A Comparison of Chinese Word Segmentation on News and Microblog Corpora with a Lexicon Based Method

Yuxiang Jia[1], Hongying Zan[1], Ming Fan[1], Zhimin Wang[2]
1. School of Information Engineering, Zhengzhou University, China
2. College of Chinese Studies, Beijing Language and Culture University, China
{ieyxjia,iehyzan,iemfan}@zzu.edu.cn,wangzm000@gmail.com

Abstract

Microblog is a new and important social media nowadays. Can traditional methods deal well with Chinese microblog word segmentation? We adopt the forward maximum matching (FMM) method and design rules to recognize words with non-Chinese characters. We focus on comparing results between news text and microblog. The lexicon based method allows us to investigate well new words emerging in microblog by comparing with lexicon words. Experimental results show that the performance on microblog outperforms that on news text under the same setup, which may be a signal that microblog word segmentation is not as hard as expected.

1 Introduction

Chinese is writtern as a sequence of characters, with no boundary between words. Word segmentation or word breaking is a task to recognize words and turn a sequence of characters into a sequence of words. Because word is the basic unit of a language, word segmentation is considered as the first step of Chinese language processing.

Extensive work has been done on Chinese word segmentation. Word segmentation methods can be divided into two categories. The first category is lexicon based method. This method needs a predefined lexicon or word list. Solely based on the lexicon, maximum matching method can be used for word segmentation. Combined with labeled corpus, statistical methods can be applied (Huang and Zhao, 2007). The other category is character tagging method (Xue, 2003). This method considers word segmentation as a character position classification problem or sequence labeling problem, and applies related machine learning models.

Supervised machine learning methods need labeled data. In order to alleviate human labeling labor and utilize large scale unlabeled data, semi-supervised (Sun and Xu, 2011) and unsupervised methods (Wang et al., 2011) are also studied.

SIGHAN has organized several bakeoff tasks for Chinese word segmentation on news corpora (Emerson, 2005; Zhao and Liu, 2010), which has greatly pushed the advancement of Chinese word segmentation. This year it turns to microblog word segmentation, in the face of the great development of microblog and social network in Chinese.

Compared with news text, microblog has more words containing non-Chinese characters, like numbers, alphabets, symbols, etc. Such words are of great number but can be classified into different types and recognized respectively based on rules. Chinese character sequences in microblog are relatively shorter than those in news text. So a traditional segmenter enhanced by a special process of non-Chinese characters may have a good performance.

In this paper, we propose a lexicon and rule based method, using forward maximum matching (FMM) method to recognize Chinese words and regular expressions to recognize words with non-Chinese characters. FMM is simple and fast implemented, and is always taken as a baseline method. Here we take FMM to compare the baseline performance on corpora of different styles.

The rest of this paper is organized as follows. Section 2 describes the word segmentation process. Section 3 gives experimental results and analysis, including comparison of different lexicons, comparison of different corpora, and comparison of experimental results. Conclusions are given in section 4.

Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 95–98,
Tianjin, China, 20-21 DEC. 2012

2 Segmentation Method

The word segmentation process is shown in figure 1. Preprocessing step combines non-Chinese character sequence as one character, just like a Chinese character.

FMM step takes forward maximum matching method for word segmentation. The maximum word length is set to be 7. The lexicons used here will be discussed in the next section.

Chinese character words are recognized in the FMM step. In the next step, with a rule based method, non-Chinese character sequences are divided into meaningful words, such as URLs, Emails, English words, numbers, etc.

In the postprocessing step, some words need to be combined to make a final word. For example, word sequence "一" (one), "九" (nine), "九" (nine), "八" (eight), "年" (year) should be combined as a word "一九九八年" (the year 1998). Other processes can also be added into this step.

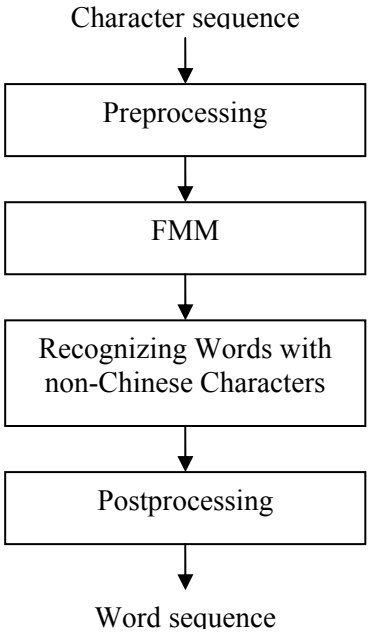

Figure 1. Word segmentation process

3 Experiments and Analysis

Several popular Chinese lexicons are compared to explore the impact of lexicons on the FMM method. Word distributions are compared between news and microblog corpora. Experimental results with respect to different metrics are compared and analyzed.

3.1 The Lexicons

The Chinese lexicons used here are as follows:

1. The Grammatical Knowledge-base of Contemporary Chinese (GKB) (Yu et al., 2003). GKB organizes words into different categories and provides comprehensive grammatical knowledge for each word. The version of GKB used here has a vocabulary of 74188 word types.

2. HowNet (HN) (Dong and Dong, 2006). HowNet encodes relations between concepts into a semantic network. It provides a definition for each concept as a combination of basic sematic units. HowNet version 2000 has a vocabulary of 55496 word types.

3. TongYiCiCiLin (CiLin) (Che et al., 2010). CiLin is a semantic lexicon. A concept is represented as a synonym set, and all concepts are organized into trees of the same height. CiLin has a vocabulary of 77457 word types.

4. Lexicon of Common Words in Contemporary Chinese (LCW) (Li et al., 2008). LCW is a list of words frequently used in various corpora, including news, literature, etc. LCW has a vocabulary of 55731 word types.

The sizes of vocabulary intersection of different lexicons are shown in table 1. We can see that the vocabularies are different greatly from each other. There are only 41419 words in common in the first three lexicons and 34540 words in common in all the four lexicons, while there are 104150 distinct words in total in the four lexicons.

	GKB	HN	CiLin	LCW
GKB	74188	43740	61780	45780
HN	-	55496	45652	37601
CiLin	-	-	77457	45612
LCW	-	-	-	55731
CGH	41419			-
CGHL	34540			

Table 1. Size of vocabulary intersection of different lexicons

3.2 Data Sets

The data sets used here are as follows:

News corpus. We choose Peking university test set of the 2nd International Chinese Word Segmentation Bakeoff as the news corpus. This corpus contains 1944 sentences and 104372 words (13148 types).

Microblog corpus. We choose the sample corpus of the bakeoff task this year as the test set,

which contains 503 sentences and 20058 words (5047 types).

Statistics about the two corpora are shown in table 2. Column names are out-of-vocabulary rate (OOVR), average word length (AWL), rate of words with non-Chinese characters (RWNC). Let the union of the above four lexicons as our lexicon (104150 word types), we can see that microblog text contains more out-of-vocabulary words and much more words with non-Chinese characters. The average word length is shorter in microblog text.

	OOVR	AWL	RWNC
News	9.61%	2.13(type)/ 1.61(token)	2.61%
Microblog	13.91%	1.79(type)/ 1.38(token)	7.98%

Table 2. Statistics of news and microblog corpora

3.3 Results

Metrics used to evaluate system performance are Precison (P), Recall (R), F1-mearsure (F1), R_{IV}, R_{OOV}. R_{IV} is the recall of in-vocabulary word, and R_{OOV} is the recall of out-of-vocabulary word.

	P	R	F1	R_{OOV}	R_{IV}
GKB_m	87.20	**91.71**	**89.40**	79.30	**96.22**
GKB_n	85.31	91.01	88.07	73.37	96.10
$CiLin_m$	87.40	90.69	89.01	**81.44**	93.95
$CiLin_n$	86.61	90.06	88.30	77.76	93.37
HN_m	83.48	88.56	85.94	58.45	94.69
HN_n	82.19	88.09	85.04	42.22	94.98
LCW_m	83.50	89.13	86.22	74.51	95.12
LCW_n	79.60	87.62	83.42	65.35	95.55
$Union_m$	**87.67**	89.49	88.57	70.44	92.56
$Union_n$	86.60	88.32	87.45	57.28	91.62

Table 3. Experimental results

Experimental results are shown in table 3. The numbers in bold indicate the highest values of each metric. GKB_m and GKB_n mean that we use GKB as the lexicon. $Union_m$ and $Union_n$ mean that we use the union of all the four lexicon as the lexicon. The subscript "m" denotes result on microblog and "n" denotes result on news corpus. We can see that the all the results on microblog outperform those on news corpus. The results of the metric R_{IV} indicate that even in-vocabulary words are better recognized in microblog. GKB and CiLin achieving better results than lexicon union shows that the lexicon is not the larger the better for FMM. Lexicon needs filtering.

The official test data contains 5000 pieces of microblog. The evaluation metrics are Precision (P), Recall (R), F1-measure (F1), number of correct sentence (CS), correct sentence rate (CSR). The lexicon for our submitted system is composed of the union of the above four lexicons and the word list of the sample data. The official result is shown in table 4.

P	R	F1	CS	CSR
89.84	90.83	90.33	1256	25.12%

Table 4. The official result

4 Conclusions

This paper proposes a simple, lexicon based method for Chinese microblog word segmentation. By comparing results on news and microblog corpora, we find that this baseline method achieves better performance on microblog corpus. This may be a signal that microblog word segmentation is not as hard as expected. In addition, lexicon based method makes it easy to investigate new words emerging in the new media. Lexicon quality is an important factor influencing the performance.

The performance can be improved by adding more rules and carefully enlarging lexicon vocabulary. This simple and labeled-corpus-free method can provide a baseline for statistical methods, which may better utilize contextual information to tackle OOV and ambiguity.

Acknowledgements

This work is partially supported by grants from the National Natural Science Foundation of China (No.60970083, No.611700163), the China Postdoctoral Science Foundation (No.2011M501184), the Postdoctoral Science Foundation of Henan Province, China (No.2010027), the Outstanding Young Talents Technology Innovation Foundation of Henan Province, China (No.104100510026), and the Open Projects Program of National Laboratory of Pattern Recognition (No.201001116). We are grateful to the bakeoff organizers who provide such a good opportunity for research on Chinese word segmentation on microblog corpora.

References

Wanxiang Che, Zhenghua Li, and Ting Liu. 2010. *LTP: A Chinese Language Technology Platform.* In Proceedings of the COLING 2010: Demonstrations, pp. 13-16.

Zhendong Dong and Qiang Dong. 2006. *HowNet and the Computation of Meaning.* World Scientific.

Thomas Emerson. 2005. *The Second International Chinese Word Segmentation Bakeoff*. In Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, pp. 123-133.

Changning Huang and Hai Zhao. 2007. *Chinese Word Segmentation: A Decade Review.* Journal of Chinese Information Processing, 21(3): 8-19.

Xingjian Li, et al. 2008. *Lexicon of Common Words in Contemporary Chinese.* The Commercial Press.

Weiwei Sun and Jia Xu. 2011. *Enhancing Chinese Word Segmentation Using Unlabeled Data*. In Proceedings of the EMNLP2011, pp. 970-979.

Hanshi Wang, et al. 2011. *A New Unsupervised Approach to Word Segmentation*. Computational Linguistics, 37(3): 421-454.

Nianwen Xue. 2003. *Chinese Word Segmentation as Character Tagging.* International Journal of Computational Linguistics and Chinese Language Processing, 8(1): 29-48.

Shiwen Yu, et al. 2003. *The Grammatical Knowledge-base of Contemporary Chinese, AComplete Specification* (*2nd edition*). Tsinghua University Press.

Hongmei Zhao and Qun Liu. 2010. *The CIPS-SIGHAN CLP2010 Chinese Word Segmentation Backoff.* In Proceedings of the CIPS-SIGHAN CLP2010, pp. 199-299.