Towards Unified Chinese Segmentation Algorithm

Fu Lee Wang, Xiaotie Deng and Feng Zou
Department of Computer Science, City University of Hong Kong,
Kowloon Tong, Hong Kong
E-mail: {flwang,csdeng}@cityu.edu.hk, phenix@cs.cityu.edu.hk

Abstract

As Chinese is an ideographic character-based language, the words in the texts are not delimited by spaces. Indexing of Chinese documents is impossible without a proper segmentation algorithm. Many Chinese segmentation algorithms have been proposed in the past. Traditional segmentation algorithms cannot operate without a large dictionary or a large corpus of training data. Nowadays, the Web has become the largest corpus that is ideal for Chinese segmentation. Although the search engines do not segment texts into proper words, they maintain huge databases of documents and frequencies of character sequences in the documents. Their databases are important potential resources for segmentation. In this paper, we propose a segmentation algorithm by mining web data with the help from search engines. It is the first unified segmentation algorithm for Chinese language from different geographical areas. Experiments have been conducted on the datasets of a recent Chinese segmentation competition. The results show that our algorithm outperforms the traditional algorithms in terms of precision and recall. Moreover, our algorithm can effectively deal with the problem of segmentation ambiguity, new word (unknown word) detection, and stop words.

1. Introduction

In information retrieval, a document is traditionally indexed by the frequency of words within documents (Rijsbergen, 1979). For English and other western languages, the segmentation of texts is trivial. Texts in those languages can be segmented into words by using spaces and punctuations as word delimiters. However, many Asian languages do not delimit the words by spaces. The absence of word boundaries poses a critical problem for information retrieval in Asian languages.

In Chinese information retrieval, segmentation of the texts is required before indexing of a document. Many segmentation algorithms have been developed for Chinese language. Typically, algorithms for Chinese segmentation fall into three categories: dictionary-based approaches, statistics-based approaches and hybrid approaches. The dictionary-based approaches segment the texts by matching the text-trunks against entries of a large machine-readable dictionary of Chinese words. (Wu & Tseng, 1993) The major shortcoming of dictionary-based approaches is identification of new words (Chen & Ma, 2002). Statistics-based approaches or hybrid approaches are proposed to solve the problem of new words detection (Chen & Ma, 2002; Sproat & Shih, 1990). However, the performances of statistics-based approaches significantly depend on the size of the training corpus. A large corpus in the same domain is not always available. Therefore, a corpus-free segmentation algorithm is more desirable.

The Chinese languages in different geographical areas are different from each other greatly. It is extremely difficult to develop an unified segmentation algorithm for different areas. The accuracies of traditional segmentation algorithms vary greatly with dataset from different areas. Web has become the largest corpus because of the explosion of Internet. Most of the online documents have been indexed by search engines. The search engines are widely used for search of relevant documents. The number of hits returned by search engines is also employed to provide statistics for some information retrieval tasks (Mihalcea & Moldovan, 1999). In Web search, if two events or two concepts co-occur in a Web page frequently, we assume that there is a strong association between these two events or concepts. In statistics-based approaches of segmentation, statistical data of a large corpus is used to calculate the association of adjacent characters and the segmentation points are determined accordingly. Web search and Chinese segmentation are similar in the sense that both of them discover the association between two objects. As a result, the Web search technique is a promising application for segmentation of Chinese texts.

The search engines maintain huge databases of the online documents. Although they do not segment the texts into proper words, they have recorded some important figures of the documents, such as the frequency of string in the document. Besides, the search engines update their databases regularly and frequently. Web data are domain independent and geographical independent. The Web together with search engines provide important resources for Chinese segmentation.

In this paper, we propose a segmentation algorithm based on the number of hits returned by search engines. The algorithm uses statistical data of adjacent Chinese characters to segment Chinese texts. Experiments on datasets for a recent international segmentation competition have been conducted to evaluate the accuracy of our algorithm. Experimental results show that segmentation algorithm by web mining outperforms the traditional segmentation algorithms significantly. Moreover, the novel algorithm is geographical area-independent, which can segment the Chinese documents from different geographical areas effectively. It can also deal with the problem of segmentation ambiguity, new word detection, and stop words.

2. Traditional Segmentation

A Chinese text appears to be a linear sequence of non-spaced ideographic characters that are called morphemes. Usually, a Chinese word consists of more than one ideographic characters and the number of characters varies. Segmentation of texts into words is essential for indexing of Chinese documents. Chinese segmentation is probably the single most widely addressed problem in the literatures on Chinese language processing (Wu & Tseng, 1993). Many segmentation algorithms have been developed. Typically, they fall into three categories: dictionary-based approaches, statistics-based approaches and hybrid approaches.

2.1. Difficulties in Segmentation

Although Chinese languages are used in different areas, they are different from each other greatly. Simplified characters are used in Mainland China, while Complex-form characters are used in other areas. Moreover, different encoding schemes are used in different areas, i.e., GB for Mainland China, BIG-5 with Hong Kong Supplementary Character Set for Hong Kong, BIG-5 for Taiwan, GBK, EUC-CN, Unicode are also used for other areas. Among the encoding systems, BIG-5 and GB are the most popular ones.

Mainland Standard

[中華人民共和國]
[People Republic of China]

ROCLING Standard and U Penn Standard

[中華]	[人民]	[共和國]
[China]	[People]	[Republic]

Figure 1: Example of Different Segmentation Standards

There are numerous difficulties in Chinese segmentation. In western languages, there are only a small number of characters. However, the Chinese language does not have a fixed number of characters. The BIG-5 encoding system in Taiwan and Hong Kong defines about 13,500 complex-form characters and the GB encoding system in Mainland China defines about 7,700 simple-form characters (Lunde, 1998). On the other hand, there are different segmentation standards in different areas (Sproat & Emerson, 2003), i.e., Mainland Standard, ROCLING Standard, University of Pennsylvania / Chinese Treebank Standard, etc. A phrase or a sentence is segmented differently in different areas. Taking the phrase "中華人民共和國" (People Republic of China) as an example (Figure 1), it is segmented as a single word in Mainland Standard, but it is segmented as three words in other standards (Sproat & Shih, 2002).

Phrase:	中國文化工業	(Chinese Cultural Industry)
Bi-gram:	[中國]	(China)
	[國文]	(Chinese)
	[文化]	(Culture)
	[化工]	(Chemical Industry)
	[工業]	(Industry)
Tri-gram:	[中國文]	(Chinese Language)
	[化工業]	(Chemical Industry)
Quad-gram:	[中國文化]	(Chinese Culture)
	[文化工業]	(Cultural Industry)

Figure 2. Example of Segmentation Ambiguity

The meaning of the phrase remains unchanged in different segmentation standards. They just reflect the fact that different standards have different identifications of word boundaries. Some consider the phrase as a single word while some consider the phrase as a composite phrase of three words, but they represent the same concept. In some cases, different segmentations of a sentence may have different meanings. The problem of different ways to segment the sentence is known as segmentation ambiguity. This problem is extremely serious in Chinese language. Take the phrase "中國文化工業" (Chinese Cultural Industry) as an example (Figure 2), every adjacent bi-gram is a valid word, and there are also some tri-grams and quad-grams in this phrase with 6 characters only.

Automatic identification of words in Chinese texts is a challenging task. Different Chinese segmentation algorithms have been proposed in the past, but none of them has been commonly accepted as a standard.

2.2. Dictionary-Based Approaches

The dictionary-based approaches are the most straightforward approaches for Chinese segmentation (Sproat & Shih, 2002; Wu & Tseng, 1993). They can be implemented based on a machine-readable dictionary. Texts are divided into trunks with n consecutive characters that are called n-grams. The n-grams are matched against the entries in a large dictionary of Chinese words to determine the segmentation points.

The major concern for dictionary-based approaches is how to deal with segmentation ambiguities. The most popular approach dealing with segmentation ambiguities is the maximum matching method (Sproat & Shih, 1990). Starting from the beginning of the text, the maximum matching method groups the longest substring that matches a dictionary entry as a word. This procedure continues until the text is completely segmented.

The major shortcoming of dictionary-based approaches is the identification of new words (also called as unknown words), which are some words that are not listed in an ordinary dictionary (Chen & Ma, 2002). Study on a 5 million words Chinese corpus with proper word segmentation has found that 3.51% of Chinese words are not listed in the most powerful machine-readable dictionary (Chen & Ma, 2002). Moreover, there will be new words generated everyday, and there is a certain delay before the new word is included in a dictionary. As a result, more advanced techniques are required for segmentation of Chinese texts.

2.3. Statistics-Based And Hybrid Approaches

Statistics-based approaches or hybrid approaches are proposed to solve the problem of unknown words. Given a large corpus of Chinese texts, the statistics-based approaches measure the statistical association of characters in the corpus. The texts are then segmented according to the associations of adjacent characters (Sproat & Shih, 1990). There are different measurements of the associations. Among them, the mutual information is the most popular approach (Yang & Li, 2005).

The statistics-based approaches are weak in dealing with stop words. Stop words are words that appear frequently in the corpus but they do not convey any significant information to the document, for example, "of", "the" etc. (Luhn, 1958). In Chinese language, the situation is very complicated because Chinese language is character based and characters are put together to form words. For example, the character "的" (of) is commonly accepted as a stop word in Chinese. However, if it put together with the character "士", then they form a word "的士" (taxi).

Therefore, no Chinese stop word list has been commonly accepted yet. Sometimes, some stop words are highly associated with another character statistically, but

they do not form a valid word. Practically, filtering of stop words is impossible because there is no well accepted Chinese stop word list available.

On the other hand, the statistics-based approaches segment the texts based on the statistical figures of a large corpus of documents. Thus, the accuracy depends on the size of the training corpus significantly. As it is well known that a histogram of the words in a reasonable size corpus for any language will reveal a Zipfian distribution (Zipf, 1949). As a result, there will be a large number of words which occur only once in a corpus. In English, it is found that 40% of the words occur just once in the 37 million words corpus (Sproat & Shih, 1990). Since Chinese language is character-based, statistics on character frequency found that 11% of the characters occur only once in the 10 million characters corpus (Sproat & Shih, 1990). Among the large number of Chinese characters, only about 6,000 characters are frequently used (Ge, Wanda & Padhraic, 1999). It will be very difficult for a segmentation algorithm to segment the text, if some characters in the text are unseen in the training corpus. Therefore, the performance of segmentation is deeply affected by the size of corpus.

Moreover, there are different vocabularies in different geographical areas. One word in different areas may carry some totally different meanings. On the other hand, an object or a concept is described using different words in different areas. The vernacular differences in their languages have a deep impact on Chinese segmentation applications. For example, the statistical figures of the corpus in one area cannot be applied to segmentation of documents in another area. In order to capture the language differences in difference areas, the corpus must include documents from different areas. Moreover, the corpus needs to be updated constantly in order to solve the problem of new words. The web data is the ideal corpus for Chinese segmentation, because it includes the documents from different areas. Moreover, most of the online documents have been indexed by search engines and the search engines will update the database frequently and regularly.

3. Pinyin Search from Web

Traditionally, Chinese language is regarded as a morphological language with ideographic characters. The sociological applications in the twentieth century attempted to Romanize Chinese orthography (Sproat & Shih, 1990). In Romanization scheme, pinyin of Chinese characters are grouped into word-sized trunks. Word boundaries are therefore clearly indicated in Romanized pinyin of Chinese texts. However, the techniques of Romanized pinyin have not yet been employed in Chinese segmentation.

Under traditional pinyin scheme, the pronunciation of each Chinese character is written as a word of pinyin syllables, and words of syllables are disjointed from each other (Figure 3). Under Romanization scheme, the pinyin of individual character are grouped into word-sized trunks. Considering the phrase "中華人民共和國" (People Republic of China) as an example (Figure 3), the pinyin is clearly segmented into words. Each English word in the example is corresponded to one Chinese word except the stop word ("of" is not translated in this example).

English:	People Republic of China					
Chinese:	中華人民共和國					
Traditional Pinyin:	Zhong	Hua	Ren	Min	Gong	He Guo
Equivalent Chinese:	中	華	人	民	共	和 國
Romanized Pinyin:	Zhonghua		Renmin		Gongheguo	
Equivalent Chinese:	中華		人民		共和國	
Equivalent English:	China		People		Republic	

Figure 3: Example of Traditional Pinyin and Romanized Pinyin in Chinese Language

The Romanization scheme of Chinese language indicates the boundaries of words clearly in the texts. It can be utilized for segmentation. However, it is difficult to find a large corpus of Chinese texts with Romanized pinyin. Therefore, no segmentation technique has been developed based on pinyin so far. As the development of Internet, the Web has become the largest corpus in the world. Some search engines store Chinese documents with Romanized pinyin. As a result, their databases provide an important resource for Chinese segmentation.

The search engines do not segment the texts into words. Documents are indexed as word trunks of all possible *n*-grams by brute force (Figure 4). They search documents purely by string matching of the query against the *n*-gram database. However, the *n*-gram may be just a sequence of characters across the word boundary. In Chinese language, if we search a permutation of any two Chinese characters, the search engines will return some hits. However, the permutation by itself may be meaningless. If we search by pinyin, the results will be totally different.

Character-Based *n*-grams Indexing

Uni-gram:	中 華 人 民 共 和 國
Bi-gram:	中華 華人 人民 民共 共和 和國
Tri-gram:	中華人 華人民 人民共 民共和 共和國
Quad-gram:	中華人民 華人民共 人民共和 民共和國
Penta-gram:	中華人民共 華人民共和 人民共和國
Hexa-gram:	中華人民共和 華人民共和國
Hepta-gram:	中華人民共和國

Romanized Pinyin-Based *n*-grams Indexing

Uni-gram:	中華	人民	共和國
	(Zhonghua)	(Renmin)	(Gongheguo)
Bi-gram:	中華人民		人民共和國
	(Zhonghua Renmin)		(Renmin Gongheguo)
Tri-gram:	中華人民共和國		
	(Zhonghua Renmin Gongheguo)		

Figure 4: Example of Indexing of Chinese Phrase

Although search engines do not segment Chinese texts during indexing, implicit segmentation can be achieved by Romanized pinyin. A certain proportion of Chinese online documents are annotated by Romanized pinyin. The search engines index all possible pinyin-based *n*-grams of pinyin instead of character-based *n*-grams of Chinese characters (Figure 4). Under Romanization scheme, the pinyin of characters are grouped together if and only if they form a word. The smallest unit is word instead of character. When you search by pinyin, the search engines match the query with the database of Romanized pinyin-based *n*-grams. Therefore, the search

engines return only those documents where the query is one valid word. Considering the previous example (Figure 4), if we search the word "華人" (Chinese People) by *n*-gram, the documents with the phrase "中華人民共和國" (People Republic of China) will be returned, because it is the second bi-gram of characters. However, if we search the pinyin "Huaren" of the word by pinyin, the documents with the same phrase will not be returned, because it is not a valid *n*-gram of pinyin. Therefore, search results from searching by pinyin are more trustworthy.

Chinese Phrase	Romanized Pinyin	Segmented Chinese	Equivalent English Phrase
今年的	jinnian de	[今年] [的]	[of] [this year]
明年的	mingnian de	[明年] [的]	[of] [next year]
去年的	qunian de	[去年] [的]	[of] [last year]
前年的	qiannian de	[前年] [的]	[of] [the year before last year]
後年的	hounian de	[後年] [的]	[of] [the year after next year]
上年的	shangnian de	[上年] [的]	[of] [previous year]
下年的	xianian de	[下年] [的]	[of] [next year]
…	…	…	…

Table 1. Example of Association of Stop Words

The presence of stop words can affect the accuracy of segmentation significantly (Sproat & Shih, 2002). The stop words sometimes are highly associated with other characters, but they do not form a valid word. In our study, we have found that character "年" (year) is usually followed by the character "的" (of). However, they do not form a valid word; they are just happened to co-occur too frequently. Because these two characters are highly associated by statistics, they will be segmented as a word under statistics-based approaches. However, word boundaries are clearly indicated in Romanized pinyin. There will be a space between two characters, and the words are segmented correctly (Table 1).

Another concern in search by pinyin is ambiguity of pronunciation. In Chinese language, some characters have different pronunciations when they are used together with other characters as a word. In order to solve the problem of ambiguity of pronunciation, we search all the combination of different pronunciations by brute force and the highest number of hits is used for our calculation.

4. Web Data Based Segmentation

We have developed a Chinese segmentation algorithm by mining the web data. The center of the algorithm is to segment the texts into words such that the number of hits of words returned from the search engines is maximized. The system searches all possible *n*-grams through search engines, and the sentence is segmented according to the number of hits matched.

Traditional statistics-based segmentation algorithms segment the texts in the way that the associations between characters of segmented words are maximized. Instead of using the probability measurement of associations between characters in the *n*-gram, we segment the texts by using the number of hits of the *n*-gram returned by search engines. The segmentation algorithm is initiated by the idea that a large number of hits will be returned by the search engines provided that the *n*-gram is a widely used word. Because the search engines index all possible *n*-grams in the documents by brute force, if an *n*-gram is a widely used word, it will occur in the documents frequently and therefore indexed by the search engines for a lot of times.

Given a sentence, our algorithm tries to maximize the total number of hits for the words extracted. The system divides the texts into trunks of *n* consecutive characters that are called *n*-grams. The system then obtains the number of hits of all possible *n*-grams. Related research has shown that about 70% of Chinese words are bi-grams (Sproat & Shih, 1990). Related research assumes that a Chinese word is of length between 1 and 4 characters (Ge, Wanda & Padhraic, 1999). Therefore, we limit *n*-grams to only bi-grams, tri-grams, and quad-grams in our experiment.

As our study shows that the pinyin-based search is more reliable, the *n*-gram is translated into pinyin by using an electronic pinyin dictionary of Chinese characters. The Romanized pinyin of all possible *n*-grams are submitted to search engines to get the number of hits by pinyin search. Concerning the detection of new words, because of the latency between the introduction of a new word and the period that the pinyin of the word is commonly used online, the system obtains the number of hits of *n*-gram by direct search of character sequence in addition to searching by pinyin. Then, the arithmetic mean of two numbers of hits is taken as the number of hits for the *n*-gram.

Our algorithm is to segment the texts such that the total number of hits is maximized. However, study in statistics shows that the number of hits is affected by the length of words. Therefore, number of hits of an *n*-gram is normalized by the length of words by multiply the number of hits with the length of words as the number of character hits. Our algorithm is to maximize the total number of character hits. To solve the problem of segmentation ambiguity, the sentence is segmented by greedy algorithm (Cormen, Leiserson & Rivest, 1989). Given a sentence, our algorithm first extracts the *n*-gram with the highest number of character hits. The algorithm continues to segment the two sub-sentences iteratively. The segmentation algorithm is shown as below:

Segmentation Algorithm:

1. *Obtaining number of hits for all possible n-gram.*
   The system gets the number of hits by pinyin search as well as the number of hits by *n*-gram for all possible *n*-grams. We calculate the number of hits of the *n*-gram by the arithmetic mean of these two numbers.
2. *Calculate the number of character hits for n-grams*
   For all *n*-grams, the system calculates the number of character hits as the product of number of hits of the *n*-gram and the number of characters in the *n*-gram.
3. *Extract the n-gram with the highest character hit*
   The system extracts the *n*-gram in the sentence with the highest number of character hits as a word. The sentence is then divided into two sub-sentences, i.e., one is the substring before the word, and one is the substring after the word.
4. *Further Extraction in Sub-sentence*
   Step 3 is iterated for each sub-sentence until there is no character to be grouped, i.e., only one character is left in the sub-sentence or the number of character count is zero for all possible *n*-grams in the sub-sentence.

Among the search engines, Google and Yahoo are the most famous and largest. Our algorithm has been implemented on top of these two search engines. The sentence "美國反擊中俄開綠燈" (United States counterattack, China-Russia turns on green light) is taken as an example to demonstrate the operation of our algorithm based on the data from Google (Table 2, Figure 5)[1].

Bi-gram	Char Hits
美國	82,000,000
國反	17,100
反擊	3,700,000
擊中	1,540,000
中俄	1,030,000
俄開	1,450
開綠	26,700
綠燈	860,000

Tri-gram	Char Hits
美國反	888,000
國反擊	2,829
反擊中	66,300
擊中俄	2,670
中俄開	15
俄開綠	0
開綠燈	277,500

Quad-gram	Char Hits
美國反擊	1,852,820
國反擊中	0
反擊中俄	60
擊中俄開	0
中俄開綠	0
俄開綠燈	0

Table 2: Statistics of *n*-grams in the Example Sentence

	美	國	反	擊	中	俄	開	綠	燈
1.	[美國]		反	擊	中	俄	開	綠	燈
2.	[美國]		[反擊]		中	俄	開	綠	燈
3.	[美國]		[反擊]		[中俄]		開	綠	燈
4.	[美國]		[反擊]		[中俄]		開	[綠燈]	
5.	[美國]		[反擊]		[中俄]		[開]	[綠燈]	

[United States] [counterattack] [China-Russia] [turns on] [green light]

Figure 5. Example of Segmentation of a Sentence

5. Experiments and Analysis

Comparison of accuracy of Chinese segmentation across systems is very difficult, because different research use different datasets and different ground-rules (Yang, Senelart & Zajac, 2003). A recent competition on Chinese word segmentation provides a set of standard corpora and measurement (Sproat & Emerson, 2003). An experiment has been set up using the same setting as the competition. Experimental results show that our algorithm outperforms the traditional segmentation algorithms.

In the competition, four datasets are provided. Each of the dataset consists of a corpus of training data with segmentation by human professionals and a document for testing. These datasets come from four different areas, namely, Hong Kong, Mainland China (Peking), Taiwan and United State of America. Therefore, the corpora are significantly different from each other. Moreover, they are significantly different in their segmentation standards. Twelve sites participated in the competition. Because of the difficulties of Chinese segmentation, only two sites took part in the open tracks for all the four corpora. In the open track, in addition to the training data for the particular corpus, they are also allowed to use any other materials. Table 3 summarizes the accuracy of system in the competition. The Precision P, Recall R, and F-score F are defined as the standard formula:

$$\text{Precision}\,(P) = \frac{\text{no. of correct segmentation points}}{\text{total no. of segmentation points by the system}}$$

$$\text{Recall}\,(R) = \frac{\text{no. of correct segmentation points}}{\text{total no. of segmentation points in standard answer}}$$

$$F-\text{Score}\,(F) = \frac{2\times P\times R}{(P+R)}$$

The accuracy of each segmentation algorithm varies from corpus to corpus. It is found that the performance on the United States corpus was generally lower than all other corpora. It is mainly because the number of new words (out-of-vocabulary) in this corpus is much higher than the others (Sproat & Emerson, 2003).

To analyze the accuracy of our algorithm, we have conducted the experiment using the same setting as the competition. Table 4 compares the accuracy of the segmentation algorithm by web mining with the accuracy of segmentation system of the participants' sites in the competition. As shown in Table 4, the segmentation algorithm by web mining is independent of search engines, because there is no substantial difference in the accuracies for the segmentation algorithm by using different search engines. It can be assumed that similar results will be obtained by crawling Web pages and collecting the statistics of Chinese n-grams by own effort rather than building our own search engine.

F-score is considered as an overall measurement for system performance (Rijsbergen, 1979). The algorithm gets a higher accuracy than the average of other systems in the first three corpora, and gets an F-score a little bit lower than the average in the Peking Corpus. On the other hand, all the winner sites can win in only one corpus. However, the new algorithm outperforms the winner sites in the Taiwan and USA corpora. In the competition, all the systems performed badly on the United States corpus because of the new word detection problem (Sproat & Emerson, 2003). However, our algorithm does not suffer from this problem. Because our algorithm retrieves the

Site Name	Taiwan Corpus			USA Corpus			HK Corpus			Peking Corpus		
	P	R	F	P	R	F	P	R	F	P	R	F
HK Polytechnic University	0.853	0.892	0.872	0.806	0.853	0.829	0.863	0.909	0.886	0.911	0.940	0.925
SYSTRAN Software. Inc.	0.894	0.915	0.904	0.877	0.891	0.884	0.860	0.898	0.879	0.869	0.905	0.886

Table 3. Summary of Results of the First International Chinese Word Segmentation Bakeoff.

	Taiwan Corpus			USA Corpus			HK Corpus			Peking Corpus		
	P	R	F	P	R	F	P	R	F	P	R	F
Average in Competition	0.874	0.904	0.888	0.842	0.872	0.857	0.862	0.904	0.883	0.890	0.923	0.906
Highest in Competition	*SYSTRAN Software*			*ICL, Beijing U Inc.*			*CKIP Ac. Sinica Taiwan*			*Microsoft Research*		
	0.894	0.915	0.904	0.907	0.916	0.912	0.954	0.958	0.956	0.956	0.963	0.959
Segmentation using Google	0.946	0.931	0.938	0.978	0.935	0.956	0.968	0.925	0.946	0.931	0.904	0.918
Segmentation using Yahoo	0.952	0.943	0.947	0.953	0.946	0.949	0.949	0.933	0.941	0.939	0.919	0.929

Table 4. Comparison of Accuracy of Segmentation Algorithm by Web Mining with the Participants in the Competition

[1] All the data in this paper are taken in September 2005.

data from the Web, and vocabulary size of online documents is so large that it can eliminate the effects of new words. In our analysis, it is found that new algorithm is good at detection of new words.

Despite the fact that new algorithm does not take any data from the training corpora, our system can get an accuracy comparable to the state-of-art segmentation algorithms. The novel algorithm is developed directly based on the number of page hits returned by search engines. More advanced techniques have been developed for segmentation. It is natural to predict that integration of those techniques to our algorithm can further improve the accuracy. As a result, our algorithm is a promising technique in Chinese segmentation.

6. Conclusion

The search engines maintain huge databases of online documents and the frequency of string in the document. Their databases are some important resource in Chinese segmentation. A novel segmentation algorithm by mining web data is proposed in this paper. This is the first algorithm developed based on the Romanized pinyin of Chinese characters. Without taking any data from the training corpora, the algorithm can achieve an accuracy as high as other state-of-art segmentation algorithms. The experimental results have shown that our algorithm is a promising technique in Chinese segmentation. Moreover, our algorithm is the first unified segmentation algorithm for documents from different Chinese areas, and it can also solve the problem of new word detection.

References

Chen K.J., Ma, W.Y. (2002). Unknown Word Extraction for Chinese Documents, *Proceedings of the 19th International Conference on Computational Linguistics COLING 2002*, Taipei, China, August 2002, pp. 169-175.

Cormen, T.H., Leiserson, C.E., Rivest, R.L., (1989). *Introduction to Algorithms*, The MIT Press, Cambridge, MA., USA.

Ge X.P., Wanda P., Padhraic S. (1999). Discovering Chinese Words from Unsegmented Text. *Proceedings of the Twenty-second International ACM Conference on Research and Development in Information Retrieval. (SIGIR'99)*, Berkeley, CA., USA, August, 1999, pp. 271-272.

Luhn H. (1958). Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, pp. 159-165.

Lunde K. (1998). *CJKV Information Processing. Chinese Japanese, Korean & Vietnamese Computing*. New York, NY: O'Reilly.

Mihalcea R., Moldovan D., (1999) A method for word sense disambiguation of unrestricted text. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)***,** College Park, MD, USA, June, 1999, pp. 152-158.

Rijsbergen C.V. (1979). *Information Retrieval*. Second Edition, London: Butterworths,.

Sproat R., Emerson T. (2003). The First International Chinese Word Segmentation Bakeoff, *Proceedings of The Second SIGHAN Workshop on Chinese Language Processing*, Sapporo, Japan, July 200, pp. 133-143.

Sproat R., Shih C. (1990). A statistical method for finding word boundaries in Chinese text, *Computer Processing of Chinese and Oriental Languages*, vol. 4, pp. 336-351.

Sproat R., Shih C. (2002). Corpus-based methods in Chinese morphology and phonology. *Proceedings of the 19th International Conference on Computational Linguistics COLING 2002*, Taipei, China, August 2002.

Wu Z., Tseng G. (1993). Chinese text segmentation for text retrieval: Achievements and problems. *Journal of the American Society for Information Science*, vol. 44, no. 9, pp. 532–542.

Yang C.C., Li K.W. (2005). A Heuristic Method Based on a Statistical Approach for Chinese Text Segmentation. *Journal of the American Society for Information Science and Technology*, vol. 56, no. 13, pp. 1438-1447.

Zipf G.K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley.