A Sequence Transformation Model for Chinese Named Entity Recognition

Qingyue Wang[1,3], Yanjing Song[2], Hao Liu[2], Yanan Cao[3(✉)], Yanbing Liu[3], and Li Guo[3]

[1] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
qingyue.wang2018@gmail.com
[2] Software Institute, Beijing Institute of Technology, Beijing, China
yanjing.song2018@gmail.com, huakaicary@gmail.com
[3] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
{caoyanan,liuyanbing,guoli}@iie.ac.cn

Abstract. Chinese Named Entity Recognition (NER), as one of basic natural language processing tasks, is still a tough problem due to Chinese polysemy and complexity. In recent years, most of previous works regard NER as a sequence tagging task, including statistical models and deep learning methods. In this paper, we innovatively consider NER as a sequence transformation task in which the unlabeled sequences (source texts) are converted to labeled sequences (NER labels). In order to model this sequence transformation task, we design a sequence-to-sequence neural network, which combines a Conditional Random Fields (CRF) layer to efficiently use sentence level tag information and the attention mechanism to capture the most important semantic information of the encoded sequence. In experiments, we evaluate different models both on a standard corpus consisting of news data and an unnormalized one consisting of short messages. Experimental results showed that our model outperforms the state-of-the-art methods on recognizing short interdependence entity.

Keywords: Named Entity Recognition · Deep learning
Sequence to sequence neural network · Conditional Random Fields

1 Introduction

Named Entity Recognition (NER) is used to accurately identify a series of entities from text, such as person, location and organization, which can be used for senior natural language processing (NLP) applications.

Most related works regard NER as a sequence tagging task. Typical statistical models include Hidden Markov Model (HMM) [13], Conditional Random Fields (CRF) [17] and etc. They still suffer from extracting effective grammatical features and templates manually. As more and more systems using neural models have achieved good performances in different NLP tasks, deep neural network on sequence tasks raises continuing concern. Collobert [6] firstly addressed the sequence tagging

© Springer Nature Switzerland AG 2018
W. Liu et al. (Eds.): KSEM 2018, LNAI 11061, pp. 491–502, 2018.
https://doi.org/10.1007/978-3-319-99365-2_44

problems in an end-to-end way, which tried to pre-process features as little as possible and designed a multilayer neural network architecture for Word Segmentation, Chunking and Named Entity Recognition. However, its performance is limited by the fixed size window of words although the neural language is closely related to the context. Most recently, Huang [20] combined a bidirectional Long Short-Term Memory [11] (LSTM) network and a CRF layer, called BiLSTM-CRF, which produced state-of-art accuracy on several NLP tagging tasks. In this model, BiLSTM uses both past and future information of input and CRF layer utilizes sentence-level tags. Ma [8] introduced an end-to-end network architecture which combines bidirectional LSTM, convolutional neural networks (CNN) and CRF, and it benefits from both word-level and character-level representations.

Unlike previous works, we regard NER as a sequence transformation task in this paper. In order to properly model this task, we propose a variety of sequence-to-sequence (seq2seq) neural network models. In the baseline seq2seq model, we use a BiLSTM encoder and a LSTM decoder to capture the context information for LSTM's good ability to solve long-term dependencies in source text. In the upgraded model, we combined the seq2seq model with a CRF layer, which can utilize the past and future tags to predict the current tag with high precision. Besides, we utilize an attention mechanism to both above models, which is conditioned on a distinct context vector for each target label, making the decoder pay more attention on current context information during predicting sequence. These models are all evaluated on a standard corpus (People's Daily news) and a short message corpus we constructed. And the experiments results showed that our model reaches a good performance especially on short dependence entity.

Our contributions can be summarized as follows. (1) We design several sequence-to-sequence models for Chinese Named Entity Recognition. As far as we know, we are among the first endeavors to resolve the NER problem in this way. (2) We explore the effectiveness of attention mechanism on our models for Chinese Named Entity Recognition. (3) We systematically compare the performance of existing models on both short messages and news texts for NER. And we show that our model based seq2seq-CRF-attention can produce state-of-the-art (or close to) F1 scores on recognizing person, location and organization.

The remainder of this paper is organized as follows. Section 2 discusses related work. Section 3 describes the model we propose in detail. Section 4 introduces the experiments and analysis results on different methods and entities. Section 5 draws conclusions finally.

2 Related Work

The earliest Named Entity Recognition method was rule-based recognition, which relied on the language experts or domain experts to specify effective grammar rules such as gazetteers, costing a lot of time and energy. For sequence labeling task, Hidden Markov Model [13] (HMM), Support Vector Machine [16] (SVM), maximum entropy

Markov models [15] (MEMMS) and Conditional Random Fields [17, 19] (CRF) once achieved good results on NER. In recent years, several neural architectures start showing great learning power. Deep Neural Network proposed by Collobert [6] introduced a radically approach trying to preprocess features as little as possible and used a multilayer neural network-based windows and sentences. Lample [7] presented a LSTM-CRF architect with a char-LSTM layer learning spelling features from supervised corpus and didn't use any additional resources. Following the idea, Dong [9] was the first to investigate Chinese radical-level representation in BiLSTM-CRF architecture and got better performance without carefully designed features. Ma [8] proposed a BiLSTM-CNNs-CRF architecture using CNNs to model character-level information. To apply neural network to natural language, word embedding [4, 10] is used to convert language tokens to vectors, which greatly help express word meaning in space and improve the performance of many NLP applications.

In machine translation, Sutskever [1] proposed an encoder and a decoder for each language. This model is jointly trained to maximize the probability of a correct translation. Hermann [3] used the similar neural encoder-decoder model in question answering and Nallapati [5] proposed a neural network model in abstractive text summarization using sequence-to-sequence. Bahdanau [2] achieved a novel neural network model based on attentional encoder-decoder model for machine translation. Inspired by this mechanism, Paulus [12] introduced a neural network model with an intra-attention and a new training method that combines standard supervised word prediction in abstractive summarization. Unlike in machine translation and speech recognition, alignment is explicit in some NLP applications. Liu and Lane [18] described their approach introducing attention to the alignment-based RNN models for joint intent detection and slot filling. Our model basically follows their idea, but we modify the model to solve NER problem.

3 Proposed Methods

We combine the seq2seq model, attention mechanism and a CRF network to form a seq2seq-Attention-CRF model, which is illustrated in Fig. 1. Given an input sentence denoted by $x = \{x_1, x_2, \ldots, x_T\}$, x_t represents the t-th character (or word) at time step t. We use a bidirectional LSTM network (BiLSTM) to encode the input sentence, and a LSTM network to decode the hidden state h_t and vector c_t from the encoder. The vector c_t computed by attention mechanism is used to capture the information of the encoded sequence. After decoding, CRF layer utilizes the probability generated from the decoder and transition matrix to predict optimal tagging sequence. The output $y = \{y_1, y_2, \ldots, y_T\}$ represents labeled sequence corresponding to the input. Here, we use the most common tagging scheme named IOB (Inside, Beginning, Outside). Next, we introduce the components of our model respectively: aligned encoder-decoder, encoder-decoder based attention and CRF layer.

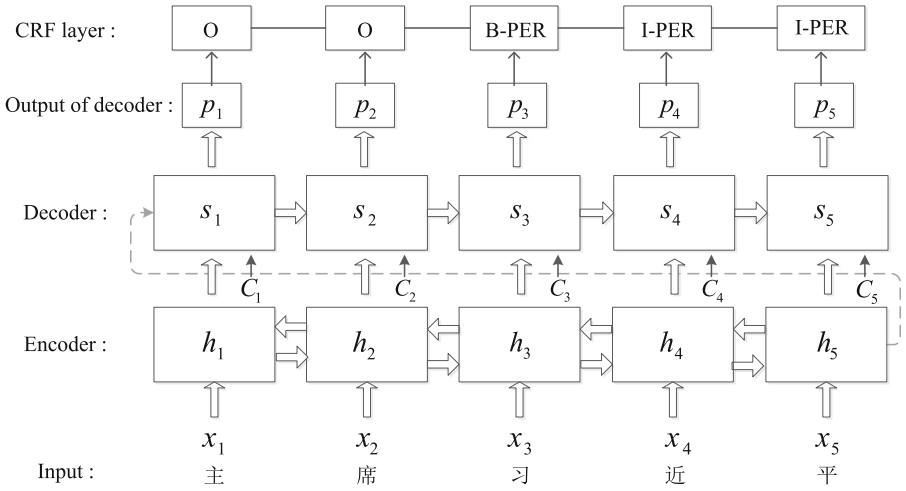

Fig. 1. Our model for Chinese Named Recognition based character-level. We use a bidirectional LSTM as an encoder, a unidirectional LSTM as a decoder, attention mechanism in seq2seq model and a CRF layer after decoding.

3.1 Aligned Encoder-Decoder

Here, we briefly describe the underlying framework, called encoder-decoder which learns to align and recognize name entity. The encoder and decoder are two separate RNNs.

On encoder side, the model reads the input sequence with a bidirectional LSTM encoder. For a given sentence x containing T characters, each character is represented as a d-dimensional vector. A LSTM computes a representation $\vec{h}_t$ of the left context of the sentence at each time t. Similarly, the right context $\vec{h}_t$ starting from the end of the sentence provides the future information of input. The final encoder hidden state h_t at each time step t is obtained by concatenating its left and right context representations $h_t = [\vec{h}_t, \overleftarrow{h}_t]$.

On decoder side, we use a unidirectional LSTM because the encoder with forward and backward LSTM has carried entire information of the sequence. We initialize the decoder hidden state with $s_0 = h_T$. At each decoding step t, the decoder hidden state s_t is equal to a function of the previous emitted label y_{t-1}, the aligned encoder hidden state h_t, and the context vector c_t:

$$s_t = f(s_{t-1}, y_{t-1}, h_t, c_t) \tag{1}$$

f is a nonlinear function and we use LSTM as f. The context vector c_t computed by attention mechanism will be introduced in next section.

3.2 Encoder-Decoder Based Attention

By allowing a model to automatically search for parts of a source sentence that are relevant to predicting a target word, attention mechanism can be spread throughout the sequence of annotations and retrieved by the decoder accordingly. Attention mechanism has shown promising results in many other NLP tasks such as machine translation, speech recognition and etc. Inspired by these works, we introduce the attention mechanism in neural machine translation which is proposed by Bahdanau [2]. The illustration of the attention mechanism in our model is shown in Fig. 2.

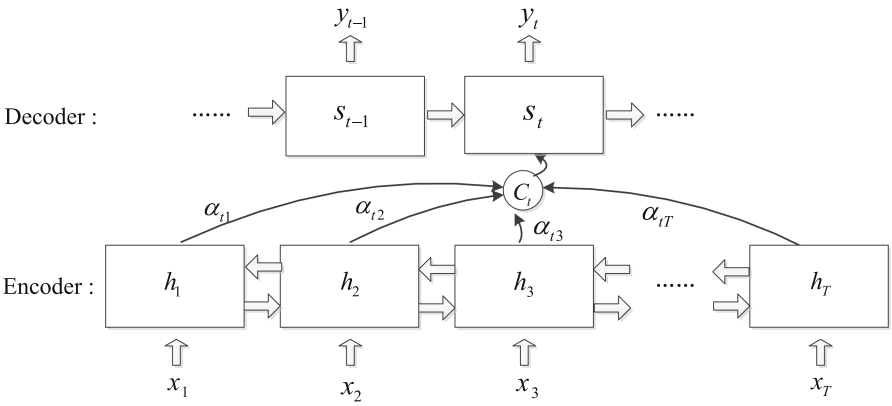

Fig. 2. The graphical illustration of the attention mechanism trying to generate the t-th predicted label given a source sentence x.

At each decoding step t, an attention function is used to attend over specific part of the encoded input sequence. The context vector c_t input to the decoder is computed as a weighted sum of the encoder hidden state h_t:

$$c_t = \sum_{j=1}^{T_x} \alpha_{tj} h_j \tag{2}$$

The weight α_{tj} of each hidden state h_j is computed by:

$$\alpha_{tj} = \frac{\exp(e_{tj})}{\sum_{k=1}^{T_x} \exp(e_{tk})} \tag{3}$$

$$e_{tj} = a(s_{t-1}, h_j) \tag{4}$$

where a is a feedforward neural network which is jointly trained with all the other components of the proposed model.

3.3 CRF Layer

It has been shown that CRF can produce higher tagging accuracy for part-of-speech (POS), chunking and NER, because it can efficiently use sentence level tag information. A CRF layer has a state transition matrix that can be trained with other parameters in the seq2seq-CRF-attention network.

We consider that the probability matrix $f_\theta([x]_1^T)$ is output by the network we proposed. The element $[f_\theta]_{[i]_t,t}$ of the matrix is the score output by the network with parameters θ, for sentence $[\mathrm{x}]_1^T$ and for the i-th tag at the t-th word. In our model, we note the new parameters as $\tilde{\theta} = \theta \cup \{[\mathrm{A}]_{i,j} \forall i,j\}$. The transition probability matrix $[A]_{i,j}$ represents the transition from the i-th tag to j-th tag. The final score of a sentence $[\mathrm{x}]_1^T$ is defined as follows:

$$s([\mathrm{x}]_1^T, [i]_1^T, \tilde{\theta}) = \sum_{t=1}^{T} ([\mathrm{A}]_{[i]_{t-1},[i]_t} + [f_\theta]_{[i]_t,t}) \tag{5}$$

To choose the optimal labeling sequence, we make use of the equation by the principle of maximum likelihood estimation. The best sequence can be computed as follows:

$$[\hat{i}]_1^T = \arg\max(s([x]_1^T, [i]_1^T, \tilde{\theta})) \tag{6}$$

4 Experiments and Result

4.1 Datasets and Evaluation

We evaluate the proposed approach on both unnormalized text (short messages) and standard text (People's Daily News). It should be noted that the short messages dataset contains more noise data including inform nicknames and wrong characters compared with news corpus. We estimate the system performance using precision (P), recall (R), F1 scores (F1) and IOB tagging scheme.

Short Messages. This corpus includes 200,000 messages, and the average text length is about 60 characters. We use 160,000 messages as a training dataset and 40,000 messages as a testing one. There are a few organization entities in this corpus, so we mainly recognized two types of entities: person and location.

People's Daily News. This dataset contains the whole 2000 year's news, and we regard the first ten months including 431289 sentences as a training dataset and the rest 98579 sentences as a testing one. In this corpus, we recognized three types of entities: person, location and organization.

4.2 Comparative Methods

We aim to evaluate the effectiveness of our proposed seq2seq models. In experiments, we use several typical statistical machine learning methods and neural network models

as comparative methods, including state-of-the-art models. Here is a brief induction to these methods.

HMM. Hidden Markov model [13] is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (i.e. hidden) states. The sequence of tokens generated by an HMM gives some information about the sequence of states so that it is especially known for their application in sequence problems such as speech and part-of-speech tagging.

CRF. Conditional Random Fields applied for labeling sequential data is a probabilistic model, and it also is the best statistical model on NER. In practice, we use the CRF++ package[1], a customizable and open source implementation of CRF, as an indispensable part of comparative methods.

BiLSTM+CRF. This method combining BiLSTM with CRF layer gains advantages of both neural network model and probabilistic model, and it also achieves state-of-the-art performance on tagging problems. We apply the same network architecture described by Huang [20] on Chinese Named Entity Recognition.

Seq2seq. This method is one of our baseline models. Since seq2seq was proposed, it has been applied in many sequence problems such as machine translation and image captioning. As a baseline model, it is designed as bidirectional LSTM encoder and unidirectional LSTM decoder.

Seq2seq+Attention. It is another model of our baselines. In experiments, we extend seq2seq model (mentioned above) with attention following the design of Bahdanau [2]. We compare this model with seq2seq to verify the effectiveness of attention on NER problem.

4.3 Implementation Details

The implementation of two LSTM follows the design in [22]. We set the number of units in LSTM cell as 200. Dropout rate 0.5 is applied to the non-recurrent connections [22] during model training for regularization. We adopt Adam optimization algorithm [21] starting with an empirical learning-rate of 0.001. Word embedding of size 256 are randomly initialized and fine-tuned during mini-batch training with batch size of 16. To avoid generating an oversized vocabulary, we delete the low-frequency character (or word) which appears less than 5 times in the corpus. The maximum norm for gradient clipping is set to 5. Our implementation is fully based on tensorflow1.4 [23]. A scrip tool[2] called "colleval.pl" is used to evaluate the performance of NER.

4.4 Results and Analysis

We use our model and all comparative methods to recognize various entities on two datasets. Tables 1 and 2 respectively show the results of recognizing person and

[1] https://www.findbestopensource.com/product/crfpp.

[2] http://www.cnts.ua.ac.be/conll2000/chunking/.

location on the Short Message corpus, while Table 3, 4 and 5 respectively show the results of recognizing person, location and organization on the People's Daily News. We analyze the results in the following.

Evaluation on Input Representation. In this paper, we both use the character-lever and word-level input for all models. From Table 1, we can find that all neural network models using character-level representation get higher F1 scores than those using word-level one, while systems using word-level achieve better results on location from Table 2. It maybe because that short messages, which tends to be spoken language and informal expressions, generally contains more noise such as nicknames and spelling mistakes. Chinese person entity is a loose internal structure that words or characters are independent of each other. Character-level avoids the interference of word meanings. Location, differing with person, has strong meanings and consists of two or more words. Although we design the encoder using BiLSTM to connect the context, the system with character-level still can't understand the interdependence and meanings of words. We can also observe that the systems using word-level perform better on both person and location for News (Tables 3 and 4). News is a kind of written language with formal expressions, which means it contains less wrong word segmentation than short messages. In other words, only accurate word segmentation is helpful to recognize entity. Similar to location, organization is usually composed of several practical significance words. However, this dependence between words becomes weaker and weaker as the length of organization increasing. This also explains why the systems with character-level outperform again on organization (Table 5).

Table 1. Proposed approaches on Short Messages for person.

Model	Character-level			Word-level		
	P	R	F1	P	R	F1
HMM	78.24	87.47	82.60	85.66	75.75	80.04
CRF	85.09	90.48	87.70	86.43	80.52	83.37
Seq2seq	92.32	92.32	92.32	89.43	88.18	88.80
Seq2seq+Attention	92.19	93.54	92.86	90.72	89.28	90.00
BiLSTM+CRF	93.99	94.60	**94.30***	91.91	90.58	**91.24***
Seq2seq+CRF	93.52	92.66	93.09	90.25	89.02	89.63
Seq2seq+Attention+CRF	92.74	94.90	93.80	92.75	89.62	91.16

Table 2. Proposed approaches on Short Messages for location.

Model	Character-level			Word-level		
	P	R	F1	P	R	F1
HMM	76.20	83.22	79.55	93.73	92.08	92.90
CRF	91.52	94.17	92.82	95.51	91.06	93.23
Seq2seq	93.54	93.12	93.33	95.95	95.89	95.92
Seq2seq+Attention	95.42	96.49	95.95	96.78	96.76	96.77
BiLSTM+CRF	96.29	96.52	96.41	97.18	96.98	97.08
Seq2seq+CRF	94.43	93.93	94.18	97.16	95.64	96.39
Seq2seq+Attention+CRF	96.41	96.55	**96.48***	97.00	97.24	**97.12***

Effectiveness of the CRF Layer. We find that CRF outperforms HMM both on character-level and word-level for recognizing all entities (From Tables 1, 2, 3, 4 and 5). CRF doesn't have the strict independence assumption of HMM so it can accommodate much context information, without its flexible features. Comparing Seq2seq +Attention model with Seq2seq+Attention+CRF model, we find that the performance with CRF is improved evidently. For example, the system (character-level) reached 94.19% F1 with CRF but only 92.70% F1 without CRF on location (Table 4). The reason is that CRF using both past and future labels avoids generating wrong tagging sequence to a great degree. Besides, it is easy to see that the improvement is more evident on character-level representation, because the models need to generate much more tags for input characters than words. So the systems with CRF using character-level can work better on sequence tagging problems.

Table 3. Proposed approaches on People's Daily News for person.

Model	Character-level			Word-level		
	P	R	F1	P	R	F1
HMM	68.45	69.76	69.10	94.35	88.81	91.49
CRF	89.43	89.84	89.63	96.10	86.71	91.16
Seq2seq	88.79	84.93	86.82	94.73	90.28	92.45
Seq2seq+Attention	95.35	94.60	94.97	96.70	96.07	96.38
BiLSTM+CRF	97.02	94.98	**95.99***	97.75	95.67	**96.70***
Seq2seq+CRF	92.46	88.98	90.68	95.23	92.05	93.61
Seq2seq+Attention+CRF	96.57	94.64	95.59	96.58	96.14	96.36

Table 4. Proposed approaches on People's Daily News for location.

Model	Character-level			Word-level		
	P	R	F1	P	R	F1
HMM	58.31	66.69	62.22	87.34	89.91	88.61
CRF	77.43	84.34	80.73	92.15	80.40	85.87
Seq2seq	82.24	77.76	79.94	92.56	89.20	90.85
Seq2seq+Attention	92.58	92.82	92.70	95.62	93.28	94.43
BiLSTM+CRF	93.41	93.32	93.36	95.82	93.48	94.64
Seq2seq+CRF	88.06	84.01	85.99	92.88	90.73	91.79
Seq2seq+Attention+CRF	94.29	94.09	**94.19***	94.95	94.68	**94.81***

Evaluation on Sequence Transformation Models. For both People's Daily News (Tables 3, 4 and 5) and Short Messages (Tables 1 and 2), the Seq2seq+Attention+CRF model outperform BiLSTM+CRF on location, but ranks only second to BiLSTM+CRF on person and organization. As we mentioned above, person entity generally contains less inside information while location contains much internal dependence between atomic words. Organization is usually longer than location, which may increase the accumulation of errors in decoding. According to above analysis, we conclude that the

Seq2seq with attention and CRF model is very good at recognizing short strong interdependent entity such as location. Even to other entities, such as person and organization, the performance of our model is much close to the best results (BiLSTM+CRF).

Table 5. Proposed approaches on People's Daily News for organization.

Model	Character-level			Word-level		
	P	R	F1	P	R	F1
HMM	47.99	50.86	49.38	77.56	81.86	79.65
CRF	50.22	71.43	58.98	80.42	84.11	82.22
Seq2seq	83.39	83.39	83.39	92.42	92.98	92.70
Seq2seq+Attention	91.79	93.68	92.72	92.47	93.16	92.81
BiLSTM+CRF	95.22	94.57	**94.89***	94.52	93.84	**94.18***
Seq2seq+CRF	91.79	87.01	89.34	92.82	90.09	91.44
Seq2seq+Attention+CRF	95.21	93.83	94.52	94.13	95.07	94.09

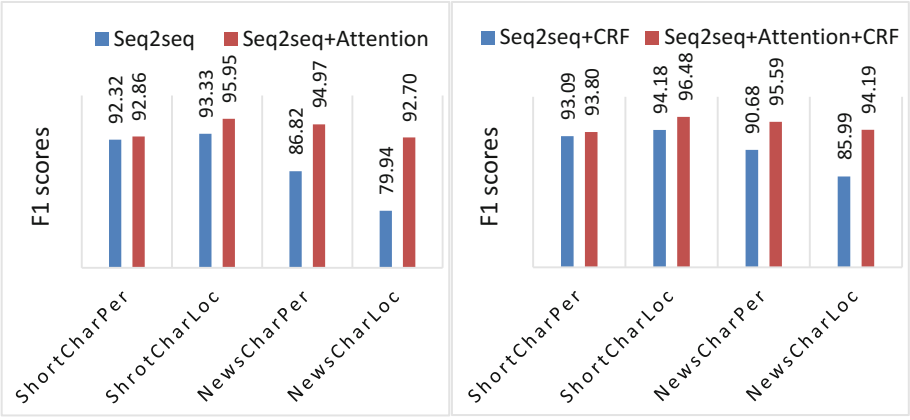

Fig. 3. The graphical illustrations of the attention mechanism using different corpus on F1 scores.

Effectiveness of the Attention Mechanism. To estimate the effectiveness of the attention mechanism, we compare the results using Seq2seq model and Seq2seq +Attention model, Seq2seq+CRF model and Seq2seq+Attention+CRF model on different corpus for person and location. The contrastive results on F1 score are shown in Fig. 3. By adding attention mechanism, the systems do really prompt the performance of NER. Besides, it is easily to see that the promotion is more obvious on News corpus. As we mentioned above, short messages includes more noise data because of its informal expression. Models using attention on informal text may gain wrong surrounding information during recognition compared with formal text sometimes. So we conclude that seq2seq can't capture enough contextual information very well which can be compensated by the attention, especially on formal text.

5 Conclusion

In this paper, we explored a sequence transformation framework for Chinese Named Entity. We also systematically compared the performance of different NER systems, showing that our model achieves a good performance especially on short interdependence entity.

There are still some problems need to be considered. Firstly, word segmentation information is helpful for formal text in NER but not for informal, which can be considered how to join word-level and character-level embedding in the further NER research. Secondly, our model is a supervised method relying on a large number corpus that is not suitable to small-labeling-data such as social media. So, it is necessary to study a semi-supervised framework or transfer learning framework for NER.

Acknowledgement. This work was supported by the National Key Research and Development program of China (No. 2016YFB0801300), the National Natural Science Foundation of China grants (No. 61602466).

References

1. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. **4**, 3104–3112 (2014)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. Computer Science (2014)
3. Hermann, K.M., Kočiský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., et al.: Teaching machines to read and comprehend, pp. 1693–1701 (2015)
4. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. **26**, 3111–3119 (2013)
5. Nallapati, R., Zhou, B., Santos, C.N.D., Gulcehre, C., Xiang, B.: Abstractive text summarization using sequence-to-sequence RNNs and beyond (2016)
6. Collbert, R., Weston, J., Bottou, L.: Natural language processing (almost) from scratch. J. Mach. Learn. Res. **12**, 2493–2537 (2011)
7. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition, pp. 260–270 (2016)
8. Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. arXiv:1603.01354v4 (2016)
9. 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
10. 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)
11. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
12. Paulus, R., Xiong, C., Socher, R.: A deep reinforced model for abstractive summarization. arXiv preprint arXiv:1705.04304v3 (2017)

13. Su, J., Su, J.: Named entity recognition using an HMM-based chunk tagger. In: Meeting on Association for Computational Linguistics, pp. 473–480. Association for Computational Linguistics (2002)
14. Borthwick, A.: A Maximum Entropy Approach to Named Entity Recognition. New York University (1999)
15. Hai, L.C., Ng, H.T.: Named entity recognition: a maximum entropy approach using global information. In: International Conference on Computational Linguistics, pp. 1–7. Association for Computational Linguistics (2002)
16. 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)
17. 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)
18. Liu, B., Lane, I.: Attention-based recurrent neural network models for joint intent detection and slot filling (2016)
19. Lafferty, J.D., Mccallum, A., Pereira, F.C.N.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Eighteenth International Conference on Machine Learning, vol. 3, pp. 282–289. Morgan Kaufmann Publishers Inc. (2001)
20. Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF models for sequence tagging. Computer Science (2015)
21. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. Computer Science. arXiv preprint arXiv:1412.6980 (2014)
22. Zaremba, W., Sutskever, I., Vinyals, O.: Recurrent neural network regularization (2014)
23. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C.: TensorFlow: large-scale machine learning on heterogeneous distributed systems (2016)