SEQUENCE LABELING OF CHINESE TEXT BASED ON BIDIRECTIONAL GRU-CNN-CRF MODEL

DI LIU, XINYI ZOU

School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China
E-MAIL: zou_xinyi@hotmail.com

Abstract:

Sequence labeling is the basis for many tasks in natural language processing (NLP). It plays an important role in tasks such as word segmentation, named entity recognition (NER), and part-of-speech (POS) tagging. The current mainstream method for sequence labeling is to combine neural network with conditional random field (CRF). The common model is usually a bidirectional RNN-CRF model, which can solve the problem that the labeling task with traditional method cannot be combined well with the context. This paper proposes a Chinese sequence labeling model based on bidirectional GRU-CNN-CRF, which can pay more attention to local features and context relationships, and has better performance in word segmentation and NER. This paper takes the corpus provided by Chinese Wikipedia as the training data set and preprocesses the text by word embedding. The data are then processed through a three-tier architecture of bidirectional Gated Recurrent Unit (GRU), Convolution Neural Network (CNN) and CRF, and finally complete the task of sequence annotation. Compared with the traditional Chinese word segmentation system, this method is more accurate. And it performs better than bidirectional GRU-CRF model on NER issues.

Keywords:

Sequence labeling; Neural network; Word segmentation; Named entity recognition; CRF

1. Introduction

The basic tasks in NLP, such as word segmentation, part-of-speech tagging and named entity recognition, are inseparable from sequence tagging, which is the key link in text processing. For Chinese text, we can regard it as a sequence of words, a sentence represents a sequence, and every word in a sentence is an element. The task of sequence labeling is to annotate every element in the sequence.

Traditional methods of sequence labeling such as support vector machine (SVM), hidden Markov model (HMM), maximum entropy model, CRF, etc. are the mainstream methods in the past. Until recently, there have been people using the HMM model to solve the NER problem [1]. However, these traditional models have defects that cannot be ignored. They rely heavily on artificial feature extraction. The results of feature extraction will directly affect the prediction results of the model. In addition, these models appear to be weak when dealing with cross-domain text [2].

In recent years, deep learning has been shown in many fields. This nonlinear model based on cognitive computing has also brought new possibilities to the NLP field. The neural network model has excellent performance in text classification, text generation and sequence labeling [3]. Huang et al. [4] proposed the Bi-LSTM-CRF model in 2015, which is an architecture for NER problems. The results show that it performs well in phrase recognition and part-of-speech tagging tasks. This model becomes a classic model structure in the field of sequence labeling. then Yao et al. [5] applied this model on the Chinese word segmentation system, which does not require prior knowledge or pre-designed models to learn the context information amazingly; Ling et al. [6] and John et al. [7] have achieved good results in the part-of-speech tagging task. With the successful performance of LSTM, the relatively simple GRU structure has gradually entered everyone's field of vision, Yang [8] and others successfully applied the Bi-GRU-CRF model to the sequence labeling task. These models again and again stunning performance allows people to see the powerful capabilities of the neural network model, the researchers tried to add CNN to the sequence labeling task. Zhai et al. [9] introduced the Seq2Seq model into the sequence labeling task, while Zhai et al. [10] compared CNN and LSTM in character-level word embedding performance and found that the CNN-based Bi-LSTM model has more computational advantages.

This paper consists of four parts. The first part will introduce the development of the sequence labeling task

and the work results in recent years. The second part will describe the composition of the three-layer model used in this paper and the functions of each layer. The third part will be the experimental results. Analyze and compare; the fourth part summarizes and forecasts. The experimental results show that we are more accurate in the Chinese word segmentation than the Jieba system. Compared with the Baidu LAC system, we performed better on the NER problem, achieving 92.5% accuracy, 90.4% recall rate, and 93.2%. The F1 value.

2. Neural Network Architecture

In this part, we will introduce the concrete composition of the three-tier model in detail.

2.1. BGRU

As mentioned earlier, LSTM and GRU belong to RNN category. Although LSTM initially dominated NLP applications, Chung and others have proved that GRU performs equally well in sequence labeling [11].

2.1.1. GRU Unit

In order to overcome RNN's inability to deal with long-distance dependence, LSTM has been proposed. GRU is a very effective variant of LSTM network. It is simpler and more effective than LSTM network, so it is also a very manifold network. Three gate functions are introduced in LSTM: input gate, forgetting gate and output gate to control input value, memory value and output value. In the GRU model, there are only two gates: the update gate and the reset gate. Its specific unit structure is shown in Fig.1.

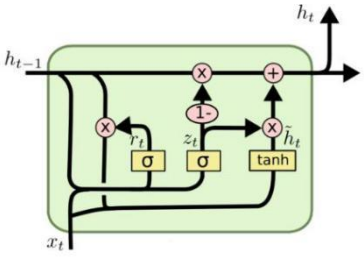

Fig.1 GRU Unit the z_t and r_t in the graph represent the update gate and the reset gate, respectively.

The update gate is used to control the extent to which the state information of the previous moment is brought into the current state. The larger the value of the update gate, the more the state information of the previous moment is brought in. Reset gates control how much information is written to the current candidate set $\tilde{h}_t$ in the previous state. The smaller the reset gates are, the less information is written to the previous state.

2.1.2. Forward Propagation of GRU

According to Fig.1, we can get the following relational expressions:

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t])$$
$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t])$$
$$\tilde{h}_t = \tan h\,(W_{\tilde{h}} \cdot [r_t * h_{t-1}, x_t])$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$
$$y_t = \sigma(W_o \cdot h_t) \quad (1)$$

Where [] denotes that two vectors are connected and * denotes the multiplication of matrix elements.

2.1.3. Bidirectional GRU

Sometimes the prediction may need to be decided by the previous input and the latter input together, which will be more accurate. Therefore, in order to make GRU network have better learning effect, we can learn both the positive and negative rules of the sequence at the same time. For input Chinese text data, we will train GRU network to learn its order and reverse order information at the same time. In this way, we can better understand the information contained in its context.

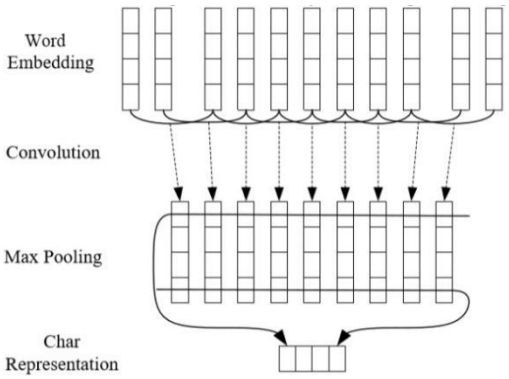

Fig.2 CNN Learning Local Features of Word Embeddings through Fixed Size Windows

2.2. CNNs

Recently, people began to apply CNNs in NLP field, and achieved some remarkable results. Although the earliest people found that the most suitable tasks for CNNs were classification tasks, such as semantic analysis, spam detection and topic classification. However, recent studies

have shown that CNNs can also be used at the character level. It can learn the vector representation at the character level and combine them with the pre-trained word vectors to complete the labeling task. In addition, Zhang [12] et al. found that the CNNs model can be used to learn characters directly without pre-training word vectors. So far, we can say that CNN can be used to process word vectors, and has a good performance. In this paper, we use CNN to expand the dimension of word vectors in RNN layer, and enhance the extraction of local features to make up for the loss of part of information in GRU layer due to the long distance in the learning process. Here, the input of CNN is still a pre-trained word vector, and the specific processing flow is shown in Fig.2.

2.3. CRF

CRF was first proposed by Lafferty et al. in 2001. The idea of CRF is derived from the Hidden Markov Model (HMM). CRF is a statistical model used to mark and segment serialized data. The model calculates the joint probability of the whole sequence given the observation sequence that needs to be marked. Distribution condition attributes of sequence markers can make CRF fit real data well. In NLP, linear chain conditional random fields are most commonly used. Assuming that the sequence X= (X_1, X_2,..., X_n) and Y= (Y_1, Y_2,..., Y_n) are random variable sequences represented by linear chains, if the conditional probability distribution P(Y|X) of random variable sequence Y constitutes a conditional random field under the condition of given random variable sequence X, that is to say, it satisfies Markovian property:

$$P(Y_i|X, Y_1, Y_2, \dots, Y_n) = P(Y_i|X, Y_{i-1}, Y_{i+1}) \quad (2)$$

P (Y | X) is called conditional random field. In the annotation problem of natural language, Y denotes markup sequence, X denotes observation sequence. Let P (Y | X) be a linear chain conditional random field, then the conditional probability of the random variable Y to y is as follows when the random variable X is x:

$$P(y|x) = \frac{1}{Z(x)} \exp(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} u_l s_l(y_i, x, i)) \quad (3)$$

The normalization factor Z (x) is:

$$Z(x) = \sum_y \exp(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} u_l s_l(y_i, x, i) \quad (4)$$

Finally, we train the CRF model by maximum likelihood estimation.

2.4. BGRU-CNN-CRF Model

Finally, combined with the above models, we propose a new hybrid model. The whole architecture is shown in Fig.3 for text waiting for sequential labeling, we first standardize the word embedding preprocessing, and then embedding the word as input of the model. First, context information is learned by bi-directional GRU network. At the same time, we use a CNN with a fixed window to learn the local features of input data. Then we merge the learning results of bi-directional GRU network and the local features of CNN learning. Finally, we access the CRF layer through the full connection layer. The experimental results show that the model performs well in sequence labeling tasks.

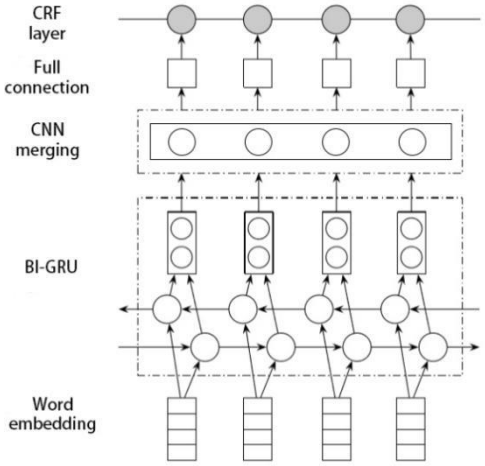

Fig.3 Bidirectional GRU-CNN-CRF model

3. Experimental Results and Analysis

In this section, we will introduce some details of the implementation of the experiment in detail. Firstly, it preprocesses the Chinese Wikipedia corpus, then compares the experimental results with the traditional word segmentation system, and finally compares the performance of NER with Baidu LAC system.

3.1. Word Embeddings

Wikipedia provides an open source raw data package, and the first thing we need to do is preprocess the corpus. Before deep learning is formally applied to NLP, traditional methods often regard vocabulary as a discrete single symbol. The coding regularity of these symbols is not strong, and it is difficult to provide the possible relationship between words. Later, with the advent of deep learning, the vector representation of vocabulary successfully solved this problem. In this paper, data preprocessing actually means

word embedding. Firstly, we use Wikipedia Extractor to extract the text part of the data, then simplify the text, remove special symbols and roughly make word segmentation. Finally, we use Word2vec model to process the corpus into the standard format of word embedding.

3.2. Word Segmentation

NLP technology encounters many problems in dealing with Chinese, one of which is ambiguity in Chinese. In many cases, this ambiguity can only be eliminated by combining the context. As a well-known Chinese word segmentation tool, the Jieba system still has great difficulty in resolving ambiguity problem, which is mainly attributed to the traditional algorithm used in the system. In this paper, the hybrid model combined with neural network can solve the ambiguity problem in word segmentation. Fig.4 shows the disambiguation of a simple sentence.

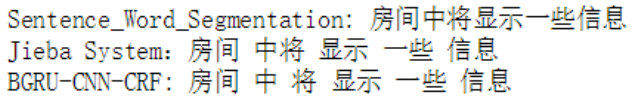

Fig.4 A Simple Demonstration of Word Segmentation

3.3. NER

Named entity recognition tasks are mainly to identify the names of people, places, organizations, time and digital expressions that appear in the text. In recent years, Baidu's Chinese natural language processing platform has performed well, while Baidu LAC system is a lightweight version of the platform and has better NER performance. We compare the hybrid model proposed in this paper with Baidu LAC system. The test set used is one hundred articles obtained on the open source website. Finally, the comparison results are shown in Table 1. Obviously, the model presented in this paper performs better on NER problem.

Table 1 The comparison of NER performance

Model	Precision	Recall	F1-Score
Baidu Lac	0.913	0.863	0.881
BGRU-CNN-CRF	0.925	0.904	0.932

4. Conclusions

In this paper, we propose a new hybrid model, which combines the neural network model with the traditional CRF algorithm, and achieves excellent performance on the related issues of sequence annotation, such as word segmentation and NER problems. But at the same time, the model also has some areas to be improved. Firstly, when CNN completes the merging operation with the output of bidirectional GRU network, the weights of the two networks can be adjusted in the merging process. Secondly, the pre-training operation of the corpus is still not flexible enough. Next, we will consider how to improve the system into an end-to-end system for more convenient use.

References

[1] L. Wang, S. Li, D. F. Wong, and L. S. Chao, "A Joint Chinese Named Entity Recognition and Disambiguation System," 2012.

[2] X. Ma and F. Xia, "Unsupervised Dependency Parsing with Transferring Distribution via Parallel Guidance and Entropy Regularization," in *Meeting of the Association for Computational Linguistics*, 2014.

[3] Y. Goldberg and G. Hirst, *Neural Network Methods in Natural Language Processing*. Morgan & Claypool Publishers, 2017, p. 309.

[4] Z. Huang, W. Xu, and K. J. C. S. Yu, "Bidirectional LSTM-CRF Models for Sequence Tagging," 2015.

[5] Y. Yao and Z. Huang, "Bi-directional LSTM Recurrent Neural Network for Chinese Word Segmentation," in *International Conference on Neural Information Processing*, 2016.

[6] W. Ling *et al.*, "Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation," pp. 1899-1907, 2015.

[7] J. Wieting, M. Bansal, K. Gimpel, and K. Livescu, "Charagram: Embedding Words and Sentences via Character n-grams," 2016.

[8] Z. Yang, R. Salakhutdinov, and W. Cohen, "Multi-Task Cross-Lingual Sequence Tagging from Scratch," 2016.

[9] F. Zhai *et al.*, "Neural Models for Sequence Chunking," 2017.

[10] Z. Zhai, D. Q. Nguyen, and K. J. a. p. a. Verspoor, "Comparing CNN and LSTM character-level embeddings in BiLSTM-CRF models for chemical and disease named entity recognition," 2018.

[11] J. Chung, C. Gulcehre, K. H. Cho, and Y. J. E. A. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," 2014.

[12] X. Zhang, J. Zhao, and Y. Lecun, "Character-level Convolutional Networks for Text Classification," 2015.