A Comparative Study of Chinese Address Segmentation Methods

Jiaqi Yu*

College of Remote Sensing and Technology, Wuhan University, Wuhan, Hubei Province, China, 430072
**Corresponding author. Email: 2019302130161@whu.edu.cn*

ABSTRACT
Nowadays, natural language processing continues to grow with its popularity in research and commercial fields. With this trend happening, researchers now put more effort into applying machine learning to achieve natural language processing. This paper concentrates on the word segmentation aspect of Chinese natural language processing, and introduces and compares Bi-LSTM-CRF model and typical toolkits for Chinese word segmentation, aiming for a better understanding of which method to choose on a limited training basis. It can be carried out that when training at a small dataset scale, Bi-LSTM-CRF model segments addresses more accurately than typical toolkits.

Keywords: *word segmentation, natural language processing, Chinese, deep learning*

1. INTRODUCTION

When using a search engine to query information with some key words, some irrelevant or unwanted content is unavoidably included in the results. Under this condition, natural language processing becomes useful, as it helps machines understand people's real meanings [1]. When inputting addresses to a graph database, this study finds it hard to do so correctly because some addresses are out of order. Those errors unavoidably make the graph database inaccurate, thus blocking following analysis. To solve this problem, this study turned to NLP, hoping to reorder the addresses after word segmentation.

This paper was trying to figure out a better method for Chinese address segmentation. The possible answers could either be the traditional dictionary-search method or the neural network-based method. Using the two training methods respectively, two models was trained and the correct segment rate was calculated.

The study offers insights for companies in mapping, smart cities, and more when doing Chinese address segmentation. Future studies might focus on the integration of graph database and address information. Since only the Bi-LSTM-CRF model was used in this study, future studies could also try more different neural network models to get a better result in address segmentation.

2. CURRENT RESEARCH PROGRESS

2.1 Chinese Word Segmentation

Chinese Word segmentation mainly consists of two main methods.

The first method, which is the traditional way, is based on word dictionary. To train such toolkit, researchers must first input sentences or phrases which are already segmented into words. After that, machine counts the times each phrase appears and stores the data in a dictionary. When doing word segment, the dictionary is being searched. Once a certain phrase is found in the dictionary, the matching process is successful and the phrase is identified. This process continues until the sentence cannot be segmented anymore. This method is most widely used and the fastest among all methods. However, when the training set isn't so large, or when the toolkit is used to segment new phrase it doesn't recorded in its dictionary, the result is normally not so ideal.

The other method is based on machine learning and statistics. Models related to this method includes SVM, CRF, deep learning, etc. Most Chinese word segment toolkits nowadays are based on this method, such as Hanlp or Jieba. These toolkits combine one or more models with word dictionary methods together, thus gains ability to segment phrases either already existed or non-existed. Recently, with the development of machine learning, one other typical model is Bi-LSTM-CRF

model. This model combines neural network with CRF model together.

This paper mainly compares the two kinds of models mentioned above, which are typical Chinese word segment toolkits and Bi-LSTM-CRF model.

2.2 Other Languages

For those languages with high usage, it is never difficult to form a training set. Normally, word segment toolkits for those languages are trained with thousands or millions of records of data. Lucene, for example, is a great toolkit belonging to Apache. Though Lucene is created for English and German word segment, it can also be used in other languages.

As for those languages not so widely used, less NLP research based on those languages has been conducted. Consequently, training data for those languages is usually not enough, not to mention the language-technical rules that create barriers between different languages. When researchers who do not speak such language are trying to create a toolkit to do word segments, they sometimes transplant existing models to achieve such a goal.

A Welsh NLP toolkit applied the typical way for most toolkits: a combination of word dictionary and machine learning[2]. A Laotian word segment toolkit was created by Chinese researchers using the LSTM model[3]. This illustrates that both models are effective when applying existing models to foreign languages. The figures below are the structures for the two toolkits.

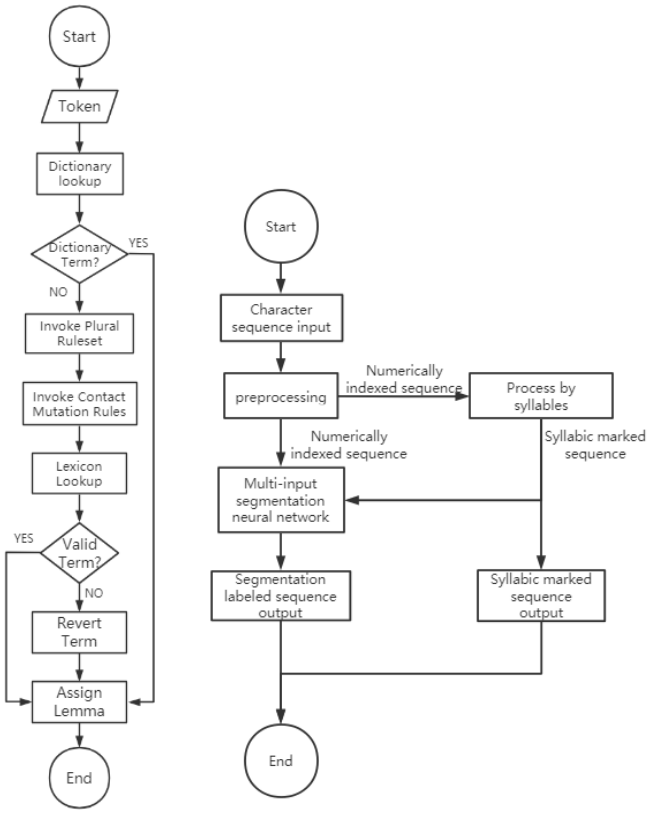

Fig1. structures for Welsh(left) and Laotian(right) NLP toolkits

3. METHOD AND EXPERIMENT

3.1 Dataset

This dataset this paper uses contains 220,000 standard addresses, which come from China's Zhejiang province civil administration department, including fieldnames such as province, city, county, town, community, street, village, door, point of interest, building, unit, floor, room. These field names are sorted in order of administrative divisions. An example is "Zhejiang Province Huzhou City Deqing Town Xiazhuhu Road Kangjie village Xintong Bridge NO.984". In this address, the highlighted characters represent regional unit, thus they are also used as a symbol of segmentation. So, the address above should be segmented as "Zhejiang Province", "Huzhou City", "Deqing Town", "Xiazhuhu Road", "Kangjie village", "Xintong Bridge", "NO.984" 7 phrases.

3.2 Models

3.2.1 Typical Toolkits

The most common toolkit models are divided into two steps. First, when a sentence is being input, the toolkit searches in the sentence for existing phrases the dictionary. If the whole sentence can be segmented this way, then exit the process. Otherwise, the toolkit calls it

the machine learning component, and tries to segment other parts of the sentence.

The word dictionary is created through reading a text that contains all the segmented phrases. The machine learning model used in this paper is word2vec. As an NLP model invented by Google in 2013, word2vec's most well-known characteristic is that it expresses all words as low-dimensional and dense vectors. In this way, similar words are closer in distance in word vector space. Also, users can easily and qualitatively measure the similarity between two words. The generated word vector has 64 dimensions. Considering the nearest 3 words, sg=1 method is adopted, namely the skip-gram. After the word vector is created, it is saved in a model for later use.

3.2.2 Bi-LSTM-CRF Model

CRF (Conditional Random Field) is the basis of many models. CRF is a Markov random field of a random variable Y given random a variable X. The linear chain random field is mainly used for sequence labeling, as shown in Fig2[4,5]. In the figure, O1-O5 is the observed sequence, while R1-R5 is the prediction sequence.

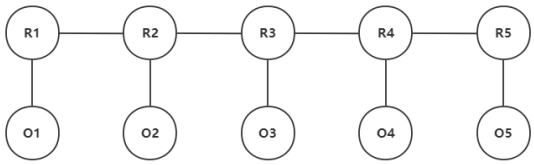

Fig2. CRF probability model diagram [6]

LSTM (long short-term memory) is a kind of RNN model adopted to solve long dependency problems in sequence annotations. There're no essential differences between RNN and LSTM, only the hidden structure in LSTM is called cell structure. The in-gate determines what percentage of data input will be kept, the oblivion-gate determines what percentage of data will be kept from the last hidden layer, and the out-gate determines what percentage of data will be output [6]. The structure of LSTM is as Fig3.

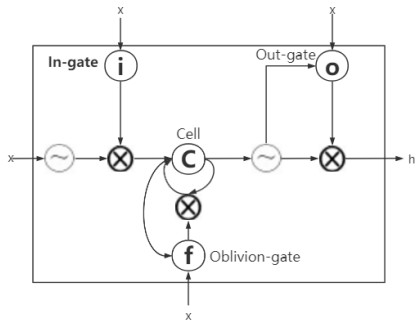

Fig3. LSTM cell structure diagram [6]

The Bi-LSTM-CRF model is an optimization of the LSTM model. It captures two separate hidden layer's past and future information by processing each sequence forward and backward respectively, and connects them to a final output. In this way, the Bi-LSTM-CRF model utilizes the context better and segments the sentence better.

The updating formula of the front to back neural network layer is:

$$h_i^{\rightarrow} = H\left(w_{xh_i^{\rightarrow}} x_i + w_{h^{\rightarrow}h^{\rightarrow}} h_{i-1}^{\rightarrow} + bh_i^{\rightarrow}\right)$$

The updating formula of the back to front neural network layer is:

$$h_i^{\leftarrow} = H\left(w_{xh_i^{\leftarrow}} x_i + w_{h^{\leftarrow}h^{\leftarrow}} h_{i-1}^{\leftarrow} + bh_i^{\leftarrow}\right)$$

The two RNN layers are superposed before input to the hidden layer:

$$P_i = W_{h^{\rightarrow}y} h_i^{\rightarrow} + W_{h^{\leftarrow}y} h_i^{\leftarrow} + b_y$$

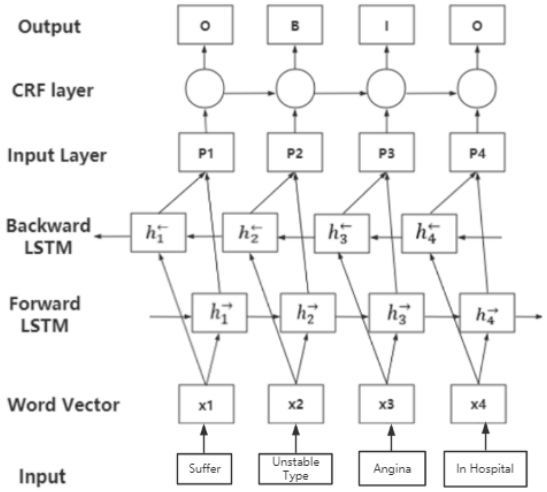

Fig4. Bi-LSTM-CRF model

3.3 Evaluation Criterion

Since this paper aims at giving a referable comparison of two models on a limited training basis, only 6000 addresses are chosen. Among the chosen data, 5000 random data are chosen as the training set, used to train word dictionary (model 1) or Bi-LSTM-CRF (model 2) models, while the other 1000 are used as testing sets.

After training the two models, each data in the testing set is segmented respectively by either model. The output is then compared with the correct answer, which is stored in the complete data set. The percentage of correct rage will then be printed and compared, helping concluding results.

4. RESULTS AND ANALYSIS

This research calculated the accuracy rates of address segmentation for the two methods distinctively, at datasets of 5,000 and 100,000 sets. The testing set is always made up of 1,000 sets. The results are shown in the table below.

Table1. Accuracy rate of different methods

Method / Dataset scale	Traditional model	Bi-LSTM-CRF model
5000 sets	70.5%	89.0%
100,000 sets	96.0%	95.5%

As is shown in the table, when the training set is not large enough, which corresponds to those languages with less usage, the Bi-LSTM-CRF model's address segmentation is more accurate than the traditional dictionary model. However, when it comes to 100,000 sets for the training set, both models work quite well in addressing segmentation.

The huge gap between the accuracy rates for Traditional model with 5,000 sets and Traditional model with 100,000 sets is probably caused by the dictionary method's characteristics. As a dictionary, it cannot process words that do not exist in it, so as the training set's scale increases, the dictionary becomes more and more adequate to segment testing set correctly.

As for the Bi-LSTM-CRF model, when the training set is not enough, it works way better than the former method. This is because that neural network is created based on computer simulation, which performs better when a few inputs perfectly match the training set. As the scale of training set increases, the prediction becomes more accurate, and can reach a limit of about 95%. Considering there might be a few errors in the training set, the two methods work equally well when the training set contains 100,000 sets.

When it comes to training speed, traditional model training takes less time (about half) than Bi-LSTM-CRF model in python.

In conclusion, the two methods have their pros and cons. Traditional model performs well when the training set is on a large scale. It takes less time than Bi-LSTM-CRF model, and maintains a good accuracy rate. However, when the training set is on a small scale, the prediction of the Bi-LSTM-CRF model is normally the more accurate one, despite it taking more time for training.

5. CONCLUSION

This study mainly focuses on a comparison between the traditional dictionary word segment model and the Bi-LSTM-CRF word segment model. With comparative experiments, the following results can be concluded.

When the training set's scale is not large enough, the Bi-LSTM-CRF model is more accurate. When the training set's scale is large enough (more than 100,000 sets), either model is very accurate. However, training the Bi-LSTM-CRF model takes twice as long as creating a dictionary model.

This study's main weakness lies in the experiment part. With no repetition, the results are subject to error. Also, only one kind of neural network model was introduced to this experiment, which leaves open the possibility of a more accurate outcome for other neural networks.

In future studies, researchers can either compare the two models in different language segmentations, or introduce other algorithms and methods to re-experiment.

REFERENCES

[1] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," Scientific Am., May 2001, pp. 34-43.

[2] D. Cunliffe, A. Vlachidis, D. Williams, D. Tudhope, "Natural language processing for under-resourced languages: Developing a Welsh natural language toolkit", 2022.

[3] Yongbin Hao, Lanjiang Zhou, Chang Liu, "An End-to-end Multi Task Method for Laotian Word Segmentation via LSTM", Sep 2021.

[4] Hang Li, "Statistical learning method" [M]. Beijing ： Tsinghua University Press,2012

[5] Kai Liu, "Named entity extraction of TCM clinical records based on conditional random field" [J] references Computer Engineering, 2014, 40(9):312-316.

[6] Yanshu Chen, "Chinese words segmentation based on Bi-LSTM neural network model" [J]. Chinese Technology academic journal, 2018,32(2):29-37.