ATLANTIS PRESS
Advances in Computer Science Research, volume 75
7th International Conference on Mechatronics, Computer and Education Informationization (MCEI 2017)

Biomedical Named Entity Recognition Based on Long and Short Term Memory Model

Youliang Huang[1, a *], Sajid Ali[2, b], Li Wang[1, c] and Renquan Liu[1, d]
[1]Information Center, Beijing University of Chinese Medicine, Beijing, China
[2]Department of Computer Science, University of Education, Lahore, Pakistan
[a]huangyl@bucm.edu.cn, [b]sajid.ali@ue.edu.pk, [c]liwangli2000@126.com, [d]rqliu@bucm.edu.cn
* the corresponding author: huangyl@bucm.edu.cn

Keywords: Long and short term memory model; Literature mining; Biomedical naming recognition; Neural network

Abstract. In view of the problem of biological entity recognition, this paper proposes an improved Long Short-Term Memory (LSTM) recognition method based on the improved bidirectional long and short term memory model. First of all, based on the improvement of re constructed corpus is used to solve the imbalance problem in the distribution of biological entities data sampling algorithm; then, by coupling the forgotten and the input threshold combination to improve the LSTM memory unit, update method to choose the reasonable use of forgotten door control unit state in memory left information improve the biological entity recognition effect. Finally, the test was carried out on the JNLPBA 2004 corpus, and the accuracy rate of 79.7% and the value of 74.1% of the F were obtained. Experimental results show that the proposed recognition method not only has better generalization ability without external assistance, but also effectively improves the recognition effect of biological entities.

Introduction

With the development of computer technology, the collection and communication of data and information become more convenient and fast, and a large amount of information is constantly emerging. People urgently need a way to find specific content quickly from massive information. The application of named entity recognition technology effectively solves the current difficulties. Named Entity (NE) is also known as the named entity, which is usually the semantic unit in Natural Language Processing research. It is also used to help people understand the basic information element [1] of text content. In real life, named entities are usually name, place name, organization name and so on. It can also be address, a meeting name, or a quantitative expression, etc. Named Entity Recognition (NER) is used to annotate [2] in real life, including entity phrases or entity phrases contained in the text. In the sixth message understanding meeting (Message Understanding Conference-6, MUC-6) first proposed a concept named entities and named entity recognition, the main goal is to identify and locate the text contained in the field of knowledge or domain experts interested in words, phrases and jargon, including personal names, place names, organization names, currency, percentage time and date, seven types of [3]. Named entity recognition is not only the key step of text mining, but also the paving for information extraction. In the field of biomedical research, a large number of biomedical texts have increased rapidly, making the biomedical text data to reach an unprecedented scale. In general, cell names, genes, proteins, drugs and diseases are all the most valuable information in biomedical texts. The application of biological entity recognition technology can help researchers to quickly find these biological entities. Biological entity recognition is to identify the molecular biology and medicine in the field of professional vocabulary in biomedical texts, such as DNA (Deoxyribonucleic Acid, DNA) cell line (Cell_line), ribonucleotide (Ribonucleic Acid, RNA) and protein (Protein), cell type (Cell_type), disease (Disease) and drugs (biological entities such as [Drug). Bio entity recognition is one of the basic tasks of biomedical text mining. It is the basic work and foreshadowing for deep research of biological entity relationship extraction, knowledge discovery and hypothesis generation.

Related Work

Biological entity recognition is an important basic work in biomedical text mining. At present, a lot of experiments have been done to summarize the methods of bioentity recognition, which can be classified into four kinds: dictionary based, rule-based, machine learning and hybrid based.The dictionary based recognition method is a partial matching or complete matching method, which is compared with the manually created biomedical dictionaries. It is the earliest identification method to find the similar or identical entity naming strings in biological texts. The rule recognition method is to create rules by domain experts, and then to recognize biological entities based on rules and the naming rules and internal and external characteristics of biological entities. Machine learning based recognition method is based on the statistics of word form, part of speech, dictionaries and other related features and parameters from the existing biological sample data set or training corpus, and then machine learning recognition model is established to complete the recognition of biological entities [4]. The hybrid recognition method is based on the advantages of two or more different recognition methods. It is used to identify biological entities and improve recognition results.

Materials and Methods

Materials. In order to promote the development of biological entity recognition, many competitions and evaluation conferences have been carried out in the world. The more important is 2004 JNLPBA (Joint Workshop on Natural Language Processing in Biomedicine and Its Applications 2004) and BioCreAtlve (Critical Assessment of Information Extraction Systems in Biology), the conference provides relevant data and establish the evaluation standard. In order to make it easier to compare with other methods, this article selects the open corpus of JNLPBA 2004, which is provided by the JNLPBA conference, [5]. JNLPBA 2004 corpus is a set of test sets consisting of 2000 MEDLINE abstracts and 404 MEDLINE abstracts.

Evaluation Metrics. In the field of biological entity recognition, the evaluation criteria of accuracy Precision (P) and recall rate Recall (R) are usually used to evaluate the of the recognition method. Accuracy refers to the ratio of correctly identified by the identified results, and the recall rate is the proportion of the identified entity to the total number of entities.

Methods. By studying the structure of LSTM, LSTM solves the problem of long term dependence through memory cells. Memory cells are similar to cells that store memories. During every state update, information changes such as additions, deletions or updates can be done through memory units. Because of the limited annotation of biomedical text, it is easy to show the phenomenon of fitting if the traditional LSTM structure is strictly based on the biological entity recognition. In LSTM, the output current of gate nodes represent forgotten memory cells retained no correlation with the current input to the memory unit, an important node in the output node representing the input gate, there is a negative correlation between the two, some means that in one dimension, the output value of the forgotten door node. The greater the value corresponding to the input and output, the door will be relatively small. Through the above analysis, this paper proposes an improved bidirectional LSTM biological entity recognition model, which mainly includes input layer, hidden layer and output layer. Its overall structure is shown in Figure 1.

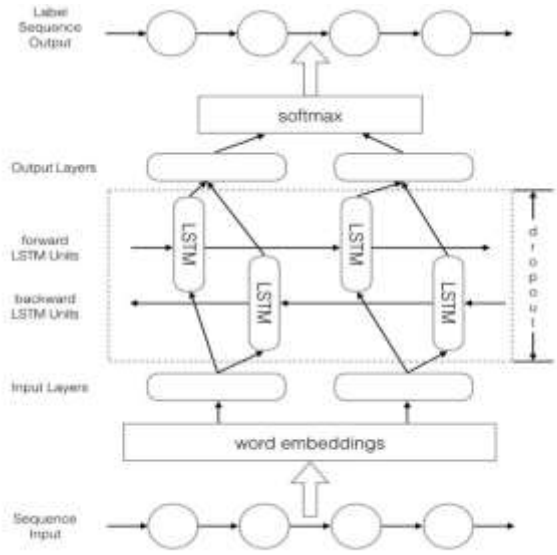

Figure 1. Schematic diagram of biological entity recognition model

The structure of the biological entity recognition model is as follows: 1) the input layer. The sentence is vectored by word2vec tool, the length of sentence vector is set to 50, and the vector dimension is set to 512. Because the number of each node in the neural network is not strictly regulated, it is usually set up according to experience. The setting of the relevant parameters of this model is based on the experimental results of the experimental section. 2) the hidden layer. In order to avoid one-way information transmission of LSTM structure, bidirectional LSTM is adopted in the process of structural design. Each layer contains two models of forward and backward propagation.3) the output layer. The probability distribution of each entity word is serialized, the number of output nodes is the number of categories, and the Softmax model is used for the output neuron activation function.The whole biological entity recognition model uses the category cross entropy as the loss function of the model, and the model is optimized by 15 rounds of iterative optimization on the training set to get the final model.

Results and Analysis

In order to make full use of the corpus and improve the generalization ability of the recognition method, all the experiments involved in this paper use 90% off cross validation. In the experiment, the JNLPBA 2004 corpus was used to train and test the recognition model. All the experimental results were measured by the average of the 90% off cross validation of P, recall (R) and F (F) in the evaluation criteria. In order to verify the different effects on the improvement of recognition results, based on JNLPBA 2004 data sets were compared to experimental LSTM network structure and two-way LSTM network structure; secondly, the introduction of memory unit improved, respectively compared the bidirectional LSTM improved LSTM and improved; finally, the introduction of network structure and improved resampling method the memory unit, respectively, by contrast experiment. The results of the contrastive experiment are shown in Table 1.

Table 1 Comparison experiment of three different grouping strategies

Method	P(%)	R(%)	F-value(%)
LSTM	70.4	62.1	65.9
Bi-directional LSTM	72.1	65.7	68.8
Improvement of LSTM	72.8	65.4	68.9
Improved bi-directional LSTM	76.3	67.2	71.5
Improved LSTM+ resampling	75.8	68.1	71.7
Methods	79.7	69.4	74.1

Table 1 summarizes the identification results of different network structures. It can be seen from the results of the table that the recognition results of the bidirectional network structure are better than the unidirectional network recognition results. Because the bidirectional network structure can use context information at the same time, one way only can only get the information of the past. After introducing the technology of resampling and improving memory unit, although the accuracy of recognition model based on LSTM and resampling method is lower than that based on bidirectional LSTM recognition model, the overall recognition F value is improved. By comparing the experimental results, it can be found that the double LSTM structure with the introduction of resampling technology and improved memory unit has the best recognition effect, and the accuracy rate P and F are 79.7% and 74.1% respectively.

Conclusion

At present, the biological entity recognition is not up to the requirement of practical application, the traditional recognition methods rely on the majority of experts in the field of knowledge and experience to construct artificial dictionaries, rules or the use of traditional machine learning model to solve the recognition process, did not fully consider the biological characteristics and biological entity corpus itself named characteristics, identification method has some limitations. In this chapter, an improved bi-directional LSTM biological entity recognition method based on the biological entity recognition problem is proposed. First, an improved upper sampling method is proposed to reconstruct the training corpus to solve the unbalance problem of the corpus of biological entity tagging. Secondly, we improve the memory unit of LSTM by combining the combination of forgetting and input threshold, and use the forgotten gate to control the state of remaining information in the memory unit to choose a reasonable update method. Finally, the experimental test on the JNLPBA 2004 corpus obtained the accuracy rate of 79.7% and the value of 74.1% of the F. The experimental results show that the model proposed in this paper is feasible and effective, and achieves good recognition effect, which not only increases the generalization ability of the recognition method, but also avoids overfitting phenomenon to varying degrees.

Acknowledgements

The authors are very grateful to the referees and anonymous reviewers for their helpful comments and suggestions. This work was supported, in part, by Beijing University of Chinese Medicine (Grant No. 2016-JYB-QNJSZX005) and Beijing University of Chinese Medicine (Grant No. 2016-JYB-LSMS-019).

References

[1] Alshaikhdeeb, Basel, and K. Ahmad. "Biomedical Named Entity Recognition: A Review." 6(2016):889.
[2] Crichton, G, et al. "A neural network multi-task learning approach to biomedical named entity recognition." Bmc Bioinformatics18.1(2017):368.

[3] Habibi, M, et al. "Deep learning with word embeddings improves biomedical named entity recognition." Bioinformatics 33.14(2017):i37.
[4] Gridach, M. "Character-level neural network for biomedical named entity recognition." Journal of Biomedical Informatics 70(2017):85.
[5] Leaman, Robert, and Z. Lu. "TaggerOne: joint named entity recognition and normalization with semi-Markov Models." Bioinformatics32.18(2016):2839.