Research on Methods for Complex Chinese Entity Recognition

Ren Mengxing, Ma Lei*, Tian Ye and Qin Hongchen
Institute of Automation Chinese Academy of Sciences

*Email: lei.ma@ia.ac.cn

Abstract. Named entity recognition (NER), as a key task of natural language processing, is of great significance in relational capture, information retrieval, knowledge mapping, machine translation, and question and answer systems. In the neural network-based Chinese named entity recognition method, the vectorized word representation only maps words into one vector, which cannot represent the above problem. In order to solve these problems, this paper proposes a Cascading model based on the character vectorization. This method uses BERT to represent the ambiguity of the word. In the agile BIGRU model, the initial entity recall is completed to increase the entity recall rate. Then the output of this layer is input into the high-level bilstm model of self-attention mechanism to further filter and improve the accuracy of the final recognition. Finally, the CRF is used to complete the output of the final recognition result. Experiments show that the model can effectively improve the accuracy, accuracy and recall rate of complex Chinese entity recognition. The f1 value has increased by about 2% overall.

1. Introduction
NER predicts the role entities in the text, which can be decomposed into two subtasks, entity boundary determination and entity category division, and can identify more categories of entities according to business needs. In 2018, Yang Pei et al.[1] introduced the attention mechanism on the basis of BiLSTM-CRF model. The attention mechanism can obtain the context representation of words in the full text. This model is applied to the task of chemical drug entity recognition[2]. By pre-training word vectors and character-level LSTM on biological texts, 90.77% of F1 is finally obtained. He and Wang, Liu, Li and others[3] compared the statistical methods of character level and word level. The research shows that the method of character-based named entity recognition generally has better performance. Xu et al.[4] used joint training of word segmentation and named entity recognition to fuse word segmentation information. The Lattice LSTM network structure proposed by Zhang et al.[5] has the best effect. This method improves the traditional LSTM unit to grid LSTM, makes explicit use of word and word order information on the basis of word model, and avoids the problem of wrong transmission of word segmentation. It achieves 93.18% F1 value in MSRA corpus. However, the problems in the above method of Chinese named entity recognition based on words are that it cannot represent the polysemy of words. In a sentence, the same word has different meanings, but in the usual method of word vector representation, the vector representation of three words is exactly the same, which is inconsistent with objective facts and cannot contain the meaning of words very well. Therefore, researchers have proposed the use of pre-training language model for word representation. In the BERT model proposed by Devlin et al.[6][7] in 2018, the Bidirectional Transformer network structure with stronger semantic ability is used to pre-train the language model. In view of the powerful
ambiguity of the BERT pre-training language model, this paper introduces the BERT pre-training language model on the Chinese named entity recognition task. Based on this, a BERT-BiGRU-BILSTM-CRF-selfAttention layered network structure model is proposed.

2. Overall Cascade Model
The cascade model structure of BERT-BiGRU-BILSTM-CRF-selfAttention is shown in Figure 1. The whole model is divided into four parts. Firstly, the semantic representation of input is obtained by BERT pre-training language model, and then the vector representation of each word in the text is obtained. Then, the word vector sequence is input into the low-level network BiGRU, and the recall rate is given priority. Then the output results are input into the high-level BILSTM model fused with self-attention mechanism to further filter and improve the accuracy of the final recognition. Finally, the final recognition results are output through CRF.

2.1. BERT model structure
In order to fuse the context on the left and right sides of the word, BERT adopts two-way Transformer as the encoder[8]. The model also innovative put forward the "Masked language model" and "the next sentence" two tasks and capture the word level and sentence level respectively, said the joint training. Masked language model is to train the depth two-way language vector, this method uses a very direct way, covering some words in the sentence, let the encoder to predict the word of the original words, and then randomly cover 15% of the words as the training sample, the model structure diagram as shown in figure 2.

The most important part of BERT is the bidirectional Transformer encoding structure. Transformer abandons the circular network structure of RNN and models a piece of text completely based on attention mechanism. And the most important module of coding unit is the self-attention part, such as formula (1), where \( Q, K \) and \( V \) are input vector matrices, and \( d_t \) is input vector dimension. New expressions of each word can be obtained by adjusting the importance (weight) of each word by using these relationships. This new representation implies not only the word itself, but also the relationship between other words and the word, so it is a more global expression than mere word vectors.

\[
\langle Q, K, V \rangle = \text{soft max} \left( \frac{QK^T}{\sqrt{d_t}} \right)V
\]  

(1)
In order to extend the ability of the model to focus on different locations and increase the "presentation subspace" of attention units, Transformer has adopted a "multi-head" mode, such as formula (2)(3).

\[
(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_n)W^o
\]

\[
\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)
\]

In addition, to address degradation in deep learning, Transformer has incorporated residual networks and layer normalization in its coding units, such as the following formula(4)(5).

\[
\text{LN}(x_i) = \alpha \times \frac{x_i - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}} + \beta
\]

\[
\text{FFN} = \max(0, xW_1 + b_1)W_2 + b_2
\]

One of the most important features in natural language processing is timing feature. To solve the problem that timing feature cannot be extracted by self-attention mechanism, Transformer adopts the method of location embedding to add timing information, as shown in formula (6) and (7). BERT’s input is the sum of word embedding, position embedding and type embedding.

\[
\text{PE}(pos, 2i) = \sin\left(pos / 10000^{2i/d_{\text{dim}}}ight)
\]

\[
\text{PE}(pos, 2i + 1) = \cos\left(pos / 10000^{2i/d_{\text{dim}}}ight)
\]

2.2. BIGRU layer

Recurrent Neural Network referred to as “RNN”, mainly for data modeling and the model of sequence, but the basic RNN cannot solve the problem of long series data dependence, GRU, LSTM as variations of RNN, is mainly used to solve the problem of long series data dependence[9], and compared with LSTM GRU helped structure more simple, less parameters, can speed up the training time. And BIGRU can better capture two-way semantic dependencies. Therefore, for complex Chinese named entity recognition, the advantages of BIGRU such as fast speed can be used to preliminarily complete rough entity recognition, in which GRU combines forgetting gate and input gate into a single update gate, while mixing cell state and hidden state. Its unit structure is shown in the figure3.

Figure 3. GRU coding Unit.

Figure 4. BiLSTM Coded Sentences.

The specific calculation process is shown in the following formula:

\[
z_i = \sigma(W_z * [h_{i-1}, x_i])
\]

\[
r_i = \sigma(W_r * [h_{i-1}, x_i])
\]

\[
h_i = \tanh(W_c * [r_i \odot h_{i-1}, x_i])
\]

\[
h_i = (1 - z_i) \odot c_{i-1} + z_i \odot h_i
\]
Where $\sigma$ is sigmoid function, $\bullet$ is the dot product, $\chi_t$ is the input vector at time $t$, $h_t$ is the hidden state, and is also the output vector, contains all valid information at time $t$ before, $z_t$ is an update gate, controls information flow to the next time, $r_t$ is a reset gate, controls information loss. Together, they determine the output of the hidden state.

2.3. BILSTM layer
LSTM is a special RNN, which can learn long-term dependence. The key of LSTM is cell state[10]. It can add or delete messages to cell state through a structure called gate, but LSTM can not encode information from back to front. BILSTM is a combination of forward LSTM and backward LSTM. It is usually used to model context information and capture bidirectional semantic dependencies. Therefore, BILSTM is used to further process entity recognition. For example, the sentence "对空警戒雷达的维修人员" is coded, and the model is shown in Figure 4.

Forward LSTM inputs "Air defense radar", "of" and "Repair personnel" in turn to get three vectors $[h_{1}, h_{2}, h_{3}]$. Backward LSTM inputs "Repair personnel", "of", "Air defense radar" get three vectors $[h_{1}, h_{2}, h_{3}]$. Finally, the forward and backward hidden vectors are joined together to form $[h_{1}, h_{2}, h_{3}, h_{1}, h_{2}, h_{3}]$.

2.4. The CRF layer
In the named entity recognition, some labels can not be continuously appear, therefore, the model cannot be used independent of $h(t)$ to do the label decision, CRF can by considering the tag between the adjacent relations to obtain the global optimal sequence, so use CRF to modeling sequence tags, CRF for a given sequence and the corresponding sequence tags define evaluation score formula (12).

$$s(x, y) = \sum_{i=1}^{n} \left( W_{i-1,i} + P_{i} x_{i} \right)$$

(12)

where, $W$ is the transformation matrix, $W_{i-1,i}$ represent the tag transfer fraction, $P_{i}$ represent the fraction of the $y_{i}$ the tag of this character, and $P_{i}$ is defined as formula (13).

$$P_{i} = W_{i} h_{t} + b_{i}$$

(13)

where, $h_{t}$ is the hidden state of input data $x_{i}$ at time $t$ of the previous layer, and the parameters are weight matrix and parameters respectively. Maximum conditional likelihood estimation is used for training CRF. For training set $\{(x_{i}, y_{i})\}$, its likelihood function is formula (14).

$$L = \sum_{i=1}^{n} \log \left( P(y_{i} | x_{i}) \right) + \frac{\lambda}{2} \| \theta \|^2$$

(14)

where, $P$ is shown in formula (15), denoting the corresponding probability from the original sequence to the predicted sequence.

$$P(y | x) = \frac{e^{s(x, y)}}{\sum_{y \in Y} e^{s(x, y)}}$$

(15)

2.5 Attention mechanism
Attention is an Attention mechanism in the decoding process[11], you only need to pay Attention to a certain part of the sentence when decoding, and you add this Attention mechanism when decoding, so in the encoding process, you don't need to encode all the information in the original sentence into a vector of fixed length. If Attention is not added, a BILSTM is used to read the whole sentence to get a vector of fixed length, namely vector $W$, which is the encoding process. Then another BILSTM takes vector $W$ as the starting point to generate each sentence in the target successively, which is the
decoding process. The disadvantage of this approach is that no matter how long the previous sentence is, it will eventually be compressed into a vector of several hundred dimensions. This means that the longer the sentence is, the more information will be lost in the resulting vector \( W \). The experimental results also indicate that the longer the input sentence is, the worse the final effect will be. After adding Attention, Attention generated the output at step \( i \) and threw \( y_i \) away. It only paid Attention to the hidden state \( h \) related to \( y_i \), as shown in the following formula:

\[
a_i = \text{Attend}(s_{i-1}, a_{i-1}, h)
\]

\[
g_i = \sum_{j=1} a_{ij} h_j
\]

\[
y_i = \text{Generate}(s_{i-1}, g_i)
\]

\( S_{i-1} \) is the state of the circulatory neural network at step \( i - 1 \), called the generator, \( a_{i \in R} \) is the weight vector of the Attention mechanism, \( g_t \) is called glimpse, and the state of the circulatory neural network at the next moment is:

\[
s_i = \text{Recurrency}(s_{i-1}, g_i, y_i)
\]

3. Experiment

3.1. Experimental environment

The experimental hardware and software environment configuration is shown in table 1:

| Project          | Environment                  |
|------------------|------------------------------|
| System           | Ubuntu 18.04.1 LTS           |
| GPU              | NVIDIA GeForce GTX 1080 Ti  |
| The hard disk    | 2T                           |
| Memory           | 32G                          |
| Python           | python 3.6.5                 |
| Tensorflow       | TensorFlow 1.14.0            |

3.2. Experimental Corpus and Labeling

This paper uses MSRA data set and Military information knowledge text, which includes three types of entities: Location, Weaponry and Person, and contains more than 50,000 statements. In this experiment, BIO mode is used to realize data annotation, B means entity start, I means entity middle and entity end, O means not entity, such as table 2 for specific entity category annotation.

| Entity Class      | Mark Code |
|-------------------|-----------|
| Person Name       | B-PER     |
| Location Name     | B-LOC     |
| Weaponry Name     | B-WEA     |
| Non-Entity        | O         |

3.3. Model evaluation index and result analysis

3.3.1. BERT model structure. Accuracy, precision, recall and f-measure are adopted as evaluation criteria in this paper. The specific definitions are as follows:
\[ \text{Precision} = \frac{T_p}{T_p + F_p} \]  
\[ \text{Recall} = \frac{T_p}{T_p + F_r} \]  
\[ F_1 = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \]

Where, True Entities (TE): Identify the correct number of entities, True Recognize (TR): Number of identified entities, True Sample (TS): The number of entities in the sample, accuracy represents the proportion of correctly identified named entities to the total number of identified named entities. Recall rate represents the ratio of the number of correctly identified named entities to the total number of named entities in the corpus, that is, the number of correctly identified entities. In order to improve the recognition effect and reduce the accidental impact of the selection of recognition methods on recall rate and accuracy rate, a comprehensive balance should be made between recall rate and accuracy rate, and the weighted average value of the two should be taken as F1 value.

3.3.2. Comparative analysis of Experimental result

Table 3. Experimental results.

| Model Name                               | Accuracy | Precision | Recall  | F1    | Loss     |
|------------------------------------------|----------|-----------|---------|-------|----------|
| BIGRU-CRF                                | 94.36%   | 89.57%    | 84.89%  | 87.17%| 1.671224 |
| Cascaded HMM                             | 92.39%   | 85.12%    | 85.83%  | 81.40%| 1.892614 |
| BERT-BIGRU-BILSTM-CRF-selfAttention      | 99.36%   | 93.56%    | 95.30%  | 94.42%| 1.471224 |

Its parameters are configured as train_batch_size = 32, eval_batch_size=4, predict_batch_size=4, num_train_epochs=16.0, learning_rate=5e-5. F1 value, accuracy rate and recall rate are used as the evaluation indexes for the experimental results, which are shown in Table 3, including the recognition rate, recall rate and F1 for person name, place name and organization name. As a whole, the entity recognition effect based on bert-bigru-bilstm-crf-selfattention model is the best, with the accuracy rate of 99.36%, accuracy rate of 93.56, recall rate of 95.30%, and F1 value of 94.42, It is better than the other two entity recognition models.

Table 4. Experimental results based on BIGRU-CRF model.

| Classification | Precision | Recall | F1   |
|----------------|-----------|--------|------|
| LOC            | 92.13%    | 90.49% | 91.30%|
| WEA            | 85.04%    | 83.94% | 84.49%|
| PER            | 95.13%    | 93.80% | 94.46%|

Table 5. Experimental results based on Cascaded HMM model.

| Classification | Precision | Recall | F1   |
|----------------|-----------|--------|------|
| LOC            | 92.41%    | 87.99% | 90.15%|
| WEA            | 87.40%    | 81.62% | 84.41%|
| PER            | 95.60%    | 94.57% | 95.08%|

Table 6. Experimental results based on BERT-BIGRU-BILSTM-CRF-selfAttention mode.

| Classification | Precision | Recall | F1   |
|----------------|-----------|--------|------|
| LOC            | 95.41%    | 95.99% | 95.70%|
| ORG            | 91.04%    | 92.94% | 91.98%|
| PER            | 98.13%    | 97.80% | 97.96%|

For the names of people, places and weaponry, through the recognition effect of three models in Table 4, 5 and 6, in general, the entity recognition effect of bert-bigru-bilstm-crf-selfattention model is significantly better than the other two models. The recognition effect of complex weaponry names in
this model is not as good as the recognition effect of place names and people names, and the recognition accuracy rate for people names reaches 98.13%, The recall rate also reached 97.80%.

4. Conclusions
In this paper, Since the improved word vector-based cascade model can efficiently extract the relevant features of words, for long and complex Chinese named entities, the cascade model first adopts a low-level neural network to initially label the ship's knowledge target entities, and secondly through fusion attention The high-level neural network of the force mechanism corrects the problematic sequence labeling. Finally, the conditional random field is used to consider the dependency between the labels, and the global optimal label sequence is obtained, and then the boundary of the named entity is accurately located.

However, this paper still has some shortcomings, such as: Due to fewer characters vector information is restricted at present Chinese named entity recognition model effect is the key, so information on how to rich character vector is also key points of future research, later will try more methods, such as model based on word on the subject of method to enrich distributed according to the amount of information, increase the ability of character vector said.

References
[1] python Core Techniques and Algorithms in Natural Language Processing Practice.
[2] RauL F.Extracting Company Names from Text[C]. In: Proceedings of the 7th IEEE Conference on Artificial Intelligence Applications.1991:29-32.
[3] Grishman R, Sundheim B.Message Understanding Conference-6: A Brief History[C]. In: Proceedings of the 16th International Conference on Computational Linguistics. 1996.
[4] Hammerton J.Named Entity Recognition with Long Short-Term Memory[C]/ Conference on Natural Language Learning at Hlt-naacl. Association for Computational Linguistics,2003.
[5] Ma X, Hovy E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF[J]. 2016.
[6] 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.
[7] Collobert R, Weston J, Bottou L, et al. Natural Language Processing (Almost) from Scratch[J]. Journal of Machine Learning Research, 2011.
[8] Santos C N D, Guimarães,Victor. Boosting Named Entity Recognition with Neural Character Embeddings[J].Computer Science, 2015.
[9] Strubell E,Verga P, Belanger D, et al. Fast and accurate entity recognition with iterated dilated convolutions[J]. arXiv preprint arXiv:1702.02098, 2017.
[10] Chinese Named Entity Recognition Method Based on BERT Embedding.
[11] Li Lishuang, Guo Yuankai. Biomedical Named Entity Recognition Based on CNN-BLSTM-CRF Model [J]. Chinese Journal of Information, 2018.