A Mixed Semantic Features Model for Chinese NER with Characters and Words

Ning Chang[1], Jiang Zhong[1,2(✉)], Qing Li[1], and Jiang Zhu[3]

[1] Chongqing University, Chongqing 400044, People's Republic of China
zhongjiang@cqu.edu.cn
[2] Key Laboratory of Dependable Service Computing in Cyber Physical Society, Chongqing University, Chongqing 400044, People's Republic of China
[3] Chengdu Library and Information Center, Chinese Academy of Sciences, Chengdu 610041, People's Republic of China

Abstract. Named Entity Recognition (NER) is an essential part of many natural language processing (NLP) tasks. The existing Chinese NER methods are mostly based on word segmentation, or use the character sequences as input. However, using a single granularity representation would suffer from the problems of out-of-vocabulary and word segmentation errors, and the semantic content is relatively simple. In this paper, we introduce the self-attention mechanism into the BiLSTM-CRF neural network structure for Chinese named entity recognition with two embedding. Different from other models, our method combines character and word features at the sequence level, and the attention mechanism computes similarity on the total sequence consisted of characters and words. The character semantic information and the structure of words work together to improve the accuracy of word boundary segmentation and solve the problem of long-phrase combination. We validate our model on MSRA and Weibo corpora, and experiments demonstrate that our model can significantly improve the performance of the Chinese NER task.

Keywords: Chinese named entity recognition · Self-attention · Mixed semantic feature · Entity boundary segmentation

1 Introduction

In recent years, named entity recognition (NER) has received a lot of attention in the field of natural language processing (NLP), and it is the basis of many

Supported by National Key Research and Development Program of China Grant 2017YFB1402400, in part by the Graduate Research and Innovation Foundation of Chongqing under Grant CYB18058, in part by the Key Research Program of Chongqing Science and Technology Bureau No. cstc2019jscx-fxyd0142, in part by the Fundamental Research Funds for the Central Universities under Grant 2018CDYJSY0055.

J. M. Jose et al. (Eds.): ECIR 2020, LNCS 12035, pp. 356–368, 2020.
https://doi.org/10.1007/978-3-030-45439-5_24

downstream NLP tasks. NER refers to the identification of entities with specific meaning in the text, usually including names of people, places, institutions, proper nouns, and so on. For English text, this problem has been studied extensively [13,20,23]. However, Chinese NER still faces challenges such as Chinese word segmentation, and it is often difficult to define what constitutes a word in Chinese.

Most methods of existing state-of-the-art models for Chinese NER are usually based on word segmentation, and train neural network and Conditional Random Field (CRF) to perform sequence labeling on word-level [17]. However the effect of the word segmentation depends heavily on the quality of the dictionaries and segmentation tools, and it's possible to lead to error propagation if the boundaries are partitioned improperly at the very start. Moreover, it can not deal with unseen words. There are also some models which recognize entities in character-level, which solve the problem of out-of-vocabulary (OOV) [19]. However, fully character-based models cannot express enough semantic information and word structure, and could lead to wrong word boundaries.

In order to take advantage of both character-level semantic information and word structure content, some models mix word embedding and its corresponding character vectors, and then feed mixed representation into neural network for NER [2,22,26]. The generic model mentioned above is shown in Fig. 1(a), these methods divide each sequence into several characters, and then represent these character vectors as a comprehensive representation through LSTM networks or other models. About word vectors, they concatenate each word vector with the representation of its corresponding characters, and then form a new multi-granularity representation of the word. In the process of generating the final word representation, the intermediate dimensional transformation may lead to original information loss. Moreover, for each word, the concatenation of the two granularity representations at the word-level does not express well the relationship between characters and words. These drawbacks affect the accuracy of Chinese word boundary segmentation and entity recognition.

In this paper, we incorporate the self-attention mechanism into the Chinese named entity recognition model to compute the weighted sum of character and word vectors, and integrate the semantic features of the two representations. Different from the previously mentioned model, our model captures character content features and the information of token structure in word level (as shown in Fig. 1(b)). Our model uses two sequences of character and word segmentation as input, and outputs the final character-based recognition tags through the attention mechanism. The model preserves the character-level semantic representation and the word tokens structure completely, and uses self-attention to assign the weight of both. Multi-granularity semantic and structural features are combined with word representation to enrich character representation and reduce the loss of original information. Moreover, the character level and the word segmentation structure are complementary to each other, and a single character can correct the word boundary error caused by the word segmentation level.

Moreover, phrases that do not appear in the prior dictionary can also be identified. As the example in Fig. 1 shows, given the sequence of "Beijing/People/Park" that is segmented using the dictionary, our model could add the segmentation structural information into character-based semantic information. When predicting tags of characters, we can determine the phrase boundary based on the comprehensive context and correctly identify "Beijing people's park" as a phrase to be marked.

We experiment with our model on MSRA and Weibo data sets, and the results show that using the self-attention mechanism to fuse two granularity semantic and structure representations in sequence context can significantly improve performance.

The contributions of our paper are as follows:

- We improve the accuracy of word boundary segmentation by combining two granularity features. Our model retains the primitiveness of character semantics and participle structures completely, and the two embedding information assist each other. Character semantics combined with word tokens structure could modify word boundary segmentation.
- We investigate a method to enhance the recognition of Chinese long phrases that do not appear in prior dictionaries. Our model uses a self-attention mechanism to integrate features of word segmentation into a character-level sequence, and predicts it in conjunction with the context of the sentence to merge the short tags into long phrases.

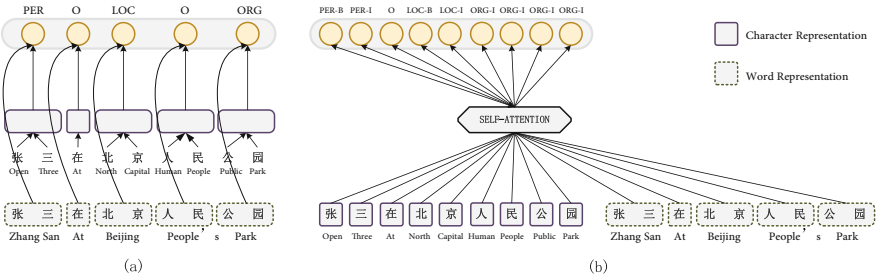

Fig. 1. Example of how previous models (a) and our model (b) combine two granularity representations of characters and words.

2 Related Work

NER. Early named entity recognition methods are based on rules and statistical machine learning such as Hidden Markov Model (HMM) [1], Conditional Random Fields (CRF) [12], and Support Vector Machines (SVM) [11]. In recent years, with the development of machine learning, more and more neural network

models are used for the NER task. Collobert et al. [4] propose a unified neural network architecture that can be used in various NLP tasks. Zhou et al. [27] formulate Chinese NER as a joint identification and categorization task. Huang et al. [10] first apply BiLSTM-CRF model to NER, and achieve the advance results at that time. The BiLSTM-CRF model is now also the benchmark model for many pieces of research. Lample et al. [13] use BiLSTM-CRF as the basic model, rely on character-based word representations learned from the supervised corpus and unsupervised word representations learned from unannotated corpora. For Chinese NER, Zhang et al. [26] investigate a lattice-structured LSTM model, utilize information on words and character sequences, and solve the problem of Chinese words boundaries. Dong et al. [6] utilize both character-level and radical-level representations based on bidirectional LSTM-CRF. Besides, incorporating the five-stroke information into the network also achieves outstanding performance [24]. [14] add gazetteer-enhanced sub-tagger on hybrid semi-Markov CRF architecture and observe some promising results. And [5] also propose a neural multi-digraph model with the information of gazetteers.

Self-attention. Vaswani et al. [21] first proposed a self-attention mechanism for machine translation to connect all positions with a constant number of sequentially executed operations, and attract great attention. Subsequently, a large number of studies begin to use the attention mechanism. Zukov et al. [29] use no language-specific features, and the model they proposed is based on RNN structure, coupled with a self-attention mechanism for NER. Yang et al. [25] propose a novel adversarial transfer learning framework and first introduce a self-attention mechanism to the Chinese NER task. And then Zhu et al. [28] propose a convolutional attention network for Chinese named entity recognition. They use a character-based CNN with local-attention and GRU with self-attention to get information from characters of the sentence.

Joint Character and Word Embedding. Some models join characters with words for sequence tagging. Lample et al. [13] feed the characters of a word into the bidirectional LSTM, and connect the final output of the forward and backward network as character representation. This character-level representation is then concatenated with its corresponding word representation. Rei et al. [18] use the same structure [13] to represent character-level representation. Instead of connecting two-level representations directly, an attention mechanism is used to calculate the weighted sum of character embedding and word embedding. Ma et al. [15] utilize CNN to compute character representation for each word, and concatenate it with word embedding before feeding into the BiLSTM network.

The main benefit of Chinese characters is they can solve the problem of phrases that is not in the dictionary, and can flexibly determine the phrase boundary. Besides, word-level modeling can provide information about the structure of common words. We propose a model based on self-attention which use the sequence-level joint representation of characters and words to take advantage of two granularity embedding.

3 Methodology

In this chapter, we will introduce our methodology in detail. Our model utilizes BiLSTM-CRF as our basic structure, and extends a self-attention mechanism to obtain the long distance dependencies of the character encoder and word encoder sequence. As illustrated in Fig. 2, the architecture of our model mainly consists of character and word embedding, Bi-LSTM network with self-attention and CRF for tagging. We will describe our method in the following sections.

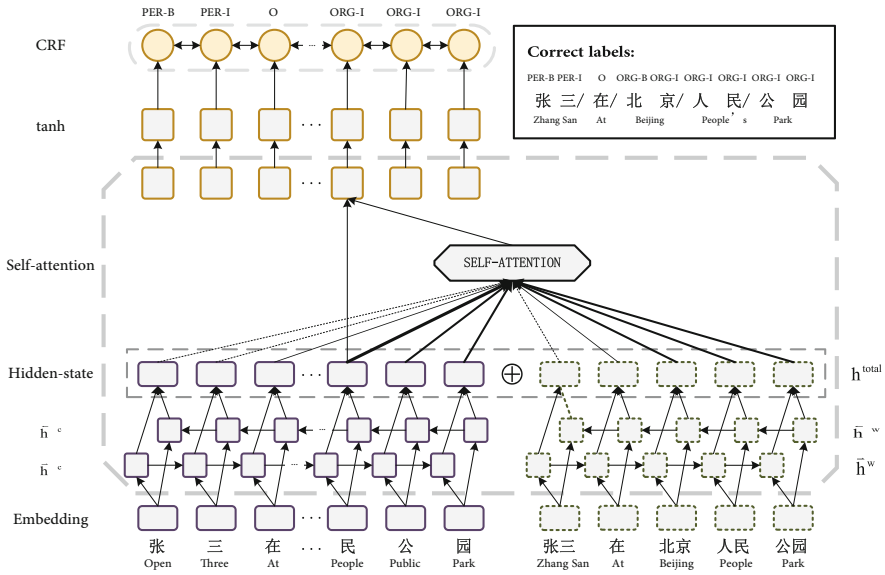

Fig. 2. The general architecture of our proposed model.

3.1 Characters and Words Representation in Sequence

Word embedding, also known as distributed word representation, can capture both the semantic and syntactic information of the words from a large unlabeled corpus. We use open source Chinese word vector corpus of Tencent AI Lab[1], which includes more than 8 million Chinese words, and each word corresponds to a 200-dimensional vector. For a sentence, we utilize *jieba*[2] to perform word segmentation. And every word is disintegrated into individual characters. Furthermore, characters in a sentence also contain the rich context of the entities, and Chinese character-based embedding could alleviate problems of long

[1] https://ai.tencent.com/ailab/nlp/embedding.html.
[2] https://github.com/fxsjy/jieba.

phrases that are not in dictionaries. Our model uses both granularity levels of embedding information to learn the mixture semantic of characters and words.

In this paper, we use Bi-LSTM [9] as our basic structure to use forward and backward information of character and word embedding. We denote the two embedding sequences separately as $[E_1^c, E_2^c, ..., E_n^c]$ and $[E_1^w, E_2^w, ..., E_m^w]$. And they are generated by a look-up layer, and are fed into two parallel Bi-LSTM structures respectively, which have the same structure, but with different parameters. The output character and the word level hidden state are represented as h^c and h^w. Join the two hidden layers to form a total hidden state (represented as $h^{total} = [h^c, h^w]$), where the front part is the characters representation, and the latter part is the words semantic feature. Then a self-attention mechanism operation is performed on the total hidden state sequence.

3.2 Multi-granularity Representation Fusion by Using Self-attention

Solve the Problem of Boundary Segmentation. In the process of word segmentation, there may be problems with word boundary errors. As shown in the example in Fig. 1, the first three characters may be incorrectly split into a person's name in the sentence, which would lead to error propagation, and cause severe bias effects on subsequent predictions. The previous general model can not solve the problem of word segmentation very well, and cause some content loss when combining the embedded information of characters and words. Our model combines two granular hidden states at the sequence level to preserve the original features intact. Also, the self-attention trains the weight information of the total sequence, and preserves the semantic information of the context characters to perform a calculation with the information of the word sequence structure. The character information will correct the error problem of word boundary segmentation, and correctly identify "Zhang San" as a person name, while the third character as a preposition.

Solve the Problem of Phrase Combination. In Chinese, long phrases are usually composed of short phrase sequences in order. For example, Beijing People's Park is composed of Beijing/ people/ park. Compared with English, there is usually no need for prepositional connections in phrases, which leads to the poor distinction of the boundaries in long phrases. It is also a difficult point in the recognition of Chinese named entities. Dictionary-based word segmentation usually divides sentences into short words. For long phrases that do not appear in the dictionary, there is currently no good solution. This paper proposes a method to improve the above problem by using two granularity semantic representations to assist each other. The model uses the self-attention mechanism to calculate the similarity on different levels of representation subspace, sequentially calculates each character with all tokens in the total sequence. This method captures the structural information of the word sequence, in order to compute similarity and correlation with character context information to further identify the combined boundaries of the long phrases.

In addition, the attention mechanism uses the weighted sum calculation to generate the output, which effectively solves the problem of the gradient disappearing. And the self-attention mechanism can be calculated in parallel, which greatly improves efficiency.

3.3 CRF for Tag Prediction

We quote a standard Conditional Random Field (CRF) layer on top of the attention layer. The CRF can use the state feature function and the state transfer function to maximize the characteristics of the text. Besides, it can consider the context information and the annotation information of adjacent words. The feature functions are defined as follows:

$$f_j\left(s, j, l_i, l_{i-1}\right) = \begin{cases} t_j\left(l_{i-1}, l_i, s, i\right) & State\ transfer\ function \\ s_j\left(l_i, s, i\right) & \text{State feature function} \end{cases} \tag{1}$$

Where s indicates the sentence we want to predict. l is the label sequence of the sentence, and l_i represents the label of i-th token. i is the current location. The state transition function defines the probability of the $(i-1)$-th token label l_{i-1} move to the label l_i of the next i-th token in the sentence s. And the state feature function indicates the probability that the current i-th token is marked as l_i.

Then we normalize the score to get the probability that the label sequence is l given the sentence s. Given all predicted tag sequences l, the probability of label sequence s is calculated as follows:

$$p(l|s) = \frac{\exp[s\,\mathrm{core}(l|s)]}{\sum_{l'} \exp\left[\mathrm{score}\left(l'|s\right)\right]} \tag{2}$$

Where l represents all possible tag sequences.

The output of the self-attention mechanism is independent of each other. Although the context information is taken into account when performing the matrix transformation, the outputs do not affect each other. Our model uses CRF for label prediction. By considering the transition characteristics between output labels, we constrain the final label and improve the accuracy of entity label prediction.

4 Experiments

4.1 Datasets

We use corpora provided by Microsoft Research Asia (MSRA) and Weibo corpus [17] extracted from Sina Weibo to experiment with the model presented in this paper. MSRA contains three entity types: Person (PER), Location (LOC) and Organization (ORG). And Weibo dataset is annotated with four types of entities (in addition to the above three entities, there is also a Geo-Political entity type, GPE). We train on both name mentions and nominal mentions in the Weibo data set. The detailed statistics of the corpora are summarized in Table 1.

We preprocess the datasets and annotate the entity type using BIO rules, which indicates Begin, Inside and Outside of a named entity.

Table 1. The statistics of datasets.

Corpus		Train set	Test set	Dev set
MSRA		46317	4376	—
Weibo	Named mention entity	957	153	211
	Nominal mention entity	898	226	198

4.2 Experimental Settings

Our experiments employ character-level precision (P), recall (R), and F1-score (F) as the evaluation criteria. We use *Jieba* for segmentation, and utilize word embedding dataset published by Tencent AI Lab to perform embedding, and the dimension of word embedding is 200, the same as character embedding.

Pytorch library is used to build our model. We train the model using an *Adam* optimizer with an initial learning rate of 0.001, and the network is fine-tuned by back-propagating. For the over-fitting and vanishing gradient problems, we employ the dropout method with a probability of 0.5. We control the length of the sentence to be 80, and the number of words after sentence segmentation to be 40. Otherwise, we would pad the shorter sequences, truncate the longer parts. Detailed hyper-parameters are listed in Table 2.

Table 2. Hyper-parameter settings.

Parameters	Values
Character embedding dim	200
Word embedding dim	200
Hidden dim	50
Optimizer	Adam
Initial learning rate	0.001
Dropout rate	0.5
Batch size	64
Epoch	40

4.3 Evaluation of Components

Considering that the character-based models are not dependent on the quality of dictionaries and are more flexible, our model would use character-based output in the subsequent experiments.

Ablation experiments are designed to verify the necessity of each part in our model and its impact on the experimental results. We gradually add each

component to the baseline architecture BiLSTM-CRF. The results are shown in Table 3.

To evaluate the effects of two embedding approaches, we perform comparison experiments on character embedding and word embedding respectively. In the third comparative experiment, the embedding for each word and its corresponding characters compose a concatenation to be the input of the Bi-LSTM layer (as [15] did).

Experimental results show that character-based model performance is better than word-based models on the two data sets. At the same time, the model using two embedding methods for prediction has a slight improvement compared with the original two models, but the effect is not obvious. By contrast, adding a self-attention mechanism can significantly improve the performance of NER.

Attention can obtain sentence context information from the long-distance relationship between tokens, overcoming the limitations of recurrent neural networks. In this model, self-attention can capture the dependence of characters and words at the same time over a long distance.

Table 3. Experiments of each component on MSRA and Weibo datasets.

Models	MSRA			Weibo
	Precision	Recall	F1-score	F1-score (overall)
Baseline (Char.Emb)	90.74	89.85	90.29	57.51
Baseline (Word.Emb)	91.21	88.60	89.89	57.02
Baseline (Char+Word.Emb)	92.79	89.93	91.34	57.63
Our model	**95.92**	**94.80**	**95.36**	**61.46**

4.4 Comparison with Previous Work

In this section, we compare our BiLSTM+Self-Attention+CRF model based on a mixture of characters and words with the previous proposed advanced models on the MSRA and Weibo data sets. The comparison results are listed in the Tables 4 and 5.

MSRA Dataset. Chen et al. [3] first apply Conditional Random Fields (CRF) for sequence tagging, and achieve 86.20% F1-score in MSRA corpus. Zhou et al. [27] formulate NER as a joint identification to recognize entity-level features, which effectively improves performance. And Cao et al. [2] also use the information of CWS for NER. Zhang et al. [24] and Ding et al. [5] add additional

features, and the latter achieve 94.4% F1-score. Zhu et al. [28] investigate a Convolution Attention Network to capture the information from adjacent characters and sentence contexts, which achieves F1-score of 92.97%. Our model utilizes self-attention on character+word hidden state and gets effective performance improvement with 95.36 F1-score.

Table 4. Results on MSRA dataset.

Models	Precision	Recall	F1-score
Conditional Probabilistic Models [3]	91.22	81.71	86.20
Joint Identification and Categorization [27]	91.86	88.75	90.28
Adversarial Transfer Learning with Self-Attention [2]	91.30	89.58	90.64
Five-Stroke based Cnn-birnn-crf [24]	92.04	91.31	91.67
Convolutional Attention Network [28]	93.53	92.42	92.97
Multi-digraph Model with Gazetteers [5]	94.6	94.2	94.4
Our model	**95.92**	**94.80**	**95.36**

Table 5. Results on Weibo dataset.

Models	Named entities	Nominal mentions	Overall
Joint Trained Embedding [16]	51.96	61.05	56.05
Word Segmentation Representation Learning [17]	55.28	62.97	58.99
BiLSTM with F-score driven [7]	50.60	59.32	54.82
Unified Model for Cross-domain and Semi-supervised [8]	54.50	62.17	58.23
Lattice Network [26]	53.04	62.25	58.79
Multi-digraph Model with Gazetteers [5]	63.1	56.3	59.5
Our model	**59.79**	**63.22**	**61.46**

Weibo Dataset. We compare our model with the latest models on Weibo corpus. Weibo-NER is in the domain of social media. Results of named mentions, nominal mentions, and the total are demonstrated in Table 5 respectively. As there are many non-standard data in social media data, such as spelling errors, and informal words, the overall result of social media corpus is lower than that of MSRA data set. We can see that the model we proposed has achieved state-of-the-art performance.

Peng et al. [16] propose joint training for embedding and achieve 56.05 F1-score. Peng et al. [17] utilize word boundary tags as features to provide richer information and improve the F1-score to 58.99%. He et al. [8] propose a unified model for cross domain and improve F1-score to 58.23% from 54.82% [7]. Zhang et al. [26] investigate a lattice network which explicitly leverages word and word

sequence information, and achieve F1-score of 58.79%. Our proposed model has a significant improvement in the named entities, which improves 1.96% compared with Ding et al. [5]. And overall performance is significantly better than other models.

From the experimental results, we can see that our model has improved on both datasets compared with previous models. On the MSRA dataset, our model has improved 0.96, and 1.96% on the Weibo dataset. Because MASA data is standard, previous studies have achieved valid results on this data set. While there are many unregistered words in the Weibo dataset, and the recognition model based on two granularity representations with self-attention can effectively improve the recognition results.

The model improves on existing approaches to reduce out-of-vocabulary and word segmentation issues by using self-attention to fuse the information of the two granularity. The word-level structure make judgment on segmentation of the common words, and character-based semantic information can make more flexible combination of phrase.

5 Conclusion

This paper incorporates self-attention mechanism into BiLSTM-CRF neural network for Chinese named entity recognition. Our model uses self-attention to capture multi-granularity information through the total sequence, which combines the semantic and structural features of characters and words to predict entity tags. We solve the problems of word boundary segmentation and long-phrase combination, and the experimental results show that our method has improved the accuracy of Chinese named entity recognition.

Future work will focus on more granular information representations, such as sentence and paragraph levels, and apply this work to specialized entity identification in a variety of areas.

References

1. Bikel, D.M., Miller, S., Schwartz, R., Weischedel, R.: Nymble: a high-performance learning name-finder. In: Conference on Applied Natural Language Processing (1997)
2. Cao, P., Chen, Y., Liu, K., Zhao, J., Liu, S.: Adversarial transfer learning for Chinese named entity recognition with self-attention mechanism. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 182–192 (2018)
3. 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)
4. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. J. Mach. Learn. Res. **12**(Aug), 2493–2537 (2011)

5. Ding, R., Xie, P., Zhang, X., Lu, W., Li, L., Si, L.: A neural multi-digraph model for Chinese NER with gazetteers. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, pp. 1462–1467. Association for Computational Linguistics, July 2019. https://doi.org/10.18653/v1/P19-1141
6. 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
7. He, H., Sun, X.: F-score driven max margin neural network for named entity recognition in Chinese social media. arXiv preprint arXiv:1611.04234 (2016)
8. He, H., Sun, X.: A unified model for cross-domain and semi-supervised named entity recognition in Chinese social media. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
10. Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF Models for Sequence Tagging. arXiv:1508.01991 [cs], August 2015
11. Isozaki, H., Kazawa, H.: Efficient support vector classifiers for named entity recognition. In: International Conference on Computational Linguistics, pp. 1–7. Association for Computational Linguistics (2002)
12. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: probabilistic models for segmenting and labeling sequence data (2001)
13. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. CoRR abs/1603.01360 (2016). http://arxiv.org/abs/1603.01360
14. Liu, T., Yao, J.Q., Lin, C.Y.: Towards improving neural named entity recognition with gazetteers. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 5301–5307 (2019)
15. Ma, X.: End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. arXiv preprint arXiv:1603.01354 (2016)
16. Peng, N., Dredze, M.: Named entity recognition for Chinese social media with jointly trained embeddings. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 548–554 (2015)
17. Peng, N., Dredze, M.: Improving named entity recognition for Chinese social media with word segmentation representation learning. arXiv preprint arXiv:1603.00786, pp. 149–155 (2016). http://aclweb.org/anthology/P16-2025
18. Rei, M., Crichton, G.K., Pyysalo, S.: Attending to characters in neural sequence labeling models. arXiv preprint arXiv:1611.04361 (2016)
19. Shao, Y., Hardmeier, C., Tiedemann, J., Nivre, J.: Character-based joint segmentation and POS tagging for Chinese using bidirectional RNN-CRF. arXiv preprint arXiv:1704.01314 (2017)
20. Shen, Y., Yun, H., Lipton, Z.C., Kronrod, Y., Anandkumar, A.: Deep active learning for named entity recognition. arXiv preprint arXiv:1707.05928 (2017)
21. Vaswani, A., et al.: Attention Is All You Need. arXiv:1706.03762 [cs], June 2017
22. Xiang, Y., et al.: Chinese named entity recognition with character-word mixed embedding. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 2055–2058. ACM (2017)

23. 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 (Volume 1: Long Papers), vol. 1, pp. 1237–1247 (2017)
24. Yang, F., Zhang, J., Liu, G., Zhou, J., Zhou, C., Sun, H.: Five-stroke based CNN-BiRNN-CRF network for Chinese named entity recognition. In: Zhang, M., Ng, V., Zhao, D., Li, S., Zan, H. (eds.) NLPCC 2018. LNCS (LNAI), vol. 11108, pp. 184–195. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99495-6_16
25. Yang, Y., Zhang, M., Chen, W., Zhang, W., Wang, H., Zhang, M.: Adversarial Learning for Chinese NER from Crowd Annotations. arXiv:1801.05147 [cs], January 2018
26. Zhang, Y., Yang, J.: Chinese NER using lattice LSTM. arXiv preprint arXiv:1805.02023 (2018)
27. Zhou, J., Qu, W., Zhang, F.: Chinese named entity recognition via joint identification and categorization. Chin. J. Electron. **22**(2), 225–230 (2013)
28. Zhu, Y., Wang, G., Karlsson, B.F.: CAN-NER: Convolutional Attention Network for Chinese Named Entity Recognition. arXiv:1904.02141 [cs], April 2019
29. Zukov-Gregoric, A., Bachrach, Y., Minkovsky, P., Coope, S., Maksak, B.: Neural named entity recognition using a self-attention mechanism. In: 2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI), pp. 652–656. IEEE (2017)