# Chinese-clinical-record Named Entity Recognition using IDCNN-BiLSTM-Highway Network

**Tinglong Tang[1, 2*], Yunqiao Guo[1], Qixin Li[1], Mate Zhou[1], Wei Huang[3] and Yirong Wu[1,2]**
[1] Hubei Key Laboratory Intelligent Vision Based Monitoring for Hydroelectric Engineering, China Three Gorges University, Yichang, 443002, China
[2] Yichang Key Laboratory of Intelligent Medicine, China Three Gorges University
Yichang, 443002, China
[3] Tianjin University of Technology
Tianjin, 300384, China
[e-mail: tangtinglong@ctgu.edu.cn]
[*]Corresponding author: Tinglong Tang

*Received March 9, 2023; revised April 4, 2023; accepted May 30, 2023; published July 31, 2023*

## Abstract

Chinese named entity recognition (NER) is a challenging work that seeks to find, recognize and classify various types of information elements in unstructured text. Due to the Chinese text has no natural boundary like the spaces in the English text, Chinese named entity identification is much more difficult. At present, most deep learning based NER models are developed using a bidirectional long short-term memory network (BiLSTM), yet the performance still has some space to improve. To further improve their performance in Chinese NER tasks, we propose a new NER model, IDCNN-BiLSTM-Highway, which is a combination of the BiLSTM, the iterated dilated convolutional neural network (IDCNN) and the highway network. In our model, IDCNN is used to achieve multiscale context aggregation from a long sequence of words. Highway network is used to effectively connect different layers of networks, allowing information to pass through network layers smoothly without attenuation. Finally, the global optimum tag result is obtained by introducing conditional random field (CRF). The experimental results show that compared with other popular deep learning-based NER models, our model shows superior performance on two Chinese NER data sets: Resume and Yidu-S4k, The F1-scores are 94.98 and 77.59, respectively.

***Keywords:*** Bidirectional long short-term memory network, Chinese named entity recognition, Conditional random field, Highway network, Iterated dilated convolutional neural network.

## 1. Introduction

Named entity recognition (NER) is a vital sub-task in natural language processing (NLP) task that seeks to recognize named entities of a given text, which serves as the foundation of other important NLP missions like relation extraction [1], event extraction [2], machine translation [3]. In addition, it is also widely used in various application fields, such as feature extraction from medical records to assist in medical decision support [4].

Traditional NER models are linear statistics-based models, for instance, the hidden Markov models (HMMs) and conditional random fields (CRFs) [5, 6, 7]. However, those models cannot efficiently model entity names in unstructured text due to feature engineering with domain-specific vocabulary and knowledge. With the rapid development of neural networks in these years, many deep learning models have been developed to accomplish NER tasks. These models are all developed based on nonlinear neural networks. These models can automatically learn word features from a large amount of corpus data instead of relying on handcrafted features developed from a specific data set, demonstrating multiple advantages over traditional NER models. Two popular network structures have been developed: deep learning based convolutional neural network (CNN), deep learning based recursive neural network (RNN) [8]. Collobert et al. [9] proposed CNN models to solve the sequence tagging problem using a fixed-size window approach. They also combined CNN with CRF structure to generate promising results on sequence tagging tasks. Strubell et al. [10] improved the network structure of CNN and proposed an iterated dilated convolutional neural network (IDCNN), which has superior ability than original CNNs for NER of a long sequence of text. On the other hand, RNNs are widely used in NLP applications since they are ideal for solving problems where a sequence is more important than individual items for language processing. Leveraging potential characteristics of long-distance dependency in a sentence, long short-term memory (LSTM), a RNN with special structure, has gradually become the mainstream approach to NLP applications. Yao et al. [11] merged RNNs with CRF for language understanding, which demonstrates the effectiveness of RNN-CRF structure for sequence tagging. Basaldella et al. [12] proposed a BiLSTM model to perform automatic extraction of key phrases. Corbett et al. [13] proposed three models for the NER task: CRF, BiLSTM and BiLSTM-CRF. They found that the BiLSTM-CRF model can implement the most advanced NER performance. Additionally, Ma et al. [14] designed an end-to-end sequence labelling system which is using a consist of CNN, BiLSTM and CRF. The system does not rely on task-specific resources, feature engineering or data preprocessing, so that it is suitable for a broad range of sequence tagging missions. Recently, Vaswani et al. [15] advanced an attention mechanism in the field of machine translation. This mechanism has attracted widespread attention from researchers in the domain of NLP [16, 17]. It has become the focus of neural network research and has achieved good results in various fields [18, 19, 20, 21]. Tang et al. [22] developed an attention-based CNN-LSTM-CRF model to recognize the Chinese entity in clinical reports, demonstrating the effectiveness of the attention mechanism.

Most of these researches mainly focused on English NER. Compared with English, Chinese has uniqueness in many aspects, such as the Chinese text having no natural boundaries like the spaces in the English text. The NER task is challenging in unstructured Chinese text and the need for Chinese-clinical-record NER applications is increasing.

To further increase the performance of Chinese NER tasks, we advance a new NER model, which is based on IDCNN combined with BiLSTM and a highway network, denoted as IDCNN-BiLSTM-Highway. Similar to other network structures, CRF is used to complete the tag prediction task. The main contributions of this paper are as follows:

1) A new NER model, IDCNN-BiLSTM-Highway, is proposed, which combines IDCNN, BiLSTM, and highway network to accomplish NER tasks.

2) An efficient approach using IDCNN before BiLSTM is adopted so that the information of a single neuron can be used to represent the characteristics of local neurons.

3) A novel highway network with a special transform gate is designed to incorporate the information from intermediate layers of embedding module and BiLSTM.

## 2. Related Work

### 2.1 Iterated dilated convolutional neural network

Yu et al. [23] advanced a dilated convolutional neural network (DCNN) to solve the problem of dense prediction in semantic segmentation. DCNN is especially suitable for dense prediction because it can improve the accuracy by extending the receptive field and ensure that the resolution or coverage is not lost. Strubell et al. [10] proposed a fast and accurate IDCNN for NER tasks. It uses DCNN instead of RNN to achieve computational advantages due to its parallel computing ability. The IDCNN model can attain high-recognition accuracy by learning neighboring information after multiple iterations of expanded convolution. As the number of network layers increases, a broad context can be incorporated to remedy the shortcomings of not being able to learn global features. He et al. [24] use a dilated-gated convolutional neural network (DGCNN) for sound event detection. It utilizes gated units that are alike to LSTM to control information flow to the next layer.

### 2.2 Bidirectional long short-term memory network

Huang et al. [25] comprehensively compared the performance of several sequence marking models which are based on LSTM, such as part of speech tagging (POS), chunking and NER. They developed a neural network model with manual characteristics to develop the recognition effect. Labeau et al. [26] explored bidirectional recurrent neural networks with CNN models (BiRNN-CNNs) for POS tagging, which can infer meaningful word representations from a raw character stream, allowing the models to exploit the morphological properties of words without using any handcrafted features or external tools. Chiu et al. [27] proposed a BiLSTM-CNN structure for NER, which is similar to BiRNN-CNNs, with the difference of using LSTM instead of RNN. In this study, we adopted this structure but we used IDCNN instead of CNN to enable the convolutional layer to automatically extract information from a wider window of an input sequence.

### 2.3 Highway network

Highway network implements multi-channel transmission of input context information and integrates information from different layers. Attention mechanisms in deep neural networks focus on the most relevant parts of the input to make decisions. Compared with the attention mechanisms, the use of the highway network can usually solve the problem of degradation, in which the model is difficult to train due to the depth of the network [28]. To smoothly change the behavior between the following two layers: a plain layer, a layer which simply passes its inputs through, a highway network utilizes two gate structures to control how much input information is transformed. Liu et al. [29] proposed an LM-LSTM-CRF model, in which the highway layer is used after the character level LSTM to map the hidden layer vector to two different vector spaces, and each space completes different tasks. We believe that this structure not only solves the degradation problem due to network depth but also provides some ideas

for cross-layer connections. In our work, we use a different gate structure instead of the routinely used gate structure in the highway network to control information flow between different layers.

## 3. Proposed Work

Conventional NER models use the output vectors of BiLSTM or the attention output obtained by their weighted sum as final results, leading to the abandonment of the information from the intermediate layers of the network. Different from conventional models, we propose IDCNN-BiLSTM-Highway to incorporate the information from the intermediate layers, which is inspired by the structure of the highway network [28].

The overall network structure of the IDCNN-BiLSTM-Highway is shown in **Fig. 1**, which is the main model for Chinese NER task. It contains 5 layers: an embedding layer, which represents each Chinese character in the sentence as a vector; an IDCNN layer, which encodes local context features of Chinese characters into vectors; a BiLSTM layer, which learns global features of a sentence through two different directions LSTM networks; a highway network layer, which enables multichannel transmission of input context information, while integrating information from different layers; and a CRF layer, which performs a label sequence prediction by considering the association of the neighbor labels. The detail of the five layers of our model are described in the next sections.

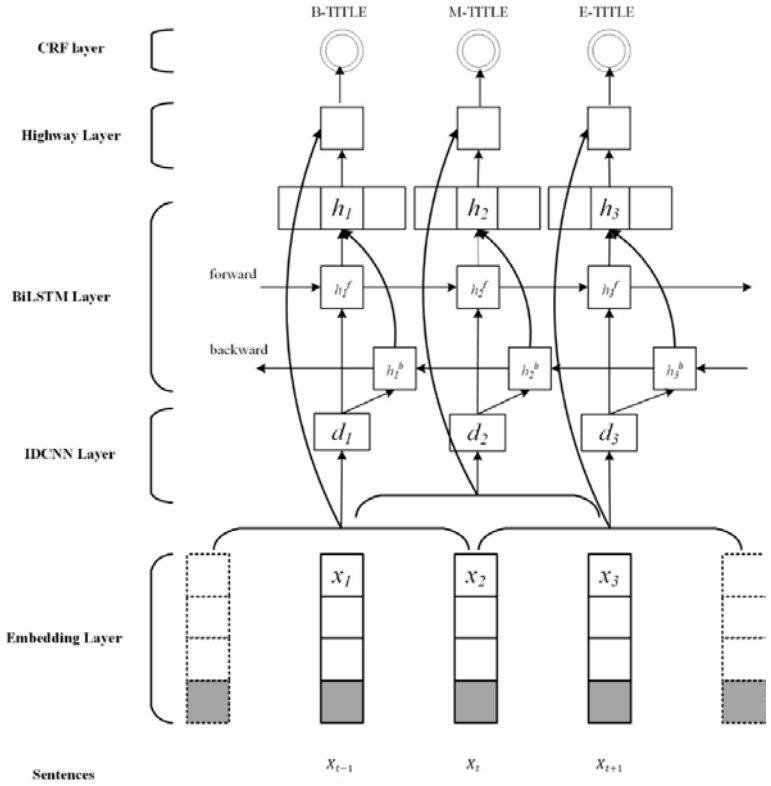


**Fig. 1.** The general structure of our proposed model. $X_{t-1}, X_t, X_{t+1}$ represent three consecutive Chinese characters in a sentence.

### 3.1 Embedding layer

This first layer is used to perform word-embedding which mapping each Chinese character to a vector by learning a distributed matrix representation of character. For a given Chinese sentence $= w_0 w_1 \cdots w_n$ , where $w_t$ stands for the $t$-th character in a sentence, then we employ a word-embedding vector $x_t$ from a pretraining word-embedding matrix instead of the corresponding character $w_t$ itself for training.

### 3.2 IDCNN layer

The second layer is used to perform iterated dilated convolutions. To avoid adjacent inputs, an effective input width is defined by skipping δ inputs at a time. Dilated convolution is applied to each input vector $x_t$, which is defined as follows.

$$d_t = W_d \bigoplus_{k=0}^{r} x_t \pm k\delta \tag{1}$$

where $\oplus$ is a vector concatenation operator, $W_d$ is an affine transformation function, which is equivalent to the CNN convolution kernel in image processing, $r$ is one-half of the sliding window width, δ is the dilation width, and $d$ is the dilated convolution output. When δ is equal to 1, the dilated convolution is equivalent to the ordinary convolution. In our study, the dilated convolution design is shown in **Fig. 2**. The dilated widths of dilated convolutions are 1, 1, and 2.

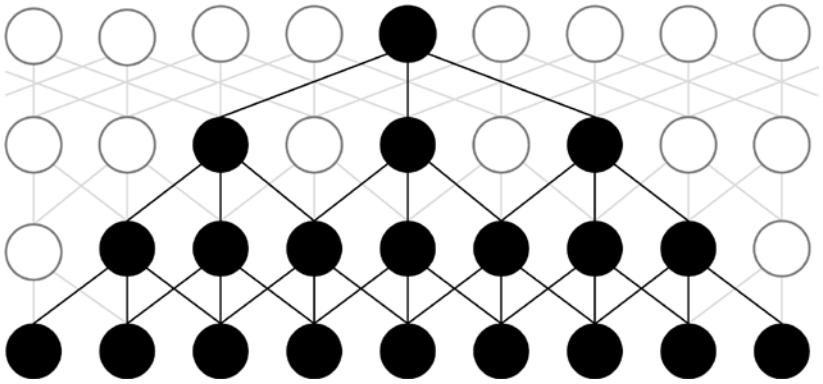

**Fig. 2.** A DCNN block with the maximum dilation width 2 and filter width 3.

### 3.3 BiLSTM layer

The third layer is used to learn sequence features through BiLSTM. In the sequence tagging task, we need to simultaneously access the past and future features in a given time frame. Therefore, we use BiLSTM to effectively utilize context information in two different directions, including past contextual (forward) and future contextual (backward). A basic LSTM unit consists of an input gate $i_t$, a forget gate $f_t$ and an output gate $o_t$; in general, at time $t$, a basic LSTM unit is described as below:

$$i_t = \sigma(W_i h_{t-1} + U_i d_t + b_i) \tag{2}$$
$$f_t = \sigma(W_f h_{t-1} + U_f d_t + b_f) \tag{3}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c h_{t-1} + U_c d_t + b_c) \quad (4)$$

$$o_t = \sigma(W_o h_{t-1} + U_o d_t + b_o) \quad (5)$$

$$h_t = o_t \odot \tanh c_t \quad (6)$$

where $\odot$is a product operator of element; $c_t$ is the memory unit, which is used to store the information from forget gate and the proportion of new information; and $W$ represent the weight matrix of hidden state $h$, $U$ represent the weight matrix of input $d$, and and $b$represent the bias, respectively. Hence, **Fig. 3** is the structure of BiLSTM, and described as below:

$$h_t^f = \overrightarrow{LSTM}(h_{t-1}^f, d_t) \quad (7)$$

$$h_t^b = \overleftarrow{LSTM}(h_{t+1}^b, d_t) \quad (8)$$

$$h_t = h_t^f \oplus h_t^b \quad (9)$$

Where $h_t^f$ , $h_t^b$ represent the hidden states of the LSTM in two different directions at position $t$, respectively, $d_t$ an output of IDCNN, and $\oplus$ denotes the vector concatenation operation.

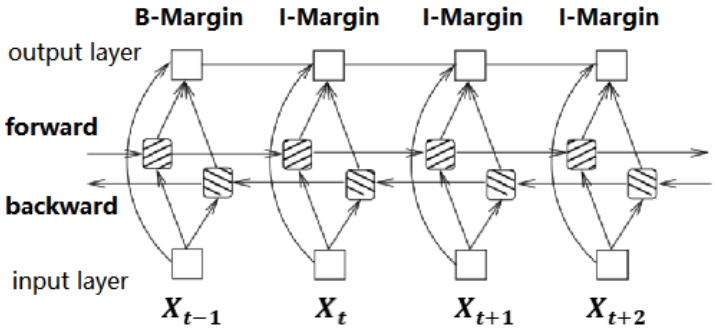


**Fig. 3.** BiLSTM network structure. $X_{t-1}, X_t, X_{t+1}, X_{t+2}$ represent four consecutive Chinese characters in a sentence.

### 3.4 Highway layer

The fourth layer is used to utilize a highway network to connect the outputs of word embedding $x$ and BiLSTM $h$. This idea of using a highway network can be described as:

$$z = h * T + x * C \quad (10)$$

where $T, C$ represent transform gate and carry gate. In this paper, after we set $C = 1 - T$[28], we obtain

$$z = h * T + x * (1 - T) \quad (11)$$

We also design a new transform gate, $T = \sigma(\tanh(W * x + U * h) * V)$, to replace routinely used $T = \sigma(W * x + b)$. Our highway structure is shown in **Fig. 4**.

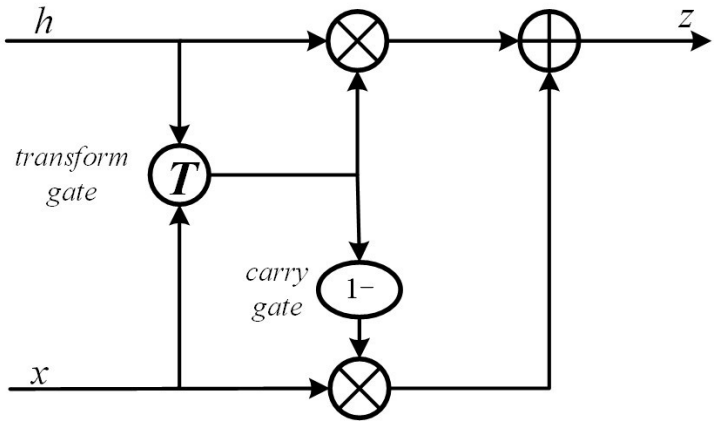


**Fig. 4.** Proposed highway network structure.

### 3.5 CRF layer

The last CRF layer is used to process the features obtained from the highway network layer. In the sequence labeling task, it is necessary to think about the correlation between the joint decoding of tags in the neighborhood in order to find the globally optimized label sequence of specified input sentence. For example, in NER with BMEO format (B, M, E, O represent the beginning, middle, end and outer of the entity, respectively) [30], M and E cannot follow O. Therefore, our model predicts sequence labels by using CRF, which can combine the probability of each independent label and the probability of state transition between labels, instead of the maximum probability calculated by the softmax function at each moment, avoiding local optimal problems.

CRF takes the output of the highway layer $z = z_0 z_1 \cdots z_n$ as an input, where $z_t$ is the output vector from the highway layer of the $t$-th character. The most likely label sequence $y = y_0 y_1 \cdots y_n$ is the output. CRF uses the transition parameter matrix $T$ and the emission parameter matrix $E$ to compute the label sequence score $S$ to make use of the correlation between adjacent labels, described as follows:

$$S(z, y) = \sum_{t=1}^{n} \left( E_{y_t, t} + T_{y_{t-1}, y_t} \right) \tag{12}$$

where $E_{y_t,t}$ is the probability of word $z_t$ with label $y_t$, and $T_{y_{t-1},y_t}$ is the probability of word $z_{t-1}$ with label $y_{t-1}$ followed by $z_t$ with label $y_t$. Thus, the conditional probability of label sequence $y$ can be calculated as:

$$p(z, y; W, b) = \frac{e^{S(z,y)}}{\sum_{\dot{y}} e^{S(z,\dot{y})}} \tag{13}$$

where $Y_z$ represents all possible label sequences for an input state $z$. $W$ and $b$ are the weight vector and bias, respectively.

Our model uses the maximum likelihood estimation. The logarithm of the maximum likelihood is expressed as:

$$L(W, b) = \sum_{t} \log p(y \mid z; W, b) \tag{14}$$

The highest conditional probability for the label sequence $y^*$ is obtained as:

$$y^* = \arg\max_{y \epsilon Y_z} p(z, y; W, b) \tag{15}$$

The Viterbi algorithm is regularly used to decode the optimal label [31].

## 4. Experiments

### 4.1 Details setting

All these experiments were performed on PCs with the following specifications:

Intel (R) core (TM) i7-8700, 16 GB RAM, CPU clock speed or frequency @3.20 GHz, and the GPU acceleration library is CUDNN 7.6.5/CUDA 10.0. The deep learning framework is based on TensorFlow and the models are based on Python 3.7.1.

We conducted experiments on two datasets: the Chinese Resume date set provided by Zhang et al. [30] and the Yidu-S4K data set. The Resume data set has a total of 8 types of entities. The Yidu-S4K data set contains 1379 Chinese clinical reports, of which the entity category is divided into 6 categories. **Table 1** shows the setting of hyper-parameters applied in our experiment.

**Table 1.** Hyper-parameter values

| Parameter | Value |
|---|---|
| embedding size | 100 |
| batch size | 20 |
| gradient clip | 5 |
| dropout | 0.5 |
| LSTM units | 100 |
| epochs | 50 |
| learning rate | 0.001 |
| optimizer | Adam |

### 4.2 Evaluation metrics

In order to understand the quality of the NER model, we must comprehensively evaluate the model, usually from two aspects: whether the entity boundary is correct; whether the entity type is marked correctly [32]. The evaluation indexes used in this paper are defined as below:

Precision rate is the proportion of the correct extraction results of NER model to all extracted results for prediction results.

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \times 100\% \tag{16}$$

Recall rate refers to the ratio of the results correctly extracted by NER model to all possible correct results for the original sample.

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \times 100\% \tag{17}$$

F1-score can be regarded as a weighted average of P and R. Only when the precision and recall are high, the F1 value will be high, which reflects the robustness of the model.

$$\begin{cases} \frac{2}{F1} = \frac{1}{P} + \frac{1}{R} \\ F1 = \frac{2*R*P}{R+P} \times 100\% \end{cases} \tag{18}$$

The F1-score jointly considers the precision and recall rate of the classification model and gets their weighted harmonic average. Generally speaking, the evaluation indexes P and R may be contradictory, so they are often considered comprehensively by calculating the F1-score. High recall rate is paid more attention to NER, but high accuracy rate is more focused on information retrieval [32, 33].

### 4.3 Evaluation results

To demonstrate the effectiveness of our NER model, we develop other deep learning-based NER models for comparison, which include BiLSTM, BiLSTM with attention mechanism (BiLSTM-Attention), BiLSTM with highway network (BiLSTM-Highway), IDCNN, IDCNN with attention mechanism (IDCNN-Attention), IDCNN with highway network (IDCNN-Highway), IDCNN-BiLSTM and IDCNN-BiLSTM with attention mechanism (IDCNN-BiLSTM-Attention). Comparison experiments are carried out on two Chinese NER data sets, Resume and Yidu-S4k, we use the metrics of P, R, and F1-score to quantify the performance of our models in the experiments.

From our experimental results, we observed and analyzed from the following aspects:

• **Effectiveness of our model.** The comparison results of our model and other basic models are shown in **Table 2**. We observe that our IDCNN-BiLSTM-Highway model achieves the highest F1-scores of 94.98% on the Resume data set and 77.59% on the Yidu-S4K data set. Our model used IDCNN before BiLSTM so that the information of individual neurons could contain features of local neurons, showed superior recognition performance than BiLSTM model and IDCNN model, increasing the F1-score from 93.45% to 94.98% on the Resume data set and achieving a 1.40% improvement on the Yidu-S4K data set. Meanwhile, we found that two models based on IDCNN-BiLSTM achieved faster convergence rate during model training.

**Table 2.** Basic results of our experiment on two data sets

| Data set | Method | P | R | F1 |
|---|---|---|---|---|
| Resume | BiLSTM | 93.28 | 93.62 | 93.45 |
| | IDCNN | 93.90 | 94.42 | 94.16 |
| | IDCNN-BiLSTM | **95.04** | 93.99 | 94.51 |
| | IDCNN-BiLSTM-H | 94.86 | **95.09** | **94.98** |
| Yidu-S4K | BiLSTM | 77.28 | 75.14 | 76.19 |
| | IDCNN | 77.46 | 76.20 | 76.82 |
| | IDCNN-BiLSTM | 78.11 | **76.47** | 77.28 |
| | IDCNN-BiLSTM-H | **78.77** | 76.44 | **77.59** |

[a] "H" denotes "Highway"

• **Effectiveness of the highway layer.** In order to prove the influence of highway layer on the performance of the model designed in this paper, we adopted the self-attention mechanism and performed comparative experiments. The detailed comparison results are given in **Table**

3. We could observe that the model using the highway layer has better performance than the basic model. On the two data sets, the effect of the highway structure is better than that of the self-attention mechanism.

**Table 3.** Comparison results of Highways and Attention on two data sets

| Data set | Method | P | R | F1 |
|---|---|---|---|---|
| Resume | BiLSTM-A[a] | 94.11 | 94.05 | 94.08 |
| | BiLSTM-H | 94.06 | 94.17 | 94.11 |
| | IDCNN-A | 93.96 | 94.48 | 94.22 |
| | IDCNN-H | **95.01** | 94.54 | 94.77 |
| | IDCNN-BiLSTM-A | 94.57 | 94.05 | 94.31 |
| | IDCNN-BiLSTM-H | 94.86 | **95.09** | **94.98** |
| Yidu-S4K | BiLSTM-A | 77.08 | 76.54 | 76.81 |
| | BiLSTM-H | 77.61 | 76.29 | 76.92 |
| | IDCNN-A | 74.96 | 75.99 | 75.47 |
| | IDCNN-H | 78.12 | **77.05** | 77.58 |
| | IDCNN-BiLSTM-A | **78.96** | 75.73 | 77.31 |
| | IDCNN-BiLSTM-H | 78.77 | 76.44 | **77.59** |

[a] "A" denotes "Attention"

• **Effectiveness of the IDCNN layer.** The model performance can be improved when an IDCNN structure is added. On the Resume data set, the performance of IDCNN-BiLSTM is 1.06% higher than that of BiLSTM in terms of F1-score (**Table 2**); the performance of IDCNN-BiLSTM-Attention is 0.23% higher than that of BiLSTM-Attention (**Table 3**); the performance of IDCNN-BiLSTM-Highway is 0.87% higher than that of BiLSTM-Highway (**Table 3**). By observing the results on the Yidu-S4K data set, we found that the performance of all models which contained with IDCNN layer has been drastically improved. (**Table 2** and **Table 3**). These conclusions indicate that the context information of Chinese characters learned from the IDCNN structure is useful for the Chinese NER task.

• **Effectiveness of the proposed transform gate.** We compare the performance of highway networks between the proposed transform gate and the conventionally used transform gate in terms of F1-score. **Table 4** lists the performance of three models that contain highway structures. We observe that on the Resume data set, the performance of the models with the proposed transform gate is improved by 0.05%, 0.92%, and 0.91% compared with the models that contain the conventionally used transform gate. On the Yidu-S4K data set, the performance is also improved by 0.66%, 0.58% and 1.07%. These results indicate that our proposed transform gate has a potential to improve recognition performance.

**Table 4.** Comparison of the performance between two different transform gates according to F1-Score on two data sets

| Dataset | Method | Conventionally [a] | Proposed [b] |
|---|---|---|---|
| Resume | BiLSTM-H | 94.06 | **94.11** |
| | IDCNN-H | 93.85 | **94.77** |
| | IDCNN-BiLSTM-H | 94.07 | **94.98** |
| Yidu-S4K | BiLSTM-H | 76.26 | **76.92** |
| | IDCNN-H | 77.00 | **77.58** |
| | IDCNN-BiLSTM-H | 76.52 | **77.59** |

[a] conventionally used transform gate

[b] proposed transform gate

### 4.4 Time Complexity

For the two data sets, we roughly evaluated the average running time of all models mentioned above. From the bar char **Fig. 5**, we can intuitively see that the processing speed of models that contained with IDCNN is the fastest, and contained with BiLSTM is very slow due to the high calculation intensity. Considering the effect of the final Chinese NER results and the increasing complexity of the model, the highway network proposed in this paper has advantages in speed compared with other models with attention mechanism. The experimental results indicate that the IDCNN-BiLSTM-Highway model we designed in this paper could get results at the fastest speed while ensuring the recognition accuracy. The introduction of IDCNN and highway gate is effective.

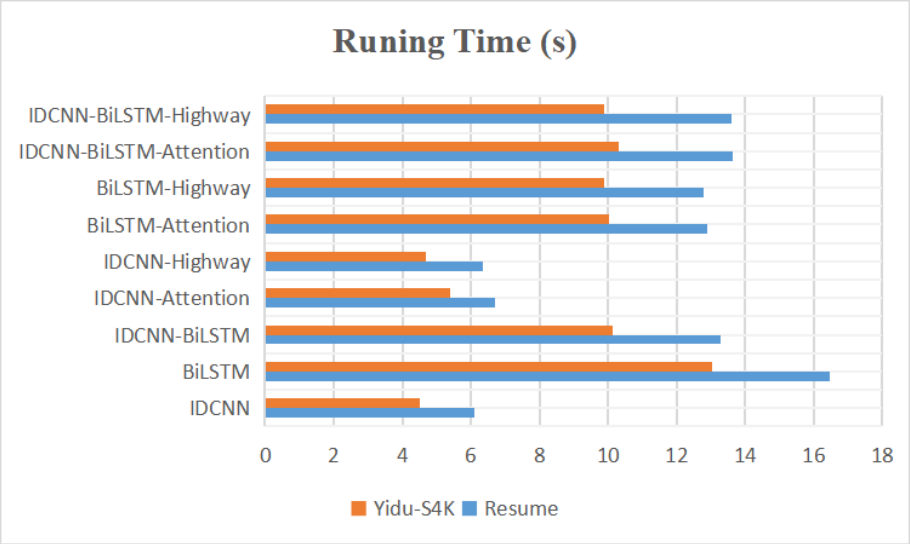


**Fig. 5.** Average running time comparison in seconds(s).

## 5. Conclusion

In this study, we develop an IDCNN-BiLSTM-Highway model to accomplish Chinese NER tasks. An IDCNN layer is added before the BiLSTM layer to allow the model to capture longer context information. A novel highway network with special transform gates is added after the BiLSTM layer to incorporate information from the embedding and BiLSTM middle layers. Through the above experiments, it is concluded that our model has higher performance than the recent deep learning based NER models on the two Chinese data sets.

In the future, we will verify the conclusions in this study with more datasets. Additionally, we plan to explore other approaches that combine context information from different network layers to improve NER performance further.

## References

[1] B. Zqga, B. Gfca, B. Ymha, L. Gang, L. Fang, "Semantic relation extraction using sequential and tree-structured lstm with attention," *Information Sciences*, vol. 509, pp. 183-192, 2020. Article (CrossRef Link)

[2] D. Li, L. Huang, H. Ji, J. Han, "Biomedical Event Extraction based on Knowledge-driven Tree-LSTM," in *Proc. of ACL*, vol. 1, pp. 1421-1430, 2019. Article (CrossRef Link)

[3] C. Su, H. Huang, S. Shi, P. Jian, X. Shi, "Neural machine translation with Gumbel Tree-LSTM based encoder," *Journal of Visual Communication and Image Representation*, vol. 71, pp. 102811, 2020. Article (CrossRef Link)
[4] B. Tang, X. Wang, J. Yan, Q. Chen, "Entity recognition in Chinese clinical text using attention-based CNN-LSTM-CRF," *BMC Medical Informatics and Decision Making*, vol. 19, no. Suppl 3, pp. 74, 2019. Article (CrossRef Link)
[5] G. Luo, X. Huang, C.Y. Lin, Z. Nie, "Joint Entity Recognition and Disambiguation," in *Proc. of EMNLP*, pp. 879-888, 2015. Article (CrossRef Link)
[6] R. Leaman, C. H. Wei, and Z. Lu, "tmChem: a high performance approach for chemical named entity recognition and normalization," *Journal of Cheminformatics*, vol. 7, no. Suppl 1, pp. S3, 2015. Article (CrossRef Link)
[7] G.T. Ngomp, S. Harispe, G. Zambrano, J. Montmain, S. Mussard, "Detecting Sections and Entities in Court Decisions Using HMM and CRF Graphical Models," *Advances in Knowledge Discovery and Management*, vol. 8, pp. 61-86, 2019. Article(CrossRef Link)
[8] T. Young, D. Hazarika, S. Poria, E. Cambria, "Recent Trends in Deep Learning Based Natural Language Processing," *IEEE Computational Intelligence Magazine*, vol. 13, pp. 55-75, 2018. Article (CrossRef Link)
[9] R. Collobert, et al., "Natural Language Processing (almost) from Scratch," *Journal of Machine Learning Research*, vol. 12, pp. 2493-2537, 2011. Article (CrossRef Link)
[10] E. Strubell, P. Verga, D. Belanger, A. McCallum, "Fast and Accurate Entity Recognition with Iterated Dilated Convolutions," in *Proc. of EMNLP*, pp. 2670-2680, 2017. Article (CrossRef Link)
[11] K. Yao, B. Peng, G. Zweig, D. Yu, F. Gao, "Recurrent Conditional Random Field for Language Understanding," in *Proc. of IEEE ICASSP*, pp. 4077-4081, 2014. Article (CrossRef Link)
[12] M. Basaldella, E. Antolli, G. Serra, C. Tasso, "Bidirectional LSTM Recurrent Neural Network for Keyphrase Extraction," in *Proc. of Digital Libraries and Multimedia Archives, IRCDL*, pp. 180-187, 2018. Article (CrossRef Link)
[13] P. Corbett, and J. Boyle, "Chemlistem: Chemical named entity recognition using recurrent neural networks," *Journal of Cheminformatics*, vol. 10, no .1, pp. 59, 2018. Article (CrossRef Link)
[14] X. Ma, E. Hovy, "End-to-end sequence labeling via bidirectional LSTM-CNNs-CRF," in *Proc. of ACL*, vol. 1, pp. 1064-1074, 2016. Article (CrossRef Link)
[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, "Attention Is All You Need," *Advances in Neural Information Processing Systems*, pp. 6000-6010, 2017. Article (CrossRef Link)
[16] A. Galassi, M. Lippi, and P. Torroni, "Attention in Natural Language Processing," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 10, pp. 4291-4308, 2021. Article (CrossRef Link)
[17] B. Zqga, B. Gfca, B. Ymha, L. Gang, L. Fang, "Semantic relation extraction using sequential and tree-structured lstm with attention," *Information Sciences*, vol. 509, pp. 183-192, 2020. Article (CrossRef Link)
[18] W. Li, F. Qi, M. Tang, Z. Yu, "Bidirectional LSTM with self-attention mechanism and multi-channel features for sentiment classification," *Neurocomputing*, vol. 387, pp. 63-77, 2020. Article (CrossRef Link)
[19] Z. Lin, M. Feng, C.N. dos Santos, M. Yu, B. Xiang, B. Zhou, Y. Bengio, "A Structured Self-Attentive Sentence Embedding," in *Proc. of ICLR*, 2017. Article (CrossRef Link)
[20] Z. Tan, M. Wang, J. Xie, Y. Chen, X. Shi, "Deep Semantic Role Labeling With Self-Attention," in *Proc. of The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pp. 4929–4936, 2018. Article (CrossRef Link)
[21] P. Verga, E. Strubell, A. McCallum, "Simultaneously Self-Attending to All Mentions for Full-Abstract Biological Relation Extraction," in *Proc. of ACL*, vol. 1, pp. 872-884, 2018. Article (CrossRef Link)
[22] B. Tang, X. Wang, J. Yan, Q. Chen, "Entity recognition in Chinese clinical text using attention-based CNN-LSTM-CRF," *BMC Medical Informatics and Decision Making*, vol. 19, no. Suppl 3, pp. 74, 2019. Article (CrossRef Link)

[23] F. Yu, V. Koltun, "Multi-Scale Context Aggregation by Dilated Convolutions," in *Proc. of ICLR*, 2016. Article (CrossRef Link)
[24] K.X. He, W.Q. Zhang, J. Liu, Y. Liu, "Dilated-Gated Convolutional Neural Network with A New Loss Function on Sound Event Detection," in *Proc. of APSIPA ASC*, pp. 1491-1495, 2019. Article (CrossRef Link)
[25] Z. Huang, X. Wei, and Y. Kai, "Bidirectional LSTM-CRF Models for Sequence Tagging," *Computer Science*, 2015. Article (CrossRef Link)
[26] M. Labeau, K. Lser, A. Allauzen, "Non-lexical neural architecture for fine-grained POS Tagging," in *Proc. of EMNLP*, pp. 232-237, 2015, Article (CrossRef Link)
[27] Chiu, Jpc, and E. Nichols, "Named Entity Recognition with Bidirectional LSTM-CNNs," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 357-370, 2016. Article (CrossRef Link)
[28] R.K. Srivastava, K. Greff, and J. Schmidhuber, "Highway Networks," in *Proc. of the Deep Learning Workshop, International Conference on Machine Learning*, Lille, France, 2015. Article (CrossRef Link)
[29] L. Liu, J. Shang, F. Xu, R. Xiang, and J. Han, "Empower sequence labeling with task-aware neural language model," in *Proc. of AAAI-32*, 2018. Article (CrossRef Link)
[30] Y. Zhang, J. Yang, "Chinese NER Using Lattice LSTM," in *Proc. of ACL*, vol. 1, pp. 1554-1564, 2018. Article (CrossRef Link)
[31] A.J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, pp. 260-269, 1967. Article (CrossRef Link)
[32] J. Cheng, J. Liu, X. Xu, D. Xia, L. Liu and V. S. Sheng, "A review of Chinese named entity recognition," *KSII Transactions on Internet and Information Systems*, vol. 15, no. 6, pp. 2012-2030, 2021. Article (CrossRef Link)

**Tinglong Tang** received the Ph.D. degree in Control Science and Engineering from Zhejiang University of Technology in 2018. He is an Associate Professor at China Three Gorges University. His research interests focus on machine learning and intelligent information processing.

**Yunqiao Guo** received the Bachelor degree in Communication Engineering from Shandong University of Technology in 2018. He is currently working toward the M.S. degree at China Three Gorges University. His research interests focus on natural language processing.

**Qixin Li** received the M.S. degree in Computer Technology from China Three Gorges University, Yichang, China, in 2021. Her research interests focus on natural language processing and deep learning.

**Mate Zhou** received the M.S. degree in Computer Technology from China Three Gorges University, Yichang, China, in 2020. His research interests focus on natural language processing and deep learning.

**Wei Huang** received the Ph.D. degree in Optical Engineering from the Institute of Modern Optics, Nankai University, Tianjin, China, in 2016. She is currently an Associate Professor with the School of Computer Science and Engineering, Tianjin University of Technology, Tianjin, China. Her research interests include inverse design and prediction of optical structure based on machine learning pattern recognition.

**Yirong Wu** received the Ph.D. degree in Computer Science from University of Wisconsin-Milwaukee, USA, in 2008. He is a Professor at China Three Gorges University. His research interests focus on natural language processing and image processing.