module with a linear layer, which is able to convert the dimensions of the representation from R^{n*m} to R^{n*k}, where k is the number of candidate tags. As shown in Fig.1,

the hidden state sequence $(H_1, H_2, \dots, H_n) \in R^{n*m}$ is converted to the hidden state sequence $(P_1, P_2, \dots, P_n) \in R^{n*k}$. Then, we utilize a CRF layer to predict the most possible tag sequence. The parameter of the CRF layer is a matrix, denoted as A, where A_{ij} represents the transfer score from the i-th tag to the j-th tag. If there is a candidate tag sequence $Y = \{y_1, y_2, \dots, y_n\}$ for the input character sequence X, then the score of Y can be estimated with Equ.(3).

$$score(X, Y) = \sum_{i=1}^{n} P_{i,y_i} + \sum_{i=1}^{n+1} A_{y_{i-1}y_i} \tag{3}$$

According to Equ.(3), the score of the whole tag sequence is equal to the sum of the scores of each position. The score of each position is obtained by two parts, one part is determined by the P_i obtained by GRU passing through a linear layer, and the other part is determined by the transfer matrix A of CRF. Softmax can obtain the normalized probability, as in Equ.(4). The model is trained by maximizing the logarithmic likelihood function, as in Equ.(5). Finally, the model utilizes the dynamic programming Viterbi algorithm to find the optimal path, and get the final tag sequence, as in Equ.(6).

$$P(Y|X) = \frac{exp(score(X, Y))}{\sum_{Y'} exp(score(X, Y'))} \tag{4}$$

$$log\, P(Y^X|X) = score(X, Y^X) - log(\sum exp(score(X, Y'))) \tag{5}$$

$$Y^* = \arg\max_{Y'} score(X, Y') \tag{6}$$

4 | EXPERIMENTS

4.1 | Data Sets

Three real-world data sets for clinical NER task are used to verify the performance of our proposed stacked neural network model. Our source code and data are publicly available on Github[†]. As the scarcity of Chinese clinical NER data, the three data sets are valuable, which are as follows:

- **CCKS-2017.** China conference on knowledge graph and semantic computing (CCKS) is organized by the technical committee on language and knowledge computing of the Chinese information processing society of China. CCKS-2017 provides 600 electronic clinical record texts, which requires to recognize the named entities including anatomical parts, independent symptoms, description of symptoms, surgery and drugs. The data set is annotated by YiDuYun (Beijing) technology co., LTD.
- **CCKS-2018.** In 2018, CCKS provided another data set for the evaluation task, which also included 600 annotated electronic clinical record texts in the same format as the previous year.
- **Hospital-BJ.** We collect a data set from Internet and find out an annotated data set from a hospital in Beijing, which contains 1,200 clinical record texts. It covers all entities in the former CCKS data sets.

4.2 | Labeling rules

Two kinds of labeling rules are applied in our following experiments, i.e., IOB [12] and IBOES respectively. For IOB rule, B means the beginning of an entity, I means the inside of an entity, O means the outside of any entity. IBOES is a more complex and complete annotation rule derived from IOB method. Besides the existing three labels in IOB rule, E means the end of an entity and S means an entity with only single character. Two examples for IOB and IBOES rules are shown in Table. 1.

4.3 | Evaluation Metrics

In our experiments, following the previous work [34,35], the common precision, recall and F_1-score are utilized to evaluate the performance. Precision is the percentage of the number of correctly recognized entities versus the number of all recognized

[†]https://github.com/ruoyuu/Stacked-Neural-Network-ZRY

TABLE 1 Labeling rules examples.

	上	腹	部	闷	痛	不	适	,	腹	部	C	T	提	示	肝	占	位
IOB	B-B	B-I	B-I	S-B	S-I	O	O	O	B-B	B-I	O	O	O	O	B-B	O	O
IBOES	B-B	B-I	B-E	S-B	S-E	O	O	O	B-B	B-E	O	O	O	O	B-S	O	O
Entity type	body			symptom					body						body		

entities, which is marked as P. Recall is the percentage of the number of correctly recognized entities versus the number of all entities in the data set, which is marked as R. F_1-score is derived from precision and recall. The three evaluation metrics are defined in Equ.(7)-(9).

$$P = \frac{TP}{TP + FP} \tag{7}$$

$$R = \frac{TP}{TP + FN} \tag{8}$$

$$F_1 - score = \frac{2PR}{P + R} \tag{9}$$

In the above equations, according to the consistency between the true tags and predicted ones, TP refers to the number of true positive instances, FP refers to the number of false positive instances and FN refers to the number of false negative instances.

4.4 | Comparison Methods

Baselines To test the performance of our model, our proposed stacked neural network is compared with six common state-of-the-art NER models. All the baselines are implemented with Keras in Python. As the baselines only contain a part of components of our proposed stacked neural network, they also could be viewed as the simplified variations of our method.

- **CRF[10]:** This method is the most traditional model for named entity recognition, where CRF directly predicts the tag sequence.
- **LSTM-CRF[16]:** This method combines LSTM and CRF together, where LSTM extracts features followed by a CRF layer to predict the tags.
- **GRU-CRF:** This method replaces the LSTM component in LSTM-CRF model with GRU.
- **BiLSTM-CRF[36]:** This method is similar with LSTM-CRF. The difference lies that this method considers both forward and backward LSTMs, instead of only one direction in LSTM-CRF model.
- **BiGRU-CRF[37]:** This method replaces the BiLSTM component in BiLSTM-CRF model with BiGRU.
- **CNN-BiLSTM-CRF[38]:** This method enhances the model of BiLSTM-CRF with CNN component, where the output of CNN is taken as the input of BiLSTM, followed by CRF to ensure the legitimacy of the predicted tags.

4.5 | Experimental parameters

In the experiments, we use pre-trained embeddings to convert the characters in clinical medical text to vector representations, whose dimension is 300[‡]. The batch size was set to 128. Both the dimensions of the BiLSTM and BiGRU are 128. We apply dropout strategy to each layer in our approach to mitigate overfitting. The dropout rate is set to 0.5. RMSprop is selected as the optimization algorithm for training.

[‡]https://github.com/liuhuanyong/ChineseEmbedding/blob/master/model/token_vec_300.bin

4.6 | Experimental Results and Analysis

We adopt six state-of-the-art NER methods as the baselines to compare with our proposed stacked neural network. All baselines and our proposed method are evaluated on the three data sets with two labeling rules respectively. The detailed experimental results are shown in Table.2-4. It is obvious that our proposed model showns a significant superiority over the baselines on all data sets under both of IOB and IBOES labeling rules, which demonstrates the powerful ability of the proposed stacked neural network model.

TABLE 2 Experimental results on CCKS-2017.

Labeling rules	IOB			IBOES		
	Precision	Recall	F_1-score	Precision	Recall	F_1-score
CRF	91.05	84.26	87.52	90.01	84.67	87.26
LSTM-CRF	95.00	89.33	92.08	91.15	87.89	89.49
GRU-CRF	94.88	90.04	92.40	91.80	87.09	89.38
BiLSTM-CRF	96.45	93.50	94.95	93.66	88.20	90.85
BiGRU-CRF	96.78	92.18	94.42	93.50	88.16	90.75
CNN-BiLSTM-CRF	97.83	94.29	96.03	**93.80**	88.04	90.83
Our model	**98.03**	**94.38**	**96.17**	93.78	**88.40**	**91.01**

TABLE 3 Experimental results on CCKS-2018.

Labeling rules	IOB			IBOES		
	Precision	Recall	F_1-score	Precision	Recall	F_1-score
CRF	95.88	92.27	94.04	93.77	92.34	93.08
LSTM-CRF	98.13	97.04	97.58	97.00	97.51	97.25
GRU-CRF	98.23	97.01	97.62	97.21	96.90	97.05
BiLSTM-CRF	98.54	98.01	98.27	98.31	98.01	98.16
BiGRU-CRF	98.38	97.88	98.13	98.30	97.71	98.00
CNN-BiLSTM-CRF	98.57	97.97	98.27	98.35	**98.15**	98.25
Our model	**98.58**	**98.11**	**98.34**	**98.49**	98.13	**98.31**

TABLE 4 Experimental results on Hospital-BJ.

Labeling rules	IOB			IBOES		
	Precision	Recall	F_1-score	Precision	Recall	F_1-score
CRF	89.86	87.06	88.44	85.41	82.97	84.17
LSTM-CRF	90.99	89.35	90.16	89.21	87.01	88.10
GRU-CRF	90.93	88.90	89.90	88.95	87.19	88.06
BiLSTM-CRF	92.64	91.29	91.96	90.48	89.75	90.11
BiGRU-CRF	92.86	90.58	91.71	92.22	89.55	90.87
CNN-BiLSTM-CRF	91.55	91.49	91.52	91.99	89.07	90.51
Our model	**93.28**	**91.55**	**92.41**	**92.37**	**90.07**	**91.21**

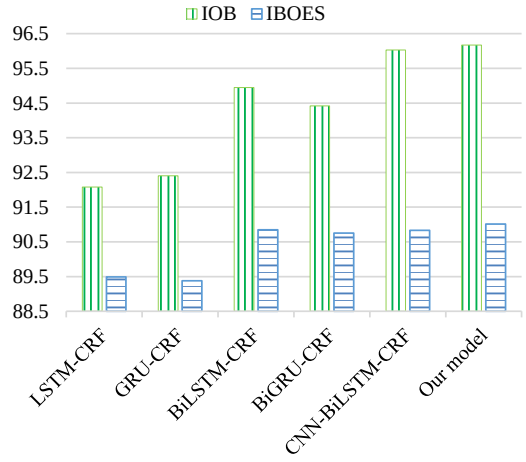

FIGURE 2 F_1-score comparison on CCKS-2017

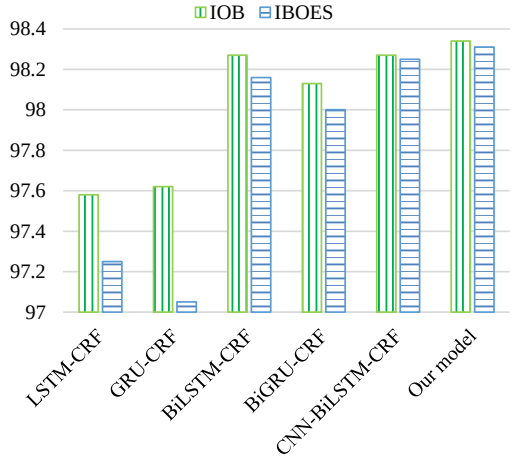

FIGURE 3 F_1-score comparison on CCKS-2018

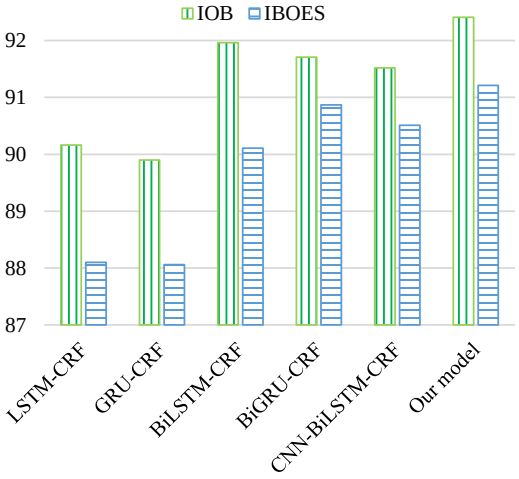

FIGURE 4 F_1-score comparison on Hospital-BJ

In order to further analysis the experimental results, considering the importance of F_1-score, we compare the baselines on F_1-score measure, as shown in Fig.2-4 [§]. According to the figures and Table.2-4, we have several observations.

First, the methods based on neural networks, i.e., LSTM and GRU related methods, outperform traditional feature-engineered method, i.e., CRF. Among all neural networks, LSTM-CRF and GRU-CRF are the weakest under different labeling rules. LSTM-CRF demonstrates 4.56%, 3.54%, 1.72% F_1-score improvement under IOB rule and 2.23%, 4.17%, 3.93% improvement under IBOES rule over CRF on each data set respectively. GRU-CRF demonstrates 4.88%, 3.58%, 1.46% F_1-score improvement under IOB rule and 2.12%, 3.97%, 3.89% improvement under IBOES rule over CRF on each data set respectively. The reasons for the gap between neural network based method and feature-engineered method may be as follows. The traditional feature-engineered method depends on the manually designed features. It is very difficult to find enough effective features, which heavily affects the performance. However, neural network based methods can automatically capture more deep and sophisticated features contained in the input sequence, which gives the natural advantages for neural works to surpass feature-engineered method. The powerful feature representation ability of neural network is beneficial for the prediction of named entities.

Second, the methods with more neural layers always demonstrate a better performance than those with less neural layer. Among the compared neural methods, CNN-BiLSTM-CRF and our method are equipped with two BiLSTM and BiGRU layers, BiLSTM-CRF and BiGRU-CRF have one bi-directional recurrent neural layer, LSTM-CRF and GRU-CRF only have one uni-directional recurrent neural layer. As shown in Fig.2-4, the performance of CNN-BiLSTM-CRF and our method are the best, BiLSTM-CRF and BiGRU-CRF are the suboptimal, LSTM-CRF and GRU-CRF are the worst. It is obvious that the performance

[§] As the performance of CRF is significantly worse than the other methods, the figures do not show it again.

of different models is consistent with their structural complexity. This is because that more neural layers and more complex structure mean the more powerful representation ability to capture and learn the features hid in the input sentences, which can provide more abstract high-level clues to label the right tags for each character. We also notice that CNN-BiLSTM-CRF is slightly worse than BiLSTM-CRF or BiGRU-CRF on the Hospital-BJ data set. However, it still outperforms them on the other two data sets. Its overall performance is better.

Third, our proposed model shows a superiority over BiLSTM-CRF and BiGRU-CRF. Our model stacks BiLSTM and BiGRU layers together to encode the input sentence, which is followed by CRF to assure the reasonableness of the labeled tag sequences. BiLSTM-CRF is a simplified variants of our model by removing the BiGRU layer from encoder module. And, BiGRU-CRF is a simplified variants of our model without BiLSTM layer. According to Table.2-4, compared with BiLSTM-CRF and BiGRU-CRF on CCKS-2017 data set, our model improves F_1-score by 1.22% and 1.75% under IOB labeling rule and 0.16% and 0.26% under IBOES labeling rule. Compared with them on CCKS-2018 data set, our model improves F_1-score by 0.07% and 0.21% under IOB labeling rule and 0.15% and 0.31% under IBOES labeling rule. Compared with them on Hospital-BJ data set, our model improves F_1-score by 0.45% and 0.7% under IOB labeling rule and 1.1% and 0.34% under IBOES labeling rule. It is obvious that the simplified BiLSTM-CRF and BiGRU-CRF are beated heavily by our model. This also demonstrates that to stack BiLSTM and BiGRU together is crucial in our model. Either BiLSTM or BIGRU layer is removed from our model, which will leads to a huge performance hurt. The reason for the superiority of our model may be that the feature capture and representation ability of our model with stacked layers is more powerful than that of the models with single layer, i.e. BILSTM-CRF and BiGRU-CRF.

Fourth, for all methods, the performances on IOB rule are better than those on IBOES rule. For example, our model with IOB rule demonstrates 5.16%, 0.03%, 1.20% F_1-score improvement over that with IBOES rule on each data set respectively. This is because that IBOES enlarges the annotated tags, i.e, the end of an entity and the entity with only single character, which means more difficult to recognize and label the entities.

Last, our proposed stacked neural network model can consistently outperform all compared baselines. For IOB labeling rule, compared to CRF, LSTM-CRF, GRU-CRF, BiLSTM-CRF, BiGRU-CRF, CNN-BiLSTM-CRF, our model improves F_1-score by 8.65%, 4.09%, 3.77%, 1.22%, 1.75%, 0.14% in CCKS-2017 data set; by 4.3%, 0.76%, 0.72%, 0.07%, 0.21%, 0.07% in CCKS-2018 data set; by 3.97%, 2.25%, 2.51%, 0.45%, 0.7% , 0.89% in Hospital-BJ data set. For IBOES labeling rule, compared to CRF, LSTM-CRF, GRU-CRF, BiLSTM-CRF, BiGRU-CRF, CNN-BiLSTM-CRF, our model improves F_1-score by 3.75%, 1.52%, 1.63%, 0.16%, 0.26%, 0.18% in CCKS-2017 data set; by 5.23%, 1.06%, 1.26%, 0.15%, 0.31%, 0.06% in CCKS-2018 data set; by 7.04%, 3.11%, 3.15%, 1.1%, 0.34%, 0.7% in Hospital-BJ data set. Compared with the feature-engineered CRF method, our method combines LSTM and GRU together, which can capture long-distance dependency features. This shows that the information of orders and long-distance dependencies are more valuable for the right prediction in NER task. Compared with the other neural methods, our method stacks two bidirectional LSTM and GRU layers to encode the input text twice, which can capture more sophisticated features and is beneficial to outperform the existing methods.

4.7 | Parameter Analysis

4.7.1 | Comparison with the Different Optimizer

TABLE 5 F_1-score performance with different optimizer.

Optimizer	Adam	Adamax	Adadelta	Adagrad	Nadam	RMSprop	SGD
CCKS-2017^{IOB}	96.13	95.81	96.07	96.00	96.10	<u>96.17</u>	80.69
CCKS-2017^{IBOES}	<u>91.01</u>	90.80	90.88	90.85	90.85	<u>91.01</u>	80.44
CCKS-2018^{IOB}	98.15	97.87	97.88	97.99	98.00	<u>98.34</u>	82.33
CCKS-2018^{IBOES}	98.17	98.02	97.78	98.09	98.14	<u>98.31</u>	82.59
Hospital-BJ^{IOB}	<u>92.56</u>	92.18	92.33	92.38	92.41	92.41	81.01
Hospital-BJ^{IBOES}	<u>91.33</u>	91.30	91.28	91.29	91.34	91.21	80.06

In the section, we conduct several experiments to explore the effectiveness of different optimizers. The detailed experimental results on F_1-score measure are shown in Table.5. For each data set, we underline the results of the best-performing optimizer. According to Table.5, Adam and RMSprop are better than the others, SGD is the worst on all data sets. Though Adam achieves

a slight improvement over RMSprop on Hospital-BJ data set, RMSprop beats Adam on all the other data sets. Though Adam is an updated version of RMSprop, RMSprop is more suitable and stable for our model on NER task.

4.7.2 | Comparison with Different LSTM and GRU Dimensions

In this section, we explore the influence of LSTM and GRU dimensions on the performance of our model on CCKS-2018 data set. To validated the effectiveness of dimensions, we set the different dimensions, i.e, 32, 64, 128 and 256, for LSTM and GRU separately to compare their F_1-score performance under IOB rule, which are shown in Fig.5.

As shown in Fig.5, the performance grows with the increasement of dimension. When the dimension is 32, the performance is worst. With the dimension increasement from 32 to 128, the performance is improved significantly and become stable. Though the performance is best when the dimension is 256, there is only a very slight improvement than 128. Considering the computational cost and complexity, we set the dimension as 128 in our model.

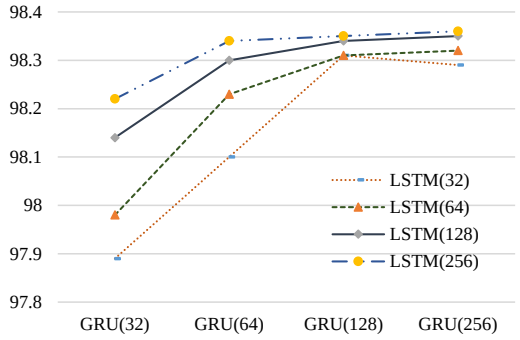

FIGURE 5 F_1-score of different LSTM and GRU parameter

5 | CONCLUSION

In this paper, we introduce a Chinese clinical named entity recognition based on stacked neural network model, which performs well on the three real-world data sets. Our model consists of three parts, i.e., input, encoder and output modules. The input module is responsible to convert input characters into their embeddings with the help of a look-up layer. The encoder module consists of the stacked neural network layer, which contains two bidirectional LSTM and GRU sub-layers to encode the clinical text twice. The output module includes a linear layer and CRF layer, which predicts the tag for each input character according to the encoding of stacked neural network layer. Extensive experimental results show that our method significantly outperforms six state-of-the-art NER methods, which demonstrate the superiority of our proposed method. For future work, we plan to investigate how to integrate more neural components into the proposed model and how to apply it on more similar tasks. We will also further study how to enrich more effective features to the model so as to achieve more higher performance.

ACKNOWLEDGEMENT

The authors would like to thank CCKS for providing the experimental data, and the valuable comments from the reviewers. The research work is partly supported by National Nature Science Foundation of China under Grant No.61502259 and National Key R&D Program of China under Grant No.2018YFC0831700.

References

1. Vikas Y, Steven B. A survey on recent advances in named entity recognition from deep learning models. In: Bender EM, Derczynski L, Isabelle P., eds. *Proceedings of the 27th International Conference on Computational Linguistics*; 2018; Santa Fe, New Mexico, USA: 2145–2158.

2. Wang Q, Zhou Y, Ruan T, Gao D, Xia Y, He P. Incorporating dictionaries into deep neural networks for the chinese clinical named entity recognition. *Journal of biomedical informatics* 2019; 92: 103-133.

3. Singh G, Bhatia P. Relation Extraction using Explicit Context Conditioning. In: Burstein J, Doran C, Solorio T., eds. *Proceedings of NAACL-HLT*; 2019; Minneapolis, Minnesota: 1442–1447.

4. Jin M, Bahadori MT, Colak A, et al. Improving hospital mortality prediction with medical named entities and multimodal learning. *arXiv preprint arXiv:1811.12276* 2018.

5. Zhang J, Li J, Wang S, et al. Category multi-representation: a unified solution for named entity recognition in clinical texts. In: Dinh P, Vincent ST, Geoffrey IW, Bao H, Mohadeseh G, Lida R., eds. *Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining*Springer. ; 2018; Melbourne, VIC, Australia: 275–287.

6. Bhatia P, Celikkaya B, Khalilia M. Joint Entity Extraction and Assertion Detection for Clinical Text. In: Korhonen A, Traum D, Màrquez L., eds. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*; 2019; Florence, Italy: 954–959.

7. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE journal of biomedical and health informatics* 2017; 22(5): 1589–1604.

8. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association* 2010; 17(5): 507–513.

9. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association* 2010; 17(3): 229–236.

10. Settles B. Biomedical named entity recognition using conditional random fields and rich feature sets. In: Nigel C, Patrick R, Nazarenko A., eds. *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*; 2004; Geneva, Switzerland: 107–110.

11. Chen Y, Wang J, Liu S, et al. Multiscale fast correlation filtering tracking algorithm based on a feature fusion model. *Concurrency and Computation: Practice and Experience* 2019: e5533.

12. Skeppstedt M, Kvist M, Nilsson GH, Dalianis H. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. *Journal of biomedical informatics* 2014; 49: 148-158.

13. Lu H, Li Y, Chen M, Kim H, Serikawa S. Brain intelligence: go beyond artificial intelligence. *Mobile Networks and Applications* 2018; 23(2): 368–375.

14. Lu H, Wang D, Li Y, et al. CONet: A cognitive ocean network. *IEEE Wireless Communications* 2019; 26(3): 90–96.

15. Xu X, Lu H, Song J, Yang Y, Shen HT, Li X. Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval. *IEEE transactions on cybernetics* 2019.

16. Gridach M. Character-level neural network for biomedical named entity recognition. *Journal of biomedical informatics* 2017; 70: 85–91.

17. Habibi M, Weber L, Neves M, Wiegandt DL, Leser U. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics* 2017; 33(14): i37–i48.

18. Gaizauskas R, Wakao T, Humphreys K, Cunningham H, Wilks Y. University of Sheffield: Description of the LaSIE Systems Used for MUC-6. In: Ralph G, Beth S., eds. *Proceedings of the Sixth Message Understanding Conference (MUC-6)*; 1995; Columbia, Maryland.

19. Lu W, Wu H, Ping J, Huang Y, Huang H. An empirical study of classifier combination based word sense disambiguation. *Ieice Transactions on Information & Systems* 2018; 101(1): 225-233.

20. Liu Z, Wang J, Liu G, Zhang L. Discriminative low-rank preserving projection for dimensionality reduction. *Applied Soft Computing* 2019; 85: 105768.

21. Xiang L, Zhao G, Li Q, Hao W, Li F. TUMK-ELM: a fast unsupervised heterogeneous data learning approach. *IEEE Access* 2018; 6: 35305–35315.

22. Wang Y. *Electronic medical record named entity recognition based on cascade condition random field*. PhD thesis. Jilin: Jilin University, 2014.

23. Finkel J, Dingare S, Nguyen H, Nissim M, Manning C, Sinclair G. Exploiting context for biomedical entity recognition: from syntax to the web. In: Nigel C, Patrick R, Nazarenko A., eds. *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*; 2004; Geneva, Switzerland: 91–94.

24. Qu C. Chinese electronic medical record named entity recognition research. master's thesis. Harbin Institute of Technology. 2015.

25. Liu Z, Lai Z, Ou W, Zhang K, Zheng R. Structured optimal graph based sparse feature extraction for semi-supervised learning. *Signal Processing* 2020: 107456.

26. Lu W, Meng F, Wang S, et al. Graph-based Chinese word sense disambiguation with multi-knowledge integration. *Computer, Materials & Continua* 2019; 61(1): 197–212.

27. Zhou Q, Wang Y, Liu J, Jin X, Latecki LJ. An open-source project for real-time image semantic segmentation. *Science China Information Sciences* 2019; 62(12): 227101.

28. Lu W, Zhang X, Lu H, Li F. Deep hierarchical encoding model for sentence semantic matching. *Journal of Visual Communication and Image Representation* 2020; In Press.

29. Zhou Q, Yang W, Gao G, et al. Multi-scale deep context convolutional neural networks for semantic segmentation. *World Wide Web* 2019; 22(2): 555–570.

30. Zhang Y, Lu W, Ou W, et al. Chinese medical question answer selection via hybrid models based on CNN and GRU. *Multimedia Tools and Applications* 2020; In Press.

31. Collobert R, Weston J. A unified architecture for natural language processing: deep neural networks with multitask learning. In: William WC, Andrew M, Sam TR., eds. *Proceedings of the 25th international conference on Machine learning*; 2008; Helsinki, Finland: 160–167.

32. Xiang L, Guo G, Yu J, Sheng VS, Yang P. A convolutional neural network-based linguistic steganalysis for synonym substitution steganography. *Mathematical Biosciences and Engineering* 2020; 17(2): 1041–1058.

33. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* 2013.

34. Strubell E, Verga P, Belanger D, McCallum A. Fast and accurate entity recognition with iterated dilated convolutions. *arXiv preprint arXiv:1702.02098* 2017.

35. Lu W. Word sense disambiguation based on dependency constraint knowledge. *Cluster Computing* 2019; 22(3): 7549–7557.

36. Huang Z, Wei X, Kai Y. Bidirectional LSTM-CRF models for sequence tagging. *Computer Science* 2015.

37. Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* 2014.

38. Li L, Guo Y. Biomedical named entity recognition based on CNN-BLSTM-CRF model. *Chinese Journal of Information* 2018; 32(1): 116-122.