Chinese Named Entity Recognition with Bert

Cheng Gong, Jiuyang Tang, Shengwei Zhou, Zepeng Hao and Jun Wang

ABSTRACT

As a basic task of NLP, named entity recognition has always been the focus of researchers. At the same time, the word vector representation which is a necessary part of many named entity recognition neural network models has been more and more important. Recently, the emergence of a new type word representation, BERT, has greatly promoted many NLP tasks. In this paper, we will use Bert to train Chinese character embedding and connect it with Chinese radical-level representations, and put it into the BGRU-CRF model. We have achieved good results in Chinese data set through a series of experiments.

KEYWORDS

Named entity recognition, Chinese, BERT, BGRU-CRF.

INTRODUCTION

The term of named entity was used in MUC-6 firstly, and the focus at the time was on information extraction, which extracted the structure information of corporate activities and defense-related activities from unstructured text such as newspapers, while the names of people, places, organizations, time, and numbers (including time, date, amount of money, and percentages) are key elements of structured information.

Named Entities Recognition (NER) aims to recognize the boundaries and categories of these entities. It mainly focuses on the identification methods of three types of proper nouns such as person name, location name and organization name.

In recent years, with the development of neural networks, the study on named entity recognition has changed from early methods which are based on lexicon and rules to the methods which is deep learning-based in recent years, and has a good performance on the neural network.

As an important part of all deep learning-based natural language processing model, word embedding undoubtedly occupies a very important proportion in named entity recognition tasks. They encode words and sentences in fixed-length dense vectors to greatly improve the ability of neural networks to process text data.

Developed by Google and announced by the end of October 2018, the BERT [1] has achieved outstanding performance in 11 NLP tasks. BERT (Bidirectional Encoder Representation from Transformers) is known as the best word embedding. BERT can further increase the generalization ability of the word vector model, and fully describe the character level, word level, sentence level and even sentence relationship characteristics.
In this paper, we use BLSTM to capture radical-level information which are inner features inside Chinese characters and bring additional information just like Dong[2] did. We connect the embedding from radicals to the embedding from Bert as the inputs to the neural network. We will use BIGRU+CRF to output the sequence label, which has less parameters and faster training speed, and also can get better results.

RELATED WORK

In the previous research, the named entity recognition task was studied as a sequence labeling problem. In the machine learning-based method, the NER common model includes the generated model HMM[3-4], the discriminant model CRF[5], etc. With the rise of the neural network, Hammerton[6] Using a forward LSTM model to solve this problem, this is the first time using neural networks to deal with named entity recognition problems. Subsequently (Lample et al.[7]; Chiu and Nichols[8]; Ma and Hovy[9]; Liu et al.[10]) and others have proposed the current mainstream English NER method, neural networks and conditional random field combination models, while they also take the character information into account in the word representation.

Unlike the English, the Chinese has no obvious boundaries. In particular, named entity boundaries are also word boundaries in English. So generally, the word segmentation process is followed by sequence tagging in Chinese. However, incorrect segmentation can lead to NER errors, which are particularly prominent in cross-domain segmentation (Liu and Zhang[11]; Jiang et al.[12]). Therefore, for Chinese named entity recognition, the word-based method is better than the word-based method. Dong C et al.[2] proposed to use LSTM to capture the morphological features of the Chinese character in the Chinese NER task by constructing a word from the radicals and spliced into the word vector before inputting it into the LSTM. Yue Zhang et al.[13] designed the lattice LSTM model used for Chinese named entity recognition, which encodes the input character sequence and the potential vocabulary of all matching lexicons. Compared to character-based methods, the model makes explicit use of word order information. The lattice LSTM does not have a word segmentation error compared to the word-based approach.

But one of the drawbacks of character-based NER is that it cannot make full use of explicit word order information, which is very useful for Chinese NER. To solve the above problems, we apply BERT to NER task. BERT can make full use of context semantic information and it is based on character training which doesn’t need word segmentation, so it can improve the quality of Chinese NER better.

MODEL

BERT

The BERT model follows the structure of GPT model and uses transformer encoder as the main model structure. Transformer abandons RNN circular network structure and models a text entirely based on attention mechanism.
The main idea of the attention mechanism used by Transformer is to calculate the relationship between each word in a sentence and all the words in the sentence. It is believed that the relationship between these words and words reflects the relevance and importance of different words in the sentence to a certain extent. Therefore, we can use these relationships to adjust the importance of each word to obtain a new expression of each word. This new representation not only implies the connotation of the word itself, but also implies the relationship between other words and the word, so it is a more global expression than a simple word vector.

In this thesis, we use the BERT-Base Chinese model provided by Google to improve the task. We input the character embedding generated by the last layer of Bert into the neural network.

CHINESE RADICAL INFORMATION

Similar to English prefixes and suffixes, in Chinese, radicals are the most basic units of Chinese characters. Radicals store the characteristics of Chinese characters, and it also has certain semantic information. For instance, the characters “ni”(you), “ta”(he), and “men”(people) all have the meanings in regard to human because of their all shared radical “danrenpang”(human), a variant of Chinese character “ren”(human).

To some extent, this kind of semantic information can make the characters with the same radicals more similar in vector space.

To get the radical information, We consult the Xinhua Dictionary to get radical compositions of many Chinese characters. For a character, we just replace it with its radical part. We can establish a radical table of all the character. So we can regard the radicals of a character as a sequence in writing order. We can use BLSTM to obtain the information of Chinese radical for the sequence.
GRU

Recurrent Neural Networks are a type of neural network commonly used in deep learning, it use sequence information and maintain this information through the middle layer, which makes it unique in processing sequence data.

The GRU used in this article is a variant of RNN, it is similar to the LSTM, it solves long-term dependency problems and remembers longer contextual information. The basic structure of the GRU consists of an update gate and a reset gate.

![Figure 2. Gated recurrent unit.](image)

Among them, \( z_t \) is an update gate, which determines how much information time enters the next time. \( r_t \) is a reset gate, which determines how much information is discarded, and the two parameters jointly determine the \( h_t \). The GRU has only two doors, so the structure is simpler than LSTM, it has fewer parameters, trains faster and takes less time. We use the following implementation.

\[
\begin{align*}
    r_t &= \sigma(W_r \cdot [h_{t-1}, s_t]) \\
    z_t &= \sigma(W_z \cdot [h_{t-1}, s_t]) \\
    \tilde{h}_t &= \tanh(W_{\tilde{h}} \cdot [r \odot h_{t-1}, s_t]) \\
    h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t
\end{align*}
\]  

(1)

where \( \sigma \) is the element-wise sigmoid function, \( h_t \) represent hidden layer vector at time \( t \), and \( W \) is the weight matrix.

BGRU-CRF

When the neural network is used alone for named entity recognition, the pre-trained word vector is input, the neural network outputs the predicted score of each label, and the label with the highest score of each character is selected as the label of the unit. However, the relationship between characters and characters cannot be determined. The output layer cannot fully utilize the label information of the header entity and the tail entity, but the CRF layer can obtain constraint rules from the training data, so that the probability of an illegal sequence appearing in the prediction of the tag sequence is greatly reduced, so the CRF layer can improve the recognition performance of the model.
At the same time, in order to make full use of the context information of each word, a Bidirectional GRU structure is used, which takes into account the sequence information in both the forward and backward of the sentence. We add CRF layer to the hidden layer of the bidirectional GRU model to restrict the output of each time, and get the bidirectional GRU-CRF model. We connect Bert embedding and radical embedding in series as the input embedding. The network structure is detailed in Figure 3.

![Figure 3. Main architecture of BGRU-CRF.](image)

For the sequence of inputs:

\[ A = (a_1, a_2, a_3, \ldots, a_n) \]  

We consider \( P \) to be the scores matrix which is output by the bidirectional GRU network. And the transition scores matrix \( Q \) is the parameter of CRF layer. For a sequence of predictions:

\[ B = (b_1, b_2, b_3, \ldots, b_n) \]

We define its score to be

\[ s(A, B) = \sum_{i=1}^{n} \left( Q_{b_i, s_i} + P_{s_i, s_{i+1}} \right) \]

The dimension of \( P \) is \( n \times k \), where \( k \) represents the number of all possible tag types, and the element \( P_{s_i, s_{i+1}} \) represents the score of the \( j \)th tag of the \( i \)th word in a sentence. And \( A \) corresponds to the transition score from the tag \( i \) to tag \( j \).

A softmax layer over all possible tag sequences yields a probability for the sequence \( B \):

\[ p(g|X) = \frac{e^{r(X,g)}}{\sum_{g \in G_x} e^{r(X,g)}} \]
In the training process, we need to maximize the logarithmic probability of the correct tag sequence:

\[
\log(p(b|A)) = s(A,b) - \log \left( \sum_{b \in B_A} e^{s(A,b)} \right)
\]

\[
= s(A,b) - \log \text{adds}(A,\tilde{b})
\]

(6)

Where \( B_A \) represents all possible tag sequences of sentence \( A \). So what we should do is to max the score function:

\[
b = \arg \max_{\tilde{b} \in B_A} s(A,\tilde{b})
\]

(7)

**TRAINING**

For the models presented, we set up a dropout which is 0.5 before input to GRU layer to avoid over-fitting, and observed that the result of the model can be improved to some extent. To update the parameters, we train our network with back propagation algorithm on each training example, we use stochastic gradient descent (SGD) algorithm, and the learning rate is 0.05 on training set. We set the character embedding size to 100, our model only uses a single layer for the forward and backward GRUs, and we set its dimensions to 200.

**EXPERIMENTS**

**TAGGING SCHEME AND EVALUATING INDICATOR**

The task of Chinese named entity recognition aims to distribute a entity tag to each Chinese character in a sentence, usually, sentences are also represented the IOB format which is the inside, outside, beginning. To consider more information, we use IOBES tagging scheme in our experiments.

In the evaluation of results, the evaluation results of this study are based on the matching between annotation results and actual results. We use Precision, Recall and F1-score to evaluate the performance of the model.

Take person name Recognition as an Example, the positive class represents the predicted result of the word is the person’s name, while the negative class indicates that the predicted result of the word is not the person's name. The true number of person's names in the positive class is TP; the number of non-person's names in the positive class is FP; the number of non-person's names in the negative class is TN; and the number of errors in the negative class is TN; the number of names predicted as negative classes is FN. The formulas for calculating Precision \( P \), Recall \( R \) and F1-score are as follows:

\[
P = \frac{TP}{TP + FP}
\]

(8)

\[
R = \frac{TP}{TP + FN}
\]

(9)
\[ F = \frac{P \times R \times 2}{P + R} \] \hspace{1cm} (10)

RESULTS

We make experiments on MSRA data set and People's daily data set they all include three types of named entities: locations, persons, organizations. The best score of F1 in two data sets are respectively 96.21% and 95.42% during 100 epochs.

In addition, we replaced BERT embedding with the embedding trained by using Word2Vec which is the mainstream embedding method now. Table I shows the score of F1 increase from 90.45% to 95.42% by using BERT. With an increase of nearly 5%, it shows that BERT can promote the task of Chinese named entity recognition very well.

![Table I. Comparison Word2Vec on MSRA.](image)

| Models                                  | F1  |
|-----------------------------------------|-----|
| Word2Vec+radical+BGRU-CRF               | 90.45 |
| BERT+radical=BGRU-CRF                  | 95.42 |

Table II presents our comparisons with other models on MSRA data set, it shows Our research achieves state-of-the-art performance with 95.42.

![Table II. Main Result on MSRA.](image)

| Models                  | P    | R    | F1   |
|-------------------------|------|------|------|
| Chen et al. (2006a)     | 91.22| 81.71| 86.20|
| Zhang et al. (2006)     | 92.20| 90.18| 91.18|
| Zhou et al. (2013)      | 91.86| 88.75| 90.28|
| Lu et al. (2016)        | –    | –    | 87.94|
| Dong et al. (2016)      | 91.28| 90.62| 90.95|
| zhang et al. (2018)     | 93.57| 92.79| 93.18|
| BERT+radical+BGRU-CRF   | 95.26| 95.57| 95.42|

![Table III. Comparison of Evaluation Indicators.](image)

|         | P    | R    | F1   |
|---------|------|------|------|
| PER     | 98.51| 97.62| 98.06|
| LOC     | 96.21| 95.83| 96.02|
| ORG     | 91.13| 93.65| 92.37|

Table III compares three indicators of person name, location name and organization name respectively. Compared with the other two indicators, the recognition rate of people's names is higher, because there are more unknown
words in people's names, and Bert plays a very important role in the recognition of unknown words.

At the same time, we have trained LSTM and GRU separately. GRU has relatively simple structure and fewer parameters, so its training time is about 15% less than LSTM. This is why we choose GRU.

CONCLUSION

This paper presents a neural network model for Chinese named entity recognition and achieves more advanced results. We use GRU blocks to learn long distance dependencies which has shorter training time and fewer parameters. At the same time, we use BERT learning context semantic information to construct character embedding. Compared with other embedding methods, BERT can handle the problem of unknown words better. At the same time, we add Chinese radical information to the character embedding, which can make better use of inner Semantic Information. Our model does not include any external information or hand-crafted features. It is a complete end-to-end model and can be applied to other domains.

REFERENCES

1. Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.
2. Dong C, Zhang J, Zong C, et al. Character-Based LSTM-CRF with Radical-Level Features for Chinese Named Entity Recognition[M]// Natural Language Understanding and Intelligent Applications. Springer International Publishing, 2016.
3. Bikel D M, Miller S, Schwartz R, et al. Nymble: a high-performance learning name-finder[C]// Proceedings of the Fifth Conference on Applied Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 1997: 194-201.
4. Bikel D M, Schwartz R, Weischedel R M. An algorithm that learns what's in a name[J]. Machine Learning, 1999, 34(1-3): 211-231.
5. McCallum A, Li W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons[C]// Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL, Stroudsburg: Association for Computational Linguistics, 2003, 4: 188-191.
6. Hammerton J. Named Entity Recognition with Long Short-Term Memory[C]// Conference on Natural Language Learning at Hlt-naacl. Association for Computational Linguistics, 2003.
7. Lample G, Ballesteros M, Subramanian S, et al. Neural Architectures for Named Entity Recognition[J]. 2016.
8. Chiu J P C, Nichols E. Named Entity Recognition with Bidirectional LSTM-CNNs[J]. Computer Science, 2015.
9. Ma X, Hovy E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF[J]. 2016.
10. Chen X, Qiu X, Zhu C, et al. Long short-term memory neural networks for chinese word segmentation[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015: 1197-1206.
11. Liu Y, Zhang Y. Unsupervised domain adaptation for joint segmentation and POS-tagging[J]. Proceedings of COLING 2012: Posters, 2012: 745-754.
12. Jiang W, Sun M, Lü Y, et al. Discriminative learning with natural annotations: Word segmentation as a case study[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2013, 1: 761-769.
13. Zhang Y, Yang J. Chinese NER Using Lattice LSTM[J]. 2018.