BiLSTM-CRF Chinese Named Entity Recognition Model with Attention Mechanism

Zhongbao Wan, Jie Xie, Wei Zhang and Zhaohua Huang

Department of Software Engineering, East China Jiaotong University, Faculty of Software, room 304, 808 Shuanggang East Street, Nanchang, Jiangxi, China
Email: 1528619012@qq.com

Abstract. In order to make up for the weakness of insufficient considering dependency of the input char sequence in the deep learning method of Chinese named entity recognition task, this paper proposes a method, which integrate Bidirectional Long Short-Term Memory (BiLSTM), attention mechanism and add the information of word vector. Firstly, the proposed model obtains the char vector feature extracted from the text corpus, which is then input to the BiLSTM model; Secondly, the attention mechanism is used to calculate the relevance between the current input char and the other input char of the BiLSTM model; Finally, the global feature is obtained according to relevance, concatenating the word vector feature, which is introduced to the Conditional random field (CRF) layer to perform the mutual constraint between tags. Thus, the classified result can be obtained. Based on the corpus of the Chinese Peoples’ Daily Newspaper in 1998, our experiments show that the proposed method can improved the performance and efficiency of named entity recognition, compared to the existing deep-learning method that combines word vector and char vector.

1. Introduction

Named entity recognition [1] is a basic and important task in natural language processing. It refers to extracting words or phrases from unstructured text. In earlier times named entity recognition task is adopted rules-based and dictionary-based methods [2]. The formulation of rules requires different experts in different fields to develop, and its reusability is poor. Based on statistical machine learning methods [3], people need to select features artificially, although improving the versatility of the method. With the enhancement of hardware computing power and the accumulation of data, the deep learning method [4] has made great progress, can obtain features automatically by large-scale training, reduce labor costs and improve extraction performance. An improved bidirectional long short memory conditional random field (Bi-LSTM-CRF) model was used for named entity recognition, combining with word and char information in the process of training which has achieved good results [5]. The Attention Mechanism [6] refers to the fact that due to the limited computing resources, it is very resource-consuming to record all the information in the entire sequence of input. The model emphasizes the influence of important inputs on the output by calculating the attention probability distribution. In the field of biomedicine a method which combines with the attention mechanism and bidirectional recurrent neural network to recognize English drug names was proposed, using the document level rather than the sentence level attention mechanism [2], in order to solve the problem of inconsistent prediction for tags in the same word. However, the efficiency of Chinese named entity recognition is not high. Because Chinese characters are hundreds of times larger than English characters, and computing resources are expensive.

Aiming at the problem that the attention mechanism of the document level in the literature [2] is not efficient and the accuracy of Chinese named entity recognition is low, this paper proposes a
sentence-level attention mechanism. For the difference between English and Chinese, adding the word vector increases the semantic information of the sentence. It can effectively solves the problem that the long short-term memory model does not consider the dependence between the input char sequences and the inability to capture the internal structure information of the sentence. Experiments show that the method can improve the training efficiency and shorten the training time.

2. Model

2.1 Model Input
Since the distributed representation of words [8] was proposed, more scholars have embedded pre-trained word embeddings into the tasks of natural language processing. The word embeddings of distributed representation has lower dimension and more dense vector. It can more accurately express the syntactic and semantic information. This paper selects Sogou news corpus of 1.89G, trains word embedding by using CBOW model, the trained word embedding is recorded as $word$. The char embedding feature contains the information of character’s radicals. All the input characters are stored in the dictionary of the lookup layer, and the lookup table is obtained in the training process by the deep neural network model. The input sequence maps each character into a vector corresponding the lookup table. The resulting character vector is recorded as $char$.

2.2 Bidirectional Long Short-term Memory Model
Due to the recurrent neural network (RNN) cannot solve the long-distance dependence problem, Hochreiter proposed the LSTM model, using the gate mechanism to improve. The LSTM model is composed of many LSTM units, each of which has a forget, input, output gate and a memory cell. As shown in figure 1, given the input vector $x_i$, the output vector $h_t$ at this moment is obtained. The detailed calculation process is shown in literature [5]. The bidirectional LSTM can leverage the information of the past and future moment for better performance. In BiLSTM layer, a forward LSTM computes a representation $\tilde{h}_t$ from left to right at every char embedding and backward LSTM computes a representation $\tilde{h}_t$ in reverse. $h_t=[\tilde{h}_t:\tilde{h}_t]$ is obtained by concatenating its left and right output of the hidden layer states, which can get the context information of the sentence.

$$e_i = \tanh(W_e h_i + b_e)$$

In the $\tanh$ layer, A matrix $e_i$ is obtained by the calculation of equation (1). Where $W_e$ refers to the weight, $b_e$ represents the bias vector.

2.3 CRF Layer
The conditional random field (CRF) is added to the model in order to add the constraint relationship between the labels, to ensure that the predicted labels are valid and to find an optimal label sequence.

$$s(X, Y) = \sum_{i=1}^{n} (T_{y_{i-1}, y_i} + e_{i, y_i})$$

$$p(Y | X) = \frac{e^{s(X, Y)}}{\sum_{\tilde{Y}} e^{s(X, \tilde{Y})}}$$

$$y^* = \arg\max_{\tilde{y}} s(X, \tilde{Y})$$

The output matrix $e_i$ is obtained by the neural network. The label transfer matrix is $T$. Given a
sentence $X = (x_1, x_2, \ldots, x_n)$, the output label sequence is $Y = (y_1, y_2, \ldots, y_n)$, $s(X, Y)$ is the score of $Y$, as calculated by equation (2). As shown in equation (3), The softmax function is used to normalize all possible output label sequences to obtain the probability of the label sequence $Y$. As in equation (4), when predicting labels the Viterbi algorithm is used to calculate the optimal sequence.

2.4 BiLSTM Model with Attention Mechanism
Integrating the attention mechanism can learn the correlation between character sequences and capture the dependencies and structures inside the sentences. We regard the correlation between the current input character and other input characters as the attention probability. The network structure that integrates the attention mechanism is shown in figure 1. In attention mechanism layer, a given attention matrix $A$ is calculated from the correlation between the current target output character and other characters in the input sentence. The matrix $A$ element $\alpha_{i,j}$ is the attention probability, which represents the relevance of vector $x_i$ of the $i^{th}$ char in the current output and vector $x_j$ of the $j^{th}$ char in a sentence.

$$\alpha_{i,j} = \frac{\exp(score(x_i, x_j))}{\sum_n \exp(score(x_i, x_n))}$$  \hspace{1cm} (5)

The function $score(x_i, x_j)$ computes the relevance between vectorized representation $x_i$ of the $i^{th}$ char and $x_j$ of the $j^{th}$ char in a sentence, it calculated through Euclidean distance.

$$score(x_i, x_j) = W_a(x_i - x_j)^T(x_i - x_j)$$  \hspace{1cm} (6)

The global vector $g_t$ is obtained by weighted summation of the output state values $h_j$. Then output $z_t$ of the attention mechanism layer is obtained through the $tanh$ activation function. The output $f_t$ is obtained by equation (9), then is transmitted to the CRF layer.

$$g_t = \sum_{j=1}^{n} \alpha_{i,j} h_j$$ \hspace{1cm} (7)

$$z_t = tanh(W_g[g_t; h_j])$$ \hspace{1cm} (8)

$$f_t = tanh(W_z[z_t; x_{word}])$$ \hspace{1cm} (9)

3. Experimental Results and Analysis

3.1 Experimental Datasets and the Evaluation Metrics
The experiment is done on a server with 16G memory and Intel Xeon L5630 processor. It used Linux operating system and TensorFlow 1.2.1. The corpus is the Chinese Peoples’ Daily Newspaper in 1998.

| Dataset   | Sentences number | characters number | words number(MB) |
|-----------|------------------|-------------------|------------------|
| Training set | 41053            | 1606361           | 12.7             |
| test set      | 4303             | 164247            | 1.3              |

This model is BIEO annotation mode, the person, location, organization name are marked as PER, LOC and ORG. The mark O represents the other label. The experiment uses precision (P), recall (R), F-score(F) as evaluation metrics of model.
Figure 1. BiLSTM-CRF model with attention mechanism structure

3.2 Parameter Setting

In order to ensure the char dimension of the experiment, this paper performs a comparative experiment. The experimental results are shown in figure 2. The F score trained by the model increases with the character vector dimension increases, which means that increasing the dimension can improve the model extraction performance, however increasing the char vector dimension consumes memory. Therefore, the final char vector dimension is determined to be 300 for subsequent comparative experiments.

The number of hidden nodes is set to 300. The training epoches is set to 40. The word vector dimension is set to 100. The learning rate, the window size and batch size are set to 0.001, 5, 64.

3.3 Experimental Result Analysis

In this paper, five sets of experiments were conducted to named entity recognition. The first experiment used the CRF model to obtain the results using the CRF++ tool. The second experiment only uses the char vector as input. The third experiment is to add the word vector feature. The fourth experiment is to add attention mechanism based on the first experiment, the fifth experiment is to combine char, word vector and attention mechanism to compare the performance.
Table 2. The comparison results of char, word vector and integrate attention mechanism

| Task method       | F-score | precision | recall |
|-------------------|---------|-----------|--------|
| Location name     |         |           |        |
| CRF model         | 89.74   | 93.17     | 86.57  |
| Char              | 88.33   | 88.52     | 88.15  |
| Char+word         | 92.76   | 93.87     | 91.69  |
| Char+ attention   | 90.46   | 91.03     | 89.91  |
| Char+word+ attention | 93.85  | 94.89     | 92.85  |
| Person name       |         |           |        |
| CRF model         | 89.50   | 93.02     | 86.24  |
| Char              | 90.04   | 91.30     | 88.82  |
| Char+word         | 92.85   | 93.86     | 91.88  |
| Char+ attention   | 91.47   | 92.63     | 90.35  |
| Char+word+ attention | 93.59  | 94.88     | 92.35  |
| Organization name |         |           |        |
| CRF model         | 89.66   | 92.23     | 87.24  |
| Char              | 87.05   | 87.18     | 86.94  |
| Char+word         | 92.97   | 93.45     | 92.50  |
| Char+ attention   | 88.92   | 89.24     | 88.61  |
| Char+word+ attention | 93.73  | 94.05     | 93.43  |
| total             |         |           |        |
| CRF model         | 90.21   | 90.55     | 89.87  |
| Char              | 87.74   | 89.45     | 86.09  |
| Char+word         | 92.06   | 93.41     | 90.74  |
| Char+ attention   | 89.86   | 90.46     | 89.26  |
| Char+word+ attention | 94.37  | 94.26     | 94.48  |

According to the comparison results in Table 2, it can be seen that compared with the traditional machine learning model, the deep learning method is improved. Relative to the recognition of person name, the attention mechanism has greatly improved the recognition of the organization name. This is mainly because the organization name contains more words and more semantic information. For the recognition of person name, the effect of adding the attention mechanism to extract the information is not as obvious as the improvement of the organization name. Because the name of the person in the training corpus is shorter and contains less semantic information. For the location name, the effect of adding attention mechanism is between the person name and the organization name. The reason is that in the training corpus, some location names are shorter, some location names are longer, and the attention mechanism is more effective for longer words. In the last four comparative experiments, the suffix information of the words has a positive influence on improving the recognition performance and the CRF layer is added to increase the constraint between the labels, ensuring that the predicted label is valid; Where two experiments add word vectors can found that the recognition effect of named entity is obvious compared to only use char vector. The word vector contains rich contextual semantic information. The experiments found that adding the word vector is better than adding the attention mechanism, but adding the attention mechanism also has improved the effect of named entity recognition and the training efficiency will be improved. It can be seen from figure 3 that as the number of training epoches increases, the F value increases steadily, which indicates that the recognition effect of the model is stable. The vertical axis of figure 3 starts from 40 and the curve between 0 and 40 is omitted. The experiments found that the model which integrates char, word vector with attention mechanism has the highest F value, followed by the model which integrates char, word vector. The model with attention mechanism is faster than model without attention mechanism, the convergence speed is faster and reaches the maximum value faster. The experiment found that attention mechanism can improve efficiency, shorten training time.
4. Conclusion

In this paper the long short-term memory model with attention mechanism is introduced for the Chinese named entity recognition task. The top level of the model uses the conditional random field to increase the constraint between the labels. The experiment compares the effect of char vector, word vector and attention mechanism in Chinese named entity recognition, which finds that F score can be improved after adding attention mechanism and it can improve efficiency. Since the independent Chinese characters still can't express the meaning of the word well, the next step is to consider the information of the radicals of the characters to improve the effect of recognition.

5. References

[1] Nadeau D, Sekine S. A survey of named entity recognition and classification[J]. Lingvisticae Investigationes, 2007, 30(1):3-26.
[2] Chiticariu L, Krishnamurthy R, Li Y, et al. Domain adaptation of rule-based annotators for named-entity recognition tasks[C]/Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2010:1002-1012.
[3] Su J. Named entity recognition using an HMM-based chunk tagger[J]. Proc Acl, 2002:473-480.
[4] Chiu J P C, Nichols E. Named Entity Recognition with Bidirectional LSTM-CNNs[J]. Transactions of the Association for Computational Linguistics, 2016(4):357-370.
[5] Lample G, Ballesteros M, Subramanian S, et al. Neural Architectures for Named Entity Recognition[C]/ Conference of the North American Chapter of the Association for Computer Linguistics:Human Language Technologies. 2016:260-270.
[6] Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need[C]//Advances in Neural Information Processing Systems. 2017:6000-6010.
[7] Luo L, Yang Z, Yang P, et al. An Attention-based BiLSTM-CRF Approach to Document-level Chemical Named Entity Recognition [J]. Bioinformatics, 2017, 34(8):1381-1388.
[8] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]. arXiv Preprint arXiv:1301.3781, 2013.