Research on Named Entity Recognition of Chinese Geographical Names and Addresses

ZhongYue Wang [1]
[1]Shandong University of Science and Technology, Qingdao 266000
Corresponding author: 1137578475@qq.com

ABSTRACT

With the rapid development of artificial intelligence and big data, informatization has become an irresistible trend of the times, so how to acquire and process geographic information has become one of the most important strategic resources. Named entity recognition is also the task of sequence labeling. Aiming at the problem of accurate identification of place names and addresses in geographic information , this paper proposes a method to first classify and segment place names and addresses, and then label them through a bidirectional cyclic long-short-term neural network and a conditional random field model . It solves the problem of classification and recognition of place names and addresses , and realizes the accurate identification of Chinese place names . The research results based on word segmentation and labeling and recognition through deep learning neural network model show that not only accurate word segmentation can be completed , but also place-name addresses can be accurately identified . The research results can be applied to the semantic measurement of place-name addresses and the place-name address labeling database. construction has great practical significance .

Keywords: natural language processing; place name address classification; place name participle; named entity recognition

1. INTRODUCTION

With the rapid development of geographic information and big data, the task of named entity recognition has become the key to current research. The named entity recognition of Chinese place names and addresses is an important task of natural language processing [1] , which is mainly to combine the names of people, Place names, institution names and entities with specific meanings are extracted and classified, and then organized into information that can be recognized and processed by computers, and then other technologies are used to analyze and understand the text. This plays a crucial role in the structuring of the text. Named entity recognition technology has a wide range of applications in the fields of information extraction, information retrieval, question answering systems and other natural language processing technologies [2] . The main difficulty in the recognition of place names is that the naming of place names is complicated, and the data of place names and addresses is very numerous. The place names of places are named with some uncommon words, so the accurate identification of place names and addresses is particularly important. The earliest research on the identification of place names and addresses was proposed by Zhao Ling in 1997 to identify place names based on the source of place names and the stability of place names [3] . In 2006, Qian Jing and Zhang Jie proposed a place name identification method based on maximum entropy [4] , using the maximum entropy for training to extract feature values, and at the same time combining the changing vocabulary and rules to identify place names, which has a good effect on place name address identification. In 2018, Zhan Binbin et al. proposed to first determine the type of place name based on the general name dictionary of place names and the knowledge base of place name structure rules, and then perform the similarity matching of the core word string of the place name, and finally obtain the matching result that meets the search expectation [5] . Afterwards, Wu Lun et al. proposed a Chinese place name recognition method based on conditional random fields [6] . By counting the features of the words used in place names, designing a feature template, and constructing a feature function according to the feature template, the named entity recognition was completed.There are mainly methods based on convolutional neural networks and cyclic neural networks. Convolutional neural networks focus on local features, while cyclic neural networks focus on global features [7].Therefore, this paper chooses to use the Bi-directional Long Short-Term Memory (BILSTM) combined with the Conditional Random Field (CRF) to carry out the research on the identification of place names and addresses, first by summarizing the address rules for classification, and then Word segmentation processing, combined with manual tagging for model training, so as to achieve accurate

Third International Symposium on Computer Engineering and Intelligent Communications (ISCEIC 2022),
edited by Xianye Ben, Proc. of SPIE Vol. 12462, 1246214 · · doi: 10.1117/12.2660791

automatic tagging and recognition of place names and addresses.

2. CLASSIFICATION AND WORD SEGMENTATION OF CHINESE PLACE NAMES

2.1 Classification of Chinese place names

The classification of place names is the primary task of place name recognition. Place names should be easy to remember and help users to associate the stable associative relationship between place names and geographic entities. Those geographical names that are easy to be associated with specific geographical entities, can vividly reflect local characteristics and reflect regional cultural characteristics, are favored by the society because of their strong pointing effect [8] . The characters used in place names are an important part of the composition of place names. It is precisely because of the complexity and particularity of the characters used in place names that it is difficult to use a unified standard to classify Chinese place names. According to different purposes and principles, different place name classification methods can be used [9] . Some are named after surnames, while others are named according to dynasties and era names, and some names are not static and constantly changing over time, and there are many place names. The official name is very different from that of the local residents, which makes identification very difficult. Therefore, this paper intends to summarize the place names as a whole by reducing the classification of place names through several major categories, and each major category contains many sub-categories to refine the classification.

According to the analysis of the geographical names data, the data are mainly divided into the following four categories:

1. Organization name 2. Living service place name 3. Special place name, etc. 4. Village name

On this basis, each class is subdivided as shown in Table 1 below:

Table 1 Classification of place names

place name classification	Refinement classification	example
Organization name	State organs	Qingdao Municipal Government, Shandong Province, etc.
	enterprise	Qingdao Haier Group, Shandong Province, etc.
	business unit	National Grid, etc.
	Social groups	China Women's Federation, etc.
	other organisations	Law firms, temples, places of worship, etc.
life service place name	car services	Auto repair shop, car dealership, etc.
	Catering and entertainment services	Restaurants, Cake Shops, Snacks, Streets, Tea Art Shops, Bars, Drink Shops, etc.
	daily service	Supermarkets, shopping malls, hardware stores, aquatic product stores, jewelry, beauty salons, bathing, massage, sports venues, convenience stores, furniture stores, etc.
	Travel services	Tourist attractions, landmarks, travel, train stations, airplanes, cars, etc.
special place name	place name address street	The intersection of Tongjun Road and 089 Township Road, Binzhou City, etc.
	place name address house number	No. 579 Qianwangang Road, Huangdao District, etc.
	Special place name expressway toll station	Binzhou South Toll Station (G25 Changshen Expressway entrance) , etc.
Name of the village	village name	Linjia Village Xiaodongjia Shilibao, etc.

According to the above-mentioned custom classification method, the data preprocessing of place names can have a general direction. Through the open API of AutoNavi, about 20,000 pieces of address data in Shandong Province can be crawled for this article. The classification of place names in this article refers to AutoNavi map. The categories of poi in the classification collect data. Among them, the main categories are: automobile service, shopping service, catering service, medical care, scenic spot, commercial residence, road, company, government agency. data to form the original corpus of Chinese place name and address named entity recognition.

2.2 Participle of Chinese place names

The word segmentation of Chinese place names is the primary task of place name labeling, and the accuracy of word segmentation affects the accuracy of recognition. The existing word segmentation methods mainly include stuttering word segmentation, Hamp word segmentation, etc. Python's stammering word segmentation, its word segmentation function is powerful and easy to install [10] , Bao Shuguang introduced an unconventional data dictionary index table, which greatly improved the word segmentation algorithm. Speed [11]. In this paper, stuttering word segmentation is used in combination with a custom dictionary for raw data preprocessing.

As a good word segmentation tool for python, stuttering word segmentation mainly supports three modes of word segmentation:

1. Precise mode: accurately segment sentences, which is very suitable for tasks such as text follow-up analysis

2. Full mode: It is to separate all the words in the sentence that can form a word. The speed is very fast but there is ambiguity.

3. Search engine mode: After the segmentation of the precise mode, the longer words after segmentation are segmented again to improve the recall rate.

This paper chooses to use the precise mode of stuttering word segmentation and combines custom dictionary for word segmentation. Before the word segmentation, the introduction of a custom dictionary includes the administrative divisions of Shandong Province, and the custom dictionary is imported before the word segmentation. For Chinese addresses, the idea of this article is to build a custom dictionary through Jieba to segment theinput address into counties and districts. At this level, the remaining parts after the segmentation are then processed for labeling and subsequent processing to improve the accuracy of named entity recognition and labeling. We make subsequent annotations on the part of the segmented reservation that can represent the address category.

2.3 Construction of Chinese place name annotation set

Named entity recognition is the task of sequence labeling in natural language processing, that is, through data preprocessing, text is input into the computer to identify entities and categories and label them. However, the current annotation systems are all aimed at natural language processing.According to the principle of place name classification in Table 1, and reading a large amount of data at the same time, this paper studies the relationship between address components, and constructs an annotation set from the perspective of address elements. According to the manual labeling method of the label set, the BMES labeling method is selected to label the segmented addresses. The labeled data is used to construct the experimental data of place name and address labeling according to the ratio of 8:1:1 of the training set, validation set and evaluation set.

The custom annotation set is shown in Table 2:

Table 2 Address annotation set

place name classification	callout	example
Organization name	org	Mountain B-org East M-org Division M-org Technology M-org Big M-org Learning E-org City B-org Field M-org Supervisor M-org Administration M-org Bureau E-org Green B-org Steel M-org Smelting M-org Steel M-org Plant E-org Etc.
life service place name	ser	Shun B-ser up to M-ser steam M-ser repair E-ser Fu B-ser Kee M-ser Heavy M-ser Qing M-ser Small M-ser Noodle E-ser Dynamic B - ser veins M - ser era _ _ _
special place name	exc	King B-exc Home M-exc Female M-exc Aunt M-exc Village M-exc No. 249 E-exc G15 B-exc Shen M-exc Sea M-exc High M-exc Speed E-exc etc.
Name of the village	vil	View B-vil Lao M-vil Village E-vil etc.

3. NAMED ENTITY RECOGNITION MODEL

The model is detailed as shown in Figure 1:

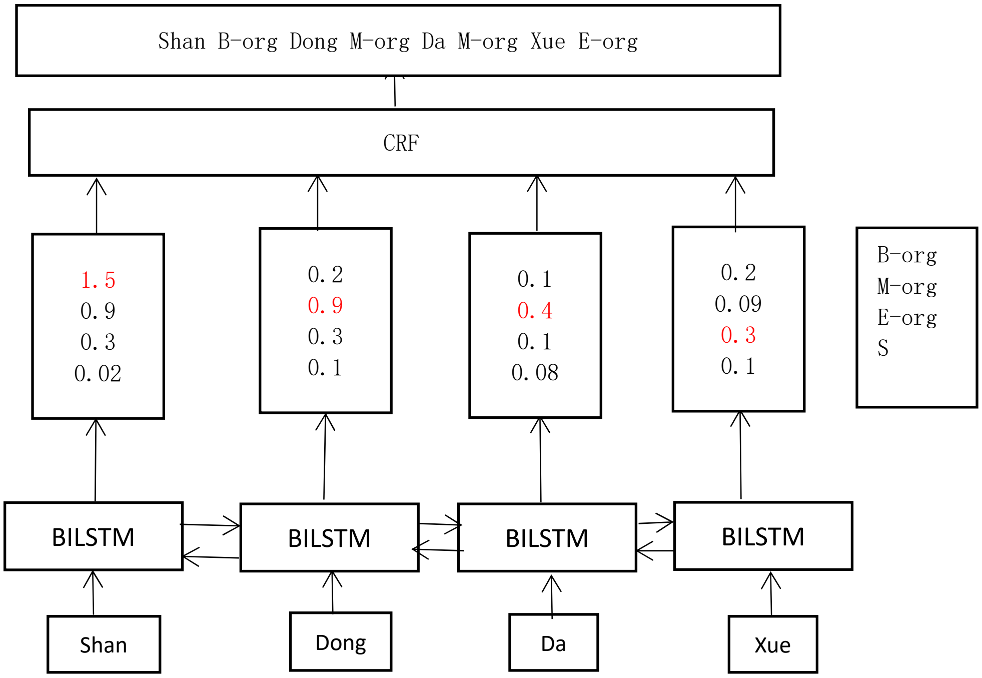

Figure 1 Detailed model diagram

3.1 BILSTM model

BILSTM is an extension of the cyclic neural network model. The cyclic neural network model is based on previous experience or just to solve the current or follow-up problems. It can be regarded as a cyclic copy of multiple ordinary neural networks. Through this, the model obtains the above results. information. With the development of recurrent neural networks, the problem of untargeted long-distance dependencies has been shown. Therefore, the BILSTM model has effectively solved the long-distance dependency problem, and can also better capture bidirectional dependencies.

Sometimes the prediction may need to be determined by several previous inputs and several subsequent inputs, so bidirectional LSTM is introduced, which combines the information of the input sequence in the forward and backward directions, which will more accurately learn long-distance dependency information.

3.2 Conditional Random Field (CRF)

Named entity recognition plays an important role in various natural language processing applications, extracting key information from large amounts of unstructured text data. Named entity recognition is the task of identifying and classifying named entities into predefined categories for a given text.CRF uses the Viterbi decoding algorithm. The Viterbi decoding algorithm has a state transition matrix. In CRF, a state transition matrix will be initialized first. This matrix will be actively learned along with the training data. Get the highest score and choose the best solution. The loss function of the CRF layer includes the real path and the total score of all possible paths. The optimal solution is the ratio of the real path to all paths. The larger the ratio, the higher the accuracy. The optimal path score is decoded by the Viterbi algorithm. result. The state transition matrix is continuously updated by continuously training the training data. With this matrix, errors will be reduced. For example, when marking Shandong University, there will be B-org, East M-exc, large M-org, and E-org errors. East should be M-org. Adding CRF will reduce such errors and achieve constraints. These constraints solve the problem that the BILSTM model cannot make the labeling reach the optimal sequence, eliminating the need to manually construct CRF features, which is of great significance for improving the accuracy of the model.

3.3 Detailed explanation of the experimental process

First, the experimental data is crawled through the AutoNavi API, and the AutoNavi data is selectively selected according to the place name classification table 1. There are four categories of about 5,000 pieces of data for each type, a total of 20,000 pieces of data, and the experimental data is preprocessed and cleaned. Then perform stuttering word segmentation on the experimental data to retain the core part, manually label the experimental data according to the labeling set, divide it according to the 8:1:1 training set, testing machine, and evaluation set, input the BILSTM+CRF model for training and testing, and use LSTM at the same time The model and the CRF model were tested, and the experimental results were compared.

4. Experimental results and analysis

4.1 Experiment Evaluation Criteria and Results

The performance of the model is evaluated by the F1 value, the precision rate P, and the recall rate R.

Accuracy rate P = (number of correct predictions / total number of predictions)

Recall rate R = (number of correct predictions/total number of annotations)

*F1= 2 * precision rate * recall rate / (precision rate + recall rate)*

The experimental results are shown in Table 3 below:

Table 3 Evaluation of experimental results

Model classification	P /%	R /%	F1 /%
CRF	81.07 %	78.6 %	79.8 %
LSTM	86.5 %	84.2 %	85.3 %
BILSTM+CRF	90.1 %	88.7 %	89.4 %

4.2 Analysis of experimental results

Through the analysis of the experimental results, it can be seen from Table 3 that the BILSTM+CRF model used in this paper has greatly improved the P value, R value and F1 value of place name address recognition, and the accuracy of labeling is improved compared with the CRF model and the LSTM model. Obviously, the dimension of the word vector in the construction model is 64, 128, and 256 dimensions. It is better to choose 128 dimensions for the word vector through experiments. The optimal effect is achieved when the parameter value is 0.2. It is very useful for the task of place name recognition through a custom word segmentation dictionary and a custom label set. Compared with the label set based on part-of-speech Easier tasks that can be based on place names have better results.

5. CONCLUSION

It can be seen from the experimental results that the accuracy of place name recognition through the BILSTM+CRF model is greatly improved than that of the CRF model and the LSTM model, but the data selected in this experiment is mainly the data of Shandong Province, because the Chinese place names Naming regions has its own culture and habits, which may lead to ambiguity in identification. Therefore, follow-up work should give more consideration to the Chinese place name standard and add more place name naming information. The preliminary work through a large number of manual annotations is effective, but it takes too much time and needs to be improved. In this paper, the classification of place names is divided into several categories for easy labeling, and it is much easier for the model to recognize and learn, but there are some special addresses such as addresses with long address names. The processing effect is not good, and the model is difficult to distinguish. On the basis of this model, it is of great significance to construct a place name address labeling database, address feature extraction and place name address matching.

REFERENCE

[1]Xu Bing, Shi Shaoqing, Chen Chao.Research on Chinese Address Matching Based on Natural Language[J].Electronic Design Engineering,2020,28(16):7-10.

[2] He Yujie, Du Fang, Shi Yingjie, et al . A Review of Named Entity Recognition Based on Deep Learning [J]. Computer Engineering and Applications, 20 21, 57 (11): 21-36 .

[3] Zhao Ling.A Brief Talk on Place Name Recognition[J].Henan Public Security Journal ,1997,(1):49-50.

[4] Qian Jing, Zhang Jie, Zhang Tao.Recognition method of Chinese names and place names based on maximum entropy[J].Small Microcomputer System . 2006 , (09):1761-1765.

[5] Zhan Binbin, Zhao Ying, Zhao Tingting, et al . Matching Algorithm for Classification and Recognition of Place Names [J]. Beijing Surveying and Mapping, 2018, 32(04): 484-487.

[6] Wu Lun, Liu Lei, Li Haoran, Gao Yong. Recognition method of Chinese place names based on conditional random fields [J]. Journal of Wuhan University (Information Science Edition), 2017, 42(2): 150-156.

[7] He Yujie, Du Fang, Shi Yingjie, et al . A Review of Named Entity Recognition Based on Deep Learning [J]. Computer Engineering and Applications . 2021,57(11):21-36.

[8] Liu Lian'an. Classification of place names and the factors affecting the vitality of place names [J]. China Place Names . 2020 , (2):4-5.

[9] Yin Junke. A Brief Discussion on the Research on Regional Place Names [J]. The Theory of Chinese History and Geography, 2003(03):67-71 .

[10] Shi Guoju. Research on Chinese word segmentation technology based on Python [J]. Wireless Internet Technology, 2021,18(23):110-111.

[11] Bao Shuguang. Optimization implementation of Chinese word segmentation algorithm based on data dictionary [J]. Modern Information Technology, 2022,6(7):80-8 4 .