Mining Web Data for Chinese Segmentation

Fu Lee Wang
Department of Computer Science, City University of Hong Kong, Kowloon Tong, Hong Kong, People's Republic of China. E-mail: flwang@cityu.edu.hk

Christopher C. Yang
Department of Systems Engineering and Engineering Management, William M. W. Wong Engineering Building, The Chinese University of Hong Kong, Shatin, Hong Kong, People's Republic of China. E-mail: yang@se.cuhk.edu.hk

Modern information retrieval systems use keywords within documents as indexing terms for search of relevant documents. As Chinese is an ideographic character-based language, the words in the texts are not delimited by white 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 most search engines have problems in segmenting 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 of search engines. On the other hand, the Romanized pinyin of Chinese language indicates boundaries of words in the text. Our algorithm is the first to utilize the Romanized pinyin to segmentation. It is the first unified segmentation algorithm for the Chinese language from different geographical areas, and it is also domain independent because of the nature of the Web. 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 problems of segmentation ambiguity, new word (unknown word) detection, and stop words.

Introduction

As a result of the fast growth of online Chinese documents, the research activities in Chinese information retrieval have become very active. In information retrieval, a document is traditionally indexed by the frequency of words within documents and the corpus (Rijsbergen, 1979; Salton & Buckley, 1988). 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 punctuation as word delimiters. However, many Asian languages, such as Chinese, Japanese, Korean, Vietnamese, and Thai, do not delimit the words by spaces. The absence of word boundaries poses a critical problem for information retrieval in Asian languages (Wu & Tseng, 1993).

In Chinese information retrieval, segmentation of the texts is required before indexing of a document. Many segmentation algorithms have been developed for the Chinese language. Typically, algorithms for Chinese segmentation fall into three categories: dictionary-based approaches (Wu & Tseng, 1993), statistics-based approaches (Leung & Kan, 1996; Teahan, Wen, McNad, & Witten, 2000; Yang & Li, 2004; Yang, Luk, Yung, & Yen, 2000), and hybrid approaches (Li, Huang, Gao, & Fan, 2004; Nie, Jin, & Hannaan, 1994). The dictionary-based approaches segment the texts by matching the text-trunks against entries of a large machine-readable dictionary of Chinese words. The major shortcoming of dictionary-based approaches is identification of new words (Chen & Bai, 1998; Chen & Ma, 2002). Statistics-based approaches or hybrid approaches are proposed to solve the problem of new word detection (Chen & Bai, 1998; Chen & Ma, 2002; Li, Huang, Gao & Fan, 2004; 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 for training of the segmentation algorithm. Therefore, a corpus-free segmentation algorithm is more desirable.

The Web has become the largest corpus because of the explosion of the Internet, and the online documents have covered a wide range of domains. Moreover, most of the online documents have been indexed by search engines. Among the search engines, Google (http://www.google.com) and Yahoo (http://www.yahoo.com) are the two most popular and largest

Accepted January 4, 2007

 ● Published online 17 August 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20629

in terms of number of visitors and number of documents indexed. Google, for example, had indexed 8,168,684,336 pages by September 2005 (Google, 2005). Besides, the search engines update their databases regularly and frequently. The Web together with search engines can be considered as the largest corpus in the world (Etzioni et al., 2004).

Nowadays, the search engines are widely used for search of relevant documents by humans. In addition, they are employed to provide statistics for some information retrieval tasks. Because of the huge knowledge space of online documents, many Question Answering systems based on search engines have been developed (Banko, Brill, Dumais, & Lin, 2002; Etzioni et al., 2004; Kwok, Etzioni, & Weld, 2001; Radev et al., 2001). In addition to the documents returned by the search engines, the numbers of hits returned by the search engines are also widely used in information retrieval. For example, the number of hits returned by Google search engine is employed to identify instance-concept pairs in a Web page (Cimiano, Handschuh, & Staab, 2004). A technique has been developed to solve the word sense disambiguation (Mihalcea & Moldovan, 1999). The appropriate sense of a word inside a text can be determined by the number of hits of words returned by the search engines.

In Web search, if two events or two concepts co-occur in a Web page frequently—that is, the number of hits in response to simultaneous search of two words is very large—we assume that there is a strong association between these two events or concepts (Agirre, Ansa, Hovy, & Martinez, 2002; Keller, Maria, & Olga, 2002; Mihalcea & Moldovan, 1999). In statistics-based approaches to Chinese segmentation, statistical data of a large corpus are 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.

Chinese languages are used in different geographical areas, such as mainland China, Taiwan, Hong Kong, Macau, and Singapore. The Chinese languages in different areas are very different from each other: different forms, different encoding systems, different pronunciations, different vocabularies, and so on. Therefore, it is extremely difficult to develop a unified segmentation algorithm for different areas. The accuracies of traditional segmentation algorithms vary greatly in datasets from different areas. Because of the nature of the Web, Web data are domain independent and geography independent. Therefore, Web data are an ideal corpus for segmentation of Chinese language.

The search engines maintain huge databases of the online documents. Although most of them have problems in segmenting the texts into proper words, they have recorded some important features of the documents, such as frequencies of character sequences. 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, called *n*-grams, to segment Chinese texts. This is the first algorithm based on Romanized pinyin of Chinese characters that has been developed. Experiments on datasets for a recent international segmentation competition have been conducted to evaluate the accuracy of our algorithm. Experimental results show that the segmentation algorithm by Web mining outperforms the traditional segmentation algorithms significantly in terms of precision and recall. Moreover, the experimental results show that the novel algorithm is geographical area independent; therefore, it can segment the Chinese documents from different geographical areas effectively. It can also deal with the problems of segmentation ambiguity, new word detection, and stop words.

The rest of this paper is organized as followings. Traditional Segmentation reviews the traditional algorithms in Chinese segmentation. Pinyin Search From WEB introduces search of Chinese words by Romanized pinyin using search engines. Web Data Based Segmentation proposes a segmentation algorithm by mining Web data. Experiments and Analysis presents the experiment results and their analysis. Conclusion gives a short summary.

Traditional Segmentation

A Chinese text appears to be a linear sequence of nonspaced ideographic characters that are called morphemes. Usually, a Chinese word consists of more than one ideographic character and the number of characters varies. Segmentation of Chinese texts into words is essential for indexing of Chinese documents. Chinese segmentation is probably the single most widely addressed problem in the literature on Chinese language processing (Wang, Su, & Mo, 1990; Wu & Tseng, 1993). Many Chinese segmentation algorithms have been developed. Typically, algorithms for Chinese segmentation fall into three categories: dictionary-based approaches, statistics-based approaches, and hybrid approaches.

Difficulties in Chinese 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: GB for mainland China, BIG-5 with Hong Kong Supplementary Character Set for Hong Kong, BIG-5 for Taiwan; GBK, EUC-CN, and Unicode are also used for other areas. Among the encoding systems, BIG-5 and GB are the most popular ones.

There are numerous difficulties in Chinese segmentation. In Western and Middle Eastern languages, there are only a small number of characters: 26 for the Latin alphabet, 28 for the Arabic alphabet, 33 for the Cyrillic alphabet, and so on. 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

DOI: 10.1002/asi

Mainland Standard:	[中華人民共和國]		
	[People's Republic of China]		
ROCLING Standard and U Penn Standard:	[中華]	[人民]	[共和國]
	[China]	[People]	[Republic]

FIG. 1. Different segmentation standards of the Chinese phrase "中華人民共和國" (People's Republic of China).

and the GB encoding system in mainland China defines about 7,700 simple-form characters (Lunde, 1998).

On the other hand, there is no unique segmentation standard in Chinese (Sproat & Shih, 2001). There are different segmentation standards in different areas (Sproat & Emerson, 2003; Sproat & Shih, 2001): Mainland Standard (State Bureau of Technology Supervision, 1992), ROCLING Standard (Huang, Chen, Chen, & Chang, 1997), University of Pennsylvania/Chinese Treebank Standard (Xia, 1999), and others. A phrase or a sentence is segmented differently in different areas. Taking Figure 1 as an example, the phrase "中華人民共和國" (People's Republic of China) is segmented as a single word in Mainland Standard, but it is segmented as three words in other standards (Sproat & Shih, 2001).

The meaning of the phrase remains unchanged in different segmentation standards. Differences just reflect the fact that different standards have different identifications of word boundaries. Some consider the phrase as a single word while others 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 (Gan, 1995). The sentence "馬路上生病了" is taken as an example (Figure 2). Two different segmentations of the same sentence lead to two different meanings.

Given a sentence, there may exist more than one way to segment it; this problem is known as segmentation ambiguity (Goh, Asahara, & Matsumoto, 2005; Li et al., 2003). The problem of segmentation ambiguity is extremely serious in the Chinese language. Take the phrase "中國文化工業" (Chinese Cultural Industry) as an example (Figure 3): every adjacent bigram in the phrase is a valid lexical word, and there are also some trigrams and quadgrams in this phrase of six characters only.

Chinese Text:	馬路上生病了
Segmentation A:	[馬路上] [生病] [了]
	He got sick on the road.
Segmentation B:	[馬] [路上] [生病] [了]
	The horse got sick on the road

FIG. 2. Segmentation ambiguity of the sentence of Chinese text "馬路上生病了".

Because of the special characteristics of the Chinese language, automatic identification of words in Chinese texts is a challenging task. A number of Chinese segmentation algorithms have been proposed in the literature, but none of them has been commonly accepted as a standard.

Dictionary-Based Approaches

The dictionary-based approaches are the most straightforward for Chinese segmentation (Sproat & Shih, 2001; Wu & Tseng, 1993). They can be implemented on the basis of 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 (Goh et al., 2005; Li et al., 2003). Several matching algorithms have been proposed. The most popular approach dealing with segmentation ambiguities is the maximal matching method, possibly augmented with further heuristics (Sproat & Shih, 1990). Starting from the beginning of the text, the maximal 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 unknown words): words that are not listed in an ordinary dictionary (Chen & Bai, 1998; Chen & Ma, 2002). A study of a 5 million word Chinese corpus with proper word segmentation found that 3.51% of Chinese words are not listed in the most powerful machine-readable dictionary (Chen & Ma, 2002). Similar results are also obtained for the English corpus (Sampson, 1989). Moreover, there will be new words generated every day. For the English language, a large number of new entries of words are added to the *Oxford Dictionary of English* each year (Soanes & Stevenson, 2005). In Chinese, they are facing the same problem. A lot of new words are generated each year (Language Information Sciences Research Center, 2006). However, 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.

Statistic-Based or Hybrid Approaches

Statistics-based approaches or hybrid approaches are proposed to solve the problem of unknown words (Banko et al.,

DOI: 10.1002/asi

Phrase:	中國文化工業	(Chinese Cultural Industry)
Bigram:	[中國]	(China)
	[國文]	(Chinese)
	[文化]	(Culture)
	[化工]	(Chemical Industry)
	[工業]	(Industry)
Trigram:	[中國文]	(Chinese Language)
	[化工業]	(Chemical Industry)
Quadgram:	[中國文化]	(Chinese Culture)
	[文化工業]	(Cultural Industry)

FIG. 3. Possible valid words from the phrase "中國文化工業" (Chinese Cultural Industry) .

2002; Li et al., 2004; Sproat & Shih, 1990; Yang & Li, 2003; Yang & Li, 2005). 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 between characters. Among them, the mutual information is the most popular approach (Sproat & Shih, 1990; Yang et al., 2000). The hidden Markov model is also employed in detection of segmentation points (Teahan et al., 2000).

The statistics-based approaches are weak in dealing with stop words. Stop words are words that appear frequently in the corpus but do not convey any significant information to the document, for example, *of* and *the* (Edmundson, 1968; Luhn, 1958). A lot of stop word lists have been constructed for the English language. However, the situation in the Chinese language is much more complicated because the 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, it is extremely difficult to identify stop words in Chinese texts. Currently, no Chinese stop word list has been commonly accepted as a standard. 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 so far.

On the other hand, the statistics-based approaches segment the texts on the basis of the statistical figures of a large corpus of documents. Thus, the accuracy of the segmentation depends significantly on the size of the training corpus. As is well known, a histogram of the words in a reasonable size corpus for any language will reveal a Zipfian distribution (Zipf, 1949). Zipf's law states that the frequency of a word is inversely proportional to its rank: the position of the word in the descending word list sorted by word frequency. According to the law, there will be a large number of words that occur only once in a corpus. A study of word frequency in the English language (Sproat & Shih, 1990) found that 40% of the words occur just once in the 37 million word corpus of the 1995 Associated Press newswire. For a smaller corpus, about 50% of words occur once in the 1 million words of the Brown corpus (Kucera and Francis, 1967).

Since the Chinese language is character-based, statistics have been collected on character frequency in the language (Sproat & Shih, 1990). It has been found that 11% of the characters occur only once in the 10 million character ROCLING corpus. Although there are 13,500 characters and 7,700 characters in the BIG5 and GB encoding systems, respectively (Lunde, 1998), 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. The performance of the segmentation algorithm will be deeply affected by the occurrence of rare characters in the corpus. Therefore, a large size corpus is essential for statistic-based segmentation algorithms to minimize the impact of rare characters.

Moreover, there are different vocabularies (Language Information Sciences Research Center, 2006) in different areas. Words can sometimes carry totally different meanings in different geographical areas. For example, the word "人流" in Hong Kong means "stream of people," but it means "artificial abortion" in China. On the other hand, an object or a concept is described using different words in different areas. For example, "computer" is known as "電腦" (electrical brain) in Hong Kong, "電算機" (electrical calculating machine) / "計算機" (calculating machine) in mainland China, and all these terms are used in Taiwan. The vernacular differences in the languages have a deep impact

DOI: 10.1002/asi

on Chinese segmentation applications. For example, the statistical information of the corpus in one geographical area cannot be applied to segmentation of documents in another geographical area.

In order to capture the language differences in different 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. Web data are the ideal corpus for Chinese segmentation, because they include documents from different areas. Moreover, most of the online documents have been indexed by search engines and the search engines update the database constantly to capture the introduction of new words.

Pinyin Search From WEB

Traditionally, the Chinese language is regarded as a morphological language with ideographic characters. Social movements in the twentieth century attempted to Romanized Chinese orthography (Pan, Yip, & Han, 1993; Sproat & Shih, 1990). In the Romanization scheme, the pinyin of Chinese characters are grouped into word-sized trunks. Word boundaries are therefore clearly indicated in Romanized pinyin of Chinese texts (National People's Congress, 1958; Sproat & Shih, 2001). However, the techniques of Romanized pinyin have not yet been employed in Chinese segmentation.

Under the traditional pinyin scheme, the pronunciation of each Chinese character is written as a sequence of pinyin syllables, and the pinyin of characters are separated from each other (Figure 4). Under the Romanization scheme, the pinyin of individual character are grouped into word-sized trunks. Considering the phrase "中華人民共和國" (People's Republic of China) as an example (Figure 4); the Romanized pinyin is clearly segmented into words. Each English word in the example corresponds to one Chinese word except the stop word *of*, which is not translated in this example. The Romanization scheme of the Chinese language indicates the boundaries of words clearly in the texts (National People's Congress, 1958; Sproat & Shih, 2001). It can be utilized for segmentation.

It is difficult to find a large corpus of Chinese texts with Romanized pinyin. Therefore, no segmentation technique based on Romanized pinyin has been developed so far. With the development of the Internet, the Web has become the largest corpus in the world. Some search engines have indexed documents containing Chinese texts annotated with Romanized pinyin or pure Romanized pinyin without corresponding Chinese texts. For example, if a query of Romanized pinyin "Zhonghua Renmin Gongheguo" for the phrase "中華人民共和國" (People's Republic of China) is submitted to Google, the search engine returns the documents containing the query phrase in Romanized pinyin. As shown in Figure 5, the search engine does not translate the pinyin into Chinese, because some documents returned do not contain the corresponding Chinese words. Instead, the search engine indexes the Romanized pinyin, and the Web search is conducted by string matching of Romanized pinyin. As search engines maintain a huge database of Web documents, their databases of Romanized pinyin provide an important resource for Chinese segmentation.

One of the disadvantages of search engines is that they do not understand the semantic meaning of a query. They search documents purely by string matching. For English Web search, Google has recently released the dataset of English word *n*-grams and their frequencies (Brants & Franz, 2006). The length of the *n*-grams in this dataset ranges from unigrams (single words) to five-grams (five words). If a query is submitted to a search engine, it will match the query against the *n*-gram database and return those documents that match the *n*-gram. For Chinese Web search, most search engines claim that they segment the text before indexing; for example, Google uses the segmentation technique of Basis Technology (http://www.basistech.com), Yahoo uses a self-developed segmentation technique, and Baidu (2006) (http://www.baidu.com) uses the segmentation technique of Hylanda (http://www.hylanda.com). Unfortunately, the search engines do not release their Chinese segmentation and indexing techniques. Therefore, we have no detailed information about those techniques.

Although the search engines make an effort to segment Chinese texts, study has shown that all search engines have problems in segmenting Chinese texts into words (Moukdad & Cui, 2005). The search engine returns the documents containing the sequence of characters in the query but they may not form a word in the correct segmentation (Moukdad & Cui, 2005). Figure 6 shows a typical mistake during a Chinese Web search. If we search for the phrase "中國"

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

FIG. 4. Romanized pinyin for the phrase "中華人民共和國" (People's Republic of China).

DOI: 10.1002/asi

FIG. 5. Search for Romanized pinyin for the phrase "中華人民共和國" (People's Republic of China) in Google.

Query: 中國 (China)

Result: 發展中國家 (Developing country)

Correct Segmentation: [發展中] [國家] ([Developing] [country])

FIG. 6. Example of mistakes during Chinese Web search.

DOI: 10.1002/asi

Chinese:	中華人民共和國							
Unigram:	中	華	人	民	共	和	國	
Bigram:	中華	華人	人民	民共	共和	和國		
Trigram:	中華人		華人民		人民共		民共和	共和國
Quadgram:	中華人民		華人民共		人民共和		民共和國	
Pentagram:	中華人民共		華人民共和		人民共和國			
Hexagram:	中華人民共和		華人民共和國					
Heptagram:	中華人民共和國							

FIG. 7. Indexing of all possible character-based *n*-grams of Chinese characters for the phrase "中華人民共和國" (People's Republic of China).

(China), the search engines may return documents containing "發展中國家" (Developing country). In this case, the search engines return documents containing a sequence of characters across the word boundary. On the other hand, the search engines will return some documents containing this sequence of characters in response to a query of a permutation of any two Chinese characters. However, the permutation by itself may be meaningless. It can be best explained by the fact that the search engines index documents as trunks of all possible *n*-grams of consecutive characters by brute force to facilitate the searching of documents (Figure 7). However, the *n*-gram may be just a sequence of characters across the word boundary.

If we search by Romanized pinyin, the results will be totally different. Although most search engines have problems in segmenting Chinese texts into words during indexing, implicit segmentation can be achieved by Romanized pinyin. A certain proportion of Chinese online documents are written in Romanized pinyin or in Chinese characters annotated with Romanized pinyin (Figure 5). For example, Chinese language records in the Library of Congress Online Catalog (http://catalog.loc.gov) in the United States (Library of Congress, 2004), the National Library of Australia (http://www. nla.gov.au) in Australia (National Library of Australia, 2003), and the Cambridge University Library (http://www.lib.cam.ac.uk) in the United Kingdom (Cambridge University Library, 2006), are mostly cataloged using Romanized pinyin.

Traditionally, the search engines index all possible word-based *n*-grams (Brants & Franz, 2006; Brin & Page, 1998; Hawking, 2006). As Chinese is a character-based language, the search engines traditionally index all possible character-based *n*-grams of Chinese characters. However, the pinyin syllables of characters are grouped together if and only if they form a word under the Romanization scheme. The smallest unit is a word instead of a character (National People's Congress, 1958; Sproat & Shih, 2001). If a document contains text in Romanized pinyin, the search engines index all possible word-based *n*-grams of Romanized pinyin instead of character-based *n*-grams of Chinese characters (Figure 8).

When we search a word by Romanized pinyin in search engines, they match the query with the database of word-based *n*-grams of Romanized pinyin. Therefore, the search engines return only those documents where the query is one valid word. Considering the previous example, if we search the word "華人" (Chinese People) by *n*-gram, the documents with

Chinese:	中華	人民	共和國
Romanized Pinyin:	Zhonghua	Renmin	Gongheguo
Unigram:	中華	人民	共和國
	(Zhonghua)	(Renmin)	(Gongheguo)
Bigram:	中華人民		人民共和國
	(Zhonghua Renmin)		(Renmin Gongheguo)
Trigram:	中華人民共和國		
	(Zhonghua Renmin Gongheguo)		

FIG. 8. Indexing of word-based *n*-grams of Romanized pinyin for the phrase "中華人民共和國" (People's Republic of China).

DOI: 10.1002/asi

FIG. 9. Search of "pinyin" in Google search engine.

the phrase "中華人民共和國" (People's Republic of China) will be returned, because it is the second bigram of characters (Figure 7). However, if we search the Romanized pinyin "Huaren" of the word "華人" (Chinese People) by Romanized pinyin, the documents with the phrase "中華人民共和國" (People's Republic of China) will not be returned, because it is not a valid *n*-gram of Romanized pinyin (Figure 8).

The search engines index many documents, including documents in Romanized pinyin and traditional pinyin. The presence of space in a query that is submitted to a search engine will give a different result. For example, 274,000 relevant pages are returned for the query "huaren," while 1,170,000 pages are returned for "hua ren." The numbers of pages returned are significantly different in searching with or without a space between the two Chinese characters. The search engine searches for the word, if we search by the Romanized pinyin of "huaren." However, the search engine searches for the character sequence if we search for traditional pinyin "hua ren," therefore, some document containing "hua" and "ren" across word boundaries. In Romanized pinyin, the pinyin syllables of characters are grouped into word-sized trunks (Figure 4), while the white space in between characters is removed. If we search by Romanized pinyin, the search engine returns only those documents in which the characters form a correct segmentation of words, because the word boundary is explicitly indicated under Romanized pinyin (National People's Congress, 1958; Sproat & Shih, 2001). However, if we search for traditional pinyin, the search engine returns all the documents in which the characters happen to occur together though they may not form a correct segmentation of words in the sentence.

Considering "拼音" (pinyin) as an example, the word contains two characters, "拼" (pin) and "音" (yin). There are 12 Chinese characters pronounced "pin" and 46 Chinese characters pronounced "yin." Therefore, there are a total of 506 permutations of these two sets of characters pronounced "pinyin." If we search by Chinese characters, the search engines return hits for almost all the permutations. However, most of these permutations are invalid words. They happen to co-occur in the documents, but they do not contain any meaning in themselves. However, if we search "pinyin" by Romanized pinyin in the search engines, the search engines return only "拼音" and "拚音" (Figure 9). Both of them are the correct words for "pinyin." The first is the traditional Chinese in complex-form characters used in Hong Kong and Taiwan, and the second is the simplified Chinese in simple-form characters used in mainland China. Therefore, search results from searching by Romanized pinyin are more trustworthy.

The presence of stop words can affect the accuracy of Chinese segmentation significantly (Sproat & Shih, 2001). 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 the character "年" (year) is usually followed by the character "的" (of). As shown in Table 1, when there is a character "年" (year), it is usually followed by the character "的" (of). However, they do not form a valid word; they just happen to co-occur frequently. Because these two characters are highly associated with each other by statistics, they will be segmented as a word under statistics-based approaches. However, word boundaries are clearly indicated in Romanized pinyin of Chinese texts (National People's Congress, 1958; Sproat & Shih, 2001). There will be a space between the character "年" and the character "的" Therefore, the words are segmented correctly (Table 2).

Unfortunately, Mandarin Chinese is not used in some Chinese-speaking areas, such as Hong Kong. Therefore, pinyin is not popular in Hong Kong. With the growth of the

TABLE 1. Example of high association of stopwords with other characters.

Chinese phrase	Equivalent English phrase
今年的	Of this year
明年的	Of next year
去年的	Of last year
前年的	Of the year before last year
後年的	Of the year after next year
上年的	Of previous year
下年的	Of next year
2000 年的	Of year 2000

DOI: 10.1002/asi

TABLE 2. Example of association of stopwords with other characters under Romanized pinyin.

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]

economy of China, it is natural to predict that Mandarin Chinese will become more popular in the near future. For example, the number of Mandarin speakers is increasing in recent years in Hong Kong. However, Romanized pinyin of Chinese document has not yet been widely used on the Internet. On the other hand, there is a certain latency between the introduction of a new word and the period when the pinyin of the word is widely used online. Therefore, the search results of searching by pinyin will be supplemented by search results of searching by Chinese character sequences.

There are various pinyin systems for different Chinese-speaking areas, that is, mainland China, Taiwan, and Hong Kong. Taking the phrase "華人" (Chinese People) as an example, the pinyin "huaren" is used in mainland China, while "hwaren" is also considered as an appropriate pinyin in Taiwan and "wahyan" is used in Hong Kong. Because the pinyin system in mainland China is most widely used in terms of the number of literatures and the size of the population, we focus our study on the pinyin system of mainland China. When we search these three pinyins in Google, 274,000 relevant pages are returned for the query "huaren," while 292 and 29,500 pages are returned for "hwaren" and "wahyan," respectively. Actually, some search engines consider the pinyin system of mainland China only. Considering the previous example, if we search "huaren" in Google, it suggests the translated Chinese phrase "華人" (Chinese People), but it does not suggest the translated term for either "hwaren" or "wahyan." The choice of pinyin systems may have a significant impact on the accuracy of segmentation. In the future, study will be conducted to investigate the differences of the pinyin systems.

Another concern in search by pinyin is ambiguity of pronunciation. In the Chinese language, some characters have different pronunciations when they are used together with other characters as a word. For example, the Chinese character "會" has two different pronunciations. It is pronounced "hui" when it appears in "會議" (meeting), but it is pronounced "kuai" when it appears in "會計" (accounting). In our study, we found that a Chinese character can have a maximum of four different pronunciations; for example, the character "著" can be pronounced "zhao," "zhe," "zhu," and "zhuo." In order to solve the problem of ambiguity of pronunciation, we search all the combinations of different pronunciations by brute force and the highest number of hits is used for our calculation.

Web Data Based Segmentation

We have developed a Chinese segmentation algorithm by mining the Web data. The focus 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 Chinese segmentation algorithms segment the texts in such a 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 be indexed by the search engines many times. Given a sentence, our algorithm tries to maximize the total number of hits for the words extracted. For instance, the segmentation B in Figure 10 gave a larger total number of hits than segmentation A; therefore, it is preferable for segmentation of texts.

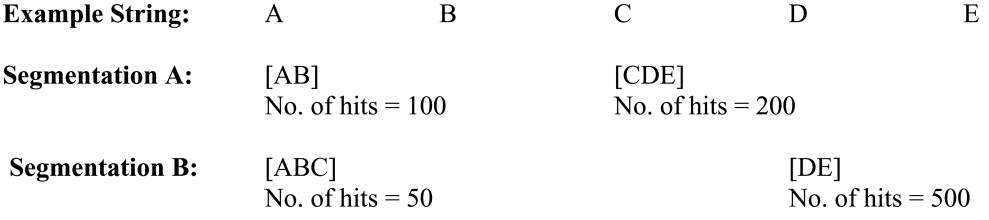

FIG. 10. Example of Chinese segmentation based on number of hits returned by search engines.

DOI: 10.1002/asi

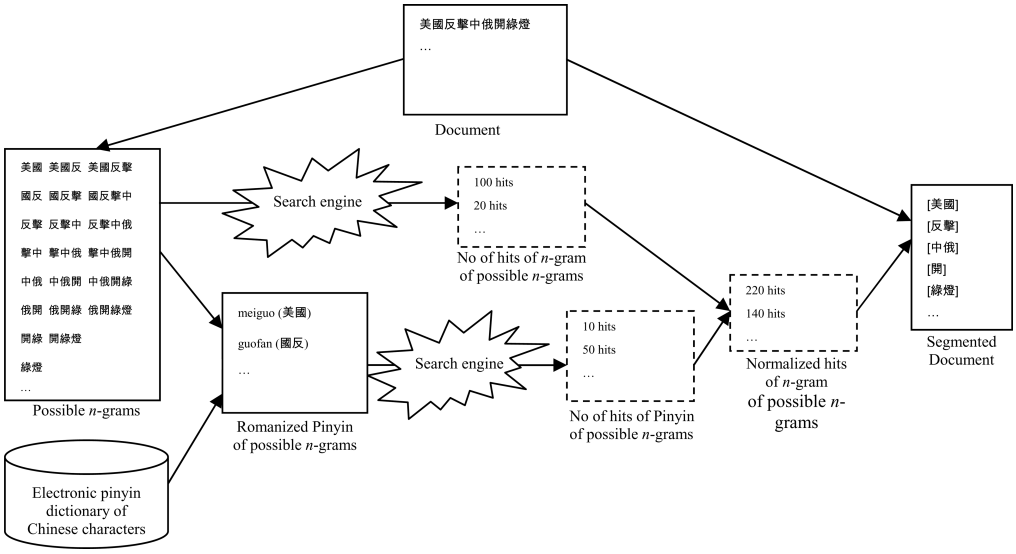

FIG. 11. Chinese segmentation through mining of Web data.

The system operation is shown in Figure 11. Given a Chinese text, the system will divide the texts into trunks of n consecutive characters that are called n-grams. Our system obtains the number of hits of all possible n-grams. Related research has shown that about 70% of Chinese words are bigrams (Sproat & Shih, 1990), and some previous Chinese segmentation systems focus on extraction of bigrams. Some research has focused on longer words (Chang & Su, 1997); another study assumes that a Chinese word is of length between one and four characters (Ge, Wanda, & Padhraic, 1999). Therefore, we limit n-grams to bigrams, trigrams, and quadgrams in our experiment.

As our study shows that pinyin-based search is more reliable than character-based search, the n-gram is translated into pinyin by using an electronic pinyin dictionary of Chinese characters. The Romanized pinyin of all possible n-grams is submitted to search engines to get the number of hits by Romanized pinyin search. Concerning the detection of new words, because of the latency between the introduction of a new word and the period when the Romanized pinyin of the word is commonly used in the Web, the system obtains the number of hits of n-gram by direct search of character sequence in addition to searching by Romanized 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. When a query is submitted to the search engine, the search engine returns all the documents containing the query words (Hawking, 2006). Studies show that the number of documents in which the query words co-occur is affected by the number of words (Wettler & Rapp, 1993; Church & Hanks, 1989). As Chinese is a character-based language, the number of documents in which the characters co-occur is affected by the length of the query, that is, the number of characters (Ma & Chen, 2003). Therefore, the number of hits of an n-gram is normalized by the length of words by multiplying 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. In order to solve the problem of segmentation ambiguity, we segment the sentence by a greedy algorithm (Cormen, Leiserson, & Rivest, 1989), to find out the local maximum of character hits. Given a sentence, our algorithm will first extract the n-gram with the highest number of character hits. The algorithm will continue to segment the two subsentences iteratively until the whole sentence is segmented.

The detailed algorithm of the segmentation algorithm is shown in the following:

Segmentation Algorithm:

1. *Obtaining number of hits for all possible n-grams.* Let the input sentence be denoted by $X = < x_1, x_2, \ldots, x_m >$. The system obtains the number of hits for all possible n-grams in the sentence, that is,

   $$n\text{-gram } \{\langle x_i, x_{i+1}, \ldots, x_{i+n-1}\rangle,\ 1 \le i \le m - n + 1, 2 \le n \le 4\}$$

   The system gets the number of hits by Romanized pinyin search $hits\#_{pinyin}$ (n-gram) as well as the number of hits by character sequence search $hits\#_{char}$ (n-gram) for all possible n-grams. The arithmetic mean of these two numbers is taken as the number of hits of the n-gram, that is, $hits\#$ (n-gram).

   $$hits\#(n\text{-gram}) = \frac{hits\#_{pinyin}(n\text{-gram}) + hits\#_{char(n\text{-gram})}}{2}$$

2. *Calculating the number of character hits for n-grams.* For all possible n-grams, the system calculates the number of character hits $char\_hits\#$ (n-gram) as the product of number of hits of the n-gram and n, which is the number of characters in the n-gram.

   $$char\_hits\#(n\text{-gram}) = \text{hits}\#(n\text{-gram}) \times n$$

3. *Extracting the n-gram with highest character hits as a word.*
   The system extracts the n-gram in the sentence with the highest number of character hits as a word. The sentence

DOI: 10.1002/asi

TABLE 3. Statistics of n-grams in the sentence "美國反擊中俄開綠" (United States counterattack, China-Russia turns on green light) from Google.

Bigram	Char hits	Trigram	Char hits	Quadgram	Char hits
美國	82,000,000	美國反	888,000	美國反擊	1,852,820
國反	17,100	國反擊	2,829	國反擊中	0
反擊	3,700,000	反擊中	66,300	反擊中俄	60
擊中	1,540,000	擊中俄	2,670	擊中俄開	0
中俄	1,030,000	中俄開	15	中俄開綠	0
俄開	1,450	俄開綠	0	俄開綠燈	0
開綠	26,700	開綠燈	277,500		
綠燈	860,000				

is then divided into two subsentences: one is the substring before the word, and one is the substring after the word. Assume that the n-gram $< x_i, x_{i+1}, \ldots x_{i+n-1} >$ has the highest number of character hits among all possible n-grams in the sentence. The sentence $< x_1, x_2, \ldots x_m >$ is segmented into three substrings as

$$< x_1, x_2, \ldots x_m > = < x_1, x_2, \ldots x_{i-1} > + < x_i, x_{i+1}, \ldots x_{i+n-1} > + < x_{i+n}, x_i + {}_{n-1}, \ldots x_m >$$

4. *Further extraction in the subsentence.*
Step 3 is iterated for each subsentence until there is no character to be grouped: that is, there is only one character left in the subsentence or the number of the character count is zero for all the possible n-grams in the subsentence, that is,

```
Segment (X){
  if length (X) < 1
    return
  else if length (X) = 1
    word list ← X
  else
    identify the n-gram with the highest character hits
    let the n-gram be W = <x_i, x_{i+1}, ..., x_{i+n-1}>
    word list ← W
    Segment (< x_1, x_2, ... x_{i-1} >)
    Segment (< x_{i+n}, x_{i+n+1}, ... x_m >)
}
```

A recent Chinese word segmentation competition has been held (Sproat & Emerson, 2003). Experiments have been conducted on the corpora used in the competition. In order to demonstrate the operation of our algorithm, one sentence is taken as an example.

Among the search engines, Google (http://www.google.com) and Yahoo (http://www.yahoo.com) are the most famous and largest. Our algorithm has been implemented on top of these two search engines. The statistical figures from Google and Yahoo are shown in Table 3 and Table 4, respectively, and the sentence is segmented on the basis of the data from Google (Figure 12) and Yahoo (Figure 13).[1]

As shown in Figure 12 and Figure 13, the sentence is segmented correctly on the basis of the number of hits returned from both search engines. The only difference is that the word "反擊" (counterattack) records a higher number of hits than the word "中俄" (China-Russia) in Google; therefore, the former word is segmented before the later on the basis of the data from Google. However, the situation in Yahoo is the opposite. The word "中俄" (China-Russia) is segmented before the word "反擊" (counterattack). If we consider the sentence as a whole, there is no major difference in their overall result.

As there is no major difference identified between these two 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 our own effort rather than building our own search engine. Experiments will be conducted in the next section to investigate whether the choice of search engines will affect the accuracy of the segmentation.

Experiments and Analysis

Comparison of accuracy of Chinese segmentation across systems is very difficult, because different studies use different datasets and different ground rules (Yang, Senellart, & 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.

TABLE 4. Statistics of n-grams in the sentence "美國反擊中俄開綠" (United States counterattack, China-Russia turns on green light) from Yahoo.

Bigram	Char hits	Trigram	Char hits	Quadgram	Char hits
美國	37,878,400	美國反	186,066	美國反擊	1,790
國反	2,770	國反擊	32	國反擊中	0
反擊	1,320,302	反擊中	4,622	反擊中俄	0
擊中	662,001	擊中俄	75	擊中俄開	0
中俄	6,250,099	中俄開	50	中俄開綠	0
俄開	657	俄開綠	0	俄開綠燈	0
開綠	546	開綠燈	95,550		
綠燈	246,018				

[1]All the data in this paper were collected in September 2005.

DOI: 10.1002/asi

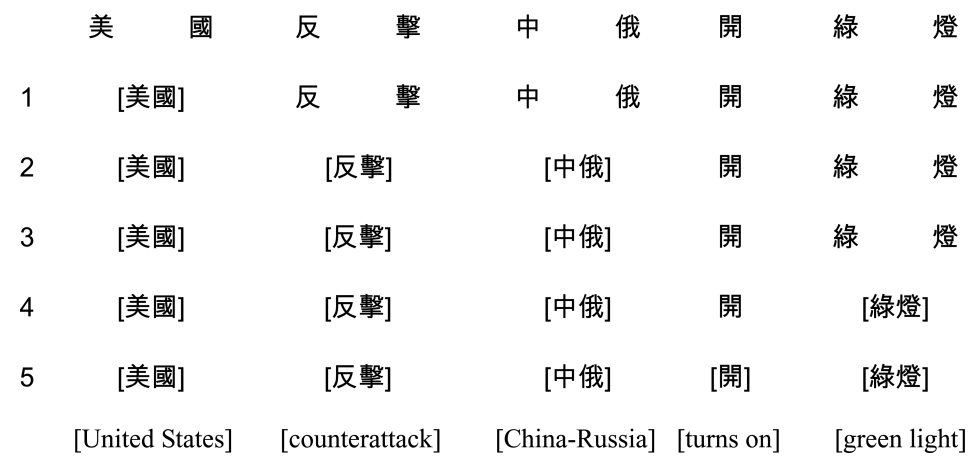

FIG. 12. Segmentation of the sentence "美國反擊中俄開綠燈" (United States counterattack, China-Russia turns on green light) based on statistical data from Google.

FIG. 13. Segmentation of the sentence "美國反擊中俄開綠燈" (United States counterattack, China-Russia turns on green light) based on statistical data from Yahoo.

In the competition, four datasets are provided. Each of the datasets consists of a corpus of training data with segmentation by human professionals and a document for testing. These datasets are from four different areas, namely, Hong Kong, mainland China (Peking), Taiwan, and the United States of America. Therefore, the corpora are significantly different. Moreover, they are significantly different in their segmentation standards.

Twelve sites participated in the competition. Because of the difficulties of Chinese segmentation, it is difficult to have a unified segmentation algorithm. 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 allowed to use any other materials: other training corpora, dictionaries, Web data, and so forth. The results of the competition are summarized in Table 5, where the Precision

TABLE 5. Summary of results of the First International Chinese Word Segmentation Bakeoff.

	Taiwan corpus			USA corpus			HK corpus			Peking corpus		
Site name	*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

Note. The experiment was conducted in September 2005. The results may vary with time base of the dynamic property of the Web.

DOI: 10.1002/asi

P, Recall R, and F-score F are defined as the standard formula:

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

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

$$F - Source(F) = \frac{2 \times P \times R}{(P + R)}$$

As shown in Table 5, we can see that the accuracy of each segmentation algorithm varies from corpus to corpus. When the performances of one participant using different corpora are compared, it is found that the performance on the United States corpus was generally lower than on the other corpora for most participants. Although the winning team in the United States corpus achieves a higher accuracy than the winning team in the Taiwan corpus, most teams have lower accuracy in the United States corpus among different corpora. Therefore, the average accuracy of the United States corpus is the lowest among all corpora (Table 6). It is mainly because the number of new words (out-of-vocabulary) in this corpus is much higher than in the others (Sproat & Emerson, 2003).

In order to analyze the accuracy of our algorithm, we have conducted an experiment using the corpora of the competition. Table 6 compares the accuracy of the segmentation algorithm by Web mining with the accuracy of the segmentation system of the participating sites in the competition.

As shown in Table 6, the t-test analysis shows that the segmentation algorithm by Web mining is independent of search engines. There is no substantial difference in the accuracies for the segmentation algorithm by using different search engines (Table 6). As F-score is considered as an overall measurement for system performance (Rijsbergen, 1979), the F-score of the segmentation by Web mining is compared with other systems. The algorithm has higher accuracy than the average of other systems in the first three corpora and an F-score a little bit lower than the average in the Peking corpus. On the other hand, the accuracy of the segmentation algorithm by Web mining is compared with the winner sites of the competition. As shown in Table 6, all the winner sites can win in only one corpus. However, the new algorithm outperforms the winner sites in the Taiwan and United States corpora. Despite the fact that the new algorithm does not take any data from the training corpora, the proposed technique achieves an accuracy level comparable to that of the state-of-art segmentation algorithms.

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 data from the Web, and vocabulary size of online documents is so large that it can eliminate the effects of new words. One-tenth documents are randomly selected from each corpus of the competition (Sproat & Emerson, 2003) to test the algorithm. The testing data contain 10,406 words and 1,126 new words, such as name of person, name of organization, and technical terms. The out-of-vocabulary rate is 10.8% in the testing data. Among the new words, only 61 (5.4%) new words appear three or more times in the testing data. The proposed technique is applied to segment the testing data. The recall of new words is calculated as the ratio of new words that can be correctly segmented by the system. In our analysis, it is found that the new algorithm is good at detection of new words. Our system records a recall of 0.693, while the participants in the competition achieved an average recall of 0.594 (Sproat & Emerson, 2003). The specialists in new word detection have reported a recall of 0.680 (Chen & Ma, 2002). Our system can obtain recall higher than that of the participants in the competition as well as the new word detection specialists. The novel technique is an effective technique in new word detection.

TABLE 6. Comparison of accuracy of segmentation algorithm by web mining with the accuracy of participants in the First International Chinese Word Segmentation Bakeoff by using the Corpus Segmentation Standard.

	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
	SYSTRAN Software			ICL, Beijing U Inc.			CKIP Ac. Sinica Taiwan			Microsoft Research		
Highest in competition	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 (corpus standard)	0.919	0.901	0.910	0.956	0.911	0.933	0.936	0.895	0.915	0.902	0.894	0.898
Segmentation using Yahoo (corpus standard)	0.932	0.921	0.926	0.945	0.923	0.934	0.923	0.901	0.912	0.911	0.898	0.904

Note. The experiment was conducted in September 2005. The results may vary with time base of the dynamic property of the Web.

DOI: 10.1002/asi

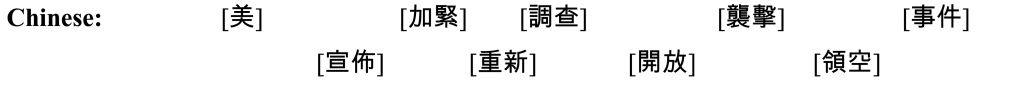

FIG. 14. Segmentation of the sentence of Chinese text 美加緊調查襲擊事件宣佈重新開放領空 (United States speed up investigation of surprise attack incident, announce open territorial airspace again.) based on the statistical data from Google.

As shown in Table 6, the average recall is higher than the average precision in all the corpora for other systems. However, our system achieves a precision higher than a recall. This problem is mainly caused by single-character words. Take the phrase "他的" as an example. Our system segments the phrase "他的" (his) as a two-character word; however, it is segmented as two one-character words as "他" (he) and "的" (of) in the competition. Similar cases occur many times in the corpora: "第五" (fifth) is segmented as "第" (number) and "五" (five), "一個" (one) is segmented as "一" (one) and "個" (unit), 個 and so on. However, they are segmented as a single word by our segmentation algorithm. Because the number of segmentation points in our system is lower than the standard answer, the recall is lower. Actually, this problem can be easily solved by statistical methods. For example, the mutual information of two characters can be calculated and only the pair of characters with mutual information higher than a threshold value is segmented as a word. However, it is not implemented in our experiment, because the authors believe that both segmentations are correct for those cases by personal judgment.

It is known that corpora from different geographical areas differ greatly. For example, the sentences in the documents from Hong Kong are usually shorter, and sentences in the documents from the Peking corpus are usually longer. We have taken one sentence from the Peking corpus as an example (Figure 14). According to the standard answer provided by the competition, the sentence is segmented correctly. However, the authors have identified the problem of compound words. As shown in Figure 14, the phrase "重新開放" is segmented as two two-character words "[重新] [開放]" ([open] [again]) by our segmentation algorithm and the standard answer provided. However, the authors believe that the phrase could be considered as one four-character word "重新開放" (reopen).

Research shows that 70% of Chinese words are made up of two characters (Sproat & Shih, 1990); these two-character words are further combined as longer words (Sproat & Shih, 2001). However, different segmentation standards may segment these words differently (Figure 1). Similar cases occur many times in the corpora. For example, the Chinese phrase "二十世紀" (twentieth century) is segmented as two words in our algorithm as "二十" (twenty) and "世紀" (century). However, it is segmented as one word in the testing corpora. Even in the standard corpora provided in the competition, unfortunately, there are inconsistencies found in the training corpora and the testing corpora provided by Taiwan and the United States (Sproat & Emerson, 2003). For example, the phrase "二十世紀 " (twentieth century) is segmented as "二十世紀 " ([twentieth century]) and "[二十] [世紀]" ([twenty] [century]) differently in the same corpus. Therefore, both segmentations should be accepted as correct answers.

According to our analysis, some of the segmentation mistakes by our algorithm are also adopted by some segmentation standards. This problem is mainly caused by the difference in the segmentation standards. Research has shown that segmentations by human professionals are consistent up to 76% only (Sproat, Shih, Gale, & Chang, 1996). The correct segmentation of these words is quite debatable. The main purpose of segmenting a Chinese text into words is to identify the words that could be utilized for information extraction in the future. As human professionals have different segmentation standards, all standards are considered as correct forms of segmentation. Therefore, we repeat the experiment by combining all segmentation standards together in favor of the machine segmentation to see how the result of machine segmentation compares with that of human segmentation. If all segmentation standards are taken into our consideration, the accuracy can be further increased (Table 7). This evaluation is to benchmark the proposed technique with human segmentation only. However, this evaluation is not intended to benchmark the proposed technique with techniques in the competition.

On the other hand, ANOVA analysis of *F*-score (Table 7) across corpora gives an *F*-ratio of 0. In other words, the performance of the proposed segmentation technique is independent of the corpus. Another experiment has been conducted to study the performance of the proposed technique in segmenting text from different geographical areas. We collected documents in these four different areas. For each area, 100 news articles were collected from major Chinese news agencies in the area for testing. The proposed technique was applied to segment news articles from different areas. As it has been shown that there is no significant difference in accuracy of the segmentation algorithm using different search engines, only Google is tested in this experiment. On the other hand, we have taken all the segmentation standards into consideration. Our system has measured similar accuracy using news articles from different areas (Table 8). The ANOVA analysis further confirms that there is no significant difference identified between the

DOI: 10.1002/asi

TABLE 7. Comparison of accuracy of segmentation algorithm by Web mining by using all segmentation standards and datasets in 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*
Segmentation using Google (corpus standard)	0.919	0.901	0.910	0.956	0.911	0.933	0.936	0.895	0.915	0.902	0.894	0.898
Segmentation using Google (all segmentation standards)	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 (corpus standard)	0.932	0.921	0.926	0.945	0.923	0.934	0.923	0.901	0.912	0.911	0.898	0.904
Segmentation using Yahoo (all segmentation standards)	0.952	0.943	0.947	0.953	0.946	0.949	0.949	0.933	0.941	0.939	0.919	0.929

Note. The experiment was conducted in September 2005. The results may vary with time base of the dynamic property of the Web.

performances of segmentation using texts from different areas. The proposed technique has a high level of accuracy during segmentation of texts from different geographical areas. Therefore, our algorithm is geographical area independent.

The new segmentation algorithm segments the texts on the basis of combined data of the *n*-gram search and Romanized pinyin search. Experiments have been conducted to evaluate the contribution of these data to the overall accuracy using Web data from Google (Table 9). Once again, we have taken all the segmentation standards into consideration. As shown in Table 9, the accuracy of segmentation by solely Romanized pinyin search is generally higher than of the segmentation solely by *n*-gram search except in the United States corpus. The t-test analysis shows that segmentation by Romanized pinyin-based search outperforms segmentation by character-based search at a 90% confidence interval. This is a strong evidence to show that the pinyin-based search is more reliable than the character-based search. The word boundary is indicated in Romanized pinyin and Segmentation by using the Web search result of Romanized pinyin takes the advantage of word boundary indication in Romanized pinyin. Therefore, it achieves a higher accuracy.

Our study found that about 70% of segmentation mistakes in segmentation solely by *n*-gram search are caused by stop words. On the other hand, the accuracy of segmentation by solely Romanized pinyin in the United States corpus is lower than in the other corpora. It is because the number of new words in this corpus is much higher than in the others, and we found that segmentation solely by Romanized pinyin search is weak in detecting new words. As a result, results by both *n*-gram search and Romanized pinyin search are important to segmentation. As shown in Table 9, the accuracy of segmentation by *n*-gram search and segmentation by Romanized pinyin search is significantly lower than that of segmentation by combining two search results. The new algorithm combines the data and can effectively segment the texts into words and record a higher level of accuracy.

TABLE 8. Accuracy of segmentation algorithm by Web mining using news articles from different areas.

Areas of news articles	*F*-score
Taiwan	0.938
United States	0.949
Hong Kong	0.943
Mainland China	0.939

Segmentation ambiguity is one of the major problems in Chinese segmentation (Goh et al., 2005; Li, Gao, Huang, & Li, 2003). Techniques have been developed to deal with segmentation ambiguity. Kit and colleagues reported a precision of 0.903 (Kit, Pun, & Chen, 2002), and Goh reported an accuracy of 94.3% (Goh et al., 2005). One experiment has been conducted to measure the effectiveness of the proposed algorithm to deal with segmentation ambiguity. A dataset of 200 sentences with 1,523 segmentation ambiguities was collected for evaluation. The proposed technique was applied to segment the sentences. The experimental result showed that the proposed technique achieves a precision of 0.931. As our technique can achieve a level of accuracy comparable to that of a system that specializes in ambiguity resolution, the proposed technique is a promising technique to deal with segmentation ambiguity.

As discussed in Statistic-Based or Hybrid Approaches, the stop word problem is another major problem in segmentation (Sproat & Shih, 2001). Detailed analysis of the segmentation result also showed that 96.1% of stop words are segmented correctly by our system. However, to the best knowledge of the authors, there is no literature that reports the accuracy of detecting stop words for segmentation techniques. Traditionally, the segmentation systems remove stop words by matching the texts with a stop word list and then report the final result only (Tsai, Sung, & Hsu, 2003). They usually do not report the percentage of stop words that have been segmented correctly. As a result, we cannot compare our results with those of the other system. Because our system measures a higher level of accuracy, it can be concluded that our algorithm can deal with the stop word problem effectively.

DOI: 10.1002/asi

TABLE 9. Accuracy of segmentation algorithm by Web mining using different Web data.

	Taiwan corpus			USA corpus			HK corpus			Peking corpus		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
n-Gram only	0.693	0.676	0.684	0.722	0.703	0.712	0.732	0.719	0.725	0.687	0.667	0.677
Romanized pinyin only	0.761	0.753	0.757	0.652	0.647	0.649	0.803	0.775	0.789	0.812	0.783	0.797
n-Gram + Romanized pinyin	0.946	0.931	0.938	0.978	0.935	0.956	0.968	0.925	0.946	0.931	0.904	0.918

Experimental results show that the segmentation algorithm achieves an accuracy level as high as that of the state-of-art segmentation algorithms in the segmentation competition. Moreover, the novel algorithm is geographical area independent and can segment Chinese documents from different geographical areas effectively. It can also deal with the problems of segmentation ambiguity, new word detection, and stop words. In conclusion, the proposed algorithm is a promising technique in Chinese segmentation.

Finally, this study is intended to investigate the prospect of segmenting Chinese texts by Web mining. We have shown that there is no major difference identified between different search engines; we assume that similar results will be obtained by crawling Web pages and collecting the statistics of Chinese *n*-grams by our own effort rather than building our own search engine. However, the time complexity is another important issue. It is difficult to predict the time required for an algorithm that is developed on the baisis of search engines from another party. In order to have an accurate estimation of time cost, it is desirable to build one's own database for storing the statistical information on Chinese *n*-grams. Moreover, a preprocessed database of *n*-grams can speed up the segmentation.

Conclusion

Search engines maintain huge databases of online documents and frequencies of character sequences in the documents. Their databases are important resources in Chinese segmentation. A novel segmentation algorithm by mining Web data is proposed in this paper. This is the first algorithm 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 that of other state-of-art segmentation algorithms.

There are many different ways to improve the accuracy of the proposed Chinese segmentation algorithm. The following are some directions suggested for future research:

- Our system considers *n*-grams up to quadgrams only. However, there are some words in the testing corpus that contain more than four characters, for example, "賓西法尼亞州" (Pennsylvania). Our system fails to extract those words. A higher value of *n* could be chosen to solve the problem.
- There is segmentation ambiguity of the compound words in different segmentation standards. An algorithm is urgently needed for combining the words into a longer word.
- We simply use the arithmetic mean of two returned numbers of hits as the number of hits for the *n*-gram. We will investigate the relationship between these two numbers in the future.
- Our algorithm tries to maximize the number of character hits returned by the search engines. However, different formulations have been proposed in the past for the association between characters, for example, mutual information and the hidden Markov model. It is worthwhile to investigate the application of those formulations in Web data.
- A greedy algorithm has been used in our algorithm, and different matching algorithms have been proposed previously. In the future, we will study the effect of different algorithms.
- In the future, we will build an *n*-gram database by crawling Web pages and collecting the statistics of Chinese *n*-grams.
- There are pinyin systems in different areas. Study should be conducted to investigate the differences of the pinyin systems.

The novel algorithm is directly based on the number of hits returned by search engines. Advanced techniques have been developed for segmentation. It is natural to predict that integration of those techniques with our algorithm can further improve accuracy. As a result, our algorithm is a promising technique in Chinese segmentation. The proposed algorithm is the first unified segmentation algorithm for documents from different Chinese areas. Moreover, the new algorithm can solve the problem of new word detection.

References

Agirre, E., Ansa, O., Hovy, E., & Martinez, D. (2002). Enriching very large Ontologies using the WWW. In Proceedings of the First Workshop on Ontology Learning (OL'2000), Berlin, Germany, August 2000, CEUR Workshop Proceedings Series, Vol. 31, Sun SITE Central Europe (CEUR), Aachen, German, (pp. 73–77). Held in conjunction with the 14th European Conference on Artificial Intelligence ECAI'2000, Berlin, Germany.

Baidu. (2006). Retrieved from http://www.baidu.com

Banko, M., Brill, E., Dumais, S., & Lin, J. (2002). AskMSR: Question answering using the Worldwide Web. In Proceedings of 2002 AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases, Palo Alto, California, USA, March 2002, American Association for Artificial Intelligence, Menlo Park, California (pp. 7–8).

Brants, T., & Franz, A. (2006). Web 1T 5-Gram Version 1, Linguistic Data Consortium, Philadelphia. Retrieved from http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13

Brin, S., & Page, L., The anatomy of a large-scale hypertextual Web search engine. In Proceedings of the Seventh International World Wide Web Conference (WWW7), Brisbane, Australia, April 1998 (pp. 107–117). Amsterdam, Netherland: Elsevier Science B.V.

DOI: 10.1002/asi

Cambridge University Library. (2006). Chinese classification system. Retrieved from http://www.lib.cam.ac.uk/mulu/class.html

Chang, J.S., & Su, K.Y. (1997). An unsupervised iterative method for Chinese new lexicon extraction. International Journal of Computational Linguistics and Chinese Language Processing, 2(2), 97–148.

Chen, K.J., & Bai, M.H. (1998). Unknown word detection for Chinese by a corpus-based learning method. International Journal of Computational Linguistics and Chinese Language Processing, 3(1), 27–44.

Chen, K.J., & Ma, W.Y. (2002). Unknown word extraction for Chinese documents. In Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002), Taipei, Taiwan, August 2002, Association for Computational Linguistics, Morristown, New Jersy, USA, (vol. 1, pp. 1–7).

Church, K., & Hanks, P. (1989). Word association norms, mutual information, and lexicography. In Proceedings of 27th Annual Meeting of the Association for Computational Linguistics (ACL-89), Vancouver, British Columbia, Canada, June 1989, Association for Computational Linguistics, Morristown, New Jersey, USA, (pp. 76–83).

Cimiano, P., Handschuh, S., Staab, S. (2004). Towards the self-annotating Web. In Proceedings of the 13th World Wide Web Conference (WWW2004), New York, USA, May 2004 (pp. 462–471). New York, USA: ACM Press.

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

Edmundson, H. (1968). New method in automatic extraction. Journal of the ACM, 16(2), 264–285.

Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.M., Shaked, T., Soderland, S., Weld, D.S., Yates, A. (2004). Web-scale information extraction in KnowItAll (preliminary results). In Proceedings of the 13th World Wide Web Conference (WWW2004), New York, USA, May 2004 (pp. 100–109). New York, USA: ACM Press.

Gan, K.W. (1995). Integrating word boundary identification with sentence understanding. Unpublished doctoral dissertation, National University of Singapore.

Ge, X.P., Wanda, P., Padhraic, S. (1999). Discovering Chinese words from unsegmented text. In Proceedings of the 22nd Annual Inernational ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99) Berkeley, California, USA, ACM Press, New York, USA, (pp. 271–272).

Goh, C.L., Asahara, M., & Matsumoto, Y. (2005) Chinese word segmentation by classification of characters. Computational Linguistics and Chinese Language Processing, 10(3), 381–396.

Google (2005). Retrieved from http://www.google.com

Hawking, D. (2006). Web search engines: Part 2. Computer, 39(8), 88–90.

Huang, C.H., Chen, K.J., Chang, L.L., & Chen, F.Y. (1997). Segmentation standard for Chinese natural language processing. International Journal of Computational Linguistics and Chinese Language Processing, 2(2), 47–62.

Keller, F., Maria, L., & Olga, O. (2002). Using the Web to overcome data sparseness. In Jan Hajic and Yuji Matsumoto (Eds.). In Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, Philadelphia, Pennsylvania USA, July 2002, Association for Computational Linguistics, Morristown, New Jersey, USA, (Vol. 10, pp. 230–237)

Kit, C., Pan, H., Chen, H. (2002). Learning case-based knowledge for disambiguating Chinese word segmentation: a preliminary study. In Proceedings of the first SIGHAN Workshop on Chinese Language Processing, Taipei, Taiwan, September 2002, Association for Computational Linguistics, Morristown, New Jersey, USA, (Vol. 18, pp. 1–7).

Kucera, H., & Francis, W. (1967). Computational analysis of presentday American English. Providence, RI: Brown University Press.

Kwok C.C.T., Etzioni, O., & Weld, D.S. (2001). Scaling question answering to the Web. In Proceedings of the 10th International World-Wide Web Conference (WWW10), Hong Kong, May 2001 (pp. 150–161). New York, USA: ACM Press.

Language Information Sciences Research Center. (2006). The most common Chinese new words in 2004. Hong Kong: City University of Hong Kong.

Leung, C.H., & Kan, W.K. (1996). A statistical learning approach to improving the accuracy of Chinese word segmentation. Literary and Linguistic Computing, 11(2), 87–92.

Li, H.Q., Huang, C.N., Gao, J.F., & Fan, X.Z. (2004). The use of SVM for Chinese new word identification. In Proceedings of the First International Joint Conference on Natural Language Processing (IJCNLP-04), Sanya City, Hainan Island, China, March 2004, Lecture Notes in Computer Science Series, Vol. 3248, Springer, 2005, Springer, Berlin, Germany, (pp. 723–732).

Li, M., Gao, J., Huang, C., & Li, J. (2003). Unsupervised training for overlapping ambiguity resolution in Chinese word segmentation. In Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, Sapporo, Japan, July 2003, Association for Computational Linguistics, Morristown, New Jersey, USA. (vol. 17, pp. 1–7).

Library of Congress. (2004). New Chinese romanization guidelines. Retrieved fromhttp://www.loc.gov/catdir/cpso/romanization/ chinese.pdf

Luhn, H. (1958). Automatic creation of literature abstracts. IBM Journal of Research and Development, 2(2), 159–165.

Lunde, K. (1998). CJKV information processing: Chinese, Japanese, Korean & Vietnamese computing. New York: O'Reilly.

Ma, W.Y., & Chen, K.J. (2003) A bottom-up merging algorithm for Chinese unknown word extraction. In Annual Meeting of the ACL, Proceedings of the Second SIGHAN workshop on Chinese language processing, Sapporo, Japan, 31–38.

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

Moukdad, H., & Cui, H. (2005). How do search engines handle Chinese queries? Webology, 2(3), Article 17. Retrieved from http://www.Webology.ir/2005/v2n3/a17.html

National Library of Australia. (2002). Kinetica Chinese Japanese Korean (CJK) service report. Retrieved from http://www.nla.gov.au/librariesaustralia/cjk/events/cjkreport2002.html

National People's Congress. (1958). Han yu pin yin fang an 漢語拼音方案, Beijing Author.

Nie, J.Y., Jin, W., Hannaan, M.L. (1994). A hybrid approach to unknown word detection and segmentation of Chinese. In Proceedings of International Conference on Chinese Computing (ICCC-94), Singapore, June 1994, National University of Singapore, Singapore, (pp. 326–335).

Pan, W., Yip, P.C., & Han, Y.S. (1993). Studies in Chinese word-formation. Taipei: Student Book Company.

Radev, D.R., Qi, H., Zheng, Z., Blair-Goldensohn, S., Zhang, Z., Fan, W., Prager, J. (2001). Mining the Web for answers to natural language questions. In Proceedings of the tenth International Conference on Information and Knowledge Management (CIKM 2001), Atlanta, Georgia, USA, November 2001 (pp. 143–150). New York, USA: ACM Press.

Rijsbergen, C.V. (1979). Information retrieval (2nd ed.). Glascow: Department of Computer Science, University of Glasgow.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing and Management, 24, 513–523.

Sampson, G. (1989). How fully does a machine-usable dictionary cover English text? Literary and Linguistic Computing, 4, 29–35.

Soanes, C., Stevenson, A. (2005). Oxford Dictionary of English, Oxford, UK: Oxford University Press.

Sproat, R., Emerson, T. (2003). The first international Chinese word segmentation bakeoff. In Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, Sapporo, Japan, July 2003, Association for Computational Linguistics, Morristown, New Jersey, USA. (vol. 17, 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, 4, 336–351.

Sproat, R., Shih, C. (2001). Corpus-based methods in Chinese morphology and phonology. 2001 Linguistic Society of America Institute (LSA 2001), University of California – Santa Barbara, California, USA. Retrieved from, http://compling.ai.uiuc.edu/rws/newindex/notes.pdf

DOI: 10.1002/asi

Sproat, R., Shih C., Gale, W., Chang, N. (1996). A Stochastic Finite-State Word-Segmentation Algorithm for Chinese, Computational Linguistics, 22(3), 377–404.

State Bureau of Technology Supervision. (1992). Contemporary Chinese languages word-segmentation specification for information processing. Beijing: Author.

Sun, M.S., Shen, D.Y., T'sou, B.K. (1998). Chinese word segmentation without using lexicon and hand-crafted training data. In Proceedings of the 17th International Conference on Computational Linguistics (COLING 1998), Montreal, Quebec, Canada, August 1998, Association for Computational Linguistics, Morristown, New Jersey, USA, (vol. 2, pp. 1265–1271).

Teahan, W.J., Wen, Y.Y., McNad, R., & Witten I. (2000). A compression-based algorithm for Chinese word segmentation. Computational Linguistics, 26(3), 375–393.

Tsai, J.L., Sung, C.L., Hsu W.L. (2003) Chinese word auto-confirmation agent. In Proceedings of Fifth Conference on Computational Linguistics and Speech Processing (ROCLING XV), Hsinchu, Taiwan, September 2003, Association for Computational Linguistics and Chinese Language Processing, Taipei, Taiwan, (pp. 175–192).

Wang, Y.H., Su, H.J., & Mo, Y. (1990). Automatic processing of Chinese words. Journal of Chinese Information Processing. 4(4), 1–11.

Wettler, M., Rapp, R., (1993) Computation of word associations based on the co-occurrences of words in large corpora. In Proceedings of the First Workshop on Very Large Corpora (WVLC-1), Columbus, Ohio, June 1993, Association for Computational Linguistics, Morristown, New Jersey, USA, (pp. 84–93).

Wu, Z., & Tseng, G. (1993). Chinese text segmentation for text retrieval: Achievements and problems. Journal of the American Society for Information Science, 44(9), 532–542.

Xia, F. (1999). Segmentation guideline. Chinese Treebank Project. Technical Report, University of Pennsylvania. Retrieved from http://morph/ldc.upenn.edu/ctb/

Yang, C. C., & Li, K. W. (2003). Segmenting Chinese unknown words by heuristic method. In Proceedings of the International Conference on Asia Digital Libraries, Malaysia, December 2003, Lecture Notes in Computer Science, vol. 2911, Springer, Berlin, Germany, (pp. 510–520).

Yang, C. C., & Li, K. W. (2004). Error analysis of Chinese text segmentation using statistical approach. In Proceedings of ACM/IEEE Joint Conference on Digital Libraries, Tucson, Arizona, USA, June 2004 (pp. 256–257). New York, USA: ACM Press.

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, 56(13), 1438–1447.

Yang, C.C., Luk, J., Yung, J., & Yen, J. (2000). Combination and boundary detection approach for Chinese indexing. Journal of the American Society for Information Science, Special Topic Issue on Digital Libraries, 51(4), 340–351.

Yang, J., Senellart, J., Zajac, R. (2003). SYSTRAN'S Chinese word segmentation. In Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, Sapporo, Japan, July 2003, Association for Computational Linguistics, Morristown, New Jersey, USA. (vol. 17, pp. 180–183).

Zipf, G.K. (1949). Human behavior and the principle of least effort. Cambridge, Massachusetts, USA: Addison-Wesley.

DOI: 10.1002/asi