# Method For Extracting Named Entities in Policy Documents Based on BERT-BiLSTM-CRF

**Liu Wei[1]*, Liu Peng[2], Wu Shang-ze[3], Liu Zhi-yuan[4], Zhu Hai-ming[5].**

[1]School of Informatics Xiamen University(National Demonstrative Software School)
4221 Xiang'an South Road,Xiamen University
Xiamen,Fujian 361000
1084227037@qq.com

## Abstract

Due to the differences in policy documents in different regions, companies need to accurately extract useful information from the policy documents,which can be achieved by naming entities Recognition implementation.In order to solve the insufficient semantic representation in the traditional named entity model and the inability of the recurrent neural network (RNN) to solve the problem of long-term dependence, this paper applies the BERT-BiLSTM-CRF model to the entity recognition of policy documents.The model first enhances the semantic representation of words by using the BERT pre-training language model, then uses the bidirectional long short-term memory network (BiLSTM) to obtain the forward and backward semantic information of the sentence, and finally enters the classification result into the conditional random field CRF to identify the global maximum Excellent sequence. Experimental results prove that BERT-BiLSTM-CRF has better performance than other models on policy document data.

## Introduction

Different regions have issued a large number of policy documents for the introduction and management of enterprises, which contain a lot of important information for enterprises, such as subsidy conditions, loan policies, project application conditions, etc.. With the development of society and technology, more and more enterprises have the need to interpret policy documents.Since policy documents are mostly semi-structured and unstructured, their analysis and processing and data mining are severely restricted.The purpose of this article is to use named entity recognition technology to analyze and study policy documents, and to automatically identify and classify valuable information in policies.Named entity recognition (NER) is a research hotspot of natural language processing (NLP), which aims to discover and identify proper nouns and meaningful words in natural text(Nadeau and Sekine 2007).

In recent years, deep learning has made significant progress in the fields of NLP and image recognition, and a large number of researchers have also applied deep learning to named entity recognition(Habibi et al. 2017).Named entity recognition methods based on deep learning all need to convert text information into serialized vectors through word embedding methods(Li et al. 2020).However, the current word embedding methods such as Word2Vec(Mikolov et al. 2013), etc., have the problem of not being able to deal with the ambiguity of Chinese characters(Hu et al. 2018), for example, "疾" may indicate noun disease in different contexts, and it can also indicate that adjectives are fastIn response to this problem, many scholars have proposed different contextual word embedding methods, such as ELMO (embeddings from language models) method(Kiros, Salakhutdinov, and Zemel 2014) and OpenAI-GPT (generative pre-training)(Xia et al. 2020) method and soon. However, the current language representations of word vector embedding methods combined with context are all one-way, and it is impossible to obtain the semantic information before and after at the same time.

Recently, Jacob Devlin et al. Proposed BERT (Bidirectional Encoder Representations from Transformers) pre-trained language model(Devlin et al. 2018), which contains a deep two-way Transformers network, which can better extract sentence features. This paper introduces the BERT model into the NER task of the policy document, and uses the BERT-BiLSTM-CRF model to identify the five types of entities predefined in the policy document: enterprise scale, subsidy amount, declaration conditions, honors, and enterprise types. Experiments have proved that using the BERT pre-training model to construct word embeddings effectively improves the accuracy of named entity recognition. The F1 value obtained in this paper on the policy document data set marked by itself is 94.72%.

## Related work

NER is a sequence labeling task. The specific implementation methods can be divided into three categories:

- Based on rule methods(Riaz 2010). According to the predefined rules in linguistics, the participation of language experts is often required. The specific implementation can use regular expressions, but regular expressions cannot be exhaustive(Eftimov, Seljak, and Koroec 2017), so this also limits the development of rule-based methods.
- Based on statistical methods(Zhang, Pan, and Zhang

*With help from the AAAI Publications Committee.

2004). Use the original or manually labeled corpus for training, and the corpus can be labeled without the participation of experts in a specific field. The commonly used statistical models are Hidden Markov Model (HMM), Conditional Random Field (CRF) and MEMM algorithm.

- Based on neural network methods. Deep learning uses multiple abstraction layers to learn data expression by multiple processing layers. A typical layer is an artificial neural network composed of forward and backward. By feeding source data to the machine, potential expressions are automatically discovered and processed by the classifier or detector. NER benefits from the nonlinear learning characteristics of deep learning. Due to nonlinearity, using neural networks can learn more complex features from the data. At the same time, DL can save a lot of time for designing features, and does not require too much expert intervention to design features, and automatically learn to express. In addition, the deep neural network-based NER model can use an end-to-end model to train and learn through the gradient descent method. This feature allows us to design more complex NER systems(LeCun, Bengio, and Hinton 2015).

The above classification actually corresponds to a development history of NER, as shown in the following Figure 1:

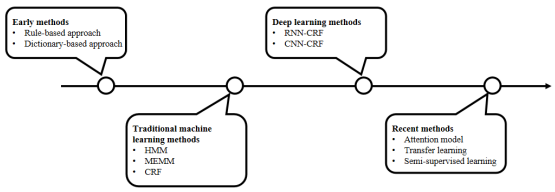


Figure 1: a development history of NER

.

## BERT-BiLSTM-CRF model

The overall structure of the BERT-BiLSTM-CRF model is shown in Figure 2. The first layer of the model is the BERT word embedding layer, which can enhance the semantic representation of sentences and obtain serialized text input; the second layer is the BiLSTM layer. The BiLSTM network can effectively solve the problem of previous methods relying on domain knowledge and feature engineering; the third layer is the CRF layer, which outputs the optimal label. Compared with other models, CRF can focus more on contextual labeling information.

### BERT model

In natural language processing, transforming unstructured text information into corresponding word vectors is a very important task in natural language processing. In recent years, academia has proposed many language models, such as one-hot, Glove, Word2Vec, etc., but the word vectors

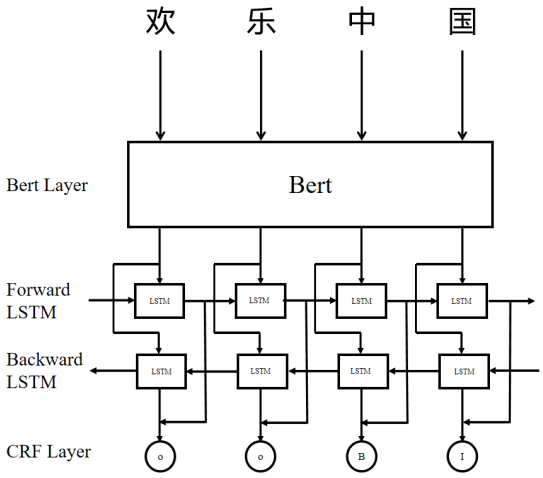


Figure 2: BERT-BiLSTM-CRF model

.

trained by these models are static vectors, and there is a problem of word ambiguity.For example, in the two sentences of ”企业名称重合” and ”商品重量”, the meaning of ”重” is completely different, but in the above language model, the word vectors generated by the two ”重” are exactly the same.

In 2018, the Google artificial intelligence team proposed the BERT pre-training language model, refreshing the best results in 11 natural language processing tasks. BERT uses the bidirectional Transformer neural network as the encoder, and the prediction of each word can refer to the input information in the front and back directions. BERT uses the ”MASK” language model for training, that is, 15% of the words in the sentence will be masked before the word sequence is input, and the original words that are masked are predicted by other unmasked words. Therefore, BERT has strong semantic acquisition and entity recognition capabilities, and can effectively solve the problem of ambiguity. The structure of the BERT model is shown in Figure 3.

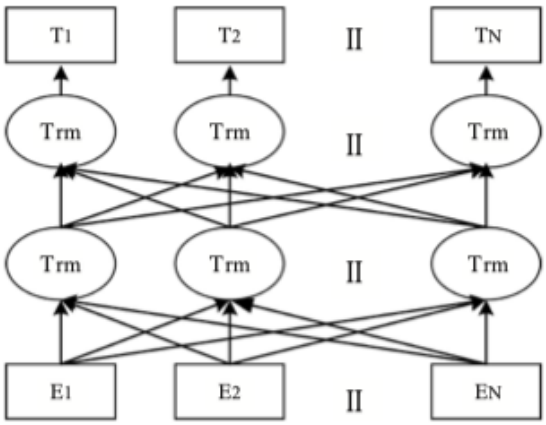


Figure 3: BERT pre-trained language model

.

Table 1: Policy entity marking symbol

| id | Entity category | Start tag | Middle tag | End tag | Training set | Validation set | test set |
|---|---|---|---|---|---|---|---|
| 1 | Enterprise size | B-Size | I-Size | E-Size | 9624 | 2879 | 1562 |
| 2 | Subsidy amount | B-SubSidy | I-SubSidy | E-SubSidy | 7124 | 1980 | 1011 |
| 3 | Declaration conditions | B-Declaration | I-Declaration | E-Declaration | 1124 | 232 | 124 |
| 4 | Honors | B-Honor | I-Honor | E-Honor | 3470 | 1546 | 1015 |
| 5 | Enterprise type | B-Type | I-Type | E-Type | 6745 | 2376 | 1789 |

The most critical part of BERT is the attention mechanism, which abandons the cyclic network structure of RNN. Its basic principle is to calculate the degree of association between each word in a sentence and all other words. The calculation formula is as follows:

$$Attention(Q, K, V) = softmax\left(\frac{QK}{\sqrt{d_k}}\right)V \quad (1)$$

Among them, Q, K, and V are all input word vector matrices.

### Bidirectional long short-term memory network (BiLSTM) model

LSTM is a special recurrent neural network (RNN), which effectively solves the problem of vanishing gradient in traditional RNN and realizes the effective use of long-distance information(Hochreiter and Schmidhuber 1997). The unit structure is shown in Figure 4. By setting up three mechanisms of forgetting gate, input gate and output gate, to selectively deal with the forgetting and transmission of information, this effectively solves the problem of gradient disappearance.

Since LSTM can only process the information before the current unit and cannot obtain the following information, this paper uses a two-layer LSTM network to obtain the forward information and the backward information of the text respectively, and stitch the information to obtain the final feature representation, which can be sufficient Capture contextual semantic information and improve the effect of named entity recognition.

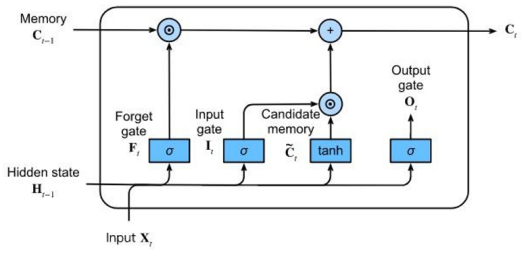


Figure 4: LSTM unit structure

.

### CRF

Lafferty proposed a linear conditional random field (CRF) model in 2001(Lafferty, Mccallum, and Pereira 2001). Considering that in the sequence labeling task, adjacent words or words need to follow certain rules, for example, the I label must be preceded by the B label, not the O label. The CRF model can reasonably consider the dependence relationship between the information and model the label sequence to obtain the optimal sequence.

The input of the CRF module is the word vector trained by the BERT and BiLSTM layers, and the sequence mark of the sentence is obtained from the input of CRF. Then use the formula $score(l \mid s) = sum_{j=1}^{m}\sum_{i=1}^{n}\lambda_j f_j(s, i, l_i, l_{i-1})$ to calculate the value of each sequence l to mark the entire sentence and obtain the score of the input sentence, where s is the sentence to be marked;i is the position of the word in the sentence;$l_i$ is the label of the current word;m is the corresponding Feature number;n is the corresponding sentence length. This paper use the above method to calculate the scores of different sentence sequences, and calculate the probability through the scores. The sequence with the largest probability value is the final output sequence.

## Experimental results and analysis

### Experimental data introduction

The data set used in this experiment is the policy-labeled data produced by my team, and it is divided into training set, validation set and test set in a 7:2:1 manner using cross-validation. The data set includes 42601 entities, with five named entity categories including enterprise size, subsidy amount, declaration conditions, honors, and enterprise type. The experiment adopts the BIOES labeling method, that is, B represents the beginning of the entity; I represents the middle part of the entity; E represents the end of the entity; O represents the text that does not belong to the entity category; S represents that a word forms a category by itself. See Table 1 for the symbol and quantity of each category of entities.

### Evaluation index

This paper uses the accuracy rate P, the recall rate R and the F1 value as the evaluation indicators for named entity recognition. The experimental indicators are defined as follows:

$$accuracyrate: \quad P = \frac{T_P}{T_P + F_P} \times 100\% \quad (2)$$

$$recallrate: \quad R = \frac{T_P}{T_P + F_N} \times 100\% \quad (3)$$

$$F1value: \quad F_1 = \frac{2 \times P \times R}{P + R} \times 100\% \quad (4)$$

Among them, $T_P$indicates the number of positive examples in the test set that were correctly predicted as positive examples; $F_P$indicates the number of negative examples in the test set that were misclassified as positive examples; $F_N$indicates the number of positive examples in the test set that were misclassified as negative cases.

### Experimental parameters and environment settings

This experiment uses the Windows 10 operating system, the CPU is Intel's I7-10750H@2.60GHZ, the GPU used is RTX3060, the video memory size is 8G, the python version is 3.7.1, and the Pytorch version is 1.7.0. The specific parameters of the experiment are shown in Table 2 below:

Table 2: Model training parameters

| Parameters | Values |
|---|---|
| max_seq_length | 150 |
| train_epochs | 100 |
| train_batch_size | 16 |
| leanrning_rate | 3e-5 |
| clip | 5 |
| drop_rate | 0.3 |
| bilstm_size | 128 |

### Analysis of results

In order to verify the performance of the BERT-BiLSTM-CRF model used in this paper, compare it with the following models: (1) CRF model (2) BiLSTM-CRF model (3) CNN-BiLSTM-CRF model.

Table 3 shows the experimental results of a variety of different models named entity recognition. It can be seen from Table 3 that the F1 value of the CRF model on the data set is 69.62%, which can identify some entities, but the recognition effect is not ideal. Method 2 uses a model combining BiLSTM and CRF for entity recognition. The F1 value of the recognition result is 88.24%. Compared with Method 1, the BiLSTM-CRF model can effectively improve the recognition accuracy. which is due to BiLSTM-CRF model can not only combine contextual information, but also consider the dependency between the tags before and after the sentence. Through experiments, it is concluded that the CNN-BiLSTM-CRF model of Method 3 has no significant improvement in effect compared with the model of Method 2.

The BERT-BiLSTM-CRF model introduces the BERT pre-training model on the basis of the method 2 model. It can be seen from the table that the F1 value of the recognition result is 6.48% higher than that of the method 2. Experimental results show that the introduction of BERT can effectively improve the accuracy of named entity recognition, and significantly improve the recall rate and accuracy. The experimental results show that the BERT model can better combine context for feature extraction. In summary, the BERT-BiLSTM-CRF model used in this article has a better effect on the named entity recognition of policy documents than the previous model.

## CONLUSION

This paper uses the BERT-BiLSTM-CRF model to realize the named entity recognition of policy documents, and uses the BERT pre-training model to replace the static word vector generated by the traditional method with the dynamic word vector trained in the large-scale corpus, which effectively solves the traditional word embedding The method has the problem of ambiguity, and makes the semantic representation more accurate. The F1 value of this model in the corpus of policy documents marked by the team reached 94.72%. Compared with other models, it has a better recognition effect. It can better complete the task of identifying the named entities of policy documents, and initially meet the needs of enterprises for the identification of named entities in policy documents.Since the experimental data has only more than 500 policy documents and only 5 categories, there are problems such as low quality of the data set annotation, fewer types of entities, and imbalance in the number of entities. Therefore, we will obtain more policy document data in the later period to enrich the identification types of the model and further promote the practical application of the model.

Table 3: Model performance comparison

| id | Model name | Accuracy rate P(%) | Recall rate R (%) | F1 value (%) |
|---|---|---|---|---|
| 1 | CRF model | 72.22 | 65.45 | 69.62 |
| 2 | BiLSTM-CRF model | 85.28 | 75.83 | 88.24 |
| 3 | CNN-BiLSTM-CRF model | 87.97 | 87.62 | 90.02 |
| 4 | BERT-BiLSTM-CRF model | 95.17 | 94.33 | 94.72 |

## References


Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Eftimov, T.; Seljak, B. K.; and Koroec, P. 2017. A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. *PLoS ONE*, 12(6).

Habibi, M.; Weber, L.; Neves, M.; Wiegandt, D. L.; and Leser, U. 2017. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Hu, K.; Wu, H.; Qi, K.; Yu, J.; Yang, S.; Yu, T.; Zheng, J.; and Liu, B. 2018. A domain keyword analysis approach extending Term Frequency-Keyword Active Index with Google Word2Vec model. *Scientometrics*, 114(3): 1031–1068.

Kiros, R.; Salakhutdinov, R.; and Zemel, R. S. 2014. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. *Computer Science*.

Lafferty, J.; Mccallum, A.; and Pereira, F. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proc. 18th International Conf. on Machine Learning*, 282–289.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *nature*, 521(7553): 436–444.

Li, J.; Sun, A.; Han, J.; and Li, C. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, (26): 3111–3119.

Nadeau, D.; and Sekine, S. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1): 3–26.

Riaz, K. 2010. Rule-based named entity recognition in Urdu. In *Proceedings of the 2010 named entities workshop*, 126–135.

Xia, Q.; Huang, H.; Duan, N.; Zhang, D.; Ji, L.; Sui, Z.; Cui, E.; Bharti, T.; and Zhou, M. 2020. XGPT: Cross-modal Generative Pre-Training for Image Captioning.

Zhang, L.; Pan, Y.; and Zhang, T. 2004. Focused named entity recognition using machine learning. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, 281–288.