Document-Level Named Entity Recognition by Incorporating Global and Neighbor Features

Anwen Hu, Zhicheng Dou(✉), and Ji-rong Wen

Beijing Key Laboratory of Big Data Management and Analysis Methods,
Renmin University of China, Beijing 100872, China
{anwenhu,dou,jrwen}@ruc.edu.cn

Abstract. State-of-the-art named entity recognition models mostly process sentences within a document separately. Sentence-level named entity recognition is easy to cause tagging inconsistency problems for long text documents. In this paper, we first propose to use the neural network to encode global consistency and neighbor relevance among occurrences of a particular token within a document. We first encode sentences within a document independently by a sentence-level BiLSTM layer, then we design a document-level module to encode the relation between occurrences of a particular token. In our document-level module, we use CNN to encode global consistency features and apply BiLSTM to model neighbor relevance features. We further apply a gate to effectively fuse these two non-local features and use a CRF layer to decode labels. We evaluate our model on the CoNLL-2003 dataset. Experimental results show that our model outperforms existing methods.

Keywords: Document-level · Named entity · Global · Neighbor

1 Introduction

Named entity recognition (NER) is usually a basic step in many natural language processing (NLP) tasks, such as calculating reputation for entities [19] and relation extraction [16]. Due to the success of recurrent neural network (RNN) and its variants in modeling sequential data, many RNN-based neural network models [3,6,8,12,15] were proposed for NER. Most of these models are designed for sentence-level named entity recognition: they treat sentences in a document independently during training or predicting. This is easy to cause that an identical entity in two separated sentences might be classified as different entity types, which is called tagging inconsistency problem. For the example given in Fig. 1, a sentence-level model named BiLSTM-CNN-CRF [15] successfully recognized the first "Wenchang" as a 'LOCATION' (a city in China). However, it misclassified the second one appearing in the second sentence, from 'LOCATION' to 'PERSON' due to its ambiguous local context.

Q. Zhang et al. (Eds.): CCIR 2019, LNCS 11772, pp. 79–91, 2019.
https://doi.org/10.1007/978-3-030-31624-2_7

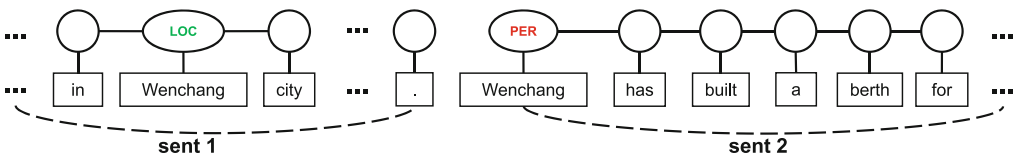

Fig. 1. An example of the label consistency problem within a document in the CoNLL-2003 English dataset.

Document-level NER has the potential to solve the tag inconsistency problem in sentence-level NER. For the example in Fig. 1, the first sentence explicitly tells Wenchang is a city. Using the context of "Wenchang" in the first sentence could help recognize the entity type of the same token in the second sentence. Many manually designed non-local features [2,5,9,10] were proposed to utilize context information in entire documents. To reduce reliance on feature engineering, some studies [14,20,22,24] proposed using neural networks to model the relation across sentences. For example, a global self-attention mechanism [24] was introduced to find useful contextual information across sentences based on semantic relevance.

In this paper, we propose a novel neural network leveraging global consistency and neighbor relevance for document-level NER. Our model uses BiLSTM to encode sentences within a document independently. Then, we apply a document-level module to model the relation between occurrences of a particular token across sentences. According to the statistic on CoNLL-2003 English dataset [21], we find more than 80% sequences of occurrences of a particular token refer to an identical entity. Thus, we use a CNN layer to learn global consistency features among all occurrences of the current token. Besides, when our humans are confused about a concept during reading, we will look for clearer contents in the document from near to far. To imitate this human's habit, we use a BiLSTM layer in the document-level module to learn neighbor relevance features from adjacent occurrences. The global consistency feature encodes how the token appears in the entire document. The neighbor relevance feature encodes the context of nearby occurrences. To decide how much information of these two features should be introduced respectively, we fuse these two features by a gated fusion module. At last, we use a CRF layer to decode labels for each sentence.

We evaluate our model on the CoNLL-2003 dataset. Experimental results show that adding either document-level feature can significantly improve the F1 score, and our gated fusion model obtains the best recognition quality.

The main contributions of this paper are:

- We propose a novel method for document-level named entity recognition. We use a sentence-level BiLSTM layer to encode sentences dependently and then use a document-level module to generate document-level features.
- We introduce two kinds of automatically learned document-level features: a global consistency feature extracted by CNN and a neighbor relevance feature encoded by BiLSTM. These features are fused by a gating mechanism.
- Experimental results confirm that our proposed method outperforms the state-of-the-art sentence-level and document-level NER models.

2 Related Work

2.1 Sentence-Level NER

Many statistics-based models, like HMM [13] and CRF [11], were effectively employed in sentence-level NER. With the development of the neural network, many RNN-based methods with CRF as the decoding layer [8,12] achieved better performance than statistic models. Besides, character-level information encoded by CNN [3] or BiLSTM [12] was proven to be significant for NER quality. BiLSTM-CNNS-CRF model [15] was truly end-to-end, which didn't require feature engineering, task-specific resources or data-processing. We propose to add a document-level module to the architecture of BiLSTM-CNNS-CRF to introduce non-local information within the entire document.

2.2 Document-Level NER

To make use of non-local information, many manually designed non-local features [2,5,10,18] were utilized for statistic-based methods and exhibited promising results. Some studies aimed to design global features to make occurrences of a particular token within a document labeled consistently, such as Init-Caps of Other Occurrences (ICOC) [2] and Entity-majority feature [10]. Similarly, in this paper, we use a CNN based global vector to introduce a consistent document-level representation for all occurrences of a particular token within a document. Besides, the context aggregation feature [18] was proposed for the case that identical tokens may not have identical label assignments. For example, "Australia" can be labeled as 'LOC', and "The bank of Australia" should be labeled as 'ORG'. In this work, BiLSTM based neighbor vector is different for each occurrence, and we incorporate local context representations and non-local representations. Both points avoid all occurrences of a particular token are labeled as the same entity type.

With the development of the neural network, there were also some neural network based methods that didn't rely on manually designed features. ID-CNN [20] iteratively applied 'block' (a stack of dilated convolutions [23]) with the same parameters several times to encode document-level features. Att-BiLSTM-CRF [14] used an attention mechanism to find useful context information within a document for the chemical named entity. Both these two work encoded the whole long sequence concatenated by sentences within a document. NER reasoner [22] was designed as a multi-layer architecture, where each layer could utilize context information of entities predicted by the last layer with a candidate pool. Global-ATT [24] is the most relevant work to ours. They used BiLSTM to encode sentences within a document independently and used a self-attention mechanism to find useful context information from all occurrences of a particular token. The main difference between Global-ATT [24] and our model is how to generate reliable non-local features. Unlike using the self-attention mechanism to focus on semantically similar occurrences, our model generates non-local features by incorporate global consistency features and neighbor relevance features.

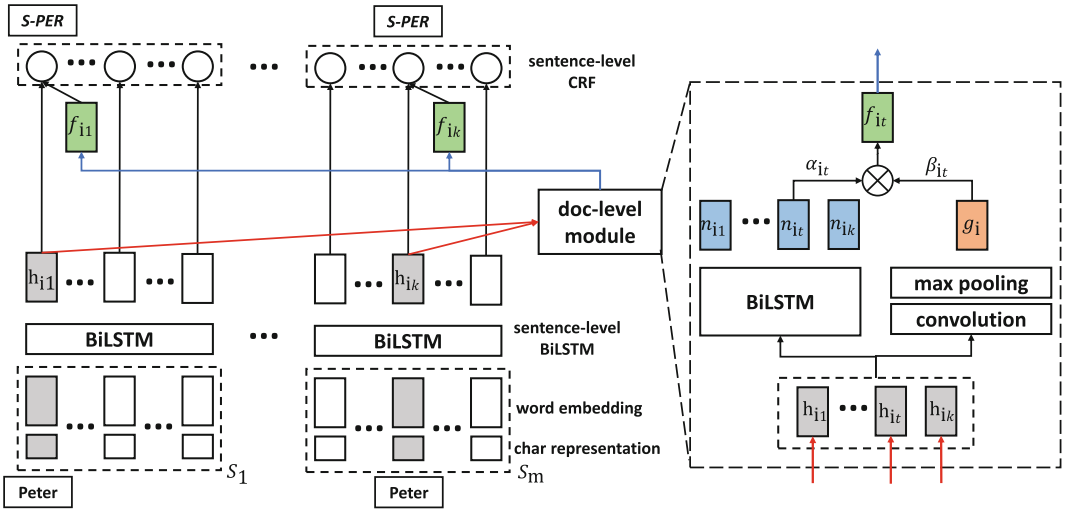

Fig. 2. The architecture of our model. $\mathbf{S} = (S_1, ..., S_{m-1}, S_m)$ is a list of sentences within a document. $\mathbf{h_i} = (h_{i1}, h_{i2}, ..., h_{ik})$ is a list of sentence-level BiLSTM outputs for occurrences of token 'Peter'. $\mathbf{n_i} = (n_{i1}, n_{i2}, ..., n_{ik})$ is a list of neighbor representations. g_i is a global representation. $\mathbf{f_i} = (f_{i1}, f_{i2}, ..., f_{ik})$ is a list of fused representations.

3 Our Document-Level NER Model: GNG

We propose a novel neural network architecture that fuses **G**lobal consistency features and **N**eighbor relevance features with a **G**ate mechanism, namely GNG. We design a document-level module to take into account context information from all occurrences of a particular token within a document. The architecture of our proposed model is shown in Fig. 2. We first encode sentences within a document independently in the sentence-level BiLSTM layer. Then for each token, we collect local contextual representations of its occurrences as the input of the document-level module. The document-level module returns a fused document-level feature vector for each occurrence. We concatenate sentence-level contextual representation and fused document-level feature vectors to new hidden states. At last, we apply a sentence-level CRF layer to decode the label sequences. We will introduce the details of each component in the remaining part of this section.

3.1 Sentence-Level Bi-directional LSTM

The Long Short-Term Memory Network (LSTM) [7] is a variant of the recurrent neural network (RNN) designed to learn long-term dependencies. LSTM is composed of a memory cell and three gates to control how much information to forget and to pass on to the next time step. However, LSTM only takes information from the past and ignore future information. Bidirectional LSTMs combine two LSTMs in two directions, one in the forward direction and the other in the backward direction. For each sentence, our sentence-level BiLSTM is fed a sequence of token representations which are concatenated by word embeddings

Table 1. Statistics of entity type consistency on the CoNLL-2003 dataset. **Consistent** and **Inconsistent** refer to counts and percentages of consistent tag sequences and inconsistent tag sequences.

Dataset	Consistent	Inconsistent
Train	19,460 (87.2%)	2,868 (12.8%)
Development	4,804 (86.4%)	753 (14.6%)
Test	4,734 (88.2%)	636 (12.8%)

and CNN based char representations [15]. At each time, BiLSTM concatenates two hidden states $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ to the output:

$$h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \tag{1}$$

h_t aggregates context information of each token within a sentence, so we call h_t as local context feature/representation.

3.2 Document-Level Module

After the sentence-level BiLSTM layer, we apply a document-level module to aggregate local context features from all occurrences of a particular token within a document. For each token x_i, we collect its occurrences as $\mathbf{u_i} = (u_{i1}, u_{i2}, ..., u_{ik})$, where k is the count of occurrences for token x_i within a document. Then, according to the positions of these occurrences, we obtain a list of local context representations for $\mathbf{u_i}$:

$$\mathbf{h_i} = (h_{i1}, h_{i2}, ..., h_{ik}) \tag{2}$$

where $h_{it}(1 \leq t \leq k)$ is obtained by Eq. (1). h_{it} means the local context feature vector of the t^{th} occurrence of token x_i. $\mathbf{h_i}$ is the input of the document-level module for token x_i.

Our document-level module can be divided into two parts. We use a CNN layer to extract a global consistency feature for all occurrences. And for each occurrence, we use a BiLSTM layer to capture another neighbor relevance feature from its adjacent occurrences. These two non-local features can be used alone or fused to a new feature.

CNN in Doc-Level Module. In most cases, different occurrences of a particular token in a document are labeled as the same entity type. The consistency statistic on the CoNLL-2003 shown in Table 1 supports this intuition. Table 1 shows the counts and percentages of consistent tag sequences and inconsistent tag sequences within a document. Consistent tag sequence means all occurrences of a particular token within a document are labeled as the same tag. For example, (B-PER, B-PER, B-PER) for token "Peter" is consistent, (B-PER, E-PER,

B-PER) or (B-PER, B-ORG, B-PER) for token "Peter" is inconsistent. Obviously, consistent tag sequences are much more than inconsistent tag sequences. So, in most cases, when the local context of a token is not sufficient for models to classify its entity type, information from other occurrences can provide some help. Based on this idea, we first introduce a global consistency feature for all occurrences of a particular token within a document. After getting sentence-level Bi-directional LSTM outputs $\mathbf{h_i}$, we apply a CNN with the max-pooling layer to extract the global consistency representation g_i. Its calculation is defined as:

$$g_i = \text{CNN}(\mathbf{h_i})$$

where $\mathbf{h_i}$ is obtained by Eq. (2). CNN() refers to the convolution and the max-pooling layer in the document-level module. g_i is shared by all occurrences of a particular token within a document.

BiLSTM in Doc-Level Module. In cases where not all occurrences of a particular token are labeled as the same entity type, it's inappropriate to introduce an identical global feature for all occurrences of a particular token. While reading an article, if we humans have doubts about a certain concept in a sentence, we will first look for neighbor occurrences of the concept to get more context information across sentences. Based on this human habit, we introduce a neighbor relevance feature for each occurrence. Note "neighbor" means neighbor occurrences rather than neighbor context or neighbor sentences. Besides, news editors always organize articles in chronological order. So we think descriptions of an entity in an article can be seen as a time sequence. For example, there is an article about the Bank of Japan governor "Matsushita" in the CoNLL-2003 dataset. It is first mentioned that Matsushita's view on the yen was quoted in Japan's leading economic daily. Then it states the effects of his comments. Next, it says that Matsushita further expressed his point of view in the following interview. Therefore, to encode the sequence composed of local context representations with timing characteristics like this, we apply another BiLSTM layer to learn neighbor relevance representation $n_{it}(1 \leq t \leq k)$ for each occurrence.

$$\mathbf{n_i} = \text{BiLSTM}(\mathbf{h_i}),$$
$$\mathbf{n_i} = (n_{i1}, n_{i2}, ..., n_{ik})$$

where $\mathbf{h_i}$ is obtained by Eq. (2). n_{it} means neighbor relevance representation for the t^{th} occurrence of token x_i.

Gated Fusion. The influence of the global consistency representation and the neighbor relevance representation may be different in different cases, so we propose a gated fusion to fuse these two features. For each occurrence of a particular token, we get its global feature and neighbor feature. Then based on its local

context feature, we compute the weights of the two non-local features and get a new fused feature. The two weights are calculated as follows:

$$\alpha_{it} = \sigma(W_\alpha(\tanh(W_{\hat{g}} g_i + b_{\hat{g}}) \oplus h_{it})),$$
$$\beta_{it} = \sigma(W_\beta(\tanh(W_{\hat{n}} n_{it} + b_{\hat{n}}) \oplus h_{it})),$$

where h_{it} is obtained by Eq. (1). g_i is global consistency representation for token x_i, n_{it} is neighbor relevance representation for the t^{th} occurrence of token x_i. $W_{\hat{g}}$, $W_{\hat{n}}$, W_α, W_β are weight matrices. $b_{\hat{g}}$, $b_{\hat{n}}$ are bias vectors. $\oplus$ is the concatenating operation, σ is the sigmoid function. α_{it} and β_{it} is the weight vector for global representation and neighbor representation, respectively. The gated fusion is defined as:

$$f_{it} = \alpha_{it} g_i + \beta_{it} n_{it}$$

where f_{it} is the fused document-level feature for the t^{th} occurrence of token x_i.

3.3 CRF Layer

To consider the correlations between labels in neighborhoods, we apply the sentence-level Condition Random Fields (CRF) layer [11] to decode the best label sequence for each sentence independently.

For a sentence $\mathbf{x} = (w_1, w_2, ..., w_l)$ (l is the number of tokens), we concatenate its local context representations and document-level representations and get new hidden states $\hat{\mathbf{h}_\mathbf{x}} = (\hat{h}_1, \hat{h}_2, ..., \hat{h}_l)$. Then, we reduce the dimension of vectors in $\hat{\mathbf{h}_\mathbf{x}}$ to the number of distinct tags e with two fully connected layers. We convert new low dimensional hidden states to a score matrix $\mathbf{P} \in \mathbb{R}^{l \times e}$. P_{ij} refers to the score of the j^{th} tag for the i^{th} token in the sentence. For any possible predicted sequence $\mathbf{y} = (y_1, y_2, .., y_l)$, its probability is defined as:

$$p(\mathbf{y}|\mathbf{x}; W^t) = \frac{\exp \psi(\mathbf{x}, \mathbf{y})}{\sum_{\tilde{\mathbf{y}} \in \mathcal{Y}_\mathbf{x}} \exp \psi(\mathbf{x}, \tilde{\mathbf{y}})}$$

where $\psi(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{l} P_{i,y_i} + \sum_{i=0}^{l} {W^t}_{y_i, y_{i+1}}$, and W^t is a learned transition matrix, ${W^t}_{y_i, y_{i+1}}$ represents the transition score from the tag y_i to the tag y_{i+1}. $\mathcal{Y}_\mathbf{x}$ is a set of all possible tag sequences. During training, we maximize the log-probability of the ground-truth tag sequence $\hat{\mathbf{y}}$. The loss function is defined as:

$$loss = -\log(p(\hat{\mathbf{y}}|\mathbf{x}; W^t)),$$

Decoding is searching for the tag sequence which obtains the maximum score:

$$\mathbf{y}^\star = \arg\max_{\tilde{\mathbf{y}} \in \mathcal{Y}_\mathbf{x}} \psi(\mathbf{x}, \tilde{\mathbf{y}})$$

4 Experiments

4.1 Dataset

We perform experiments on the CoNLL-2003 English dataset [21], which is taken from the Reuters Corpus comprised of news stories between August 1996 and August 1997. CoNLL-2003 English dataset contains four different types of named entities: persons (PER), organizations (ORG), locations (LOC), and miscellaneous names (MISC). The statistics of the dataset are shown in Table 2. We use the BIOES tagging scheme, which has been proven better than standard BIO2 by previous studies [12,18].

Table 2. Statistics of the CoNLL-2003 dataset. **#token**, **#sentence** and **#document** refer to counts of tokens, sentences and documents respectively.

Dataset	#token	#sentence	#document
Train	20,3621	14,987	946
Test	46,435	3,684	231
Development	51,362	3,466	216

4.2 Baselines

To verify the effectiveness of GNG, we compare our model with several state-of-the-art NER models. These models can be divided into two groups: sentence-level models and document-level models.

Sentence-Level Models

- **LSTM-CRF** [12], which uses a BiLSTM layer to extract character-level representation and uses another BiLSTM layer to encode sentences.
- **BiLSTM-CNN-CRF** [15], which uses CNN to extract morphological information from characters of words, and a BiLSTM layer to encode sentences.
- **BiLSTM-CNN** [3], which extracts character features with a CNN layer and encodes sentences with a BiLSTM layer. Besides, it uses lexicons as a form of external knowledge.
- **Parallel-RNNs** [6], which is a parallel LSTM model for NER.

Document-Level Models

- **Two-stage CRF** [10], which designs three features corresponding to a function of aggregate statistics of the output of the first CRF at the document level, namely Token-majority features, Entity-majority features and Superentity-majority features.

- **Ratinov09** [18], which uses three non-local features (context aggregation, two-stage prediction aggregation, extended prediction history) and external knowledge to improve the performance of perceptron based NER system.
- **Att-BiLSTM-CRF** [14], which concatenates sentences within a document to a sequence, and uses an attention layer to extract global information.
- **ID-CNN** [20], which uses Iterated Dilated Convolutional Neural Network to handle a very long sequence concatenated by sentences of a document.
- **Global-ATT** [24], which applies a self-attention mechanism on occurrences of a particular token within a document to generate global representation.
- **NER reasoner** [22], which is a multi-layer architecture, where each layer makes use of named entities recognized by the last layer.

Other Baselines. To analyze the contribution of each component in our document-level module, we also experiment with using each document-level feature alone or concatenating two document-level features directly.

- **GNG-DOC**: Previous sentence-level methods randomly shuffled sentences in the dataset during training. GNG needs complete document information, so we shuffle documents randomly during training. To compare the impact of training strategies, we train our basic sentence-level model GNG-DOC like GNG. The architecture of GNG-DOC is the same as BiLSTM-CNN-CRF.
- **GNG-LSTMN**: This model only uses global consistency representations extracted by the CNN layer in the document-level module.
- **GNG-CNNG**: This model only uses neighbor relevance representations encoded by the BiLSTM layer in the document-level module.
- **GNG-GATE**: This model concatenates global consistency representations and neighbor relevance representations, other than using a fusion gate.

4.3 Parameter Setting

We perform experiments with conventional Glove 100-dimensional embedding [17] or word embeddings produced by pre-trained language models named bert-base [4] and flair [1]. The optimizer is stochastic gradient descent (SGD) with batch size 2 and momentum 0. 9. Word length in character-level CNN is set to 64, sentence length in sentence-level BiLSTM is set to 130, both of which are slightly bigger than the maximum in the CoNLL-2003 dataset, and we apply zero-operation as necessary. We find entity tokens which appear more than 20 times within a document in the CoNLL-2003 are very rare. Thus, the maximum length of a list consisting of occurrences of a particular token is set to 20. If the number of occurrences of a particular token within a document is bigger than 20, its document-level information will be ignored.

4.4 Results and Discussion

In this paper, we use standard F1-score (F1) as the evaluation metrics. We conduct each experiment 4 times and report its mean. Experimental results are shown in Tables 3 and 4. We find:

(1) **GNG outperform existing sentence-level models and document-level models**. GNG-DOC achieves comparable performance with LSTM-CRF. It shows that for the sentence-level model, there is no obvious difference between shuffling all sentences and shuffling documents like GNG during training. So, the improvement of GNG in F1 has nothing to do with our training strategies. Our document-level module can indeed significantly improve the NER quality based on the sentence-level model.
(2) **All components in document-level modules are important**. Both GNG-CNNG and GNG-LSTMN underperform GNG, which indicates both global consistency features and neighbor relevance features are essential in our document-level module. Besides, GNG-GATE underperforms our GNG, which shows our fusion gate can fuse these two non-local features more effectively.

Table 3. F1 scores of different approaches on the test set of CoNLL-2003. ‡ marks the neural model. ∗ marks model which uses external resources. Our models use the glove as default word embedding. **F1** refers to F1-score.

Model	F1
LSTM-CRF‡	90.94
BiLSTM-CNN-CRF‡	91.21
BiLSTM-CNN‡∗	91.62
Parallel-RNNs‡	91.48
Two-stage CRF	87.24
Ratinov09∗	90.57
Att-BiLSTM-CRF‡	90.49
ID-CNN‡	90.65
Global-ATT‡	91.43
NER reasoner‡	91.44
GNG-DOC‡	90.92
GNG-LSTMN‡	91.76
GNG-CNNG‡	92.05
GNG-GATE‡	91.78
GNG‡	**92.12**

Table 4. F1 scores of GNG-DOC and GNG on CoNLL-2003 with different word embeddings. **bert-base** [4] and **flair** [1] are word embeddings produced by pre-trained language models.

Model	glove	bert-base	flair
GNG-DOC	90.92	90.76	92.64
GNG	**92.12**	**91.45**	**92.96**

(3) As shown in Table 4, with either **bert-base** [4] or **flair** [1] as initialized word embedding, GNG outperforms GNG-DOC. Our document-level module learns features across sentences, which are overlooked in these state-of-the-art sentence-level language models. Thus our document-level feature can further improve NER quality at the base of word embeddings produced by these pre-trained language models.

5 Conclusion

In this paper, we propose a novel neural network named GNG that incorporates global consistency feature and neighbor relevance feature for document-level named entity recognition. GNG encodes sentences within a document independently, and utilizes a document-level module to model relations between occurrences of a particular token. In the document-level module, there is a CNN layer to learn global consistency and a BiLSTM layer to encode neighbor relevance. A gate mechanism is further used to fuse these two non-local representations. GNG achieves the state-of-the-art result on the CoNLL-2003 English dataset.

Acknowledgments. This work was supported by National Natural Science Foundation of China No. 61872370, National Key R&D Program of China No. 2018YFC0830703, and the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China No. 2112018391.

References

1. Akbik, A., Blythe, D., Vollgraf, R.: Contextual string embeddings for sequence labeling. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 1638–1649 (2018)
2. Chieu, H.L., Ng, H.T.: Named entity recognition with a maximum entropy approach. In: Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, 31 May–1 June 2003, pp. 160–163 (2003)
3. Chiu, J.P.C., Nichols, E.: Named entity recognition with bidirectional lstm-cnns. TACL **4**, 357–370 (2016)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
5. Finkel, J.R., Grenager, T., Manning, C.D.: Incorporating non-local information into information extraction systems by Gibbs sampling. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, pp. 363–370 (2005)
6. Gregoric, A.Z., Bachrach, Y., Coope, S.: Named entity recognition with parallel recurrent neural networks. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), vol. 2, pp. 69–74 (2018)
7. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735

8. Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF models for sequence tagging. CoRR abs/1508.01991 (2015)
9. Kazama, J., Torisawa, K.: A new perceptron algorithm for sequence labeling with non-local features. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2007, pp. 315–324 (2007)
10. Krishnan, V., Manning, C.D.: An effective two-stage model for exploiting non-local dependencies in named entity recognition. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, ACL 2006, Sydney, Australia, 17–21 July 2006 (2006)
11. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: probabilistic models for segmenting and labeling sequence data (2001)
12. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. In: The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016, pp. 260–270 (2016)
13. Leek, T.R.: Information extraction using hidden Markov models. Master's thesis, University of California, San Diego (1997)
14. Luo, L., Yang, Z., Yang, P., Zhang, Y., Wang, L., Lin, H., Wang, J.: An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. Bioinformatics **34**, 1381–1388 (2017)
15. Ma, X., Hovy, E.H.: End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers, ACL 2016, Berlin, Germany, 7–12 August 2016 (2016)
16. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: ACL/IJCNLP, pp. 1003–1011. The Association for Computer Linguistics (2009)
17. Pennington, J., Socher, R., Manning, C.: Glove: global vectors for word representation. In: Conference on Empirical Methods in Natural Language Processing, pp. 1532–1543 (2014)
18. Ratinov, L., Roth, D.: Design challenges and misconceptions in named entity recognition. In: Proceedings of the Thirteenth Conference on Computational Natural Language Learning, CoNLL 2009, Boulder, Colorado, USA, 4–5 June 2009, pp. 147–155 (2009)
19. Seghouani, N.B., Bugiotti, F., Hewasinghage, M., Isaj, S., Quercini, G.: A frequent named entities-based approach for interpreting reputation in Twitter. Data Sci. Eng. **3**(2), 86–100 (2018)
20. Strubell, E., Verga, P., Belanger, D., McCallum, A.: Fast and accurate entity recognition with iterated dilated convolutions. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, 9–11 September 2017, pp. 2670–2680 (2017)
21. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003-Volume 4, pp. 142–147. Association for Computational Linguistics (2003)

22. Yin, X., Zheng, D., Lu, Z., Liu, R.: Neural entity reasoner for global consistency in ner. arXiv preprint arXiv:1810.00347 (2018)
23. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. In: ICLR (2016)
24. Zhang, B., Whitehead, S., Huang, L., Ji, H.: Global attention for name tagging. In: Proceedings of the 22nd Conference on Computational Natural Language Learning, pp. 86–96 (2018)