A Contrastive Study of Chinese Text Segmentation Tools in Marketing Notification Texts

Xianwei Zhang[1, 3], Peng Wu[1], Jiuming Cai[2] and Kun Wang[2]

[1] School of Computer Science, Xi'an Shiyou University, Xi'an, China.
[3] School of Computer Science and Technology, Xidian University, Xi'an, China.
[2] Email: xwzhang@xsyu.edu.cn.

Abstract. It is necessary to analyze and mining marketing notification texts because there are various commercial information. The base of the operation is Chinese word segmentation. The speed and accuracy of word segmentation have important influence on the subsequent texts mining. We compared accuracy, recall, and F-value of four open-source Chinese word segmentation tools (Ansj, HanLP, Word and Jieba) on the third-party datasets. Then, we compared the segmentation speed of the four tools on one million marketing notification texts. Finally, we segmented 5, 000 marketing notification texts artificially. We evaluated the performance of these segmentation tools by the results of artificial segmentation, which are known as evaluate standard. The experiments show the Base mode of the Ansj is the fastest. The HanLP is a best segmentation tool for balancing speed and accuracy of word segmentation. After adding a custom dictionary, the effect of word segmentation has been significantly improved.

1. Introduction

Marketing notification texts contains various texts which can guide users to consume, discount notices, etc. We can extract important commercial promotion information from the marketing notification texts. Enterprises can analyse user behaviour, improve marketing conversion rate, and provide corresponding marketing activities and customer services for different customer groups by data mining of the texts.

Firstly, we need to extract the characteristics of the texts so that to analyse and mine the texts. The first step of text feature extraction is to segment the texts. This paper compares the performance of different word segmentation tools in marketing notification texts, which including the segmentation speed, the word segmentation accuracy, the recall rate and the *F* value on large-volume data.

Each word in Chinese consists of different number of Chinese characters. There is no obvious separation mark between word and word. The Chinese word segmentation machines need to divide the Chinese string into a reasonable words sequence. Due to the special structure and format of Chinese, Chinese word segmentation technology has many difficulties. The key issues include the identification of unregistered words and the judgment of ambiguous words [1]. A good word segmentation system should contain at least three parts: Chinese dictionary, statistical system and large-scale corpus. One of the implementation methods of word segmentation system is to employ the dictionary [2]. This method match the strings in Chinese text based on natural language rules making use of dictionary, such as the Bi-direction Matching method. The method is simple to implement which deal bad with ambiguous and unregistered words. The statistical system is suitable for the recognition of new words and the resolution of ambiguous words [3]. Among them, the resolution of ambiguous words can be realized by counting the mutual information, t- information of Chinese characters [4, 5], or using the

Markov assumption to simplify the N-gram model [6]. As for as unregistered words, they can be identified by machine learning. For example, firstly, a large-scale corpus is used to train the hidden Markov model [7]. Then, a directed acyclic graph is constructed. Finally, the shortest path is obtained by Viterbi. We obtain the sequence of words by dividing the words according to above path. The CRF [8] is the best-performing word segmentation method whose actual word segmentation effect also depends on a good training model.

The Ansj is improved from the Chinese Academy of Sciences' Ictclas [9], which provide precise word segmentation (To Analysis), user-defined dictionary priority strategy word segmentation (Dic Analysis), new word discovery function word segmentation (Nlp Analysis) and minimum particle word segmentation (Base Analysis). Among them, To Analysis has achieved a good balance between ease of use, stability, veracity and word segmentation efficiency. Dic Analysis is most suitable if the word segmentation has a high requirement for the user-defined dictionary. Nlp Analysis can identify unregistered words. However, the word segmentation efficiency of Nlp Analysis is lower than other modes and the word segmentation speed is unstable. Base Analysis guarantees basic word segmentation function, such as minimize word segmentation granularity, the word segmentation speed is fast. However, it has limited effect on new words and long words.

HanLP is a toolkit that contains a large number of models and algorithms for processing natural language. The HanLP provides the Viterbi algorithm, Dijkstra shortest path method, Speed Tokenizer and NLP word segmentation. The standard word segmentation model of HanLP is based on the Viterbi algorithm under hidden Markov model [10, 11], which is the most balanced mode of speed and effect in HanLP. The Dijkstra shortest path method is slower than the Viterbi algorithm on the speed of word segmentation. However, the debugging information of the Dijkstra is more abundant. The Speed Tokenizer adopt Aho Corasick automaton [12] which combines Double Array Trie [13]. The Speed Tokenizer has high performance and supports generics and persistence. NLP word segmentation will recognize all named entity and label all of their part-of-speech. However, the word segmentation speed of NLP word segmentation is slower than others.

The Word is a java Chinese distributed word segmentation component which provides 10 dictionary-based word segmentation algorithms. The Word employ N-gram model [14] to eliminate ambiguous words and identify unregistered words.

The Jieba implements efficient word graph scanning based on the trie tree structure. The Jieba generates a directed acyclic graph (DAG) to represent all possible word-cutting results. In the word graph, the dynamic programming algorithm is used to find the best word segmentation combination. The Viterbi algorithm based on hidden Markov model is also used to recognize unregistered dictionary.

2. Experiments

2.1. Data Source

The third-party dataset used in this paper is the dataset which is provided by the Chinese Language Processing Group of the International Society for Computational Languages (ACL) in 2005. There are 14, 432 training samples and 14, 432 test samples in the dataset. The specific corpus are jointly provided by the CKIP, the City University of Hong Kong, the Peking University and the Microsoft Research Institute.

The experiments also used one million marketing notification texts as test for word segmentation efficiency.

2.2. Experimental Process

In this paper, we test the third-party dataset by respectively using the java version of the four word segmentation tools which include Jieba-1.0.3, Word-1.4, Ansj-5.1.3 and HanLP-1.5.3. Test objects include the four operation modes of the Ansj, the four operation modes of the HanLP, the ten operation modes of the word and the one operation mode of the Jieba. The details are as shown in Table 1.

Table 1. The operation modes of the Ansj, the HanLP, the word and the Jieba

Word segmentation tool	Ansj	HanLP	Word	Jieba
Mode	Base Analysis	Viterbi	Maximum Matching	Search
	To Analysis	Nlp	Reverse Maximum Matching	
	Dic Analysis	Speed Tokenizer	Minimum Matching	
	Nlp Analysis	Dijkstra	Reverse Minimum Matching	
			Bidirectional Maximum Matching	
			Bidirectional Minimum Matching	
			Bidirectional Maximum Minimum Match	
			Full Segmentation	
			Minimal Word Count	
			Max Ngram Score	

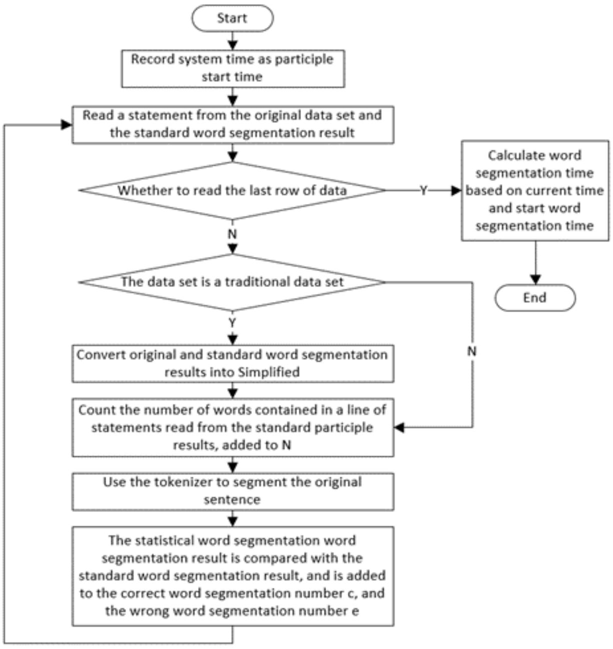
Start
Record system time as participle start time
Read a statement from the original data set and the standard word segmentation result
Whether to read the last row of data
Y
Calculate word segmentation time based on current time and start word segmentation time
End
N
The data set is a traditional data set
Y
N
Convert original and standard word segmentation results into Simplified
Count the number of words contained in a line of statements read from the standard participle results, added to N
Use the tokenizer to segment the original sentence
The statistical word segmentation word segmentation result is compared with the standard word segmentation result, and is added to the correct word segmentation number c, and the wrong word segmentation number e

Figure 1. The flow chart of word segmentation

The process of word segmentation is depicted in figure 1. Firstly, we need to prepare two text data. One of them is the original texts of the undivided words, which contain one complete sentence in each line. Another is the word segmentation file, which contain the standard results of the word segmentation corresponding to the undivided words in each line. Next, we import the two files into project and run the test program.

Secondly, we read a sentence from the undivided data file, at the same time, obtain the corresponding word segmentation result from the standard word segmentation file. There are traditional Chinese in the training dictionary of the word segmentation tool, so it is necessary to simplify the traditional text data. Then we add the number of the standard word segmentation results to the total number of word segmentation N.

Thirdly, it is time to segment word. We compare the word segmentation results of the algorithm with the standard word segmentation results. Next, we add the number of correct word segmentation to the total number of correct word segmentation c and add the number of incorrect word segmentation to the total number of incorrect word segments e. At this point, we have completed the word segmentation test of the on line.

Fourthly, we check if have read the last line of the file. If so, we record the total number N, the correct number of word segmentation c , the incorrect number of word segmentation e and end time of the word segmentation. We obtain time of test by the beginning time and the end time of word segmentation. Otherwise, we should read the next line of data from the file and repeat the operations of step 2 and step 3.

After the test of the word segmentation on 5, 000 data, we measure the word segmentation time of various algorithm on one millions marketing notification texts. The method of measurement is to read the data line by line from the file for word segmentation and then, statistics the beginning time of the word segmentation, the end time of the word segmentation.

3. Experiment Results and Analysis

3.1. Evaluation Results Segmentation

In this experiment, the accuracy rate p and the recall rate r and F values were used to evaluate the word segmentation results. Calculated as follows:

$$p = \frac{c}{c+e} \tag{1}$$

$$r = \frac{c}{N} \tag{2}$$

$$F = \frac{2 \times p \times r}{p+r} \tag{3}$$

Where N is the total number of words in the standard word segmentation? The c is the correct number of word segmentation. The e is the incorrect number of word segmentation. The higher values of the p and the r represent the more accurate of word segmentation. However, the p and the r may be mutually constrained, which would not have reach higher values at the same time. The F is an evaluation index, which combines the p and the r. The F comprehensively reflect the effect of word segmentation. The higher value of the F shows the better effect of word segmentation.

3.2. Word Segmentation Statistics

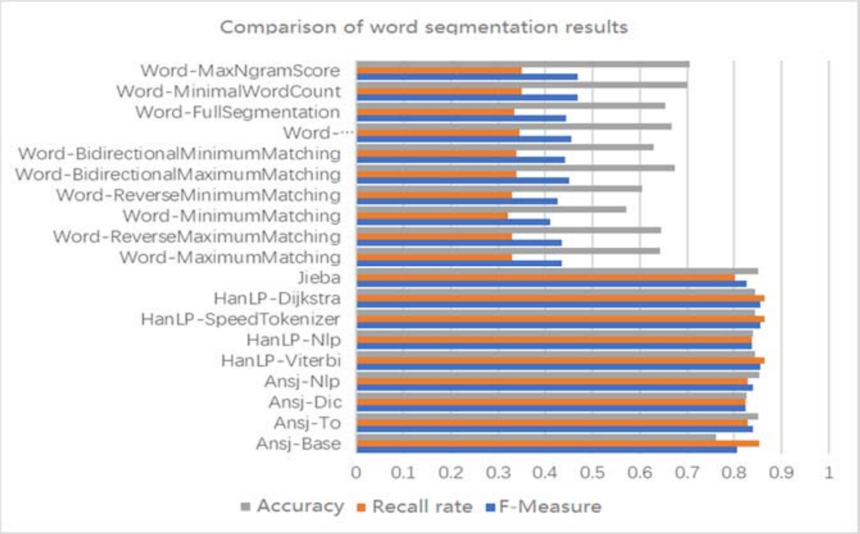

Figure 2. Comparison test results of the word segmentation tools on the third-party data set.

What we can see from the figure 2 is that the Speed Tokenizer model, the Dijkstra model and the Viterbi model of the HanLP, the Base model of the Ansj have high recall rate. The Token Streams who have high-accuracy are the Jieba, the Nlp mode and to model of Ansj. The Viterbi model, the Dijkstra model and the Speed Tokenizer mode of the HanLP have higher value on the$F$.

3.3. Statistics the Efficiency of Word Segmentation

We selected one million marketing notification texts to test the word segmentation speed of the word segmentation algorithms in those experiments as shown in figure 3. Since test time of the Word is more than fifteen minutes, it is not shown in figure 3.

From figure 3, we can see that from high to low, the word segmentation tool which has the fastest word segmentation speed respectively is: the Base mode and the To mode of the Ansj, the Speed Tokenizer mode and Viterbi mode of the HanLP and the Jieba.

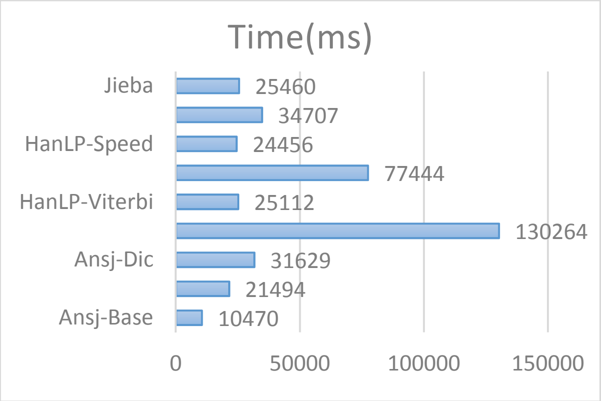

Figure 3. Comparison the word segmentation speed of the word segmentation tools.

4. Artificial Comparison

The effect of word segmentation is affected by the content of the word segmentation. Therefore, if we want to apply the word segmentation algorithm to the marketing notice text, it is necessary to comparison the word segmentation performance on the text with different word segmentation tools. We selected 5, 000 marketing notification texts and regarded the artificial word segmentation results as the standard result set in those experiments.

There are some special characters in the marketing notification text usually, such as ip, url, quantifier, time and date, etc. The special characters cannot be understood by the word segmentation tools. Therefore, we need to deal with the original text and eliminate the influence of special characters on the word segmentation before the segmentation.

Pre-process:

1. Remove all line breaks and carriage return symbol;
2. Synonymous substitution the specific format text. Ip address is replaced with “IP”. url is replaced with “URL”. All digits representing time (for example 12:00) are replaced with “TIME”. All digits representing the date (for example 2017-01-01) are replaced "DATE. All quantifiers (for example, 25kg) are replaced with “NUM ". In the process of subsequent word segmentation, the replaced items can be correctly identified by the word segmentation tools, which avoid the adverse effects of special characters on the word segmentation.

We tested the 5, 000 processed texts using the word segmentation tools. We statistics the correct number and the incorrect number of word segmentation, the accuracy, the recall and the F by comparing the results of artificial word segmentation with the results of word segmentation tools as shown in figure 4.

The results show that the HanLP and the Nlp mode of the Ansj have higher recall and F. The word segmentation tools who have higher accuracy are the HanLP, the Nlp mode of the Ansj and the Jieba.

The experimental data contain many merchant and corporate names. The merchant and corporate names have 4 to 6 Chinese characters in a single text whose average length is about 40 to 50 Chinese characters. The merchant and corporate names account for nearly 10% of the total text. Due to the particularity of the experimental data, the word segmentation effect is generally not ideal. In order to improve the accuracy of the word segmentation, a custom dictionary is added to the word segmentation tools.

The accuracy and the time of word segmentation are considered synthetically. We adopt the Viterbi shortest path algorithm of the HanLP. We segmented 5000 data mentioned above again after adding more than 6, 000 merchant and corporate names to the custom dictionary as shown in figure 5. We can see from the figure 5, the accuracy rate is 86.02%, the recall rate is 87.43%, and the F value is 86.72%. There is a significant improvement comparing the previous word segmentation effect whose accuracy is 69%, recall is 80% and F value is 75%.

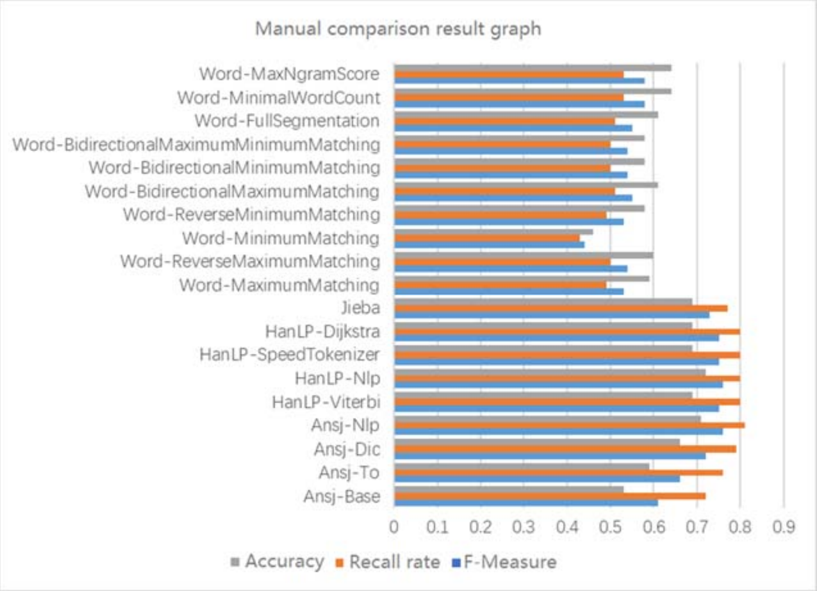

Figure 4. The results of artificial compression

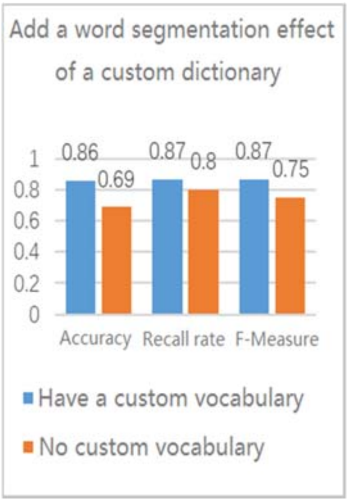

Figure 5. The word segmentation effect after adding the custom dictionary.

5. Conclusion

In this paper, we tested the performance of four open source Chinese word segmentation tools using the evaluation data of the International Chinese Language Processing Competition. The word segmentation effect of the four Chinese word segmentation tools was considered in terms of the word segmentation speed, the accuracy, the recall and the F. As can be seen from the experimental results, word segmentation results of the Word is not precise enough. Word segmentation result of the Jieba, the Ansj and the HanLP are more outstanding. The highest accuracy of word segmentation is the Nlp mode use of the Ansj, followed by the Nlp mode of the HanLP. The word segmentation effect of other word segmentation algorithms is very close. The effect of the non-Nlp mode of the HanLP is slightly better than the non-Nlp mode of the Ansj and the Jieba.

What's more, we tested word segmentation speed of each word segmentation algorithm on the one million marketing notification texts in those experiments. The experimental results show that the Word takes too long time. The Nlp mode of the HanLP and the Nlp mode of the Ansj take a little longer. The Jieba and the To Analysis mode of the Ansj are faster than the non- Nlp mode of the HanLP, whose word segmentation speed are very close. The Base mode of the Ansj is the fastest.

Finally, we selected 5, 000 marketing notice texts and removed special characters. We used the artificial word segmentation results as the evaluation criteria of word segmentation. The word segmentation results of each algorithm are compared with the artificial word segmentation results. The experimental results show that the NLP mode of the Ansj, the Speed Tokenizer mode of the HanLP, the Nlp mode of the HanLP, the Viterb mode of the HanLP 's and Dijkstra mode of the HanLP have more higher accuracy. We selected the Viterbi mode of the HanLP to the marketing notification project considering the accuracy and the speed of word segmentation. We design a custom dictionary for the text in the project. The experimental results show that the *F* of the word segmentation is increased 11% and the effect of word segmentation is significantly improved after adding the custom dictionary.

6. Acknowledgement

This work is supported by the Open topic of Key Laboratory of information transmission and distribution technology of China Electronic Technology Group Corporation fifty-fourth Research Institute (No. KX172600026), the Special research project of shaanxi education department (No. 16JK1596).

7. References

[1] Zhang L and Weiran X. research on Chinese word segmentation techniques. Software, 2012(12):103-8

[2] Zhiyan C, Xiaojie L and Shuhua Z, etc. bi-direction maximum matching method based on hash structural dictionary. *Computer science*, 2015 (S2): 49-54.

[3] Jun Z, Zhonghua Z and Wei Z. method of Chinese words rough segmentation based on improving maximum match algorithm. *Computer Engineering and Applications*, 2014 (2):124-8.

[4] Lin S. Research On Chinese word segmentation algorithm and its improvement. *Computer Knowledge and Technology*, 2017, 31: 199-200.

[5] Dongxu H and Baobao C. approaches to domain adaptive Chinese segmentation model. *Chinese Journal of Computers*, 2015, 38(2): 272-81.

[6] Jie D and Jinghui Z. research on Chinese word segmentation algorithm based on N -gram model. *Fujian Computer*, 2017, 33(5): 110.

[7] Jian W and Junni Z. applications of statistical models in Chinese text mining. *Journal of Applied Statistics and Management*, 2017, 36(4): 609-19.

[8] Adams R, Saleheen N and Thomaz E, et al. hierarchical span-based conditional random fields for labelling and segmenting events in wearable sensor data streams. *International Conference on Machine Learning*, 2016: 334-43.

[9] ICTCLAS. Chinese word segmentation tool. *http://ictclas.org, 2018.6*

[10] Qingfu W. research on Chinese word segmentation based on hidden Markov model. *Wuxian Hulian Keji*, 2016 (13): 106-7.

[11] Zhixin G and Linhui X. text classification based on hidden Markov model and semantic fusion. *Computer Applications and Software*, 2017, 34(7): 303-7.

[12] Qu J, Zhang G and Fang Z, et al. a parallel Aho-Corasick algorithm with Non-deterministic Finite Automaton Based on OpenMP. *ACN*, 2015: 52-55.

[13] Xu L, Zhang Q and Wang D, et al. research of Chinese segmentation based on MMSeg and Double Array TRIE. *Advanced Materials Research. Trans Tech Publications*, 2011, 225: 945-8.

[14] Yufeng D and Fei J. research on Chinese new word recognition in specialized field based on N-Gram. *Data Analysis and Knowledge Discovery*, 2012, 28(2): 41-7.