IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 31, 2019, accepted September 10, 2019, date of publication September 19, 2019, date of current version October 2, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2942433

LSTM-CRF Neural Network With Gated Self Attention for Chinese NER

YANLIANG JIN[1], JINFEI XIE[1], WEISI GUO[2,3], (Senior Member, IEEE), CAN LUO[1], DIJIA WU[1], AND RUI WANG[1]

[1]Key Laboratory of Specialty Fiber Optics and Optical Access Networks, Joint International Research Laboratory of Specialty Fiber Optics and Advanced Communication, Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China
[2]School of Engineering, University of Warwick, Coventry CV4 7AL, U.K.
[3]The Alan Turing Institute, London NW1 2DB, U.K.

Corresponding author: Yanliang Jin (wuhaide@shu.edu.cn)

This work was supported in part by the NSFC, China, under Grant 61771299, in part by the National Key Research and Development Program of China under Grant 2018YFB2101303, in part by the Key Laboratory of Specialty Fiber Optics and Optical Access Networks, Shanghai University, under Grant SKLSFO2012-14, in part by the Key Laboratory of Wireless Sensor Network and Communication, Shanghai Institute of Microsystem and Information Technology, in part by the Shanghai Education Committee, Chinese Academy of Sciences, in part by the Shanghai Science Committee under Grant 12511503303, Grant 14511105602, and Grant 14511105902, in part by the H2020 under Grant 778305, and in part by the Innovate U.K. under Grant 10734.

ABSTRACT Named entity recognition (NER) is an essential part of natural language processing tasks. Chinese NER task is different from the many European languages due to the lack of natural delimiters. Therefore, Chinese Word Segmentation (CWS) is usually regarded as the first step of processing Chinese NER. However, the word-based NER models relying on CWS are more vulnerable to incorrectly segmented entity boundaries and the presence of out-of-vocabulary (OOV) words. In this paper, we propose a novel character-based Gated Convolutional Recurrent neural network with Attention called GCRA for Chinese NER task. In particular, we introduce a hybrid convolutional neural network with gating filter mechanism to capture local context information and a highway neural network after LSTM to select characters of interest. The additional gated self-attention mechanism is used to capture the global dependencies from different multiple subspaces and arbitrary adjacent characters. We evaluate the performance of our proposed model on three datasets, including SIGHAN bakeoff 2006 MSRA, Chinese Resume, and Literature NER dataset. The experiment results show that our model outperforms other state-of-the-art models without relying on any external resources like lexicons and multi-task joint training.

INDEX TERMS Chinese NER, gating mechanism, highway neural network, self-attention.

I. INTRODUCTION

Named entity recognition (NER) plays a critical role in the field of natural language processing (NLP). This task aims to extract and categorize entities with specific meanings in the unstructured text, such as person (PER), location (LOC), and organization (ORG), etc. NER is one of the most widely used and key technologies in information extraction. Also, it is the chief work of NLP tasks such as relation extraction [1], event extraction [2], and question answering system [3]. Therefore, it has a high value of utility to conduct in-depth research on the NER task.

Compared with the NER of Indo-European languages represented by English, Chinese NER task is more complicated. There are apparent inflections in English (singular, plural, tense, etc.), but Chinese lacks these inflections. Besides, Chinese has the problems of fuzzy word boundaries, complex entity structure, and various forms of expression, which make the Chinese NER task more difficult. Therefore, the correct identification of named entities in Chinese text is of great significance for subsequent Chinese information processing tasks.

At present, a mature method to solve the NER task is to model the NER problem into a sequence labeling problem. The standard method of existing state-of-the-art models for English NER can effectively capture context feature information by using BiLSTM-CRF models [4]–[7]. However, there are no apparent delimiters between words in Chinese sentences, and we usually perform word segmentation before the sequence is fed into the word-based model. Each segmented

The associate editor coordinating the review of this manuscript and approving it for publication was Qichun Zhang.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

word is mapped to fixed-length word representation. Then we use the word-level sequence labeling model, which is the same as the method of dealing with English NER.

However, entity boundaries are associated with segmentation results. The performance of subsequent NER task is limited by the incorrect NER labeling, which results from segmentation errors. Moreover, many named entities are considered as OOV words in the word-based model because of the large number of Chinese words. Besides, after the word segmentation of the word-based model, the parameter size of the embedding layer is significantly increased, which result in data sparsely problems and lead to overfitting. Let's take "南京市长江大桥 (Nanjing Yangtze River Bridge)" as a typical example. Due to the limitations of Chinese linguistic features, the boundaries of characters (words) are often ambiguous. The same sentence may have distinct segmentation after performing word segmentation. For different word segmentation granularity, the sequence "南京市长江大桥" can be divided into "南京市 (Nanjing City)/ 长江大桥 (Yangtze River Bridge)" and "南京 (Nanjing City)/ 市长 (Mayor)/ 江大桥 (Daqiao Jiang)" respectively. As shown in Table 1, after performing word segmentation, we may get completely different recognition boundary information, which leads to two distinct sequence labeling outcomes. The word-based models cannot judge right or wrong, which results in incorrect entity recognition.

TABLE 1. Examples of word segmentation.

Sentence	南京市长江大桥						
Result1	Nanjin City			Yangtze River Bridge			
	南	京	市	长	江	大	桥
	B-LOC	I-LOC	E-LOC	B-LOC	I-LOC	I-LOC	E-LOC
Result2	Nanjin		Mayor		Daqiao Jiang		
	南	京	市	长	江	大	桥
	B-LOC	E-LOC	B-TITLE	E-TITLE	B-PER	I-PER	E-PER

Recently, studies have shown that character-level representation can avoid many of the listed above problems. And researchers found word-based models underperform character-based models in deep learning-based Chinese NER task [8]–[11]. Due to the polysemy and polymorphism of Chinese characters, the NER based on the pure character only focuses on the per-character information for losing the latent word and word sequence information. For this problem, it is worth exploring how to effectively integrate segmentation information into character-based models for better semantic understanding.

To overcome the shortcomings of the traditional character-based models, we propose a new neural network, called the GCRA, to improve the performance of Chinese NER task. Firstly, for the embedding layer, we apply the label segmentation vector softly concatenating into character embeddings. It uses word sequence information indirectly. So the model can not only avoid the problem of error propagation caused by word segmentation error but also achieve excellent results based on character and word information. Next, the character representation is fed into the hybrid gated convolutional layer to carry out detailed feature extraction and generate implicitly local feature information connection. Further, the highway neural network [12] is utilized to refine the hidden representation of BiLSTM. Finally, the self-attention layer is employed to capture context-related information in different multiple subspaces, which can better understand the sentence structure and ultimately improve the performance.

In this paper, we compare our model with the state-of-the-art methods on three datasets, including SIGHAN bakeoff 2006 MSRA, Chinese Resume, and Literature NER dataset. The three datasets come from news domain, social media domain, and literature domain respectively.

The main contributions of this paper can be summarized as follows:

- We propose a novel neural model called the GCRA model for Chinese NER task. The model can not only exploit local context features effectively but also capture the global dependencies of the whole character sequence.
- We design a character-level hybrid gated convolutional neural network which combines the dilated gated convolution with the standard gated convolution. It can effectively generate local feature information connection and avoid gradient vanishing during training.
- We conduct our experiment on various Chinese NER datasets in different domains. The experimental results demonstrate that our model outperforms other state-of-the-art models without using any external lexicon resources and multi-task joint training.

The remainder of the paper is organized as follows. Section II reviews the related work on Chinese NER. Section III presents the main idea of the proposed GCRA model. Section IV demonstrates the experimental results and analysis. Section V concludes our works.

II. RELATED WORK

Significant research has devoted to the NER task. The NER system in early was mainly based on rules and dictionaries, which has the shortcoming of poor expansibility and absent ability in finding OOV words. With the advent of statistical machine learning, the NER task is abstracted into a sequence labeling problem. Traditional sequence labeling models extensively utilized Hidden Markov Models (HMM) [13] and Conditional Random Fields (CRF) [14] in the NER task. However, all these models are heavily relying on feature engineering and external resources.

In recent years, deep learning has provided a new approach to solve the problems of natural language processing, which has attracted considerable critical attention. Given the shortcomings of feature engineering, deep learning is proposed as a useful tool for automatic learning, distributed representation of words, and deep feature extraction. Deep neural networks are used in deep learning to replace the artificial feature engineering model of traditional machine learning. To address the NER problems in the English field, models based on neural network demonstrate their excellent

performance in identifying entities. Collobert *et al.* [15] proposed CNN-CRF model to extract the depth feature for sequence labeling tasks automatically. Huang *et al.* [16] proposed a bidirectional LSTM-CRF network structure for sequence tagging task. But their models use the feature connection tricks to combine the hand-crafted spelling features and context features with word embeddings as the input vectors to the neural network. Lample *et al.* [4] presented a bidirectional LSTM-CRF architecture which combines word-level features with character-level features, and they applied another LSTM layer to generate character-level features. Similarly, Ma and Hovy [5] conducted the character Convolutional Neural Network (CNN) to extract English character-level features based on the LSTM-CRF network structure. Chiu and Nichols [6] reported a hybrid of bidirectional LSTM and CNNs structure, which automatically detects word-level and character-level features.

The development of Chinese NER research is relatively late, and the related research is more difficult because of the particularity of Chinese word information. Some researchers also consider Chinese NER task as a character sequence labeling problem and take advantage of external data to compensate for insufficient annotated corpus resources. In particular, in Collobert *et al.* [15], Passos *et al.* [17], Huang *et al.* [16], and Luo *et al.* [18], the researchers leveraged lexicon features to improve performance. Peters *et al.* [19] pre-trained a neural bidirectional language model to augment word representations by introducing character-level knowledge.

The existing research indicates that the character-based methods are considered as an empirically better choice than word-based methods [8]– [11]. However, the character-based NER models carry only a limited amount of character information and cannot fully exploit latent word and word sequence information. To solve this problem, some researchers have studied how to better leverage word-level information for Chinese NER task. Some proposed to use segmentation information as soft features for NER task [20], [21]. Peng and Dredze [22] and Cao *et al.* [23] designed a multi-task learning model for joint learning Chinese NER tagging and Chinese word segmentation task simultaneously. Zhang and Yang [24] integrated latent word-level information into a character-based LSTM-CRF model by identifying candidate lexicon words from the sentence using a lattice-structured LSTM model. Zhang *et al.* [25] investigated a dynamic meta-embeddings method and applied it to Chinese NER task. They utilized the attention mechanism to combine features of both character and word granularity in the embedding layer. In the work of Zhu and Wang [26], they proposed a Convolutional Attention Network model, which used word segmentation vector as soft features to improve Chinese NER model performance. Their work precluded any external word embeddings and lexicon resources dependencies.

In our work, we enhance the input representation by utilizing the segmentation label vector concatenating into character embeddings directly. Besides, we design the hybrid gated convolution layer and gated self-attention network, which can effectively alleviate gradient vanishing during training and capture depth detailed feature. Experiments on several series of datasets show that our proposed GCRA model can significantly improve the performance of the Chinese NER task.

III. MODELS

As with most named entity recognition methods, our work also turns NER task into a sequence labeling problem. To eliminate the effects of word segmentation error propagation, we utilize character-level BiLSTM-CRF as our basic structure and apply the BIOES tagging scheme for the Chinese NER task. The overview architecture of our proposed model is shown in Figure 1. The model mainly consists of five layers: embedding layer, hybrid gated convolution layer, highway BiLSTM layer, gated self-attention layer, and CRF decode layer. Each part of our proposed model will be presented in detail in the following sections.

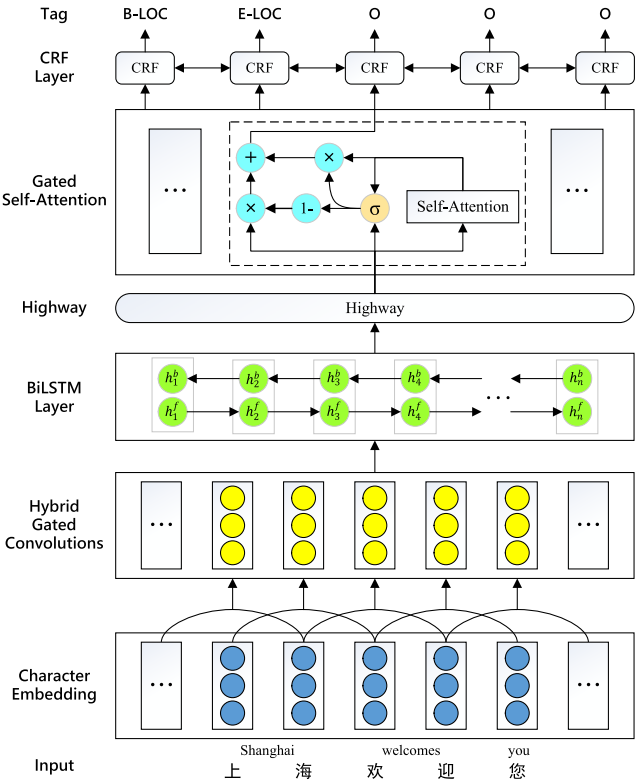


FIGURE 1. The whole architecture of our proposed GCRA model.

A. EMBEDDING LAYER

Most research shows that applying segmentation as soft features for character-based Chinese NER models can lead to improved performance [20], [21]. In this work, we concatenate the segmentation label vector into character embedding for augmenting the input representation. The word segmentation information is represented by BIOES scheme. Formally, in the Chinese NER task, we denote a input sentence as

$X = \{x_1, x_2, \ldots, x_n\}$, where x_i represents the *ith* character in the sentence X. Then, we map discrete characters into the distributed feature representations on the embedding layer. The input representation for each character is embedded in distributional space as $\boldsymbol{x}_i^{\mathrm{c}}$:

$$\boldsymbol{x}_i^c = \left[\boldsymbol{e}^c\left(x_i\right) \oplus \boldsymbol{e}^s\left(seg\left(x_i\right)\right)\right] \tag{1}$$

where $\boldsymbol{e}^{\mathrm{c}}$ and $\boldsymbol{e}^{\mathrm{s}}$ denote a pre-trained character embedding lookup table and a BIOES scheme segmentation label embedding lookup table, respectively. And the $\oplus$ is the connection operator. The formula $seg\left(x_i\right)$ represents the segmentation label of each character x_i which is given by a word segmenter.

B. HYBRID GATED CONVOLUTION LAYER

We use hybrid gated convolutions to extract local feature information connection and context information. As shown in Figure 2, it has two separate blocks. The left block is the dilated gated convolution block, which consists of two layers of dilated convolution and a gated filtering mechanism. It is similar to the highway network. The right block is the normal gated convolution block, which has a standard convolutional layer with gated linear units [27]. We splice the two separate outputs together as the final output of the hybrid gated convolution layer.

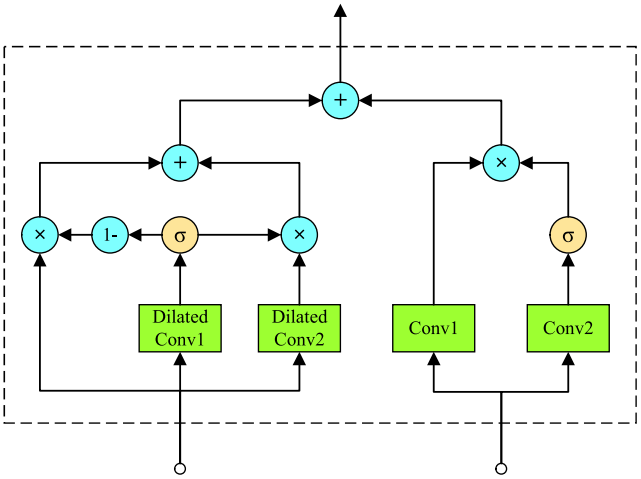


FIGURE 2. The architecture of hybrid gated convolution layer.

1) NORMAL GATED CONVOLUTION

Dauphin *et al.* [27] have shown that the gating mechanism can improve the performance of language modeling tasks. The gated convolutional network adds a gating switch to control the information flow. These gates can alleviate gradient vanishing during training since there is a convolution without any activation function. For the embedding output X, the gated convolution layer output can be expressed as:

$$Y_g(X) = (X * W + b) \otimes \sigma(X * V + c) \tag{2}$$

where $*$ denotes convolution operator, W and V with b and c denote kernels and biases respectively, which are parameters to be learned. σ represents the sigmoid function, and $\oplus$ means the element-wise product between the matrix.

2) DILATED GATED CONVOLUTION

Strubell *et al.* [28] applied the iterated dilated convolutions to expand the receptive fields, which have better capacity than traditional CNNs for NER task. To enable the CNN model to capture farther distances without increasing the model parameter number, we use a dilated convolution. In normal CNN, each kernel window consists of adjacent inputs, whereas dilated convolutions define wider effective input width by introducing dilation between inputs. Given a 1-D convolutional filter $w = \{w_{-r}, w_{-r+1}, \ldots, w_r\}$ of a widow size $l = 2r + 1$ and the input sequence $X = \{x_1, x_2, \ldots, x_n\}$. With dilation d, the convolutional operator is applied to each token x_t with output c_t is defined as:

$$c_t = \sum_{k=-r}^{r} w_t \cdot x_{t+k*d} + b \tag{3}$$

where dilation $d = 1$ is equivalent to a normal convolution.

Wu *et al.* [29] proposed gated linear dilated residual network for reading comprehension task, which mainly consists of dilated convolution and gated linear units with the residual connection. For our dilated gated convolution block, we also use dilated convolution instead of normal convolution to extend the receptive field. But for gated filtering mechanism, it is more similar to the highway network. We combine the residual connection and gated convolutional neural network to achieve selective multi-channel transmission of information. We use $C(X)$ to represent the output of the dilated convolution. The final dilated gated convolution block output can be expressed as:

$$Y_{dg}(X) = X(1-\theta) + C_2(X) \otimes \theta \tag{4}$$

$$\theta = \sigma\left(C_1(X)\right) \tag{5}$$

where X is the input of this layer, $C_1(X)$ and $C_2(X)$ mean different dilated convolution output respectively. σ represents the sigmoid function, and $\otimes$ denotes the element-wise product between the matrix. After comparing to experimental results with the different dilated rate, we use two-layer dilated gated convolutions with dilated rate 1 and 2. So the output of a hybrid gated convolution is as follows:

$$g = \left[Y_g(X) \oplus Y_{dg}(X)\right] \tag{6}$$

where the $\oplus$ represents the connection operator.

C. HIGHWAY-LSTM LAYER

Hochreiter and Schmidhuber [30] proposed LSTM to solve gradient vanishing and exploding of traditional recurrent neural network. The key role is to utilize adaptive gating mechanism and the memory cell. A typical LSTM cell structure is depicted in Figure 3. Each LSTM units is composed of a loop connected memory cell c_t, a forget gate f_t, an input gate i_t, and an output gate o_t. The input gate i_t is used to control the input signal flow. The output gate o_t is used to control the

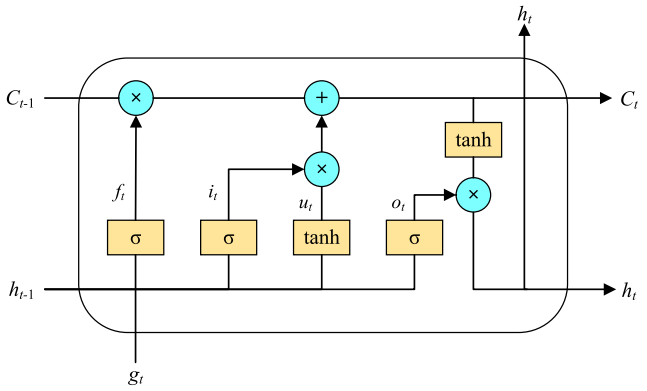


FIGURE 3. The architecture of LSTM layer.

signal strength flowing to the next unit, and the forget gate f_t is used to control the cell state before forgetting. Defining $\boldsymbol{g} = [g_1, g_2, \ldots, g_n]$ outputted by CNN layer as input. Then, the LSTM units at step t could be expressed as:

$$i_t = \sigma\left(W_i g_t + U_i h_{t-1} + b_i\right) \tag{7}$$
$$f_t = \sigma\left(W_f g_t + U_f h_{t-1} + b_f\right) \tag{8}$$
$$o_t = \sigma\left(W_o g_t + U_o h_{t-1} + b_o\right) \tag{9}$$
$$u_t = \tanh\left(W_u g_t + U_u h_{t-1} + b_u\right) \tag{10}$$
$$c_t = i_t \otimes u_t + f_t \otimes c_{t-1} \tag{11}$$
$$h_t = o_t \otimes \tanh\left(c_t\right) \tag{12}$$

where W_i, U_i, W_f, U_f, and W_o, U_o represent the weight matrix of input gate i_t, forget gate f_t, and output gate o_t, respectively. The parameters b_i, b_f, b_o are bias vectors of the input gate, forget gate, and output gate, respectively. The parameters W_u, U_u, and b_u are the weight matrix and bias vectors of new memory content u_t. The h_t is LSTM hidden state, σ is the element-wise sigmoid function. tanh is a hyperbolic tangent function, and the operator $\otimes$ denotes the element-wise multiplication.

The unidirectional LSTM only retains information from the past sequence of vectors, because the hidden state flow is passed from the front to back. To leverage the past and future sequence information, we use a bidirectional LSTM to capture the context features for sentence. So the hidden state of BiLSTM is as follows:

$$h_t = \left[\overrightarrow{h_t} \oplus \overleftarrow{h_t}\right] \tag{13}$$

where $\overrightarrow{h_t} \in R^{d_h}$ and $\overleftarrow{h_t} \in R^{d_h}$ are the hidden states of the forward and backward LSTM at position t, respectively. The $\oplus$ represents the connection operator.

Highway network allows information to pass through layers of the deep neural network at high speed, which effectively slows down the problems of the gradient. In this paper, we use the highway network to control the information flow with an adaptive gate network. The overview architecture of highway-LSTM is illustrated in Figure 4. The output of

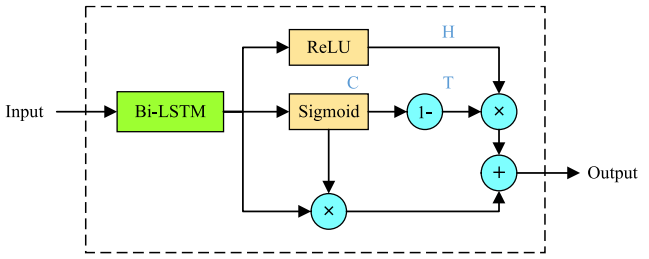


FIGURE 4. The architecture of highway network layer.

highway-LSTM layer is calculated as follows:

$$tg = \sigma\left(W_g h + b_g\right) \tag{14}$$
$$z = tg \otimes f\left(W_h h + b_h\right) + (1 - tg) \otimes h \tag{15}$$

where σ is the element-wise sigmoid function, $\otimes$ is the element-wise product, and f is the rectified linear unit. The W_g, W_h and b_g, b_h represent the weight matrix and bias vectors, respectively. The tg denotes the transform gate, which controls how much information is converted and passed to the next layer. And the $(1 - tg)$ is called carry gate, which allows the input to be passed to the next layer directly. Therefore, the highway network input h and output z require to be the same shape.

D. GATED SELF-ATTENTION LAYER

Self-Attention is a mechanism of attention that relates different locations of a single sequence to calculate an interactive representation of the sequence. Recent evidence suggests that it performs well on a variety of tasks, such as machine translation [31], semantic role labeling [32], and relation extraction [33]. Inspired by these works, we utilize the multi-head self-attention mechanism to capture the global sequence information from multiple subspaces and exploit the inner features contained in the text. Attention is essentially a mapping function consisting of many Queries and Key-values. For self-attention, we use the highway-LSTM output $\mathbf{Z} = [z_1, z_2, \ldots, z_n]$ to initialize $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$. The scaled dot-product attention could be calculated as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{16}$$

where $\boldsymbol{Q} \in R^{n \times 2\,d_h}$, $\boldsymbol{R} \in R^{n \times 2\,d_h}$ and $\boldsymbol{V} \in R^{n \times 2\,d_h}$ denote query matrix, keys matrix and value matrix respectively. $\sqrt{d_k}$ equals the dimension of hidden units of BiLSTM, and plays a regulating role, controlling the inner product of $\mathbf{Q}$ and $\mathbf{K}$ not too large.

The essence of multi-head attention is to perform multiple self-attention calculations, which can make the model capture more features from different representation subspaces. The multi-head attention mechanism will linearly project the $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ through the parameter matrix without the sharing of parameters, and then perform the scaled dot-product attention. This process repeats for m times in parallel, and

finally splices the results and linearly projects to get the new representation. The final result of S could be expressed as:

$$head_i = \text{Attention}\left(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V\right) \quad (17)$$

$$S = \mathbf{W}_s \otimes (head_1 \oplus \ldots \oplus head_m) \quad (18)$$

where $\mathbf{W}_i^Q \in R^{2\,d_h \times \widetilde{d}_k}$, $\mathbf{W}_i^K \in R^{2\,d_h \times \widetilde{d}_k}$, $\mathbf{W}_i^V \in R^{2\,d_h \times \widetilde{d}_k}$ and $\mathbf{W}_s \in R^{2\,d_h \times 2\,d_h}$ represent the projection matrix, and $\widetilde{d}_k = 2\,d_h/m$. The $\otimes$ is the element-wise product and $\oplus$ represents the connection operator.

The tag of each position in the sentence has different degrees of dependence on the context. We introduce a gating mechanism to generate a representation combining context features and self-features. The gated output representation can be expressed as:

$$P = (\sigma(z) \otimes S) \oplus ((1 - \sigma(z)) \otimes z) \quad (19)$$

where σ is the sigmoid function, $\otimes$ is the element-wise product, and $\oplus$ represents the connection operator.

Finally, we carry out a fully connected layer to compute the probability scoring matrix. It can be described as:

$$O = W_p P + b_p \quad (20)$$

where $W_p \in R^{|k| \times 4\,d_n}$ and $b_p \in R^{|k|}$ are the trainable parameters. $|k|$ denotes the number of output labels, and n is the length of the input sequence. O is the output probability matrix, whose size is $n \times k$.

E. CRF LAYER

In the NER sequence labeling task, there is a strong dependency between the tags of adjacent characters. For example, the I-PER (I-person) tag should be followed by a B-PER (B-person) tag or I-PER tag. Also, the I-LOC (I-location) tag cannot appear behind the B-PER tag or S-PER (S-person) tag. Therefore, instead of making independent tagging decisions using the output of the fully connected layers, we utilize CRF to inference the entity tags outputs of a sequence jointly. The CRF can express this dependence and add some constraints to the final predicted tag sequence effectively.

The CRF layer is trained to predict the most possible tag sequence $y = \{y_1, y_2, \ldots, y_n\}$ for a given sentence $X = \{x_1, x_2, \ldots, x_n\}$. The score of the tag sequence can be calculated as:

$$s(X, y) = \sum_{i=1}^{n} \left(O_{i,y_i} + T_{y_i, y_{i+1}}\right) + T_{y_0, y_1} \quad (21)$$

where O_{i,y_i} represents the score of the $y_i th$ tag of the ith character x_i in the sentence. T is a transition score matrix, which denotes the scores of transition from tag i to tag j. y_0 and y_{n+1} in the formula represent the start and end tags of a sentence, and we add them to the possible tag sets. Therefore, T is a square matrix of size k+2. Then, the probability of the ground-truth label sequence y is defined as:

$$p(y|X) = \frac{e^{s(X,y)}}{\sum_{\tilde{y} \in Y_x} e^{s(X,\tilde{y})}} \quad (22)$$

where $\tilde{y}$ denotes an arbitrary label sequence, and Y_x is the set of all possible output label sequences for the input X.

In decoding, we use the Viterbi algorithm [34] to predict the best path that obtains the highest scoring mark sequence:

$$\bar{y} = \underset{\tilde{y} \in Y_x}{\text{argmax}}\ s(X, \tilde{y}) \quad (23)$$

Given a set of manually labeled data $\{(x_i, y_i)\}|_{i=1}^N$, we add L2 regularization to the negative log-likelihood loss for training. The specific loss function is as follows:

$$L(\theta) = -\Sigma_{i=1}^N \log\left(P\left(y_i|x_i\right)\right) + \lambda\|\theta\|^2 \quad (24)$$

where λ is the L2 regularization hyper-parameter, and θ denotes the parameter set. For training, we minimize the loss function L through shuffled mini-batches stochastic gradient descent method with the Adam update rule.

IV. EXPERIMENTS

In this section, to evaluate the effectiveness of the proposed GCRA model, we compare our model with previous state-of-the-art methods on different Chinese NER datasets. We will describe the details of different datasets, settings of parameters, and results of our experiments.

TABLE 2. Detailed statistics of datasets.

Dataset	Type	Train	Dev	Test	Entity Type
MSRA	Sentences	46.4k	-	4.4k	3
	Chars	2169.9k	-	172.6k	
	Entities	74.8k	-	6.2k	
Literature	Sentences	24.2k	1.9k	2.8k	7
	Chars	1044.9k	86.4k	119.5k	
	Entities	133.1k	10.5k	16.2k	
Resume	Sentences	3.8k	0.46k	0.48k	8
	Chars	124.1k	13.9k	15.1k	
	Entities	1.34k	0.15k	0.16k	

A. DATASETS

We evaluate our proposed model on three Chinese NER datasets, which include MSRA NER dataset [35], Literature NER dataset [36], and Chinese resume dataset [24]. Table 2 provides detailed statistic information for each dataset.

- **MSRA** dataset comes from SIGHAN 2006 shared task for Chinese NER [35]. This dataset is news in simplified Chinese, which contains three annotated named entity types: PER (Person), ORG (Organization) and LOC (Location). The development set is not available in the MSRA dataset. Therefore, we sample 10% data of training set as the development set.
- **Literature** dataset is annotated from hundreds of Chinese literature articles, which contain seven entity types: Thing, Person, Location, Time, Metric, Organization, and Abstract. The training, development, and test sets have been divided on the Literature dataset.
- **Resume** dataset consists of resume of senior executives from listed companies in the Chinese stock market, which contains eight types of named entities:

CONT (Country), EDU (Educational Institution), LOC, PER, ORG, PRO (Profession), RACE (Ethnicity Background), and TITLE (Job Title).

B. EXPERIMENTAL SETTINGS

We adopt BIOES tagging scheme where each character in the corpus is labeled as one of B (Begin), I (Inside), O (Outside), E (End), and S (Single). Studies have suggested that BIOES scheme is remarkably better than BIO scheme since BIOES can get more detailed position information [37].

TABLE 3. Hyper-parameter values.

Parameter	Value	Parameter	Value
char emb dim	300	biLSTM layer	1
CNN window size	3	biLSTM hidden	300
CNN hidden	300	highway gate bias	-1
learning rate	0.001	muti-head	8
dropout	0.5	regularization	0.005

Table 3 shows the values of hyper-parameters for our model. In particular, we make our parameter selection according to the performance on the development set of datasets. We set the character embedding size, hidden sizes of CNN and Bi-LSTM to 300 dims. The sliding window size of all convolutional layers is set to 3. The highway gate bias is initialized with -1 vector. We exploited Adam [38] as the model optimization with an initial learning rate of 0.001, and the gradient norms clipped at 5.0. The projection number of self-attention m is 8. To avoid overfitting, we set the L2-norm regularization parameter as 0.005, and apply dropout to embedding layer with a rate of 0.5. The batch normalization is utilized to the outputs of the self-attention layer. For the batch size, we set the batch size of MSRA dataset as 64 and other datasets as 20, respectively. The character embeddings utilized in our proposed model are from Chinese-Word-Vectors [39], which are pre-trained on Baidu Encyclopedia corpus by Skip-Gram with Negative Sampling (SGNS).

For evaluation, same as most of the previous work, we also use the Precision (P), Recall (R), and F1 score as metrics to evaluate the recognition effectiveness of the model.

C. EXPERIMENT RESULTS

We compare our experimental results with previous state-of-the-art methods on MSRA dataset, Literature dataset, and Chinese Resume dataset, respectively. Besides, we propose two baselines and a GCRA model. In Table 4, 5, and 6, we use the Baseline to represent the BiLSTM + CRF model and Baseline + Highway to indicate Highway-LSTM + CRF model. The best experiment results in tables are in bold.

1) MSRA DATASET

Table 4 gives the experimental results conducted on the MSRA dataset. The first block is the results of previous models for Chinese NER on MSRA dataset. Chen *et al.* [40], Zhang *et al.* [41], and Zhou *et al.* [42] who employed the statistical model with rich hand-crafted features. Lu *et al.* [43] introduced multi-prototype embeddings features to Chinese NER task and Dong *et al.* [44] exploited neural LSTM-CRF with radical features in Chinese character. Yang *et al.* [45] proposed a five-stroke based CNN-BiRNN-CRF model for Chinese NER task by considering the semantic information as well as n-gram features. Cao *et al.* [23] used Adversarial Transfer Learning and self-attention to joint train Chinese NER task with Chinese word segmentation for better performance. Zhang and Yang [24] constructed a lattice LSTM structure to exploit word information in character sequence with incorporate lexicon information into the neural network. Although the model achieves state-of-the-art F1-score of 93.18%, it leverages external lexicon data, and the result may be affected by the quality of the lexicon. Zhang *et al.* [25] investigated a dynamic meta-embeddings method and applied it to Chinese NER task. Zhu and Wang [26] proposed a Convolutional Attention Network model to improve Chinese NER model performance and preclude word embedding and additional lexicon dependencies.

TABLE 4. Experimental results on MSRA dataset.

Models	P	R	F1
Chen et al. 2006 [40]	91.22	81.71	86.20
Zhang et al. 2006 [41]	92.20	90.18	91.18
Zhou et al. 2013 [42]	91.86	88.75	90.28
Lu et al. 2016 [43]	-	-	87.94
Dong et al. 2016 [44]	91.28	90.62	90.95
Yang et al. 2018 [45]	92.04	91.31	91.67
Cao et al. 2018 [23]	91.30	89.58	90.64
Zhang and Yang 2018 [24]	93.57	**92.79**	**93.18**
Zhang et al. 2019 [25]	90.59	91.15	90.87
Zhu and Wang 2019 [26]	93.53	92.42	92.97
Baseline	92.02	90.70	91.36
Baseline+Highway	92.25	92.19	92.22
GCRA Model	**93.71**	92.46	**93.08**

The second block in Table 4, we list the results of baselines and our proposed model. Our baseline model achieves an F1-score of 91.36% using only character embedding and soft-word information. We add a highway network for purifying the hidden representation of Bi-LSTM, and the experimental results show that the Baseline + Highway model has surpassed most of the previous methods. Compared with the state-of-the-art model proposed by Zhang and Yang [24], our character-based model gives a highly competitive accuracy of 93.71% without external lexicon data and multi-task joint training. Compared with state-of-the-art result among the character-based models proposed by Zhu and Wang [26], our GCRA model achieves higher F1-score of 93.08% to the character-based on the MSRA dataset.

2) LITERATURE DATASET

Table 5 shows the comparative results on the Literature dataset. Xu *et al.* 2018(a) [36] employed bi-directional LSTM for Chinese Literature NER, and Xu *et al.* 2018(b) [36] used CRF with the features template, which includes unigram and bigram features. The first two rows in the first block clearly show that CRF achieves better performance than

bi-directional LSTM, which probably attributed to the feature template. Zhang *et al.* 2019(a) [25] proposed DME-SUM model, which applied dynamic meta-embeddings method to combine the character and word vectors. Zhang *et al.* 2019(b) [25] presented DME-attention based model, which implemented two attention layers to integrate character and word information with a combination method of element-wise summation.

The results of our baselines and proposed models are listed in the second block of Table 5. Our baseline Bi-LSTM + CRF achieves an F1-score of 72.79%, and adding a highway network can improve F1-score to 73.48% which better than previous methods. Compared with the state-of-the-art model proposed by Zhang *et al.* 2019(b) [25], our GCRA model outperforms the state-of-the-art model without using external data and leads 1.26% increment of F1-score.

TABLE 5. Experimental results on Literature dataset.

Models	P	R	F1
Xu et.al. 2018(a) [36]	70.52	62.36	66.19
Xu et.al. 2018(b) [36]	**77.73**	65.91	71.33
Zhang et.al. 2019(a) [25]	73.38	71.19	72.58
Zhang et al. 2019(b) [25]	74.45	71.67	73.03
Baseline	74.18	71.46	72.79
Baseline+Highway	74.83	72.17	73.48
GCRA Model	75.34	**73.27**	**74.29**

3) RESUME DATASET

Table 6 shows the comparative results on the Chinese Resume dataset. The result in the first three rows of the first block respectively represents the char-based LSTM model, the word-based LSTM model, and the Lattice model proposed by Zhang and Yang [24]. Zhu and Wang 2019(a) [26] used BiGRU + CRF model and Zhu and Wang 2019(b) [26] leveraged CNN-BiGRU + CRF model for the Chinese Resume NER. Zhu and Wang 2019(c) [26] presented a Convolutional Attention Network model and achieves F1-score of 94.94% for Resume dataset.

TABLE 6. Experimental results on Chinese Resume dataset.

Models	P	R	F1
Zhang and Yang 2018(a) [24]	94.53	94.29	94.41
Zhang and Yang 2018(b) [24]	94.07	94.42	94.24
Zhang and Yang 2018(c) [24]	94.81	94.11	94.46
Zhu and Wang 2019(a) [26]	93.71	93.74	93.73
Zhu and Wang 2019(b) [26]	94.36	94.85	94.60
Zhu and Wang 2019(c) [26]	95.05	94.82	94.94
Baseline	94.07	94.48	94.28
Baseline+ Highway	94.96	94.79	94.87
GCRA Model	**95.75**	**95.33**	**95.54**

In the second block of Table 6, the results show that our proposed baseline + highway model achieves highly competitive F1-score of 94.87%. We can observe that our proposed character-based GCRA model outperforms the previous methods and achieves the state-of-the-art F1-score of 95.54% for Chinese Resume dataset, which demonstrates the effectiveness of our proposed model.

D. RESULTS ANALYSIS

With the introduction of the gating mechanism, our model can effectively avoid gradient vanishing during training and achieve the selective multi-channel transmission of information. Shown by Table 4, 5, and 6, we can observe that the baseline + highway model gains significant improvement in F1-score compared with the baseline model. It indicates that the gate network can perform more detailed feature extraction and learn more complicated dependencies. Our proposed GCRA model outperforms previous methods on Chinese Literature and Resume dataset and gives highly competitive results on MSRA dataset without utilizing any external resources. Compared with the baseline model, our proposed GCRA model lead 1.72%, 1.5%, and 1.26% noticeable improvements on MSRA dataset, Literature dataset and Resume dataset, respectively. It demonstrates that the effectiveness of our proposed model for Chinese NER task, which will better understand a sentence and achieve better recognition effect. However, the overall performance on Literature NER dataset is relatively low. And previous methods all get higher precision and lower recall. The lower recall rate means a lot of unknown entities cannot be recognized. It may be explained by the reason that there are various rhetorical devices and a large number of ambiguous cases in Chinese literature text. Nevertheless, the remarkable improvement on Literature NER dataset suggests that our proposed model can efficiently handle the problem of unknown entities.

V. CONCLUSION

In this paper, we propose a new model (GCRA) for Chinese NER task, which utilizes the gated filtering to refine the hidden representation and avert the problems of the gradient. In our model, we apply hybrid gated convolutions and highway-LSTM, and gated self-attention mechanism to learn the inner features of the sentence and capture the context information from multiple subspaces. Compared with previous state-of-the-art methods, the experiments on three datasets demonstrate that our proposed model can achieve better performance. Furthermore, our model does not depend on any external resources and domain-specific knowledge. Thus, it can be easily extended to other sequence labeling tasks, such as Chinese Word Segmentation and Part-of-Speech Tagging.

In the future, we will consider using transfer learning to integrate the knowledge of other NLP tasks in Chinese named entity recognition task to improve performance.

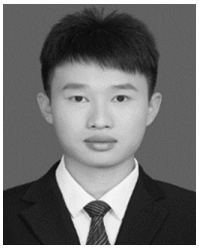
REFERENCES

[1] M. Miwa and M. Bansal, "End-to-end relation extraction using LSTMs on sequences and tree structures," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Berlin, Germany, vol. 1, Aug. 2016, pp. 1105–1116.

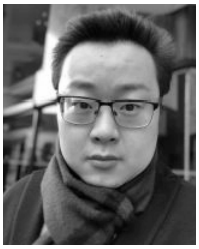
[2] Y. Chen, L. Xu, K. Liu, D. Zeng, and J. Zhao, "Event extraction via dynamic multi-pooling convolutional neural networks," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics, 7th Int. Joint Conf. Natural Language Process.*, vol. 1, 2015, pp. 167–176.

[3] X. Yao and B. Van Durme, "Information extraction over structured data: Question answering with freebase," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, Baltimore, Maryland, vol. 1, Jun. 2014, pp. 956–966.
[4] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics Hum. Lang. Technol.*, San Diego, CA, USA, Jun. 2016, pp. 260–270.
[5] X. Ma and E. Hovy, "End-to-end sequence labeling via Bi-directional LSTM-CNNs-CRF," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Berlin, Germany, vol. 1, Aug. 2016, pp. 1064–1074.
[6] J. P. C. Chiu and E. Nichols, "Named entity recognition with bidirectional LSTM-CNNs," *Trans. Assoc. Comput. Linguistics*, vol. 4, pp. 357–370, Dec. 2016.
[7] L. Liu, "Empower sequence labeling with task-aware neural language model," in *Proc. 32nd AAAI Conf. Artif. Intell. (AAAI),*, New Orleans, LA, USA, Feb. 2018, pp. 5253–5260.
[8] J. He and H. Wang, "Chinese named entity recognition and word segmentation based on character," in *Proc. 6th SIGHAN Workshop Chin. Lang. Process.*, Hyderabad, India, Jan. 2008, pp. 128–132.
[9] Z. Liu, C. Zhu, and T. Zhao, "Chinese named entity recognition with a sequence labeling approach: Based on characters, or based on words?" in *Proc. Int. Conf. Intell. Comput.*, Changsha, China, Aug. 2010, pp. 634–640.
[10] H. Li, M. Hagiwara, Q. Li, and H. Ji, "Comparison of the impact of word segmentation on name tagging for chinese and japanese," in *Proc. 9th Int. Conf. Lang. Resour. Eval. (LREC)*, Reykjavik, Iceland, May 2014, pp. 2532–2536.
[11] X. Li, Y. Meng, X. Sun, Q. Han, A. Yuan, and J. Li, "Is word segmentation necessary for deep learning of Chinese representations?" in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, Jul. 2019, pp. 3242–3252.
[12] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 2, Dec. 2015, pp. 2377–2385.
[13] G. Zhou and J. Su, "Named entity recognition using an HMM-based chunk tagger," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, Philadelphia, PA, USA, Jul. 2002, pp. 473–480.
[14] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Conf. Mach. Learn.*, Williamstown, MA, USA, Jun./Jul. 2001, pp. 282–289.
[15] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *J. Mach. Learn. Res.*, vol. 12, pp. 2493–2537, Aug. 2011.
[16] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," 2015, *arXiv:1508.01991*. [Online]. Available: https://arxiv.org/abs/1508.01991
[17] A. Passos, V. Kumar, and A. McCallum, "Lexicon infused phrase Embeddings for named entity resolution," in *Proc. 18th Conf. Comput. Natural Lang. Learn. (CONLL)*, Ann Arbor, MI, USA Jun. 2014, pp. 78–86.
[18] G. Luo, X. Huang, C.-Y. Lin, and Z. Nie, "Joint entity recognition and disambiguation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Lisbon, Portugal, Sep. 2015, pp. 879–888.
[19] M. Peters, W. Ammar, C. Bhagavatula, and R. Power, "Semi-supervised sequence tagging with bidirectional language models," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, Jul. 2017, pp. 1756–1765.
[20] H. Zhao and C. Kit, "Unsupervised segmentation helps supervised learning of character tagging for word segmentation and named entity recognition," in *Proc. 6th SIGHAN Workshop Chin. Lang. Process.*, Hyderabad, India, Jan. 2008, pp. 106–111.
[21] N. Peng and M. Dredze, "Named entity recognition for chinese social media with jointly trained embeddings," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Lisbon, Portugal, Sep. 2015, pp. 548–554.
[22] N. Peng and M. Dredze, "Improving named entity recognition for Chinese social media with word segmentation representation learning," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics, (ACL)*, Berlin, Germany, vol. 2, Aug. 2016, pp. 149–155.
[23] P. Cao, Y. Chen, K. Liu, J. Zhao, and S. Liu, "Adversarial transfer learning for chinese named entity recognition with self-attention mechanism," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Brussels, Belgium, 2018, pp. 182–192.
[24] Y. Zhang and J. Yang, "Chinese NER using lattice LSTM," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (ACL),*, Melbourne, VIC, Australia, vol. 1, Jul. 2018, pp. 1554–1564.
[25] N. Zhang, F. Li, G. Xu, W. Zhang, and H. Yu, "Chinese NER using dynamic Meta-Embeddings," *IEEE Access*, vol. 7, pp. 64450–64459, 2019.
[26] Y. Zhu and G. Wang, "CAN-NER: Convolutional attention network for Chinese named entity recognition," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics Hum. Lang. Technol.*, Minneapolis, MN, USA, vol. 1, Jun. 2019, pp. 3384–3393.
[27] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70. Sydney, NSW, Australia, Aug. 2017, pp. 933–941.
[28] E. Strubell, P. Verga, D. Belanger, and A. McCallum, "Fast and accurate entity recognition with iterated dilated convolutions," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Copenhagen, Denmark, Sep. 2017, pp. 2670–2680.
[29] F. Wu, N. Lao, J. Blitzer, G. Yang, and K. Weinberger, "Fast Reading Comprehension with ConvNets," 2017, *arXiv:1711.04352*. [Online]. Available: https://arxiv.org/abs/1711.04352
[30] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Dec. 2017, pp. 6000–6010.
[32] Z. Tan, M. Wang, J. Xie, Y. Chen, and X. Shi, "Deep semantic role labeling with self-attention," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 4929–4936.
[33] P. Verga, E. Strubell, and A. McCallum, "Simultaneously self-attending to all mentions for full-abstract biological relation extraction," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics Hum. Lang. Technol.*, New Orleans, LA, USA, vol. 1, Jun. 2018, pp. 872–884.
[34] G. D. Forney, Jr., "The Viterbi algorithm," *Proc. IEEE*, vol. 61, no. 3, pp. 268–278, Mar. 1973.
[35] G.-A. Levow, "The third international Chinese language processing bakeoff: Word segmentation and named entity recognition," in *Proc. 5th Workshop Chin. Lang. Process.*, Sydney, NSW, Australia, Jul. 2006, pp. 108–117.
[36] J. Xu, J. Wen, X. Sun, and Q. Su, "A discourse-level named entity recognition and relation extraction dataset for chinese literature text," 2019, *arXiv:1711.07010*. [Online]. Available: https://arxiv.org/abs/1711.07010
[37] L. Ratinov and D. Roth, "Design challenges and misconceptions in named entity recognition," in *Proc. 13th Conf. Comput. Natural Lang. Learn.*, Boulder, CO, USA, Jun. 2009, pp. 147–155.
[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.
[39] S. Li, Z. Zhao, R. Hu, W. Li, T. Liu, and X. Du, "Analogical reasoning on chinese morphological and semantic relations," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics, (ACL)*, Melbourne, VIC, Australia, vol. 2, Jul. 2018, pp. 138–143.
[40] A. Chen, F. Peng, R. Shan, and G. G.-Z. Sun, "Chinese named entity recognition with conditional probabilistic models," in *Proc. 5th Workshop Chin. Lang. Process.*, Sydney, NSW, Australia, Jul. 2006, pp. 173–176.
[41] S. Zhang, Y. Qin, J. Wen, and X. Wang, "Word segmentation and named entity recognition for SIGHAN bakeoff3," in *Proc. 5th Workshop Chin. Lang. Process.*, Sydney, NSW, Australia, Jul. 2006, pp. 158–161.
[42] J. Zhou, W. Qu, and F. Zhang, "Chinese named entity recognition via joint identification and categorization," *Chin. J. Electron.*, vol. 22, no. 2, pp. 225–230, Apr. 2013.
[43] Y. Lu, Y. Zhang, and D. Ji, "Multi-prototype Chinese character embedding," in *Proc. 10th Int. Conf. Lang. Resour. Eval. (LREC)*, May 2016, pp. 855–859.
[44] C. Dong, J. Zhang, C. Zong, M. Hattori, and H. Di, "Character-based LSTM-CRF with radical-level features for chinese named entity recognition," in *Proc. Int. Conf. Comput. Process. Oriental Lang.*, Kunming, China, Dec. 2016, pp. 239–250.
[45] F. Yang, J. Zhang, G. Liu, J. Zhou, C. Zhou, and H. Sun, "Five-stroke based CNN-BiRNN-CRF network for chinese named entity recognition," in *Proc. CCF Int. Conf. Natural Lang. Process. Chin. Comput.*, Hohhot, China, Aug. 2018, pp. 184–195.

YANLIANG JIN received the B.S. and M.S. degrees in electrical engineering from Xidian University, Xi'an, China, in 1997 and 2000, respectively, and the Ph.D. degree in communication and information system from Shanghai Jiao Tong University, in 2005. He is currently an Associate Professorship with the School of Communication and Information Engineering (SCIE), Shanghai University (SHU). He has published more than 30 journal/conference papers. His research interests include mobile ad hoc networks (MANETs), wireless sensor networks (WSNs), wireless multimedia sensor networks (WMSNs), wireless broadband access, and signal processing.

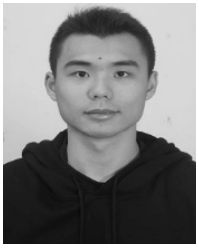


JINFEI XIE was born in Jiangxi, China, in 1996. He received the B.S. degree in communication engineering from the Hunan University of Science and Technology, Hunan, China, in 2018. He is currently pursuing the M.S. degree in communication and information system with the School of Communication and Information Engineering, Shanghai University. His current research interests include deep learning, information extraction, and natural language processing.



WEISI GUO (S'07–M'11–SM'17) received the M.Eng., M.A., and Ph.D. degrees from the University of Cambridge. He is currently an Associate Professor with the University of Warwick. He has published more than 110 articles. He is the PI on over 2.0 million pounds of research grants from EPSRC, H2020, and Innovate U.K. He is also the Coordinator of the H2020 Project: Data-Aware-Wireless-Networks for the IoE. His research has received several international awards (IET Innovation 15, Bell Labs Prize Finalist 14, and Semi-Finalist 16), and a Turing Fellowship from The Alan Turing Institute.

CAN LUO was born in Hunan, China, in 1994. He received the B.S. degree from the School of Electrical and Automation, East China Jiaotong University, Jiangxi, China, in 2017. He is currently pursuing the M.S. degree in communication and information system with the School of Communication and Information Engineering, Shanghai University. His current research interests include text classification, named-entity recognition, and keyword extraction.



DIJIA WU was born in Guangdong, China, in 1995. He received the B.S. degree in electronic information engineering from Shanghai Electric Power University, Shanghai, China, in 2018. He is currently pursuing the M.S. degree in circuit and systems with the School of Communication and Information Engineering, Shanghai University. His current research interests include text classification and relationship extraction in natural language processing.

RUI WANG received the B.S. and Ph.D. degrees in electronics and information engineering from Xidian University, Xi'an, Shaanxi, China, in 2004 and 2009, respectively. Since 2009, he has been with the School of Communication and Information Engineering, Shanghai University, where he is currently an Associate Professor. His research interests include sensor networks, geometric algebras, and multimedia signal processing.