Five-Stroke Based CNN-BiRNN-CRF Network for Chinese Named Entity Recognition

Fan Yang[1], Jianhu Zhang[1], Gongshen Liu[1](✉), Jie Zhou[1],
Cheng Zhou[2], and Huanrong Sun[2]

[1] School of Electric Information and Electronic Engineering,
Shanghai Jiaotong University, Shanghai, China
{417765013yf,zhangjianhu3290,lgshen,sanny02}@sjtu.edu.cn
[2] SJTU-Shanghai Songheng Content Analysis Joint Lab, Shanghai, China
{zhoucheng,sunhuanrong}@021.com

Abstract. Identifying entity boundaries and eliminating entity ambiguity are two major challenges faced by Chinese named entity recognition researches. This paper proposes a five-stroke based CNN-BiRNN-CRF network for Chinese named entity recognition. In terms of input embeddings, we apply five-stroke input method to obtain stroke-level representations, which are concatenated with pre-trained character embeddings, in order to explore the morphological and semantic information of characters. Moreover, the convolutional neural network is used to extract n-gram features, without involving hand-crafted features or domain-specific knowledge. The proposed model is evaluated and compared with the state-of-the-art results on the third SIGHAN bakeoff corpora. The experimental results show that our model achieves 91.67% and 90.68% F1-score on MSRA corpus and CityU corpus separately.

Keywords: CNN-BiRNN-CRF network
Stroke-level representations · N-gram features
Chinese named entity recognition

1 Introduction

Named entity recognition (NER) is one of the fundamental tasks in the field of natural language processing (NLP). It plays an important role in the development of information retrieval, relation extraction, machine translation, question answering systems and other applications. The task of NER is to recognize proper nouns or entities in the text and associate them with the appropriate types, such as the names of persons (PERs), organizations (ORGs), and locations (LOCs) [1]. Many researchers regard named entity recognition as a sequence labeling task. Traditional sequence labeling models are linear statistical models, including hidden markov model (HMM) [2], support vector machine (SVM) [3], maximum entropy (ME) [4] and conditional random field (CRF) [5,6].

M. Zhang et al. (Eds.): NLPCC 2018, LNAI 11108, pp. 184–195, 2018.
https://doi.org/10.1007/978-3-319-99495-6_16

With the development of word embedding technologies, neural networks have shown great achievements in NLP tasks. Character-based neural architecture has achieved comparable performance in English NER task [7–10]. Character-level features such as prefix and suffix could be exploited by the convolution neural network (CNN) or bidirectional long short-term memory (BiLSTM) structure, thus helping to capture deeper level of semantic meanings.

Compared with English NER, it is more difficult to identify Chinese entities, due to the attributes of Chinese words, such as the lack of word boundaries, the uncertainty of word length, and the complexity of word formation [11]. Inspired by the character embedding of English, some researches apply radical-level embedding of Chinese to improve the performance of word segmentation [12,13], part-of-speech (POS) tagging [14], or Chinese NER [15]. However, the acquired semantic information varies from the character splitting methods, which may lead to incomplete or biased results. Cao et al. [16] propose a stroke n-gram model by splitting each character into strokes, and each stroke is represented by an integer ranging 1 to 5. However, this method will bring about ambiguous information in the case of different characters with same type of strokes. As shown in Table 1, characters '天 (sky)' and '夫 (Husband)' are encoded into the same representation, so as to '由 (Reason)' and '申 (State)'. Therefore, it is essential to explore a general and effective way of semantic information extraction.

Recognizing entity boundaries is one of the most challenging problems faced by Chinese NER task, since entity boundary ambiguity can lead to incorrect entity identification. According to the word formation rules, many ORG entities contain the LOC and PER entities inside them. For example, '中国海军 (Chinese Navy)' is a ORG entity with nested LOC entity '中国 (China)'. Although the recognition of person and location names has achieved good results, the poor recognition of ORG entities still hinders the overall performance of named entity recognition [15]. Therefore, we need to pay more attention to the boundaries identification of ORG entities.

To address above difficulties and challenges, a five-stroke based CNN-BiRNN-CRF network (CBCNet hereafter) is proposed to optimize Chinese NER task. The main contributions of our model are summarized as follows:

1. A customized neural model is presented for Chinese NER, which conduces to the entity boundaries identification and entity ambiguity elimination, especially for the identification of ORG entities.

Table 1. Comparison of two character encoding methods.

Characters	Stroke n-gram [16]		Wubi method (Ours)	
	Decomposition	Code	Decomposition	Code
天 (Sky)	一一丿丶	1134	一大	GDI
夫 (Husband)	一一丿丶	1134	夫一一丶	GGGY
申 (State)	丨㇕一一丨	25112	曰丨	JHK
由 (Reason)	丨㇕一一丨	25112	由一乙一	MHNG

2. We propose the five-stroke based representations, and integrate them into character embeddings to form the final inputs. Then the convolution neural network with different size of filters is employed to simulate traditional n-gram model, which helps to identify the prefix and suffix of Chinese named entities.
3. We conduct experiments on the third SIGHAN bakeoff NER task with two Chinese annotated corpora. Experimental results show that our model achieves comparable performance over other existing Chinese NER architectures.

2 Neural Model for Chinese NER

2.1 Overview of Proposed Architecture

Our CBCNet is built upon the character-based BiLSTM-CRF architecture, as shown in the Fig. 1. Instead of using the original pre-trained character embeddings as the final character representations, we construct a comprehensive character representation for each character in the input sentence. Firstly, we incorporate stroke embeddings into original character embeddings, to construct a comprehensive character representation for each character in the input sentence. Then, a convolutional layer and a pooling layer are applied to generate n-gram features contained character representations. After that, we feed character embeddings into a BiLSTM-CRF layer, to decode and predict the final tag sequence for the input sentence.

2.2 Stroke Embedding

In this paper, we use five-stroke character model input method (or Wubi method)[1] to encode input characters. Wubi method is an efficient encoding

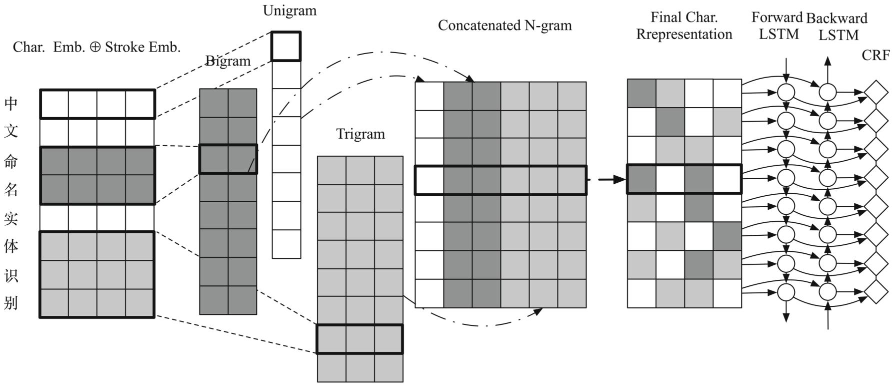


Fig. 1. The overall architecture of proposed CBCNet.

[1] https://en.wikipedia.org/wiki/Wubi_method.

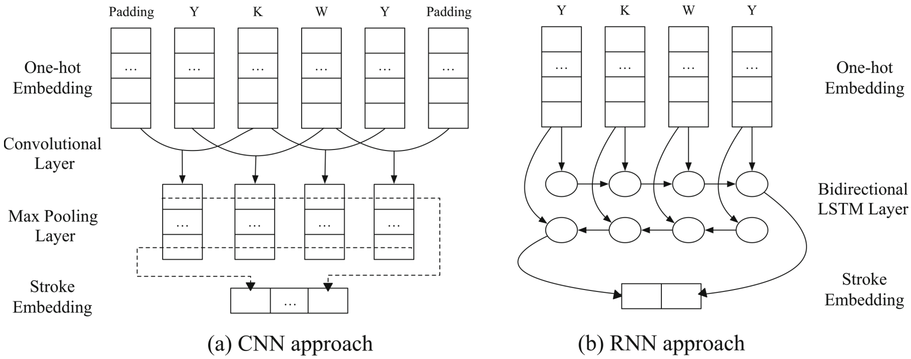
Table 2. Five regions with corresponding strokes.

Regions	ASDFG	HJKLM	QWERT	YUIOP	XCVBN
Strokes	Horizontal (一)	Vertical (丨)	Left-falling (丿)	Right-falling (丶)	Hook (乙)

system which can represent each Chinese character with at most four Roman letters (keys), according to the structure of each character. In the rule of Wubi method, 25 keys are divided into five regions, and each region is assigned with a certain type of strokes, as shown in Table 2. Then, each key stands for a certain type of components that is similar to the basic stroke in its own region. Since "Z" is a wild card, it is not listed in the table. By this means, every Chinese character could be encoded with a Wubi representation. For example, ('中', '文', '命', '名', '实', '体', '识', '别') which means ('Chinese', 'named', 'entity', 'recognition') in English, can be encoded into ('KHK', 'YYGY', 'WGKB', 'QKF', 'PUDU', 'WSGG', 'YKWY', 'KEJH'). Moreover, unlike previous work [16], Wubi method is able to distinguish words with similar structure, as shown in Table 1.

Santos and Zadrozny [17] introduce a convolutional approach with character embeddings, to extract the most important morphological information from English words. A BiLSTM layer is employed to capture semantic information of Chinese characters by Dong et al. [15]. In this paper, we compare the two methods of stroke-level feature extraction, and select the most effective way of representation. Figure 2 depicts the convolutional approach [17] and recurrent approach [15] respectively of generating stroke embeddings of character ' '识' ' (Recognition).

For example, given a character x_i encoded with a sequence of Q Roman letters $\{r_1, r_2, ..., r_Q\}$ $(0 \leq Q \leq 4)$ under the look-up table of Wubi method[2], we first transform each letter r_q into a one-hot embedding. In the case of the



(a) CNN approach (b) RNN approach

Fig. 2. Stroke embedding of character "识" (Recognition) with two approaches.

[2] https://github.com/yanhuacuo/98wubi-tables.

character with less than four Roman letters, we will randomly generate the initial embedding to ensure that each character has a four-dimensional stroke-level representation. During the training of the model, the stroke embeddings are continuously updated. Thus, the final stroke-level representation s_i of character x_i is defined as follows:

$$s_i = f(r_1, r_2, ..., r_Q) \tag{1}$$

where f is the function of a CNN or an RNN approach.

Pre-trained character embeddings are proved to be efficient over randomly initialized embeddings, since the former contain more contextual information. Therefore, we apply word2vec [18] to train our character-level embedding on Chinese Wikipedia backup dump. We concatenate the stroke embedding s_i to the character embedding v_i as the final representation c_i for each character:

$$c_i = v_i \oplus s_i \tag{2}$$

where $\oplus$ is a connection operator, and i indicates the ith character x_i in the sentence S.

2.3 Convolutional Layer

Inspired by Chen et al. [19], which use CNN to simulate a traditional discrete feature based model for POS tagging. In this paper, we use CNN to extract local information and n-gram features for the input characters, and we model different n-gram features by generating different feature map sets.

For example, $c_i \in R^d$ is the d-dimensional character representation corresponding to the ith character x_i in the sentence S. For a sentence with length l could be represented as:

$$c_{1:l} = c_1 \oplus c_2 \oplus ... \oplus c_l \tag{3}$$

where $c_{i:j}$ denotes the connection of $c_i, c_{i+1}, ..., c_j$.

The input of the network each time is an $l \times d$ matrix for one sentence. In order to ensure the integrity of the character, the width of convolutional filters is consistent with the dimension of the character representations. The convolution of the input matrix with filters are determined by the weights $W_k \in R^{kd \times N_k}$ and bias $b_k \in R^{N_k}$, where N_k is the number of k-gram feature maps. The k-gram feature map m_i^k can be generated from the combination of character representations $c_{i-\lfloor \frac{k-1}{2} \rfloor : i+\lceil \frac{k-1}{2} \rceil}$ according to the following formula:

$$m_i^k = tanh(W_k^T \cdot c_{i-\lfloor \frac{k-1}{2} \rfloor : i+\lceil \frac{k-1}{2} \rceil} + b_k) \tag{4}$$

The length and the order of character representations are maintained by padding zero to the input in the marginal case. The feature map sets matrix is $m \in R^{l \times \sum_{k=1}^{K} N_k} = \{m_1, m_2, ..., m_K\}$, where K is the maximum value of k in k-gram.

m_i is the concatenation of $m_i^k \in R^{l \times N_k}$:

$$m_i = m_i^1 \oplus m_i^2 \oplus ... \oplus m_i^K \tag{5}$$

Then, max-over-time pooling is applied to progressively reduce the spatial size of the representation and keep the most important features. Thus, the output sentence representations $c^* \in R^{l \times d} = \{c_1^*, c_2^*, ..., c_l^*\}$ can be generated after d-max pooling operation, where c_i^* is:

$$c_i^* = dmax\,\{m_i\} \tag{6}$$

2.4 Bidirectional LSTM Layer

The recurrent neural network (RNN) can effectively obtain the sequence information of the texts. As a special kind of RNN, the long short-term Memory (LSTM) network could not only solve the long-distance dependence problem of the sequence, but also effectively deal with the vanishing gradient or exploding gradient problem of RNN [15]. In order to make effective use of contextual information for a specific time frame, we use a bidirectional LSTM (BiLSTM) architecture. Thus, each hidden state h_t of BiLSTM can be formalized as a concatenation of the hidden states of forward and backward LSTMs:

$$h_t = \overrightarrow{h_t} \oplus \overleftarrow{h_t} \tag{7}$$

where $\overrightarrow{h_t}$ (or $\overleftarrow{h_t}$) can be generated from the multiplication of output gate result o_t and input character representation c_t^* at the specific time frame of t, and the calculation of o_t can refer to previous works [15].

2.5 CRF Layer

The linear CRF can obtain a globally optimal tag sequence by considering the relationship between adjacent tags. By combining BiLSTM layer and CRF layer, BiLSTM-CRF layer is able to efficiently make use of contextual features as well as sentence-level tag information. Given an input sentence $X = \{x_1, x_2, ..., x_l\}$, we consider the score matrix P, which is the output of BiLSTM layer, and the transition score matrix A. Thus, for a sequence of prediction results $y = \{y_1, y_2, ..., y_l\}$, the score of the sentence S along with a path of labels could be defined as:

$$Score(X, y) = \sum_{i=1}^{l} A_{y_i, y_{i+1}} + \sum_{i=1}^{l} P_{i, y_i} \tag{8}$$

where $A_{y_i, y_{i+1}}$ indicates the possibility of the transition from ith label to $i + 1$th label for a pair of consecutive time steps, P_{i, y_i} is the score of the ith label of the ith input character.

For the decoding phase, Viterbi algorithm [20] is used to generate optimal tag sequence y^*, when maximizing the score $Score(X, y)$:

$$y^* = \underset{y \in Y_x}{\arg\max} \quad Score(X, y). \tag{9}$$

where Y_x represents all the possible label sequences for sentence S.

3 Experimental Results and Analysis

3.1 Tagging Scheme

In this paper, we adopt the BIO (indicating Begin, Inside and Outside of the named entity) tagging set used in the third SIGHAN bakeoff [1], which followed by tags PER, ORG and LOC that denote persons, organizations and locations respectively. For instance, B-PER and I-PER denote the begin, inside part of a person's name respectively. O means that the character is not included in a named entity. We employ character-level precision (P), recall (R), and F1-score (F) as the evaluation metrics, same as the previous works [15].

3.2 Training

We use Tensorflow library to implement our neural network. Table 3 illustrates the hyper-parameters of all the experiments on different datasets. We train our network with the error back-propagation, and the network parameters are fine-tuned by back-propagating gradients. Adagrad algorithm [21] is used as the network optimizer. To accelerate network training on GPU, we adopt bucketing strategy [14], which usually implemented in seq2seq model, for the input sentences. That is, the sentences with similar lengths are grouped into the same buckets, and the sentences in the same buckets are padded into the same length. In order to reduce over-fitting during network training and improve the accuracy of neural network model, we apply dropout technique [22] before the BiLSTM layer with a probability of 0.5.

Table 3. Hyper-parameter settings.

Parameters	Details	Parameters	Details
Character embedding size	$d_c = 64$	Optimizer	Adagrad
Stroke embedding size	$d_s = 30$	Initial learning rate	$\alpha = 0.2$
Number of feature map sets	$K = 5$	Decay rate	0.05
Number of k-gram feature maps	$N_k = 100$	Dropout rate	0.5
LSTM dimensionality	$h = 200$	Batch size	10

Table 4. The statistics of NER training and testing corpora.

Copora	Training			Testing		
	Sentences	NEs	ORGs/LOCs/PERs	Sentences	NEs	ORGs/LOCs/PERs
MSRA	43907	75059	20584/36860/17615	3276	6190	1331/2886/1973
CityU	48169	112361	27804/48180/36377	6292	16407	4016/7450/4941

Sentences: Number of sentences; NEs: Number of named entities;
LOCs/PERs/ORGs: Number of location/person/organization names.

3.3 Dataset and Preprocessing

We conduct the experiments on the two corpora from the third SIGHAN bake-off [1]. The MSRA corpus is simplified Chinese character and the CityU corpus is traditional Chinese character, and both of them are in the CoNLL two column format. We convert CityU corpus simplified Chinese, so that the look-up table for pre-trained character embeddings, Wubi encoding method and other resources are compatible in the experiments. Table 4 shows the statistics of NER training and testing corpora of MSRA and CityU.

3.4 Evaluation of Different Components

We incrementally add each component on the BiLSTM-CRF with character embedding model, which is the baseline architecture in our comparison, to evaluate the impact of every component on the performance of our model. In general, experimental results given in Table 5 shows that "Ours" with stroke-level information and n-gram features significantly outperforms the baseline model, achieving 91.67% and 90.68% F1-score on MSRA and CityU separately.

To evaluate the effectiveness of two stroke-level feature extraction approaches, we test on the RNN and CNN separately, denoted as "+ Stroke Emb.(RNN)" and "+ Stroke Emb.(CNN)". Table 5 illustrates that CNN performs a little better than RNN. The reason is that CNN is an unbiased model, which treats fairly to each input in the same windows. While RNN is a biased model, in which later inputs are more dominant than earlier inputs. Thus, we

Table 5. Evaluation of different components on two corpora in F1-scores (%)

Variant	MSRA			CityU		
	P	R	F	P	R	F
BiLSTM-CRF (Char. Emb.)	90.98	89.97	90.47	91.32	88.49	89.88
+ Stroke Emb.(RNN)	92.36	90.18	91.25	91.61	88.77	90.16
+ Stroke Emb.(CNN)	92.34	90.31	91.31	91.57	88.96	90.24
+ CNN + Pooling	92.30	90.68	91.48	91.65	89.07	90.34
Ours	92.04	91.31	91.67	91.87	89.53	90.68

apply CNN approach to extract stroke-level features in our model. Moreover, compared with baseline, we could obtain relatively high performance in two corpora by incorporating stroke embeddings into character embeddings.

From Table 5 in the row "+ CNN + Pooling", we can see that notable improvement is achieved by combining 1-gram to K-gram (we set $K = 5$ for all the experiments.) features. This phenomenon is consistent with the length of the named entities, which is about two to five characters. It verifies that our CBCNet can efficiently identify entity boundaries and eliminate entity ambiguity by introducing n-gram features. Besides, the results imply that the n-gram features have a greater impact on model performance than stroke-level information.

Furthermore, it can be observed that the optimization of our model on the MSRA corpus is more significant than that on CityU corpus. The reason is that CityU corpus contains some English named entities, which are not sensitive to our two adopted measures.

3.5 Comparison with Previous Works

We compare proposed CBCNet with the reported results of several previous works on the third SIGHAN bakeoff corpora, as shown in the Tables 6 and 7. Each row represents the results of a NER model, including F1 scores for each entity category (PER-F, ORG-F, LOC-F) as well as the total precision (P), recall (R), and balanced F1-score (F).

Table 6. Evaluation of different models on MSRA corpus (%)

Model	MSRA					
	F-ORG	F-LOC	F-PER	P	R	F
CRF+Word.Emb. [5]	83.10	85.45	90.09	88.94	84.20	86.51
CRF+Char.Emb. [6]	81.96	90.53	82.57	91.22	81.71	86.20
Knowledge based [4]	85.90	90.34	96.04	92.20	90.18	91.18
Feature templates [23]	86.19	91.90	90.69	91.86	88.75	90.28
BiLSTM-CRF+Radi.Emb. [15]	87.30	92.10	91.77	91.28	90.62	90.95
Ours	**88.42**	**92.31**	**91.96**	**92.04**	**91.31**	**91.67**

Table 7. Evaluation of different models on CityU corpus (%)

Model	CityU					
	F-ORG	F-LOC	F-PER	P	R	F
CRF+Char.Emb. [6]	-	-	-	92.66	84.75	88.53
Knowledge based [4]	81.01	93.06	91.30	92.33	87.37	89.78
Ours	**83.99**	**93.76**	**91.63**	**91.87**	**89.53**	**90.68**

Conditional random fields (CRF) is one of the most popular and effective models for sequence labeling tasks [5,6]. Zhou et al. [5] used a word-level CRF-model with hand-crafted features incorporated, which won the first place in the third SIGHAN bakeoff Shared Tasks with 86.51% F1-score in MSRA corpus. Chen et al. [6] utilized character embeddings as the inputs to CRF model with 86.20% F1-score in MSRA corpus and 88.53% in CityU corpus, while the improvement of performance is still limited. By incorporating hand-crafted features [5,23] or knowledge bases [4] could relatively improve the performance. However, these methods may result in an inefficiency or failure when processing large-scale corpora or the datasets in other fields.

Dong et al. [15] implemented Chinese radical embedding into BiLSTM-CRF framework, achieving good performance. However, some characters cannot be split into radicals, leading to the failure of semantic feature extraction. Moreover, they ignore the characteristics of entities word formation. Compared with [15], we consider both the semantic information by using stroke embedding and contextual information by employing CBCNet model. In terms of MSRA corpus, experimental results indicate that proposed stroke-based CBCNet outperforms the best deep learning work [15] by +0.72 F1-score and best reported work [23] by +0.49 F1-score. For CityU corpus as shown in Table 7, compared with other works, our model obtains the best performances with 90.68% F1-score. Moreover, our model achieves significant improvement on ORG entities thanks to the extraction of n-gram features.

4 Related Works

Named entity recognition is a fundamental NLP task and studied by many researchers. Remarkable achievements have been made in the field of English NER through a variety of methods. An end-to-end architecture is implemented in [24], with a BiLSTM-CNNs-CRF model. Yang et al. [8] use transfer learning for jointly training the POS and NER tasks. Then, Liu et al. [9] enhance the neural framework by introducing a task-aware language model. Xu et al. [25] propose a FOFE-based strategy, which regards NER as a non-sequence labeling task. However, these approaches could not be simply transplanted into Chinese NER systems, due to the characteristics of Chinese named entities (NEs), which do not have word boundaries or case sensitivity.

BiLSTM-CRF architecture has been used in many Chinese sequence labeling problems. Peng and Dredze [26] adopt the model for jointly training Chinese word segmentation and named entity recognition. By extracting semantic information, the performance of labeling could be further improved. Dong et al. [15] combine radical embeddings with character embeddings in bidirectional LSTM-CRF model for Chinese NER, and show the efficiency. For the task of Chinese word segmentation and POS tagging, this fundamental structure also shows great performance as shown in [14]. He et al. [12] indicate that the performance of Chinese word segmentation could be boosted largely when tying subcharacters and character embeddings together.

5 Conclusions

In this paper we have presented a novel model for Chinese NER by considering the semantic information as well as n-gram features, without involving handcrafted features or domain-specific knowledge. The empirical study shows the effectiveness of each components of our architecture. Experiments on two different corpora from the third SIGHAN bakeoff also indicate that our model achieves outstanding performance over other approaches. In the future, we would like to extend our model to other sequential labeling tasks, such as jointly learning the Chinese word segmentation, POS tagging and NER.

Acknowledgement. This research work has been funded by the National Natural Science Foundation of China (Grant No. 61772337, U1736207 and 61472248), the SJTU-Shanghai Songheng Content Analysis Joint Lab, and program of Shanghai Technology Research Leader (Grant No. 16XD1424400).

References

1. Levow, G.A.: The third international Chinese language processing bakeoff: word segmentation and named entity recognition. In: Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pp. 108–117 (2006)
2. Fu, G., Luke, K.K.: Chinese named entity recognition using lexicalized HMMs. ACM SIGKDD Explor. Newslett. **7**, 19–25 (2005)
3. Li, L., Mao, T., Huang, D., Yang, Y.: Hybrid models for Chinese named entity recognition. In: Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pp. 72–78 (2006)
4. Zhang, S., Qin, Y., Wen, J., Wang, X.: Word segmentation and named entity recognition for SIGHAN Bakeoff3. In: Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pp. 158–161 (2006)
5. Zhou, J., He, L., Dai, X., Chen, J.: Chinese named entity recognition with a multiphase model. In: Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pp. 213–216 (2006)
6. Chen, A., Peng, F., Shan, R., Sun, G.: Chinese named entity recognition with conditional probabilistic models. In: Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pp. 173–176 (2006)
7. Chiu, J.P., Nichols, E.: Named entity recognition with bidirectional LSTM-CNNs. arXiv preprint arXiv:1511.08308 (2015)
8. Yang, Z., Salakhutdinov, R., Cohen, W.W.: Transfer learning for sequence tagging with hierarchical recurrent networks. arXiv preprint arXiv:1703.06345 (2017)
9. Liu, L., et al.: Empower sequence labeling with task-aware neural language model. arXiv preprint arXiv:1709.04109 (2017)
10. Dong, C., Wu, H., Zhang, J., Zong, C.: Multichannel LSTM-CRF for named entity recognition in Chinese social media. In: Sun, M., Wang, X., Chang, B., Xiong, D. (eds.) CCL/NLP-NABD -2017. LNCS (LNAI), vol. 10565, pp. 197–208. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-69005-6_17
11. Duan, H., Zheng, Y.: A study on features of the CRFs-based Chinese named entity recognition. Int. J. Adv. Intell. **3**, 287–294 (2011)

12. He, H., et al.: Dual long short-term memory networks for sub-character representation learning. In: Latifi, S. (ed.) Information Technology - New Generations. AISC, vol. 738, pp. 421–426. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-77028-4_55
13. Yu, J., Jian, X., Xin, H., Song, Y.: Joint embeddings of Chinese words, characters, and fine-grained subcharacter components. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 286–291 (2017)
14. Shao, Y., Hardmeier, C., Tiedemann, J., Nivre, J.: Character-based joint segmentation and POS tagging for Chinese using bidirectional LSTM-CRF. arXiv preprint arXiv:1704.01314 (2017)
15. Dong, C., Zhang, J., Zong, C., Hattori, M., Di, H.: Character-based LSTM-CRF with radical-level features for Chinese named entity recognition. In: Lin, C.-Y., Xue, N., Zhao, D., Huang, X., Feng, Y. (eds.) ICCPOL/NLPCC -2016. LNCS (LNAI), vol. 10102, pp. 239–250. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-50496-4_20
16. Cao, S., Lu, W., Zhou, J., Li, X.: cw2vec: learning Chinese word embeddings with stroke n-gram information. (2018)
17. Santos, C.D., Zadrozny, B.: Learning character-level representations for part-of-speech tagging. In: Proceedings of the 31st International Conference on Machine Learning, pp. 1818–1826 (2014)
18. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
19. Chen, X., Qiu, X., Huang, X.: A feature-enriched neural model for joint Chinese word segmentation and part-of-speech tagging. arXiv preprint arXiv:1611.05384 (2016)
20. Forney, G.D.: The Viterbi algorithm. Proc. IEEE **61**(3), 268–278 (1973)
21. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. **12**(Jul), 2121–2159 (2011)
22. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
23. Zhou, J., Qu, W., Zhang, F.: Chinese named entity recognition via joint identification and categorization. Chinese J. Electron. **22**(2), 225–230 (2013)
24. Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. arXiv preprint arXiv:1603.01354 (2016)
25. Xu, M., Jiang, H., Watcharawittayakul, S.: A local detection approach for named entity recognition and mention detection. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 1237–1247 (2017)
26. Peng, N., Dredze, M.: Improving named entity recognition for Chinese Social Media with word segmentation representation learning. arXiv preprint arXiv:1603.00786 (2016)