An End-to-End Named Entity Recognition Model for Chinese

Cheng Gong¹,*, Jiuyang Tang¹ and Zhen Li¹

¹Science and Technology on Information Systems Engineering Laboratory, National University of Defence Technology, Changsha, China, 410072.
*gcclgl95@126.com

Abstract. Named entity recognition is an important basic task in natural language processing. This paper proposes a named entity recognition method for end-to-end and efficient deep-loop neural networks. Using BERT to train sub-vectors as raw input enables the model to obtain more comprehensive text information, and at the same time, the BiLSTM network is focused on the attention mechanism, so that the network pays more attention to the key information in the text and ignores the redundant information to improve the recognition efficiency of the model. Finally, the relationship between any two tags is captured by the CRF layer, and the entire sentence is decoded and predicted. Experiments show that the method performs well on MSRA corpus.

1. Introduction

With the rapid update of the Internet, the traditional search method cannot meet the user's demand for information and knowledge. The main reason behind this is that the current search engine and other systems can only provide users with fragmented knowledge, such as a separate URL, a piece of text, a picture, or a video, they do not provide the connection between different knowledge, it is difficult to build a complete knowledge system for users. To this end, a knowledge map led by Google knowledge graph came into being.

In the process of building knowledge maps, in addition to structured knowledge, a large amount of knowledge is hidden in unstructured text. Named Entity Recognition (NER), refers to the process of automatically identifying the naming and referencing items of an entity from unstructured text and indicating its category, which is an urgent problem to be solved. A named entity is a research subject identified by a named entity. It generally consists of three major categories (entity class, time class, and numeric class) and seven subclasses (person name, place name, institution name, time, percentage, currency, and data). Judging whether a named entity is correctly identified includes two aspects: whether the boundary of the entity is correctly classified; whether the Entity type is correctly labeled.

In order to reduce the dependence on linguistic knowledge and construct complex feature engineering, the research of named entity recognition has been extended from traditional statistical methods to deep learning methods. By constructing neural networks to learn the potential relationships within the text to extract relevant entity references. Different from English named entity recognition, Chinese named entity recognition has the following nodus:

- Compared with English entities, Chinese entities lack obvious boundaries and inherent definite articles, and also there is no obvious morphological spelling change in Chinese.
- The rapid development of the Internet has led to the emergence of many emerging entities, making identification more difficult.
- Chinese entities do not have a fixed format, it will appear a lot of noise when identifying.
Therefore, how to extract more effective features in Chinese text in a complex language environment has important research value. In this regard, this paper proposes to add a self-attention mechanism based on the existing BiLSTM+CRF sequence labeling model, which can reduce the dependence on external information and capture the internal correlation of data or features, at the same time, this paper also uses BERT [1] to train the word vector as the original input, which has a certain improvement on the recognition result.

2. Related Work
In the early NER research, it was a mainstream method to manually construct finite rules and then search for strings that match these rules from the text. The most representative of these is the DL-CoTrain method proposed by Collins et al. [2], which pre-defines the seed rule set Decision List, and then performs unsupervised training iterations on the set according to the corpus to get more rules and the final rule set, the method have a classification accuracy of more than 91% for three categories of named entities (person, place, and institution names). Although the rule-based method can obtain a better recognition effect on a specific corpus, it requires a large number of rules, and it is too feasible to manually formulate these rules.

With the rise of machine learning in the NLP field, NER research has gradually turned to machine learning. Classic machine learning classification models such as HMM [3]-[4], ME [5], CRF [6], and SVM [7] have been successfully used to serialize annotations of named entities, and have achieved good results. In recent years, people's attention has shifted from the field of machine learning to the field of deep learning based on neural network models. Lample et al. [8] proposed LSTM and two neural network models based on transformation. At the same time, the features were obtained from the labeled corpus and the unlabeled corpus, and the best NER effects were obtained in all four languages. At present, the study of entity extraction based on neural network structure mainly focuses on using the attention mechanism to improve the model effect and model construction for a small number of labeled training data. Based on the RNN+CRF model structure, Rei et al. [9] use the attention mechanism to improve the splicing process of word vector and character vector, so that the model can dynamically use word vector and character vector information. In this paper, we add the attention mechanism to the LSTM output hidden layer.

3. Model
In this paper, the named entity problem is solved as a sequence labeling problem. The model is mainly divided into the following layers: word embedding layer, BiLSTM layer, Multi-headattention layer, CRF layer.

3.1. The word embedding layer
Compared with English, Chinese text lacks a separator like a space character to indicate the boundary identifier of the word boundary. Therefore, Chinese named entity recognition tasks are often accompanied by Chinese word segmentation, but incorrect word segmentation will lead to errors in subsequent named entity recognition tasks. This will have an important impact on the efficiency of Chinese named entity recognition, and the use of words as the basic unit will lose the displayed vocabulary information, which will also have an impact on the efficiency of the task.

Therefore, this paper uses the BERT [1] model to train Chinese text, generate dynamic word representation to improve the efficiency of entity recognition, the BERT model is released by google, it follows the structure of GPT model, using Transformer encoder as the main model structure, it can continually overlaps the attentional mechanism layer and the normal nonlinear layer to obtain the final textual representation after we input the text.

This paper uses the BERT-BaseChinese model provided by Google to get the dynamic word vector, and input it into the neural network as an embedded layer.
3.2. BiLSTM layer

In recent years, LSTM (Long Short-Term Memory) network has been widely used in Long sequence dependent learning. The traditional RNN network has the problem of gradient disappearance, and it is difficult to process long sequence data, but LSTM network can solve this problem well. The main cell structure is shown as figure 1:

![Figure 1. The main cell structure of LSTM](image)

As can be seen from the above figure, there is an additional hidden state at each sequence index position $t$. This hidden state is generally called cell state ($C_t$). At the same time, compared to the RNN network, the LSTM network adds three door structures at each sequence index position, namely the forgetting gate, the input gate and the output gate.

- Input gate: Determine what new information is stored in the cell state ($C_t$);
- Forgotten gate: Decide which information to discard from ($C_t$);
- Output gate: Decide which information to output.

The formula for this network is as follows:

$$
\begin{align*}
    i_t &= \sigma(W_i h_{t-1} + U_i x_t + b_i) \\
    \overline{C}_t &= \tanh(W_c h_{t-1} + U_c x_t + b_c) \\
    C_t &= f_t \otimes C_{t-1} + i_t \otimes \overline{C}_t \\
    f_t &= \sigma(W_f h_{t-1} + U_f x_t + b_f) \\
    o_t &= \sigma(W_o h_{t-1} + U_o x_t + b_o) \\
    h_t &= o_t \otimes \tanh(C_t)
\end{align*}
$$

(1)

In the formula: $X_t$ represents the input at time $t$; $W_i, W_c, W_f, W_o, U_i, U_c, U_f, U_o$ respectively represent weight matrix; $b_i, b_f, b_c, b_o$ represent offset items; $\otimes$ is a dot product; $\sigma$ is the Sigmoid activation function. One-way LSTM can only use one side of text information, so this article uses bidirectional LSTMs to take advantage of contextual information. At the same time, this article connects the output of all previous layers and the initial input of the neural network as the input of the next layer. This connection method can maximize the information flow, and there will be no problems such as gradient disappearance and gradient explosion with the increase of neural network layers.

3.3. Self-attention layer

Although BiLSTM can fully consider the global information, it can not highlight the role of local key information. For example, “Mayor Li Ming inspected the school”, in the name entity recognition, the word “school” is of little significance to identify “Li Ming”. The term "mayor" is important for identification. Therefore, this paper adds attention mechanism in the BiLSTM hidden layer, which
makes the model focus on finding related information, removing redundant and redundant information, and configuring different weights to improve the quality of named entity identification.

Attention mechanism is a selection mechanism used to allocate limited information processing capabilities. It is characterized by selective attention to certain important information and correspondingly ignoring other information received at the same time. The performance corresponding to text processing is to assign higher weights to important words and smaller weights to other words. The nature of the attention function \( attention(Q, K, V) \) can be described as a query to a series of key values. The calculation of the attention mechanism is given below:

\[
attention(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right) V
\]  

(2)

Figure 2. Flow chart of self-attention mechanism

The attention layer essentially encodes the sequence \( Q \) of \( mn \times d_k \) into a new sequence of \( n \times d_V \). The self-attention mechanism is a special case in the attention mechanism which is shown in figure 2. It only needs one sequence to calculate its representation. So the \( Q \), \( K \) and \( V \) are the same sequence, \( attention(X, X, X) \), where \( X \) is the input sequence. The self-attention mechanism is characterized by ignoring the distance between each word, directly calculating its dependencies, learning the internal structure of the sentence, and obtaining long-term dependence.

3.4. CRF layer

The existence of the CRF layer is based on the fact that the tags are not isolated from each other, but have a certain correlation. The CRF layer can find the optimal output sequence while considering the dependencies between consecutive tags, and also prevent the occurrence of an illegal sequence. The final label prediction result consists of two parts, one is the output network \( P \) from the previous network, and the second is the transition probability \( T \) between the labels output by the CRF layer. The score of the sequence is:

\[
S = \sum_{t=1}^{n} P_{Y_t \mid Y_{t-1}} + \sum_{t=0}^{n} T_{Y_t \mid Y_{t+1}}
\]  

(3)

Where \( T_{Y_t \mid Y_{t+1}} \) is the transfer feature matrix used to store the probability of transition between all tags

3.5. Overall model

The model proposed in this paper is shown in the following figure 3. It is mainly divided into four parts, embedded module, BiLSTM module, Self-attention module, CRF module. The input is the BERT word vector trained by BERT-Base Chinese model provided by Google.
For more efficient identification, this article uses multiple layers of BiLSTM, we input the word vector as the original input to the first layer of BiLSTM, and splicing the output of the first layer of BiLSTM with the original input as the input of the second layer. According to this method, after multiple layers of extraction, the final output is a deep meaning, syntactic and semantic feature representation, and then use the zoom point product attention to learn the structural features inside the sentence, capturing long-term dependencies:

$$ S = Attention(h, h, h) $$ (4)

Finally, the feature vector is input to the CRF layer to obtain a prediction label.

4. Experimental design and results analysis

4.1. The data set
The data set used in this paper is SIGHAN 2006 Bakeoff - 3 Chinese evaluation data set MSRA, which mainly contains three types of entities, namely person name, place name, organization name, and its specific parameters are included in table 1:

| Type                  | Train | Test  | validation |
|-----------------------|-------|-------|------------|
| Number of sentences   | 20.8K | 2.3K  | 4.6K       |
| Number of characters  | 1000.0K | 112.1K | 223.8K     |
4.1.1. tagging scheme and evaluating indicator. This article uses the BIOES tagging scheme instead of BIO, because the BIOES policy can more clearly define the boundaries of the entity, in the scheme B represents the starting character of the entity and I represents the intermediate character of the entity. O represents a non-entity character, E represents the end character of the entity, S represents an entity having only one character. The experiment uses the Precision value $P$, the recall value $R$ and the $F_1$ value to evaluate the recognition results. The $F_1$ value can accurately evaluate the model performance. The calculation formula is as follows:

$$P = \frac{\text{Number of entities correctly identified}}{\text{Number of entities identified}} \times 100\%$$  \hspace{1cm} (5) \\

$$R = \frac{\text{Number of entities correctly identified}}{\text{Number of entities in the sample}} \times 100\%$$  \hspace{1cm} (6) \\

$$F_1 = \frac{2 \times R \times P}{P + R}$$  \hspace{1cm} (7)

4.2. Parameter settings

The input of BiLSTM is a tensor of $100 \times \text{max}_\text{sen}_\text{len} \times 200$, where the first dimension represents batch size, the second dimension represents the largest sentence length in the batch size sentence, is a variable, and the third dimension represents the dimension of the hidden layer.

We set Dropout to 0.5 and use it in the softmax layer to prevent overfitting. To speed up the convergence of the gradient descent, we use a Gaussian distribution whose mean is zero and variance is one to initialize the model parameters. The optimization function is (the stochastic gradient decent, SGD) algorithm, the learning rate is set to 0.015, the step size is set to 0.05, and the $d_k$ in self-attention is set to 4. We find that the model can achieve the highest efficiency when the BiLSTM layer number is 6 in the same conditions after testing.

4.3. Results analysis

The following table 2 is the comparison between the model in this paper and other model:

| Number | Model | Precision | Recall | F1 |
|--------|-------|-----------|--------|----|
| 1      | Chen et al | 91.22 | 81.71 | 86.2 |
| 2      | zhang et al | 92.2 | 90.18 | 91.18 |
| 3      | Zhou et al | 91.86 | 88.75 | 90.28 |
| 4      | BiLSTM+CRF(None) | 86.21 | 83.54 | 84.84 |
| 5      | BiLSTM+CRF(Seg) | 88.16 | 85.42 | 86.77 |
| 6      | BiLSTM+CRF(Self-All) | 91.13 | 90.05 | 90.59 |
| 7      | BiLSTM+CRF(Self-All+BERT) | 94.57 | 94.23 | 94.4 |

The first three items use external knowledge and artificial features. The fifth item is the model with word segmentation simultaneously. We train the input word vectors by Word2vec and BERT respectively. The sixth and seventh items in the table above are the experimental results. From the table, we can see that the model with self-attention has higher $F_1$ value than other models, which shows that the self-attention mechanism has greatly improved the performance of the model. Among them, the use of Word2vec can improve the performance of the model. The $P$ value of the self-attention model with the original input of the word vector trained by Word2vec is slightly lower than that of Chen et al., because Chen uses the artificially designed features. From the results of the sixth and seventh experiments, it can be concluded that BERT, as a dynamic vector, greatly improves the efficiency of named entity tasks compared with Word2vec.
5. Conclusion
This paper proposes an end-to-end, high-efficiency model for Chinese named entity tasks. We evaluate it on the MASA dataset, and the model performs more efficient than the previous LSTM model. We use BERT training word vector as the original input to understand the meaning of Chinese phrases more comprehensively, and can improve the problem of Chinese new words and unclear boundaries. At the same time, this paper adds self-attention mechanism to the model, so that the model can be given The weighting method focuses on the key information in the sentence, ignoring other redundant information, and has certain benefits for improving the performance of the model. Our model does not use any artifacts or domain-specific features, so the model can be easily transferred to other areas.

Reference
[1] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.
[2] Collins M, Singer Y. Unsupervised models for named entity classification[C]// Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 1999:100—110.
[3] Bikel D M, Miller S, Schwartz R, et al. Nymble: a High-Performance Learning Name-finder[J]. Anlp, 1998:194—201.
[4] Bikel D M. An algorithm that learns what’s in a name[J]. Machine Learning, 1999, 34.
[5] Borthwick A. A Maximum Entropy Approach to Named Entity Recognition[J]. Ph. D. Thesis New York University, 1999.
[6] 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.
[7] Isozaki H, Kazawa H. Efficient support vector classifiers for named entity recognition[C]// Proceedings of the 19th International Conference on Computational Linguistics. Stroudsburg:Association for Computational Linguistics, 2002, 1: 1-7.
[8] Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition[C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2016: 260-270.
[9] Rei M, Crichton G K O, Pyysalo S. Attending to Characters in Neural Sequence Labeling Models [C]. In COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan. 2016: 309–318.