ICIC Express Letters
Part B: Applications
Volume **8**, Number **2**, February **2017**

ICIC International ©2017 ISSN 2185-2766
pp. 313–318

RESEARCH ON CHINESE SEPARABLE WORDS BASED ON LARGE-SCALE CORPUS

Yuxi Zhao[1], Yanhui Gu[1,2], Min Gu[1], Xuri Tang[3], Junsheng Zhou[1,2] and Weiguang Qu[1,2,*]

[1]School of Computer Science and Technology
Nanjing Normal University
No. 1, Wenyuan Road, Qixia District, Nanjing 210046, P. R. China
*Corresponding author: wgqu@njnu.edu.cn

[2]Jiangsu Research Center of Information Security and Privacy Technology
Nanjing 210097, P. R. China

[3]Foreign Language School
Huazhong University of Science and Technology
No. 1037, Luoyu Road, Hongshan District, Wuhan 430074, P. R. China

Received July 2016; accepted October 2016

Abstract. *This paper proposes a method to extract all separable words in order to construct a separable wordlist, which is used to improve the precision of existing syntactic analysis systems. The method extracts all the two-character separable words from a corpus compiled from 1991 to 2004 Xinhua news and designs rules to improve the precision of separable words extraction to 79.32%. Based on the observation that about 14.49% syntactic relations between the separable words are wrongly analyzed in existing syntactic analysis systems (e.g., Language Technology Platform, LTP), the separable wordlist is used to improve the precision of the syntax relation annotation.*
Keywords: Separable words, Syntactic analysis, Natural language processing

1. **Introduction.** Chinese separable words refer to the combination of two characters that are not always continuous in their usage, such as "shuo1hua4 (speak)" and "xi3zao3 (shower)". Liu [1] quoted the opinion of Chen – these words are not individual words, but often contain other ingredients between them.

Since 1980s, people began to direct their research to separable words. Zhao and Zhang [2] investigate the internal structures and the relationships of separable words, putting forward four criteria for separable word identification. Compared with previous research, this method is more scientific, but it still has its deficiencies; Liu [3] proposes a theory of "mode-block" which is able to analyze the main structures of common separable words; Wang and Wang [4] propose that light verbs hypothesis plays a key role in understanding the sentences which have separable words. They provide to view separable words from the perspective of light verbs and summarize five kinds of free structure types for separable words.

In recent years, more scholars began to focus their studies on the identification of separable words. Wang et al. [5] analyze the separable words based on large-scale corpus. After summarizing the modes of separable words, they incorporate the modes into the rule system; according to the degrees of dispersion, Ren and Wang [6] used statistical analysis to summarize separable words and analyzed the major forms of separation of separable words. Based on the results, we can calculate the frequency of separable words; Zhou and Hu [7] have done lots of research on analyzing the characteristics of the extensive forms of verb-object and coordinate compound separable words. They pointed out that separable wordlists should be applied in Chinese information processing system and special marks

should be used to distinguish them from other types of words. Xu [8] implemented a recognition algorithm which is based on the common characteristic between separable words and long-distance phenomena in Chinese. Feng [9] also systematically researched the various forms of separable words. He created a separable wordlist and a corpus in which separable words are tagged and then he designed a system which can recognize separable words automatically.

In China, the research of separable words received an increasing attention. The field is gradually extended to the field of Chinese information processing. However, internationally, the research of separable words is mainly conducted from the perspective of language information processing. From this point of view, the annotation of separable words belongs to the annotation of multi-word combination, and it is mainly concerned with the noun phrase; Justeson and Katz [10] use regular expressions to extract noun phrases; Dagan [11] identifies multi-word forms according to the frequency of combination of the training corpus; Magnini et al. [12], Diab and Resnik [13] and Cruys and Apidianaki [14] make use of different models to deal with words.

The rest of this paper is organized as follows: Section 2 describes the method of automatic identification; Section 3 describes the results of analysis in Section 2; Section 4 summarizes the application and usage of the method; Section 5 concludes the paper.

2. **Automatic Identification Methodology.**

2.1. **Wordlist construction.**

We will not use the separable words from "modern Chinese dictionary" as a wordlist. Instead, we use all the words appearing in the news texts. The corpus used in the present research is compiled from Xinhua news ranging from 1991 to 2004, which contains more than 300 million word tokens. The number of word types is more than 620,000. We extract all the two-character words whose frequency is higher than 10. The wordlist consists of 54,166 two-character words.

2.2. **Potential separable words identification.**

2.2.1. *Text pretreatment.* This module contains two sub-modules:

(1) Sentence segmentation. This sub-module is essential because the two elements of a word may be split by punctuations. We should start a newline once we meet a punctuation.

(2) Word segmentation and part-of-speech tagging. For example, in the sentence "yi3jing1/d shang4/v le/u dang4pu4/n (already go to the pawnshop)", if we could not correctly segment the words, we may mistake "shang4dang4 (be fooled)" as a potential separable word.

2.2.2. *Algorithm description.*

(1) Input pre-processed text, output the list in which the elements are sentences;

(2) Input each sentence, and output the list in which the elements are words;

(3) Combine all individual words from the list. For example, from the sentence "ta1/r de/u xing2wei2/n bang1/v le/u wo3/r yi2ge4/m da4/a mang2/n (his behavior plays an important role in my work)", we can get 21 combinations, such as "ta1de", "ta1bang1" and "ta1le", ..., "bang1mang2", ..., "da4mang2";

(4) Compare them with the wordlist, getting "bang1mang2", "da4mang2" as potential separable words.

2.3. **Separable words selection.**

Selecting separable words is based on the construction patterns of the words, as is given in Table 1. The matching rules are: (1) The part of speech is v or vn; (2) The construction patterns are v+n, a, v+v+v, n+a and n+v. If one candidate string matches the rules above, we will treat it as a separable word.

Table 1. Construction pattern of correct separable words

Correct words	The property of the elements	Correct words	The property of the elements
ban4shi4(v)	v+n	bang1mang2(v)	v+v, v+a
cao1xin1(v)	v+n, n+n	chang4ge1(v)	v+n
chao3jia4(v)	v+v	chu1guo2(v)	v+n
chu1li4(v)	v+n	chuan3qi4(v)	v+n
da3zhang4(v)	v+n	dang1bing1(v)	v+n
tan2hua4(v, vn)	v+n	fang4jia4(v)	v+a
jian4mian4(v)	v+n	kai1hui4(v)	v+n
kan4bing4(v)	v+n	qi2ma3(v)	v+n
qing3jia4(v)	v+a	shang4ke4(v)	v+n
tiao4wu3(v)	v+n, v+v	xi3zao3(v)	v+v
xia4xue3(v)	v+n	xia2yu3(v)	v+n
you2yong3(vn)	v+n	zhao4xiang4(v)	v+n
shou3ruan3(v, a)	n+a	shou3sheng1(v)	n+a
tian1liang4(v)	n+v	xin1tiao4(v, vn)	n+v

Table 2 gives some examples, which are taken from the Xinhua news in 1996. In the table, it is obvious that only "la1shou3" and "ting1shuo1" match the two rules. While the other words such as "bu4yi1" and "yi3wei2" match the rule (1), and they do not match rule (2). So we cannot ascertain that they are separable words.

Table 2. Construction pattern of part potential separable words

PSW	The property of the elements	PSW	The property of the elements
yi3wei2(v)	p+v	zai4nei4(u)	p+f
jin4nian2(t)	a+q	you3ren2(r)	v+n
jiu4shi4(d)	d+v	yi1zhong1(j)	m+f
zhe4ge4(r)	r+q	dao1shi2(d)	p+ng
duo1nian4(m)	m+q	you3ming2(a)	v+q
zhe4ci4(r)	r+q	jin4lai2(d)	a+f
ting1shuo1(v)	v+v	yi2dui4(m)	m+p
bu4yi1(v)	d+m	yin1er2(c)	p+c
zai4chang3(a)	p+q	dui4hao4(vd)	p+q
nian2jian1(f)	q+f	la1shou3(vn)	v+n

Our experiment dataset comes from Xinhua news from 1991 to 2004. The result of the number of potential separable words and the number of separable words extracted from the corpus are shown in Table 3. Among them, PSW refers to the potential separable words that are confirmed to be right approving through Section 2.2.2. SW refers to the separable words which match the two rules mentioned in Section 2.3.

3. **Results Analysis.**

3.1. **Evaluation.** Because the size of the corpus is huge, it is not possible to proofread the corpus. Thus to evaluate the recall is not possible. This paper is mainly focused on the accuracy of results. According to the appendix in Dr. Wang's paper, we proofread all the separable words we selected, which are written down as proofread separable words.

Table 3. The results we extract

Years	The number of PSW	The number of SW
1991	8,107	323
1992	7,530	408
1993	5,618	316
1994	6,822	332
1995	5,612	295
1996	4,988	267
1997	6,084	301
1998	5,548	299
1999	6,300	324
2000	5,083	261
2001	5,390	304
2002	5,597	274
2003	3,135	183
2004	6,003	339

The calculation formula is as follows:

$$\text{Accuracy} = \frac{\#\text{proofread separable words}}{\#\text{separable words}} \tag{1}$$

By analyzing the corpus from 1991 to 2004, we selected separable words whose frequency is more than 10 for analysis. The separable words and the corresponding proofread separable words (PFSW) are given in Table 4. Among them, we acquire 1676 proofread separable words among 2,113 no-repeat separable words in all news and the accuracy reached to 79.32%.

Table 4. Accuracy analysis

Years	SW	PFSW	Accuracy (%)
1991	323	264	81.73
1992	408	321	78.67
1993	316	243	76.89
1994	332	269	81.02
1995	295	241	81.69
1996	267	208	77.90
1997	301	235	78.07
1998	299	248	82.94
1999	324	243	75.00
2000	261	195	74.71
2001	304	238	78.28
2002	274	203	74.08
2003	183	124	67.76
2004	339	259	76.74

3.2. **Separable words types.** We roughly divided these 1676 words into three different types.

(1) **Verb-Object pattern**: This kind of words appears 1,611 times, and the number is the largest, 96.12% in all.

(2) **Verb-Complement pattern**: This kind of words appears 63 times, 3.76% in all.

(3) **Subject-Predicate pattern**: This kind of words is less than other types of words, only twice, 0.12% in all. The two words we extracted is: "xin1ruan3 (softhearted)" and "lian3hong2 (blush)".

In the remaining 437 words, 246 words are wrong results, while for others, as Wang's paper does not cover all separable words, there are 191 correct separable words but we cannot judge according to his paper.

4. **Syntactic Analysis System Accuracy.** In natural language processing (NLP), we need to use syntactic analysis systems to analyze sentences. However, separable words pose as confusion in the process. According to the separable wordlist we built, we can correct the wrong results of the syntactic analysis system, so the efficiency of the syntactic analysis system can be improved.

In Section 3.2, we summarized several types about the property of the elements of words. Among them, the Verb-Object pattern and Verb-Complement pattern contain more than 99% of the whole separable words. So according to these two forms we can get the accuracy of the syntactic analysis system. In other words, we need to select words whose relationship between two elements is VOB or CMP and the relationships of the words must be marked correctly.

We choose the separable words whose frequency is more than 10 of Xinhua news in 1996, summing up all the sentences in which these words occur. Then we choose 1048 sentences, and filter 511 sentences that are no-repeated. Here, we use language technology platform (LTP) to analyze these 511 sentences.

Table 5. The types of sentences

Segment	Type	Mark	Number	Ratio (%)
Wrong	VOB, CMP	Do not discuss	54	10.57
	others	Wrong	15	2.94
	ATT		8	1.57
Right	ATT		38	7.44
	There is no relationship		13	2.54
	VOB	Right	372	72.79
	CMP		11	2.15

As Table 5 shows, there is no need to discuss the first kind of sentences. The error is caused by the wrong word segmentation, so we cannot use our separable wordlist to solve this kind of error. Among the remaining 457 sentences, row 2 to row 5 show errors which are 74 sentences in total, and row 6 to row 7 show correctness which is 383 sentences in total. The accuracy that we calculate is 74.94%. According to the separable wordlist we built, we can revise the errors found in row 2 to row 5. So the accuracy after correction can reach to about 89.43%.

5. **Conclusions.** In this paper, we construct a separable wordlist according to the algorithm proposed in the paper, which is used to improve the precision of existing syntactic analysis system. There are still some deficiencies in our work, given as below.

a. Lack of standard corpus. As the quantity of authority corpus is not too much, and there are lots of errors, it may result in deviation in dealing with the corpus.

b. How to deal with the unknown words. In recent years, there are many words which had not been marked due to the fashion of network. So how to deal with these kinds of unknown separable words is also a research direction in the future we have to consider.

Acknowledgment. This work is partially supported by Chinese National Fund of Natural Science under Grant 61272221, Jiangsu Province Fund of Social Science under Grant 12YYA002, and Natural Science Research of Jiangsu Higher Education Institutions of China under Grant 14KJB520022, 15KJA420001. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] B. Liu, *Research on Automatic Recognition of Separable Words Based on Corpus*, Hebei University, 2015.

[2] S. H. Zhao and B. L. Zhang, The determination of separable words and the property of separable words in Chinese, *Language Teaching and Research*, no.1, pp.40-51, 1996.

[3] X. Liu, *The "Mode-Block" Theory Analysis in VOB Type of Modern Chinese Words in Chinese*, Sichuan Normal University, 2013.

[4] G. S. Wang and J. Wang, Research the modern Chinese from light verbs in Chinese, *Journal of Central China Normal University (Humanities and Social Sciences)*, vol.50, no.2, pp.101-105, 2011.

[5] H. F. Wang, S. Li and T. J. Zhao, Processing of "LIHECI" in Chinese-English machine translation, *Journal of the China Society for Scientific and Technical Information*, vol.18, no.4, pp.303-307, 1999.

[6] H. B. Ren and G. Wang, The analysis of split word in modern Chinese based on the large-scale corpus, *Linguistic Sciences*, vol.4, no.6, pp.75-87, 2005.

[7] W. H. Zhou and J. Q. Hu, The processing strategies of separable words in Chinese information processing in Chinese, *The Paper of the Three Gorges University (Humanities and Social Science Edition)*, vol.32, no.6, pp.39-41, 2010.

[8] J. S. Xu, *Research of the Separable Words and Long-Distance Match in Chinese*, Master Thesis, Harbin Institute of Technology, 2003.

[9] X. H. Feng, *The Automatic Identification in the Form of Extension in Modern Chinese in Chinese*, Master Thesis, Beijing Normal University, 2009.

[10] J. S. Justeson and S. M. Katz, Technical terminology: Some linguistic properties and an algorithm for identification in text, *Natural Language Engineering*, no.1, pp.9-27, 1995.

[11] I. Dagan, Similarity-based models of word cooccurrence probabilities, *Machine Learning*, vol.34, pp.43-69, 1998.

[12] B. Magnini, C. Strapparava, G. Pezzulo et al., The role of domain information in word sense disambiguation, *Natural Language Engineering*, vol.8, no.4, pp.359-373, 2005.

[13] M. T. Diab and P. Resnik, An unsupervised method for word sense tagging using parallel corpora, *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, pp.255-262, 2002.

[14] T. V. de Cruys and M. Apidianaki, Latent semantic word sense induction and disambiguation, *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics*, pp.1476-1485, 2011.