International Core Journal of Engineering Volume 6 Issue 1, 2020
ISSN: 2414-1895 DOI: 10.6919/ICJE.202001_6(1).0028

Named Entity Recognition Based on Character-level Language Models and Attention Mechanism

Chaoju Hu[1, a], Junchao Cheng[1, b]

[1]Department of Computer, North China Electric Power University, Baoding 071003, China.

[a]1574222456@qq.com, [b]2423549638@qq.com

Abstract

As a basic task in the field of natural language processing, named entity recognition plays an important role in text data processing tasks. Extracting features from the original text can be considered as the first step in the identification of named entities, but on this basic issue, traditional research still stays at the coarser granularity of words. Unlike traditional research, this paper focuses on finer granularity—character-level named entity recognition research. In order to fully extract the character-level feature representation from the character-level language model, this paper uses CNN and BiLSTM to perform feature extraction together, and introduces the attention mechanism to achieve more effective combination of character features and word features, then combines with BiLSTM-CRF to construct a complete end-to-end deep learning model (At-BiLSTM-CNNs-CRF). The experimental results show that its recognition ability exceeds most deep learning models.

Keywords

Named entity recognition,Attention mechanism,Character-level language models, Natural language processing.

1. Introduction

With the development of computer technology and the popularity of the Internet, a large amount of text data has been generated on the network. How to extract knowledge from this massive amount of text data is a hot topic in the field of natural language processing. Named Entity Recognition (NER) is an important task in information extraction and information retrieval. Its purpose is to identify and classify the components of the named entity in the text, so it is sometimes called named entity recognition and classification. As a basic task in the field of natural language processing, named entity recognition plays an important role in text data processing tasks, whose output tags can provide effective help for a series of downstream tasks such as entity relationship extraction, knowledge map construction and machine translation.

Named entity recognition is generally considered by scholars as a sequence labeling problem. Early researchers mainly used statistical learning methods to conduct research, but in addition to the factors of the model itself, the final performance relied heavily on artificially designed text feature representation. With the rise of deep learning, the ability to automatically extract feature representations from raw data has been favored by researchers, resulting in many excellent neural network-based named entity models, and constantly performing better than the traditional statistical learning models. However, to this day, researchers studying named entity still remain at the coarser granularity of words in extracting features from original text, especially in English named entity recognition. This is especially true due to the nature of the English language. Only a few researchers have noticed a finer granularity than words—Character-level Language Models (CLMs)[1] .

ISSN: 2414-1895 DOI: 10.6919/ICJE.202001_6(1).0028

This paper focuses on fine-grained named entity recognition research. In order to fully extract the character level feature representation from CLMs, this paper uses Convolutional Neural Networks (CNN) and Recurrent Neural Network (RNN) to perform feature extraction. And we also introduce attention mechanism to make the character features and word features more effectively combined to construct a more efficient word vector representation that combines the semantic information and morphological information of the word and uses it as the input. With the mainstream named entity recognition framework (BiLSTM-CRF) [2], we construct a end-to-end deep learning model (At-BiLSTM-CNNs-CRF) without any feature engineering. The experimental results show that its recognition ability exceeds most deep learning models.

2. The Model

This paper proposes the overall architecture of the At-LSTM-CNNs-CRF model as shown in Figure 1. In this section, we will introduce the main components of the model from top to bottom in the order of Figure 1.

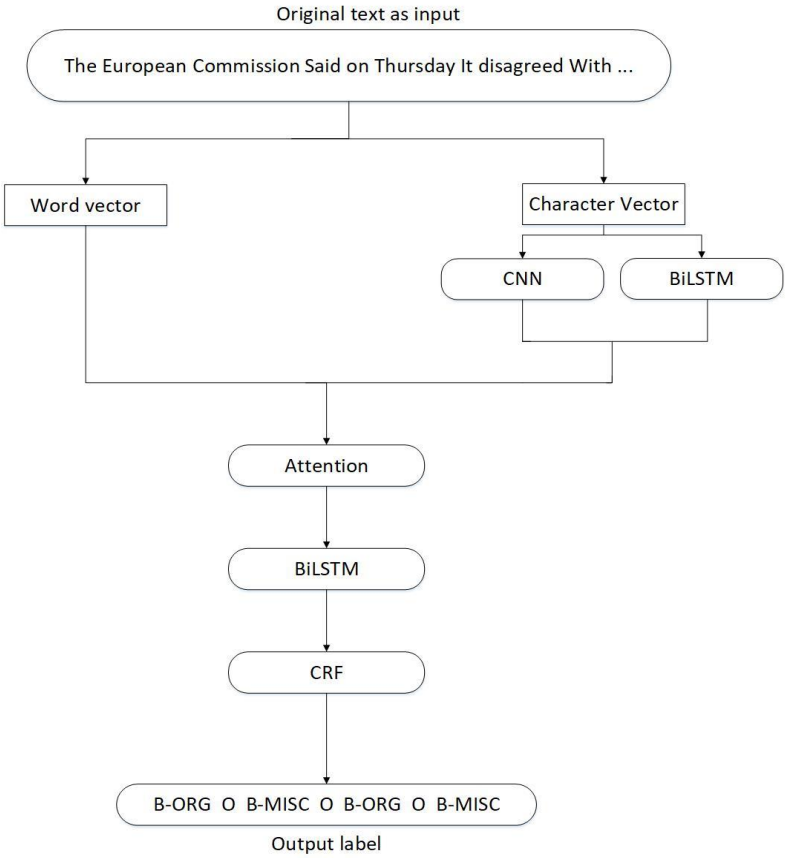

Figure 1. Model overall architecture

2.1 Word vector and character vector

At present, in the research of named entity recognition, the word vector representation method has become the mainstream technology, which is a distributed representation method. Word vector technology can learn the semantic and grammatical information of English words from large-scale unmarked text. Compared with the representation obtained by the traditional bag of words (BOW), the word vector representation has the characteristics of low dimension, denseness. The mainstream word vector technology is the word2vec technology proposed by Mikolov et al.[3]. And Pennington et al.[4] added the word sharing matrix information based on it, and proposed the GolVe model.

The main difference between a character vector and a word vector is that the captured information is different. The character vector is more likely to capture the morphological information of the word while the word vector tends to capture the grammatical and semantic information of the word. In

ISSN: 2414-1895 DOI: 10.6919/ICJE.202001_6(1).0028

addition, the addition of character vectors can effectively solve the headache of unregistered words. The study by Yu et al.[1] proves that the morphological information of words is of great help to the identification of named entities.

2.2 CNN module

The study by Ma et al. [5] demonstrates the powerful ability of CNN to extract morphological information from words. As shown in Figure 2, for each character vector sequence corresponding to each word, we use a one-dimensional convolutional neural network (Cov1D) with a window size K of one filter and a corresponding global pooling layer extracts the corresponding character feature representation. For the character feature representation obtained in this part, we mark it as m_1. The padding on both sides of the word depends on the convolution window, and since the global maximum pooling is used, the dimensions depend on the number of filters L.

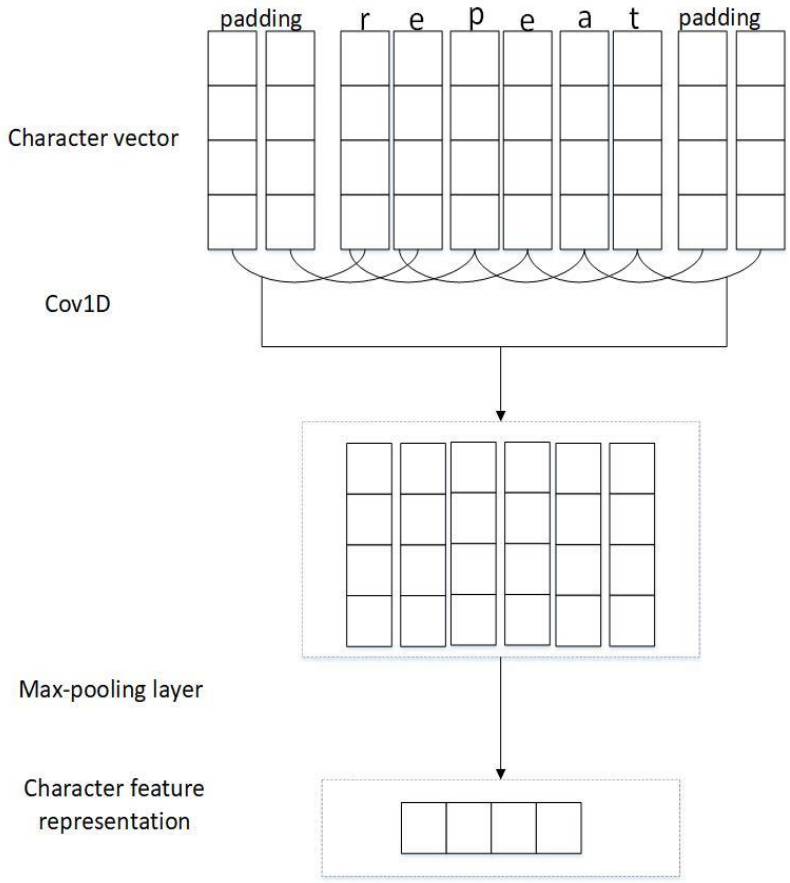

Figure 2. CNN module

2.3 BiLSTM module

Recurrent neural networks (RNNs) are a series of neural networks that can operate on sequential data. In theory, RNN's powerful learning ability can learn long-term dependencies, but it is not in fact. Moreover, when learning long sequences, RNN has gradient disappearance and gradient explosion problems. To solve these problems, we use the Long and Short Time Memory Neural Network (LSTM) designed by Hochreiter et al. [6]. LSTM solves this problem by merging memory cells through four gate mechanisms. Figure 3 is a typical LSTM cell structure diagram.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202001_6(1).0028

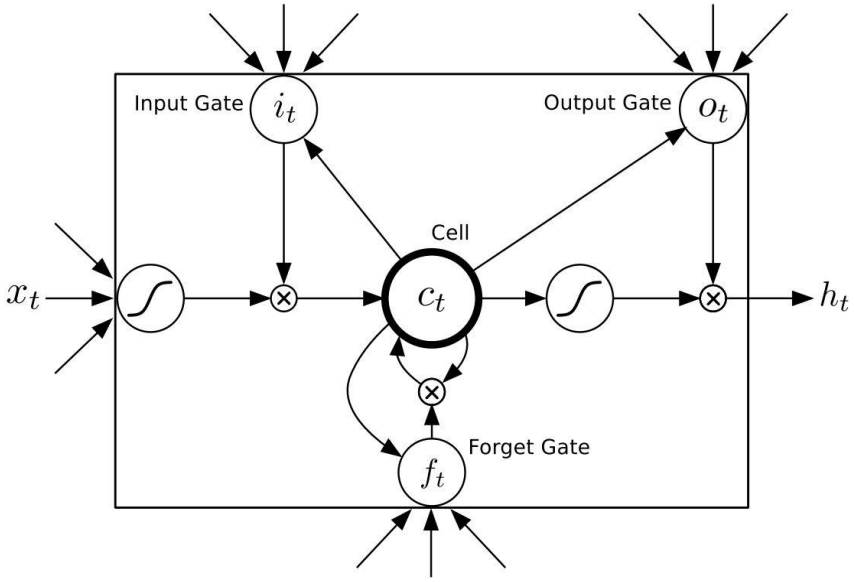

Figure 2. LSTM unit structure diagram

The specific implementation of LSTM is as follows:

$$\mathrm{i}_t = \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right) \tag{1}$$

$$c_t = \left(1-i_t\right)\odot c_{t-1} + i_t \odot \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right) \tag{2}$$

$$o_t = \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o\right) \tag{3}$$

$$h_t = o_t \odot \tanh\left(c_t\right) \tag{4}$$

Where σ is the Sigmoid function, $\odot$ is the inner product multiplication of the vector, W is the inter-layer weights, b is the bias, i, o and c are respectively the corresponding input gates, output gates, and memory cells.

For sequence data, “future” knowledge is as important as “past” knowledge, and one-way LSTM can only extract “past” knowledge, so people add a reverse LSTM extraction to the “future” knowledge, which constructs the BiLSTM module. In the model proposed in this paper, the BiLSTM module undertakes two different tasks depending on the input. For the input sequence that is the sequence of character vectors corresponding to the word, BiLSTM extracts the morphological feature representation of the word. For the sentences consisting of a sequence of word vectors, BiLSTM extracts the context information for each word and converts it into a named entity score vector, which is then passed to the final CRF layer for annotation.

2.4 Attention mechanism

The attention mechanism, originally proposed by Bahdanau et al. [7] and applied to the field of machine translation, has now become an important concept in the field of natural language processing. Rei et al. [8] used a character-level attention mechanism instead of a simple splicing method in their model, mainly to increase the flexibility of the model. The attention mechanism can make the model dynamically determine during the training process, forming a new word vector feature representation for each word, how much information comes from the original word vector, and how much comes from the character-level feature representation. For example, words with regular suffixes can share features at some character level, and irregular words can store exception information in word vectors. Further, for words that have fewer occurrences, we can achieve the purpose of maximizing the use of character features. In the model of this paper, we use the same method to calculate the weights of different vectors using a two-layer fully connected network. The specific implementation is as follows:

$$z = \sigma\left(W_z^{(3)}\tanh\left(W_z^{(1)}r^{word} + W_z^{(2)}m\right)\right) \tag{5}$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202001_6(1).0028

$$x^{word} = z \cdot r^{word} + (1-z) \cdot m \quad (6)$$

Where W is the weight of the two-layer network for calculating the weight, σ is the Sigmoid function and x^{word} is the new word vector corresponding to the generated word.

3. Experimental design and results analysis

3.1 The data set and evaluation criteria

In order to ensure the uniformity of the variables, all the models in the experiment were trained and evaluated on the same English data set provided by the CoNLL2003 shared task [9]. The data set consists of Reuters news from August 1996 to August 1997 and manual annotations. In order to facilitate the researcher, the data set has pre-defined the training set, the verification set and the test set to solve the problem of non-uniformity of the test corpus.

About the evaluation of experimental results, this paper is based on the matching between the labeling results and the experimental results. We specifically take the accuracy rate (Precesion, P), recall rate (Recall, R) and F1 (F1-score) values as the evaluation criteria, where the F1 value is used to assess the overall performance of the model.

3.2 The optimization

This paper uses the Keras framework to launch experiments under the Ubuntu 18 operating system. In the training process, we used the Nadam optimizer provided by Keras, and the initial learning rate was set to 0.0105. In order to further improve the performance of the model, we use two optimization methods, one is Gradient Clipping, its parameter is set to 5.0; the other is to prevent over-fitting, using the regularization method of dropout. The dropout rate is set to 0.5. During the training process, the training is stopped when the verification set F1 value reaches the maximum, which is achieved by the Keras callback function.

3.3 Experimental results and analysis

We compare our model with several models proposed by the predecessors, and the results are shown in Table 1

Table 1. Experimental results

The models	P/%	R/%	F1/%
BiLSTM	87.23	84.85	86.02
BiLSTM-CRF	90.25	88.44	89.34
LSTM-BILSTM-CRF	90.98	89.53	90.25
BiLSTM-CNNs-CRF	91.09	90.25	90.67
At-BilSTM-CRF	91.4	90.15	90.77
Our model	91.5	90.29	90.98

As can be seen from the above table, the BiLSTM model performs the worst, and the BiLSTM-CRF performs better than that. This shows that the CRF classifier is more suitable for the named entity recognition task than the softmax classifier, which verifies the validity of the CRF module. Compared with the two, the LSTM-BILSTM-CRF model and the BiLSTM-CNNs-CRF model perform better in terms of accuracy, recall rate and F1 value, which indicates that the addition of character features to the model can effectively improve the performance of the model. In addition, the performance of the two is not the same. Overall, the latter is better than the former. Except that CNN's ability to extract character features is better than LSTM, as mentioned by Lample et al., considering two neural

ISSN: 2414-1895 DOI: 10.6919/ICJE.202001_6(1).0028

network structures. The difference between the character characteristics information extracted by the two is not exactly the same and complementarity. The At-BilSTM-CRF model outperforms the LSTM-BILSTM-CRF model, demonstrating the effectiveness of the character-level attention mechanism. Finally, our model can be regarded as adding the character feature information extracted by the CNN module in At-BiLSTM-CRF, and our model outperforms the former. This verifies the previous point that the character feature information extracted by CNN and BiLSTM is not exactly the same. Combining all the results, our named entity recognition model achieved the best results, which shows that the model has a strong ability to identify named entities, which outperforms most deep learning models.

4. Conclusion

Unlike traditional research, this paper focuses on fine-grained named entity recognition. In order to realize the full extraction of character level feature representation from CLMs, this paper uses CNN and BiLSTM to perform feature extraction together, and introduces the attention mechanism to achieve more effective combination of character features and word features, and uses it as input and mainstream named entity recognition. The framework (BiLSTM-CRF) [3] combines to construct a complete end-to-end deep learning model (At-BiLSTM-CNNs-CRF). The experimental results show that its recognition ability exceeds most deep learning models. This study proves that the fine-grained named entity recognition model can effectively improve the recognition ability of the NER model and has certain reference value.

The next step in this paper is to explore the representation of fine-grained named entities in special-field texts, such as social texts, medical texts, and more. We believe that named entities in special fields tend to have more similarities in morphological information, and fine-grained named entity recognition will show its talents there.

References

[1] Yu, X., Mayhew, S., Sammons, M., & Roth, D. (2018). On the Strength of Character Language Models for Multilingual Named Entity Recognition. empirical methods in natural language processing.

[2] Huang, Z., Xu, W. L., & Yu, K. (2015). Bidirectional LSTM-CRF Models for Sequence Tagging. arXiv: Computation and Language,.

[3] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. arXiv: Computation and Language,.

[4] Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global Vectors for Word Representation. empirical methods in natural language processing.

[5] Ma, X., & Hovy, E. (2016). End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. arXiv: Learning,.

[6] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780.

[7] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. arXiv: Computation and Language,.

[8] Rei, M., Crichton, G. K., & Pyysalo, S. (2016). Attending to Characters in Neural Sequence Labeling Models. arXiv: Computation and Language,.

[9] Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. north american chapter of the association for computational linguistics.