Named Entity Recognition Method of Brazilian Legal Text based on pre-training model

Zhili Wang [1*], Yufan Wu [2], Pengbin Lei[3] , Cheng Peng[4]

[1]College of Electronics and Information Engineering, Shenzhen University, Shenzhen, Guangdong, 518060, China

[2]College of Electronics and Information Engineering, Shenzhen University, Shenzhen, Guangdong, 518060, China

[3]College of Electronics and Information Engineering, Shenzhen University, Shenzhen, Guangdong, 518060, China

[4]Data Science, Beijing GridSum Technology Co., Ltd., Beijing, 100000, China

[*]Corresponding author's e-mail: realchilewang@foxmail.com

Abstract. Named entity recognition (NER) is a common task in Natural Language Processing (NLP). To this end, we propose a novel approach based on pre-training model to complete the sequence labeling tasks by learning the large-scale real-world data from Brazilian legal documents. Especially, combining iterated dilated convolution[1] (IDCNN) and Bi-LSTM, we develop the scalable sequence labeling model named Sequence Tagging Model (STM) and extensive experiments validate the effectiveness of STM for NER tasks. Furthermore, compared with the IDCNN-CRF model, the experimental results show that the STM is better and the F1 score is 93.23%, which provides an important basis for NER tasks.

1. Introduction

Named Entity Recognition is aimed at recognizing mentions of rigid designators from text belonging to predefined semantic types such as location, organization and so on[2]. NER is not only an independent tool for information extraction (IE), but also plays an important role in various natural language processing (NLP) tasks, such as question answers[3], information retrieval[4], machine translation[5]. There are four main application techniques for this task design:

1) Rule-based methods, which rely on hand-made rules, so no annotated data is required;

2) Unsupervised learning method, which relies on unsupervised algorithms without manual labeled training examples;

3) Feature-based supervised learning method, which relies on supervised learning algorithms and careful feature engineering design;

4) A method based on deep learning, which automatically finds the representations required for classification or detection from the original data and inputs them end-to-end.

Especially in the past few years, quite a lot of research has applied deep learning to NER and successfully improved the latest performance. Moreover, Zhou et al.[6] proposed a neural model for extracting entities and their relationships using a pre-trained 300-dimensional word vector from Google. In addition, recent research[7] has shown the importance of pre-trained word embedding. As input, pre-trained word embeddings can be fixed or further fine-tuned during NER model training to improve the performance of the entire NER model. Currently, word embeddings including GloVe[8], BERT[9] and word2vec[10] are widely used for NER tasks. Specifically, the contribution of this paper can be summarized as follows:

- We explore to complete the sequence labeling tasks with the hybrid model based on pre-training model BERT.
- We evaluate the proposed model STM with Brazilian legal text dataset LeNER-Br[11]. The experimental results clearly validate the effectiveness of STM.

2. Data Processing

In this paper, the data set used in the experiments is from LeNER-Br, which contains a total of 10392 legal text data records. Among them, 9003 records are selected for training and verification, and 1389 data records for testing. We make statistics on the text length of each piece of data and find that the data with a length greater than 128 accounted for about 10%. To this end, we use the method of cutting by sentence, cutting the sentence with punctuation priority, and reorganizing in the original order. When the length of the reorganized sentence exceeds 128, a new data is generated and the remaining sentences are continued. The above process is completed until all sentences are assembled. This data processing method effectively solves the problem of too long text lengths, and makes full use of data information.

3. Model Architectures

We try to use the open source pre-trained model BERT, and then use the IDCNN and BiLSTM structures to build the entity extraction model. Finally, we adopt heterogeneous single-mode voting for fusion.

3.1. BERT-BiLSTM-CRF

BILSTM-CRF is a more popular named entity recognition model. We input the token vector learned by the pre-trained model BERT into the BILSTM model for further learning, so that the model can better understand the context of the text, and finally obtain the classification result of each token through the CRF layer. The BERT-BILSTM-CRF model we used is shown in Figure 1.

3.2. BERT-IDCNN-CRF

Emma Strubell et al. [1] used IDCNN for entity recognition for the first time. IDCNN improves the CNN structure by using holes (that is, zero-padding). This method captures long-distance information of long sequences of text with some local information loss. This method has better context and structured prediction capabilities than traditional CNNs. In particular, unlike LSTM, IDCNN can not only process text sequences in parallel in the GPU, but also the processing of sentences of length N requires only O (n) time complexity. The BERT-IDCNN-CRF model structure used by this team is shown in Figure 2. The accuracy of this model is comparable to that of BERT-BILSTM-CRF, but the prediction speed of the model is improved by nearly 50%.

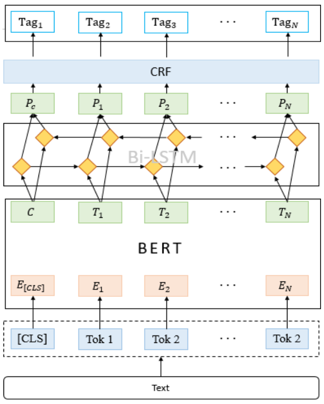

Figure 1. BERT-BiLSTM-CRF

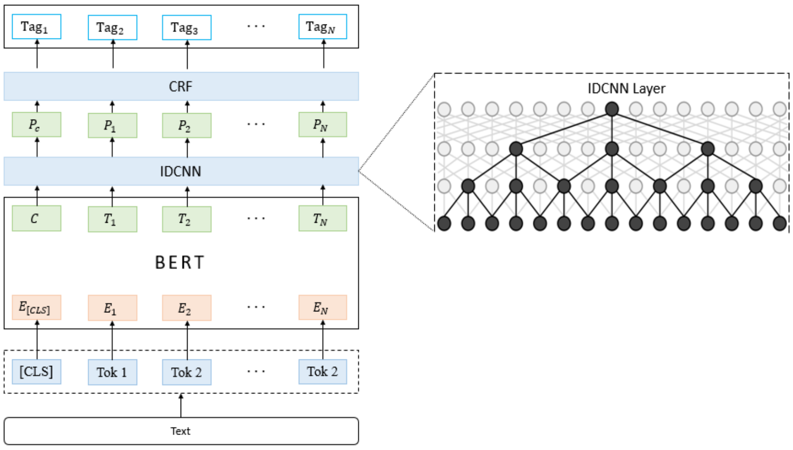

Figure 2. BERT-IDCNN-CRF

3.3. Dynamic weight fusion of multi-layer representation of BERT.

Ganesh Jawahar et al.[12] experimentally verified that the understanding of text in each layer of BERT is different. To this end, we have rewritten BERT to give weights to the 12-layer representation of BERT, and then determine the weight value through training. The initialization of weights is shown in formula (1). After weighting the representations generated by the 12 layers, the obtained representations are reduced to 512 dimensions, as shown in formula (2). Finally, the obtained representation is combined with the previous IDCNN-CRF and BILSTM-CRF models to obtain multiple heterogeneous single modes. The dynamic weight fusion structure represented by BERT multilayer is shown in Figure 3. Among them, $represent_i$ is the representation of each layer of BERT output, and α_i is the weight value of each layer of BERT.

$$\alpha_i = Dense_{unit=1}(represent_i) \tag{1}$$

$$ouput = Dense_{unit=512}\ (\ \textstyle\sum_{i=1}^{n} \alpha_i \cdot\ represent_i\ ) \tag{2}$$

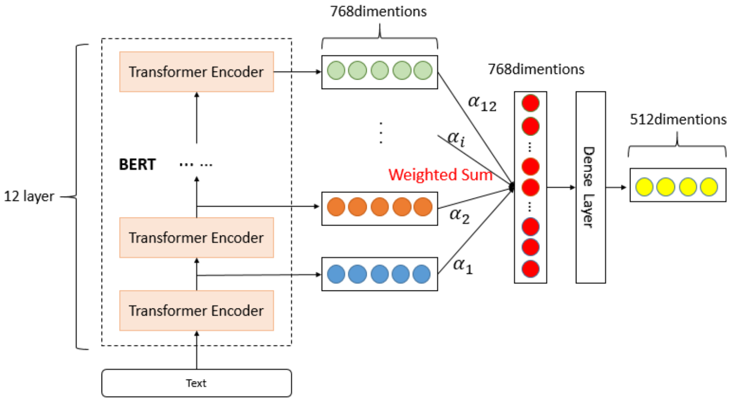

Figure. 3. Dynamic weight fusion of multi-layer representation

3.4. Model Fusion

Through the introduction in the previous sections, we have 4 types of heterogeneous models. Then we adopted a voting method to fuse the heterogeneous models to obtain the best model STM.

4. Experiments

In this section, we will evaluate the effectiveness of our purposed model. Especially, our STM will be validated on Brazilian legal text dataset LeNER-Br.

4.1. Experimental Setup

The hyperparameters setting is listed below.

- Physical calculation environment: 1080ti graphics card
- Batch size: 20
- Epochs: 20
- LSTM dimension: 128
- IDCNN dilation: [1, 1, 2]
- BERT fine-tune learning rate: 5e-4
- Learning rate of BiLSTM and IDCNN: 1e-4
- Dropout rate: 0.5

4.2. Model evaluation

For the comprehensive verification of STM, we use BERT(dynamic)-IDCNN-CRF as baseline and evaluate metrics including precision, recall rate and $\boldsymbol{F_1}$Score to evaluate model. The results are shown in Table 1, which indicates that our STM approach performs better with the ability to complete the sequence labeling tasks compared with baselines.

Table 1. Model Performance Comparison

Model	Precision	Recall rate	$\mathrm{F_1}$ Score
BERT(dynamic)-IDCNN -CRF	0.8679	0.9350	0.9002
BERT(dynamic)-BiLSTM-CRF	0.9076	0.9303	0.9188
BERT-IDCNN-CRF	0.9126	0.9444	0.9285
BERT-BiLSTM-CRF	0.9196	0.9442	0.9317
STM	0.9223	0.9428	0.9325

5. Conclusion

In this paper, we proposed a NER approach based on pre-training model to complete the sequence labeling tasks based on the analysis of LeNER-Br. Specifically, we developed the BERT-BiLSTM-CRF, BERT (Dynamic weight fusion)-BiLSTM-CRF, BERT-IDCNN-CRF, BERT (Dynamic weight fusion)-IDCNN-CRF and their hybrid model and we found that the hybrid model works best.

In the future, we will consider further improvements to the model:

- Add more text feature information, such as using dictionary information and graph neural networks to build a new entity recognition model.
- Pruning and distilling the pre-trained model BERT to reduce the time and space complexity of model.

Acknowledgments

This project was funded by National Key R&D Program of China (2018YFC0832200；2018YFC0832202).

References

[1] E. Strubell, P. Verga, D. Belanger, and A. McCallum, “Fast and accurate entity recognition with iterated dilated convolutions,” in Proc. ACL, 2017, pp. 2670–2680.

[2] D. Nadeau and S. Sekine, “A survey of named entity recognition and classification,” Lingvisticae Investigationes, vol. 30, no. 1, pp.

[3] D. Mollá, M. Van Zaanen, D. Smith et al., “Named entity recognition for question answering,” 2006.

[4] D. Petkova and W. B. Croft, “Proximity-based document representation for named entity retrieval,” in Proc. CIKM, 2007, pp.

[5] B. Babych and A. Hartley, “Improving machine translation quality with automatic named entity recognition,” in Proc. EAMT, 2003, pp. 1–8

[6] P. Zhou, S. Zheng, J. Xu, Z. Qi, H. Bao, and B. Xu, “Joint extraction of multiple relations and entities by

using a hybrid neural network," in Proc. Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Springer, 2017, pp. 135–146.

[7] Y. Shen, H. Yun, Z. C. Lipton, Y. Kronrod, and A. Anandkumar, "Deep active learning for named entity recognition," in Proc. ICLR, 2017.

[8] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[10] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*

[11] Araujo, P., Campos, T., Oliveira, R., Stauffer, M., Couto, S., Bermejo, P.: LeNERBr: a Dataset for Named Entity Recognition in Brazilian Legal Text. In: *International Conference on the Computational Processing of Portuguese (PROPOR).* pp. 313–323. L*ecture Notes on Computer Science (LNCS), Springer, Canela, RS, Brazil* (September 24-26 2018)

[12] Ganesh Jawahar, Benoît Sagot, Djamé Seddah. What does BERT learn about the structure of language?. ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Jul 2019, Florence, Italy. ffhal-02131630