Research on Named Entity Recognition Method Based on Improved LSTM-CRF Model

Yong Gan¹,², Dongwei Jia¹,² and Yifan Wang¹,²

¹ Zhengzhou Institute of Engineering and Technology, Zhengzhou, Henan, China
² School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou, Henan, China

Email: ganyong@zzuli.edu.cn; 602914623@qq.com; 925053758@qq.com

Abstract. Because the computer cannot directly understand the text corpus in the NLP task, the first thing to do is to represent the characteristics of the natural language numerically, and the word vector technology provides a good way to express it. Because Word2vec considers context and has fewer dimensions, it is now more popular words embedded. However, due to the particularity of Chinese, word2vec cannot accurately identify the polysemy of words. In this paper, a lightweight and effective method is used to merge vocabulary into character representation. This approach avoids designing complex sequence modeling architectures. For any neural network model, simply fine-tuning the character input layer can introduce vocabulary information. The model also uses the modified LSTM to bridge the enormous LSTM and the Transformer model. The interaction between input and context provides a richer modeling space that significantly improves testing on all four public datasets.

Keywords. Named entity recognition; NLP; words embedded.

1. Introduction

Online Social Networking have a large number of users and information. Whenever there are significant events in the real world, the relevant information about these events will be spread to social media [1]. Therefore, it is of great value to mine the attribute information of social media events. Social media event attribute recognition is to extract information such as the time of occurrence, the person involved, the geographical location of the occurrence, the subject to which the event belongs through the analysis of the text data of the social network so as to obtain more abundant information of the event. By identifying the attribute information of the event, public opinion monitoring and public opinion analysis can be realized, which is convenient for public decision-making institutions to obtain information and deal with it in time. Entities mainly include names, place names, and institutional names, and the length of entity characters is not fixed. At the same time, named entity recognition is also the basic link to handle other natural language processing (Natural language processing, NLP) problems [2-6]. For example, the extracted target entity may be used to construct the information of the knowledge graph system, and the translation quality of the machine translation system can be improved by identifying the unlogged words. Naming entity recognition is a typical tagging problem, and the research method can also be applied to other sequence tagging tasks. Chinese named entity recognition also has its own characteristics [7-8]. First, there is no space between Chinese characters, and the result of participle and entity recognition will affect each other. Secondly, the ambiguity of the Chinese named entity is serious, and it is easy to produce different results by different entity boundary divisions. The complexity and variety of Chinese named entity recognition tasks lead to this problem
not being completely solved [9-11].

2. Research on Identification of Named Entities

The exploration of named entity recognition started early. After many years of research, scholars have got ideal results and have been applied in some related fields. They are beginning with the concept of named entity identification Rau [12] et al. in 1991, a new chapter has been opened in this research field. They found that an unknown word in a financial news report accounted for about 8% of the text. More than 4% are company names and some organizational names. A quarter of them is still unknown words. To solve this problem, Rau and others realized the algorithm of automatically extracting company names from news corpus. By testing the corpus of more than a million words that have collected thousands of company names. More than 95% accuracy, later this study was considered to be the predecessor of named entity recognition. Until the first official use of the term named Entity (Named Entity) in the MUC session held in 1995, and set up multi-language entity identification evaluation task. The research of named entity recognition based on machine learning has experienced the development of supervision, semi-supervision, unsupervised and mixed methods. The representative work includes: named entity recognition based on hidden Markov model, maximum entropy model (Maximum Entropy,) studied by John D. Burger et al. ME, Support Vector Machines (Support Vector Machine,); and SVM [13], conditional airport (Conditional Random Fields,); and CRF [14] and AdaBoost and other supervised machine learning methods to carry out research. The Nadeau semi-supervised learning method and Liu et al. will condition the airport (Conditional Random Fields,) under the semi-supervised learning framework CRF and k-Nearest Neighbor; and KNN) the hybrid model combined with classifier studies the task of named entity recognition and verifies that the hybrid method is superior to KNN and semi-supervised learning. In order to improve the effect of Chinese automatic word segmentation. As a result, the early Chinese NER mainly focused on the identification of individual entities such as names and institutional names.

As hardware resources are abundant and deep learning technology is developing rapidly. There are also many named entity recognition models based on deep learning and machine learning methods, such as long and short time memory (Long Short Term Memory, LSTM) and CRF named entity recognition model [14], and so on. Related studies have proved that domain knowledge and artificial features in traditional named entity recognition methods can greatly improve the indicators of entity recognition, but it takes a lot of work to acquire domain knowledge and design its artificial features. It costs a lot of time and energy, and portability is also very general. Therefore, the researchers of named entity recognition have begun to use the natural language processing tools as auxiliary tools to study the recognition pathways of named entities in other natural language fields. Therefore, this paper uses the modified LSTM and discriminant model CRF combined with the word embedding of words and characters to identify the Chinese named entities for compare the experimental results of the same test data and the same experimental environment.

3. Related Technology

3.1. LSTM Layer

Long and short-term memory networks (usually referred to as “LSTM”) are a special kind of RNN. They were proposed by Hochreiter & Schmidhuber (1997) [15]. They have been improved and promoted by many people because of their long-term dependence on learning. They have performed well on a wide range of issues and are now widely used. LSTM is a special kind of RNN [16]. The disappearance or explosion of the gradient in the back propagation stage of RNN is the key to limiting its performance. LSTM can better solve the long dependence problem of sequences and reduce the loss of information compared with conventional RNN [17].

The LSTM model is mainly realized by introducing a gating mechanism consisting of five different parts: unit state, hidden state, input gate, forgetting gate, and output gate. The input gate determines how much current input will be sent to the unit state, and the forgetting gate determines how much
the previous unit state will be sent to the current unit state. The output gate determines how many unit states will be output to the hidden state. The model structure is shown in figure 1 below.

![LSTM model diagram](image)

Figure 1. LSTM model diagram.

We can see from the diagram that at t time LSTM, the input of the unit consists of three parts, which are the memory units of the previous unit. The hidden layer of the previous unit and input layer, hidden layer, and memory units constitute the output of the unit at t time. The calculation flow of the hidden layer value is as follows: first, the information output of the input gate, the output gate, and the forgotten gate is calculated, the information output of the three gates is to optimize the information of the memory unit, and then the information in the memory unit is calculated. Specific calculations such as formula:

\[
\begin{align*}
    f_t &= \text{sigmoid}(W_f \cdot x_t + U_f \cdot h_{t-1} + b_f) \\
    i_t &= \text{sigmoid}(W_i \cdot x_t + U_i \cdot h_{t-1} + b_i) \\
    o_t &= \text{sigmoid}(W_o \cdot x_t + U_o \cdot h_{t-1} + b_o) \\
    \tilde{C}_t &= \tanh(W_c \cdot x_t + U_c \cdot h_{t-1} + b_c) \\
    C_t &= f_t \ast C_{t-1} + i_t \ast \tilde{C}_t \\
    h_t &= o_t \ast \tanh(C_t)
\end{align*}
\]

Tanh is the hyperbolic tangent activation function, and the weight matrix of the forgetting gate is \(W_f, U_f\). The weight matrix of the input gate is \(W_i, U_i\). The weight matrix of the output gate is \(W_o, U_o\). The bias terms are \(b_f, b_i, b_o\). \(C_t\) Represents the state of the memory unit. \(\tilde{C}_t\) state of t moment. The main function is to update the current state of the moment, \(h_t\) is the output of the t moment, \(x_t\) input for the current moment.

Although the LSTM model theoretically solves the problem that the cyclic neural network cannot be used because of the existence of gradient disappearance in the actual experiment. We will still find that the network model of LSTM can only use the previous historical information without considering the influence of the later text on the previous text. In addition, for the whole sequence annotation task, the result is inaccurate if the context information is not fully used to predict. Some scholars have proposed a MOGRIFIER LSTM model for the problems existing in the LSTM model. The use of LSTM is generally considered to alleviate the problems of gradient disappearance and information forgetting so as to better model the long-distance semantics. Note, however, that in the LSTM the current input is \(x_t\) and previous states \(h_{t-1}\) are independent of each other, they interact only in the gate, and there is no interaction before that. Theoretically, the current input should be related to the hidden state of the previous step, but the LSTM only calculates the two and obtains the output of each gate. This may lead to the disappearance of the information before and after, thus losing the accuracy.
of the prediction. To this end, Mogrifier does not change the structure of the LSTM itself. Let the input and state first interact, expect to enhance the ability of context modeling.

The main method of Mogrifier LSTM is to alternate between $x_t$ and previous states $h_{t-1}$ before ordinary LSTM calculations, specifically:

$$x^i = 2\sigma(Q^i h^{i-1}_{prev}) \odot x^{i-2} \quad \text{for odd } i \in [1 \ldots r]$$

$$h^i_{prev} = 2\sigma(R^i x^{i-2}_{prev}) \odot h^{i-2}_{prev} \quad \text{for even } i \in [1 \ldots r]$$

The $i$ is the number of iterative rounds, the operation of the first formula is performed when the $i$ is odd, and the operation of the second formula is performed when the $i$ is even. $i=0$, the whole model degenerates into the original LSTM, and then multiplies by a constant 2, because after the sigmoid operation, its value is distributed in $(0,1)$, so that the value will be smaller and smaller. So multiply by a 2 to ensure the stability of its value. After the iteration is finished, it enters the traditional LSTM operation.

3.2. Conditional Random Field

Deep neural networks have remarkable capabilities, such as the ability to capture information from a distance, it is considered that there is a strong dependence between adjacent labels. The introduction of CRF can handle the problem of label relationship.

When LSTM and Softmax are classified, only the context of features can be learned, and the last label cannot be learned, but conditional random field can be used to learn the context of label L. So that’s the prediction of the tag. The premise of this model is that random variables form Markov random fields. The most classic method to solve the sequence label problem was put forward by Lafferty in 2011 [18]. In general, use $z = \{z_1, \ldots, z_n\}$ refers to the sequence of embedded layer input characters, $z_i$ represents the vector of the i input word, $y = \{y_1, \ldots, y_n\}$ represents the label sequence of input $z$. $Y(z)$ representing $z$ Possible label sequences. It is calculated by the following formula:

$$p(y | z; W, b) = \frac{\prod_{y_1=1}^{n} \phi(y_{i+1}, y_i, z)}{\sum_{y_1=1}^{n} \prod_{i=1}^{n} \phi(y_{i+1}, y_i, z)}$$

where $\phi(y', y, z) = \exp(W_{y', y, z}^T + b_{y', z})$ is potential function, $W_{y', y, z}$ and $b_{y', z}$ to the label $(y', y)$ Weight vector and bias value.

For CRF model training, maximum likelihood estimation is used. For the training set $\{(z_i, y_i)\}$, the likelihood estimation algorithm is given as follows:

$$L(W, b) = \sum \log p(y | x; W, b)$$

On the premise that a linear chain condition random field $\lambda$, input text sequence $x$ and all possible candidate annotation result sets $Y$ for sequence annotation task, Expected optimal label results $y^*$ can be calculated quickly and effectively by using the witerbi algorithm.

$$y^* = \arg \max_{y \in \lambda} p(y | x; \lambda)$$

3.3. Embedded Layer

Disadvantages of character-based named entity recognition is that word information is not fully utilized. In order to handle this drawback, some scholars recently proposed to incorporate word vocabulary into the character-based NER model. Furthermore, when characters match multiple words in the dictionary. This article lists and saves all the words that match the character. The weight calculated by the word frequency allows the subsequent NER model to decide which word to apply
instead of randomly selecting a word for the character. In order to realize this idea, the embedding layer of the LSTM-CRF model is modified in detail. New benchmark results are obtained on multiple open Chinese NER datasets. However, the Lattice-LSTM model architecture is quite complex. In order to introduce dictionary information, some researchers add some extra edges between input non-adjacent characters, which obviously slows down the efficiency of the model. Furthermore, it is difficult to transfer Lattice-LSTM structures to other neural network structures that are more suitable for certain specific tasks (such as convolutional neural networks and transformers).

To this end, this article uses character representation to encode dictionary information and designs as many encoding schemes as possible for dictionary matching results. Use the frequency of each character as the initial value of its weight. Since the word frequency is a constant, it can be completed during the preprocessing, so that it will not affect the efficiency of the modified LSTM, and it can also greatly speed up the calculation of the weight of each word as shown in figure 2 below.

Specifically, \( z(w) \) represents the frequency of lexical \( w \) in statistical data, and the weighted representation of lexical \( S \) is as follows:

\[
v^*(S) = \sum_{w \in S} w \cdot e^*(w)
\]

\[
Z = \sum_{w \in \text{BUMERUS}} z(w)
\]

Here the weight of all the words in the four-word sets is normalized for the overall comparison.

4. Experimental and Results Analysis

4.1. Experimental Environment
All experiments in this paper have model parameters of word vector dimension (char-dim) of 50, word vector dimension (word-dim) is 100. The number of neurons is set to 50. To solve the gradient explosion problem, clip gradients were adopted to set the gradient threshold clip to 5. To prevent overfitting, a Dropout regularization mechanism was introduced. A probability of setting the Dropout layer at 0.5. To improve model training, Optimizing the way the model updates the weights and deviation parameters. Adam algorithm is used to optimize the objective function, the initial learning rate is 0.002, attenuation rate (lr-decay) is 0.8. An epoch of 100 per cent was trained.

4.2. Data Sets and Labels
The methods were evaluated on four Chinese NER datasets, including Onto-Notes, MSRA, Compare
the model. The BIOES marking method is used instead of the BIO2 to mark the entity. Since the use of BIOES markers is better than BIO2, it can more clearly divide the boundary of the entity. The name and meaning are shown in table 1. B means that the word is at the beginning of an entity (Begin), I mean internal (inside), O means external (outside), E means that the word is at the end of an entity, S means that the word itself can form an entity (Single).

| Name   | Meaning              |
|--------|----------------------|
| B-LOC  | Begin of local name  |
| I-LOC  | Inside of local name |
| E-LOC  | End of local name    |
| B-ORG  | Begin of organization|
| I-ORG  | Inside of organization|
| E-ORG  | End of organization  |
| S      | Single               |
| O      | Outside              |

4.3. Evaluating Indicator
This paper adopts the correct rate P, recall rate R and F1 value as the evaluation indexes of named entity identification, which are the basic indexes to evaluate the experimental results in the fields of information retrieval, classification, recognition and so on. They are defined as follows:

(1) Accuracy rate is the proportion of the number of correctly retrieved entities to all retrieved entities.

\[ P = \frac{TP + TN}{TP + FN + FP + TN} \times 100\% \]

(2) Recall rate is the proportion of the number of all correctly retrieved entities to the number of entities that should be retrieved.

\[ R = \frac{TP}{TP + FN} \times 100\% \]

(3) F1 value is the harmonic mean of accuracy rate and recall rate, which represents a comprehensive consideration of accuracy rate and recall rate.

\[ F_1 = \frac{2PR}{p + R} \times 100\% \]

In the formula, TP refers to the correct prediction as a positive class, FN refers to the wrong prediction as a negative class, FP refers to the wrong prediction as a positive class, and TN refers to the correct prediction as a negative class.

4.4. Experimental Results and Analysis
In order to prove the validity of the named entity recognition model, the following experiments are carried out:

Experiment 1: Compare the performance indicators of different models in MSRA. As shown in table 2.

Experiment 2: Compare the various performance indicators of different models in OntoNotes. As shown in table 3.

It can be seen from the table that the performance of the method proposed in this paper is significantly better than the traditional model on the two data sets, and both have achieved good results.
Table 2. Performance on MSRA.

| Model          | P/%  | R/%  | F1/% |
|----------------|------|------|------|
| CRF            | 69.70| 66.12| 67.86|
| BiLSTM         | 72.69| 71.02| 71.84|
| BiLSTM+CRF     | 91.86| 88.75| 90.28|
| Lattice-LSTM   | 93.57| 92.79| 93.18|
| SoftLexicon(MLSTM) | **94.20** | **93.40** | **93.79** |

Table 3. Performance on OntoNotes.

| Model          | P/%  | R/%  | F1/% |
|----------------|------|------|------|
| CRF            | 65.59| 71.84| 68.57|
| BiLSTM         | 77.71| 72.51| 75.02|
| BiLSTM+CRF     | 76.43| 72.32| 74.32|
| Lattice-LSTM   | 76.35| 71.56| 73.88|
| SoftLexicon(MLSTM) | **84.32** | **83.12** | **83.71** |

5. Conclusion
This paper adopts the model based on the assemble of modified LSTM and discriminant model CRF and introduces word embedding and character to identify Chinese named entities. It can be directly applied to different types of entity tagging. Through the multi-group comparison experiments on the data set, the results of the system are higher than those of the traditional model. In the following work, we can further improve the model, explore multi-task learning methods, and combine more useful information with different fields. Although conditional random fields can learn the potential relationship between sequence labels, it greatly reduces the speed of the model. In the future, we will try to optimize the model with faster and more accurate decoding methods and apply the model to other sequence tagging problems.

Acknowledgments
This work is supported by the Nation Nature Science Foundation of China (NSFC), (NO. 61572445).

References
[1] Greenwood S, Perrin A, & Duggan M. 2016. Social media update 2016. Pew Research Center:11(2), 1-18.
[2] Le Q and Mikolov T 2014 Distributed representations of sentences and documents International Conference on Machine Learning pp 1188-1196.
[3] Peng M, Ma R, Zhang Q, et al. 2019 Simplify the usage of Lexicon in Chinese NER arXiv preprint arXiv:1908.05969.
[4] Li X, et al. 2020 FLAT: Chinese NER using flat-lattice transformer arXiv preprint arXiv:2004.11795.
[5] Ding R, Xie P, Zhang X, Lu W, Li L and Si L 2019 A neural multi-digraph model for Chinese NER with gazetteers Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics pp 1462-1467.
[6] Zhang Y and Yang J 2018 Chinese NER using lattice LSTM Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL) pp 1554-1564.
[7] Chen Y, Zong C and Su K Y 2010 On jointly recognizing and aligning bilingual named entities Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (Association for Computational Linguistics) pp 631-639.
[8] Mayhew S, Tsai C T and Dan R 2017 Cheap translation for cross-lingual named entity recognition Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing pp 2536-2545.

[9] Zhou J T, Zhang H, Jin D, et al. 2019 Dual adversarial neural transfer for low-resource named entity recognition Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics pp 3461-3471.

[10] Ma X and Hovy E 2016 End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics pp 1064-1074.

[11] Hochreiter S and Schmidhuber J 1997 Long short-term memory Neural Computation 1735-1780.

[12] Rau, Lisa F. 1991 Extracting company names from text. Proceedings The Seventh IEEE Conference on Artificial Intelligence Application. IEEE Computer Society pp 29-32.

[13] Gui T, Zou Y, Zhang Q, Feng M, Fu J, Wei Z and Huang X J 2019 A lexicon-based graph neural network for chinese NER Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) pp 1040-1050.

[14] Lample G, Ballesteros M, Subramanian S, et al. 2016 Neural architectures for named entity recognition Proceedings of NAACL-HLT pp 260-270.

[15] Chiu J P and Nichols E 2016 Named entity recognition with bidirectional LSTM-CNNs Transactions of the Association for Computational Linguistics 4 357-370

[16] Zhao D Y, Huang J M and Jia Y 2018 Chinese name entity recognition using Highway-LSTM-CRF Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence pp 1-5.

[17] Huang Z, Xu W and Yu K 2015 Bidirectional LSTM-CRF models for sequence tagging arXiv preprint arXiv:1508.01991.

[18] Yu K, Lin Y and Lafferty J 2011 Learning image representations from the pixel level via hierarchical sparse coding IEEE CVPR 2011 pp 1713-1720.