Chinese Named Entity Recognition in Business Domain Based on Bi-LSTM-CRF

Yanbo Li[1], Yifei Xin[2], Yunlin Fu[3]

[1]School of Electrical and Electronic Engineering, Shanghai University of Applied Technology, Shanghai, 201418

[2]College of Computer Science and Technology, Jilin University, Changchun, Jilin, 130012

[3]Marine Engineering College, Dalian Maritime University, Dalian, Liaoning, 116026

Keywords: Business Domain, Named Entity Recognition, Deep Learning, Bi-LSTM-CRF

Abstract: In order to solve the problem of disorder, disorder and fragmentation of multi-source and heterogeneous enterprise data on the current network open platform, a Bi-LSTM-CRF deep learning model is proposed to identify named entities in the commercial field. The method includes three types of named entities: enterprise full name entity, enterprise abbreviation entity and person entity. The experimental results show that the average F value of the recognition rate of enterprise full name entity, enterprise abbreviation entity and person entity is 90.85%, which verifies the effectiveness of the proposed method. It is proved that this study effectively improves the efficiency of named entity recognition in the business field.

1. Introduction

With the stable development of the national economy, China is entering a critical period of reform and opening up. The development of economic globalization and the support of national policies are bringing new opportunities and challenges to the development of domestic enterprises [1]. As an important basis for enterprises to determine their own development, inter-enterprise cooperation and government supervision, enterprise information has a significant impact on promoting economic and social development. Based on the large amount of data generated by the enterprise, it has important research value and practical significance [2].

However, the "information explosion" caused by the arrival of big data era makes it easy for people to obtain a large amount of information, but at the same time, it also brings many problems, such as complicated sources of information, disorderly data, difficult to distinguish between true and false, and so on. These problems make it difficult to understand an enterprise in an all-round way [3]. The enterprise-related knowledge that users need is usually stable and common in the industry, but this knowledge often exists in different forms, such as graphics, documents, etc., with scattered storage locations and a wide range of distributed platforms, so it is time-consuming and laborious to find and it is difficult to ensure accuracy. It is more difficult for users to read and understand this kind of information, so it is more difficult for users to read and understand, so it is more difficult for users to read and understand this kind of information, so it is more difficult for users to read and understand, so as to dig out the potential relationship between all kinds of information in enterprises, such as real-time financial announcements and news information on the Internet [4].

Under this premise, this study takes the business domain information as the research object, and proposes a Bi-LSTM-CRF named entity recognition algorithm for named entity recognition task to help users mine and organize business domain information.

2. Business entity identification

2.1 Model frame

In addition to extracting physical information from structured or semi-structured data of various corresponding websites, such as enterprise information published by government regulatory agencies, annual reports published by enterprises, and so on, there is more rich information contained in a large number of dynamic financial information, business announcements and other unstructured data.

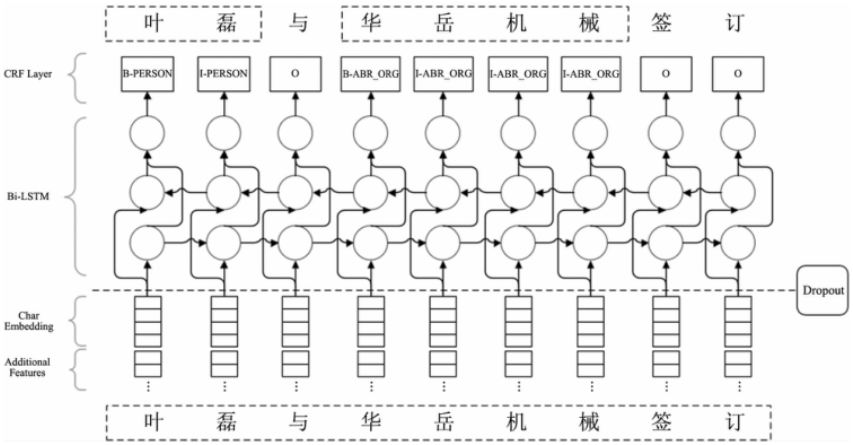

Figure 1. Bi-LSTM-CRF named entity recognition model

(1) Input: Input is the input feature layer of the model. The text of the training set is regarded as an aggregation of words, and the input of each word in the model is composed of a word vector (CharEmbedding) and an additional feature vector (AdditionalFeatures). The word vectors are word vectors trained by Word2Vec and the additional feature vectors are feature vectors formed by splicing under different feature combinations (participle features, lexical features and entity boundary features).

(2) Bi-LSTM: uses a bi-directional cyclic neural network with LSTM units to extract features from the input sequence information, and finally connects the LSTM results of the two directions and inputs them to the CRF layer.

(3) As the output layer of the model, CRFLayer: CRF generates the sequence annotation results of the text.

Between the Input and Bi-LSTM layers, a Dropout parameter is also used for regularization. Dropout prevents the common adjustment of hidden elements by randomly deactivating a certain proportion of elements in the forward calculation process, which makes the network more robust to noise.

2.2 Input feature layer

First of all, after the word segmentation of the corpus, the distributed representation is used to map each word to a shorter word vector to solve the problem that the dimension of the One-hot vector is too large. At the same time, considering that the traditional model input only considers the word-level word vector, the semantic information at the word level may be lost, so on this basis, a large number of relevant corpus information is collected from the encyclopedia website, enterprise yellow pages, Sogou thesaurus and other corpus to study and analyze the composition characteristics of enterprise names and person names in the business field. In the business field, the unique word segmentation features, part of speech features and entity boundary features of enterprise name and person name are obtained, and additional feature vectors are defined as the input supplement of the model, in order to improve the effect of model recognition.

2.3 Bi-LSTM CRF

In this paper, a neural network model is constructed to identify three kinds of entities by combining two-way long-term memory neural network and conditional random field (Bi-LSTM-CRF). Compared with the traditional machine learning, the advantage of this method is that the neural network can learn the characteristics of the data by itself, there is no need to construct complex feature engineering manually, and good results can be obtained. Because the neural network has the characteristic of autonomous learning, we can combine many different types of named entity tasks into one model, and transform different types of named entity recognition tasks into supervised multi-class sequence tagging problems. Improve the efficiency of recognition tasks.

The unidirectional LSTM neural network model can only obtain the above information of the sentence, but can not obtain the following information of the sentence. In order to make up for this deficiency, this paper uses two-way long-term neural network (BI-LSTM) model. BI-LSTM neural network model analyzes sentences forward and backward respectively, which can not only save the previous context information, but also take into account the future context information of sentences, so that it can achieve better results in the task of named entity recognition.

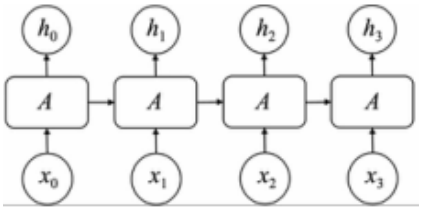

Figure 2. Neural Network structure of LSTM

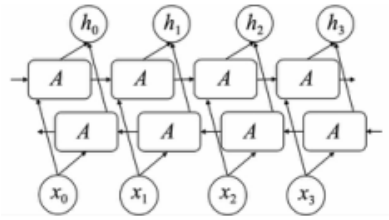

Figure 3. Neural Network structure of BI-LSTM

Table.1. Named entity experiment label quantity of all kinds of entities

Entity category	Full name of enterprise	Enterprise abbreviation	Names	Total
Dimension the number of entities	2985	3095	1139	7219

3. Experimental setting and result analysis

3.1 Data set

The experimental data come from the crawled financial announcement data of Oriental Fortune Network. Among them, 1200 pieces of text data are manually marked as the data set of the experiment, as shown in Table 1, involving 2985 enterprise full name entities, 3095 enterprise abbreviation entities, 1139 person name entities, and a total of 7219 entities. The format part of the processed data set is shown in figure 4. Using the BIO annotation mode, each word and the corresponding annotation in the text is a line, in which ORG represents the enterprise full name entity, ABR_ORG represents the enterprise abbreviation entity, PERSON represents the person entity, and O represents the non-entity. That is, in the picture, "Changsha Hagrid" represents an enterprise for short, "Yi can" and "Xu Jianjun" represent a person entity respectively, and "Hunan High-tech Venture Capital Group Co., Ltd." represents an enterprise full name entity. Finally, the

data set is divided into training set and test set according to the proportion of 7 stroke 3, and the named entity model is trained.

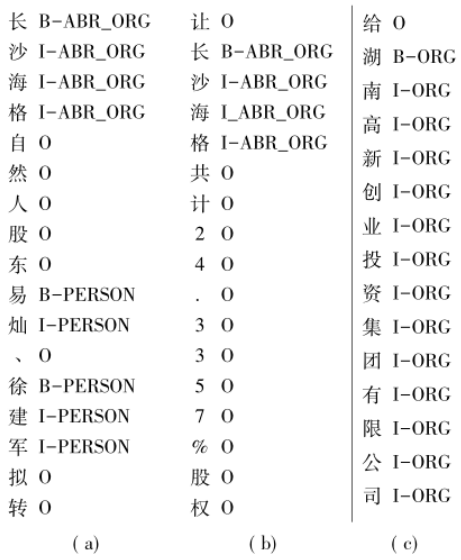

Figure 1. Named entity experiment part data set format

3.2 Experimental design

In order to find the best parameter configuration of the model, the parameter search experiment is carried out in this paper. In the process of search, the dimension of word vector is between [50100150], the number of units in each layer of LSTM is between [64128], and the number of Dropout is between [0.4, 0.5 and 0.6]. Finally, the best training parameters of the model are shown in Table 3, that is, word vector dimension is 100, word segmentation feature, part of speech feature and boundary feature vector dimension is 20, LSTM dimension is 128, dropout value is 0.5, BatchSize size is 20, learning rate is 0.001, optimization algorithm is Adam.

3.3 Analysis of experimental results

The experimental results are shown in Table 2, from which we can find that:

(1) Compared with model 1 and model 3, when only word vector input is taken into account, the named entity recognition F value of model 3 reaches 87.82%, which is significantly higher than that of model 1, which is 82.89%. And in the enterprise full name entity, enterprise abbreviation entity, comprehensive recognition effect has achieved the best, but the person name entity recognition is slightly lower than model 1, but the gap is not big. Generally speaking, the entity recognition effect of Bi-LSTM-CRF neural network model is obviously better than that of traditional CRF model.

(2) Comparing model 2 with model 3, model 2 is a CRF model which takes into account the combination of word features, word part of speech features, context word features, context part of speech features, and so on. The F value of entity recognition is 85.66%, which is 2.77% higher than that of model 1, but still 2.16% lower than that of model 3. This shows that the entity recognition effect of the CRF model considering feature combination is significantly better than that of the traditional CRF model, but it is still lower than that of the unfeatured Bi-LSTM-CRF model.

Table.2. Measurement index values of various types of entities of CRF model and Bi-LSTM-CRF model

	Model	Full name of enterprise	Enterprise abbreviation	Names	Foveral
Model 1	CRF	88.59	74.47	88.63	82.89
Model 2	CRF+Pos+Bi-Word+Bi-Word_Pos	88.79	79.67	88.12	85.6
Model 3	Bi－LSTM－CRF	88.77	86.05	87.52	87.82

4. Summary

Considering the composition characteristics of enterprise name and person name in the business field, this paper proposes a Bi-LSTM-CRF deep learning model which integrates word segmentation features, part of speech features and entity boundary features, and realizes the named entity recognition of three types of entities in the business field. Experiments verify the effectiveness of the proposed method. In the future research, we will also consider the open extraction of entities between enterprises, and cluster the extracted entity relations, and further explore the use of more complex neural network models to achieve named entity recognition in the business field.

References

[1] Tian Juan, Zhu Dingju, Yang Wenhan. A Summary of Enterprise Portrait Research based on big data platform [J]. Computer Science, 2018.

[2] Sun Zhen, Wang Huilin. Review on the research progress of named entity recognition [J]. Modern Library and Information Technology, 2010, (6): 42-47.

[3] Ralph G N. The NYU System for MUC－6 or Where's the Syntax? [C] // Message Understanding Conference, 1995.

[4] Zhou Kun. Research on named entity recognition based on rules [D]. Hefei: Hefei University of Technology, 2010.