Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 36 (2009) 3641–3651
Expert Systems with Applications
www.elsevier.com/locate/eswa

Automatic extraction of new words based on Google News corpora for supporting lexicon-based Chinese word segmentation systems

Chin-Ming Hong [a], Chih-Ming Chen [b,*], Chao-Yang Chiu [a]

[a] Department of Applied Electronics Technology, National Taiwan Normal University, Taipei 106, Taiwan, ROC
[b] Graduate Institute of Library, Information and Archival Studies, National Chengchi University, Taipei 116, Taiwan, ROC

Abstract

Chinese word segmentation is an essential step in a processing of Chinese natural language because it is beneficial to the Chinese text mining and information retrieval. Currently, the lexicon-based Chinese word segmentation scheme is widely adopted, which can correctly identify Chinese sentences as distinct words from Chinese language texts in real-word applications. However, the word identification ability of the lexicon-based scheme is highly dependent with a well prepared lexicon with sufficient amount of lexical entries which covers all of the Chinese words. In particular, this scheme cannot perform Chinese word segmentation process well for highly changeable texts with time, such as newspaper articles and web documents. This is because highly changeable documents often contain many new words that cannot be identified by a lexicon-based Chinese word segmentation system with a constant lexicon. Moreover, to maintain a lexicon by manpower is an inefficient and time-consuming job. Therefore, this study proposes a novel statistics-based scheme for extraction of new words based on the categorized corpora of Google News retrieved automatically from the Google News site to promote the word identification ability for lexicon-based Chinese word segmentation systems. Since corpora of news almost contain all words used in daily life, to extract news words from corpora of news and to incrementally add them into lexicon for lexicon-based Chinese word segmentation systems provide benefits in terms of automatically constructing a professional lexicon and enhancing word identification capability. Compared to another proposed scheme of new word extraction, the experimental results indicated that the proposed extraction scheme of new words not only more correctly retrieves new words from the categorized corpora of Google News, but also obtains larger amount of new words. Moreover, the proposed scheme of new word extraction has been applied to automatically expand the lexicon of the Chinese word segmentation system ECScanner (A Chinese Lexicon Scanner with Lexicon Extension). Currently, the ECScanner has been published on the Web to provide Chinese word segmentation service based on Web service. Experimental results also confirmed that ECScanner is superior to CKIP (Chinese knowledge information processing) in identifying meaningful Chinese words.

Keywords: Natural language processing; New word extraction; Chinese word segmentation; Information retrieval

1. Introduction

Identifying English or the other western languages texts into distinct words is natural and trivial task. By contrast, it is a very challenge and difficult task for Chinese texts, since Chinese texts consist of a string of ideographic characters without any blanks to mark word boundaries between words except for punctuation signs at the end of each sentence, and occasional commas within sentences (Chen & Liu, 1992; Foo & Li, 2004; Yeh & Lee, 1991; Zhang, Lu, & Zou, 2004). However, the word segmentation is a necessary step in processing Chinese texts, such as machine translation, Chinese text mining and information retrieval. To survey the past studies (Chen & Liu, 1992; Foo & Li, 2004; Yeh & Lee, 1991; Zhang et al., 2004), Chinese word segmentation can be categorized as three approaches including the word identification (i.e. lexicon-based identification scheme), statistical word identification,

* Corresponding author. Tel.: +886 2 29393091x88024; fax: +886 2 29384704.
E-mail address: chencm@nccu.edu.tw (C.-M. Chen).

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.02.013

and hybrid word identification schemes. The basic technique for identifying distinct words from Chinese texts is based on the lexicon-based identification scheme (Chen & Liu, 1992), which performs word segmentation process using string matching algorithms supported by a well prepared lexicon with sufficient amount of lexical entries which covers all of the Chinese words as possible. However, such a large lexicon is difficult to be constructed or maintained by manpower since the set of words is open-ended. Therefore, many used words in Chinese texts for word segmentation are often out-of-lexicon words due to insufficient amount of lexical entries so that the accuracy of Chinese word segmentation is degraded. The extraction of new words becomes a key technology for the lexicon-based Chinese word segmentation systems (Chen & Bai, 1998; Chen & Ma, 2002; Lin & Yu, 2001; Ma & Chen, 2003; Shan & Jie, 2001; Wai, Cheung, & Huang, 2004).

Moreover, the poor word segmentation results usually occur while using a general lexicon in a lexicon-based Chinese word segmentation system for specific domain texts. The best solution is to detect new words from the corpora of domain-specific and to add them into the original lexicon. With the rapid growth of Internet information, extracting new words automatically based on a large amount of collecting corpora from the Internet has become a likely task, specially from online daily Web news (Lin et al., 1998; Lu, Chien, & Lee, 2002). Currently, Google News aggregator has gathered nearly 10,000 news sources from the World Wide Web by an automatic crawler and these news sources are presented as news stories/categories in a searchable format on the Google News site. Google News uses an automatic process to pull together related headlines into a news story/category, which enables people to see many different viewpoints on the same story/category.

Therefore, this paper proposes a novel statistics-based scheme of new word extraction based on the categorized corpora of Google News automatically retrieved from the Google News site to detect new words that appear in daily Google News titles. In parallel, the proposed scheme of new word extraction is also applied to expand the lexicon of the proposed Chinese word segmentation system ECScanner (A Chinese Lexicon Scanner with Lexicon Extension, 2006) in order to improve the word identification capability. Additionally, to avoid the performance reduction of Chinese word segmentation process due to too large lexicon derived from new word extension, this study also proposes a fuzzy rule based approach to eliminate out-of-date new words based on the inferred confidence degrees of new words. The experimental results revealed that the proposed scheme of new word extraction has excellent performance in terms of the amount of extracting new words and high accuracy rate of new words. Moreover, the expanding lexicon can obviously enhance the word identification capability for the proposed lexicon-based word segmentation system ECScanner due to the reduction of unknown words.

2. The proposed scheme of new word extraction

This section aims to detail the proposed scheme of new word extraction, and is organized as follows: Section 2.1 describes why to extract new words from Google News titles, and Section 2.2 explains the detailed procedures of the proposed scheme of new word extraction. Section 2.3 proposes how to develop a Chinese word segmentation system based on the proposed scheme of new word extension, and Section 2.4 presents how to infer the life cycle of the extracted new words for eliminating out-of-date new words.

2.1. Detecting new words from Google News corpora

Google News aggregator currently accesses nearly 10,000 news sources on the Internet via an automatic crawler program. The content from these sources is presented as news stories/categories in a searchable format on the web (Google News, 2007). Without regard to political viewpoint or ideology, leading stories selected automatically by a computer algorithm are viewed as headlines on the Google News home page. Google News uses an automated process to group related headlines together into a news story/category, which, in some cases, enables people to access different viewpoints on the same story/category (Google News). News topics are updated continuously throughout the day and readers can view new stories/categories by checking the Google News website, subscribing to Google News alerts via email, or activating a really simple syndication (RSS). The Google News service is currently tailored to 22 international audiences (Google News, 2007). These related headlines grouped together provide useful information for extracting new words. Generally, the contents in news articles vary highly with time, generating new words frequently. Therefore, this study utilized a Google News story/category which has grouped related headlines together to extract likely new words from news titles.

To utilize the characteristic of new word occurring frequently in daily life news articles, this study presents a statistics-based scheme of new word extraction for the lexicon-based Chinese word segmentation systems for the promotion of word identification capability. Although inducing the appearance opportunity and position of new words from the combination of Chinese linguistic words is difficult, meaningful new words in news articles always have a high frequency of appearing in most news articles (Shan & Jie, 2001). If a word is meaningful and can be considered as a new word in Google news titles, then it should appear in numerous news events in the same news story/category, rather than being concentrated in a few news events in a news story/category (Chowdhury, 2004). Based on the property, this study proposes a statistics-based scheme of new word extraction that can assist the Chinese word segmentation system CScanner (Cscanner, 2000) in

extending its Chinese word lexicon automatically, thus speeding up its Chinese word segmentation capability.

2.2. The procedures of automatically detecting new words

First, this section explains the detailed procedures of the proposed scheme of new word extraction, which can be summarized as five steps, and described as follows:

Step 1. Counting duplicate combination frequencies of all likely *n*-gram words that appear in news titles categorized in the same Google News story/category.

Suppose W_{ni} represents the *i*th *n*-gram word extracted from a randomly selected title of Google News categorized in a Google News story/category, which collected many news articles with highly related headlines. To detect whether the W_{ni} is a potentially new word, the proposed duplicate combination frequency is first employed to perform this task. If the occurrence frequency of W_{ni} in all news titles classified in the same news story/category is equal to m, then the duplicate combination frequency of W_{ni} is represented as DCF_{ni}, and computed as follows:

$$\mathrm{DCF}_{ni} = C_2^m = \frac{m!}{2!(m-2)!} \tag{1}$$

where DCF_{ni} stands for the duplicate combination frequency of the *i*th *n*-gram word, m is the occurrence frequency of the *i*th *n*-gram word in all news titles classified in the same news story/category.

Fig. 1 shows an example of calculating the duplicate combination frequency for the bi-gram word "台灣" (Taiwan). In this example, if the occurrence frequency of the bi-gram word "台灣" (Taiwan) in all Google News titles categorized in the same Google News story/category is two, then the duplicate combination frequency of this word can be computed as $\mathrm{DCF}_{2(\text{台灣})} = \frac{2!}{2!\times 0!} = 1$. Compared to another two statistics-based schemes for new word extraction from collecting corpora (Chen & Bai, 1998; Chen & Ma, 2002), the proposed duplicate combination frequency can enlarge the occurrence frequency among likely new words, such that the threshold of occurrence frequency for determining new words can be decided more easily as well as the ambiguity of threshold can be reduced.

Step 2. Computing the difference of the duplicate combination frequency between the *i*th *n*-gram word with the *j*th $(n+1)$-gram word which contains the *i*th *n*-gram word.

To compute the difference of the duplicate combination frequency between the *i*th *n*-gram word with the *j*th $(n+1)$-gram word containing the *i*th *n*-gram word aims to judge whether the *i*th *n*-gram word should be preserved or filtered out from the set of likely new words. The difference of the duplicate combination frequency of the *i*th *n*-gram word with the *j*th $(n+1)$-gram word that contains the *i*th *n*-gram word can be computed as,

$$\begin{cases} \mathrm{DDCF}_{ni} = \mathrm{DCF}_{ni} - \sum_{j=1}^{m}(\mathrm{DCF}_{(n+1)j}) & W_{ni} \subset W_{(n+1)j} \\ \mathrm{DDCF}_{ni} = \mathrm{DCF}_{nj} & W_{ni} \not\subset W_{(n+1)j}, \quad n = 2,3,4,\ldots,k \end{cases} \tag{2}$$

where DDCF_{ni} is the difference of the duplicate combination frequency between the *i*th *n*-gram word with the *j*th $(n+1)$-gram word, DCF_{ni} stands for the duplicate combination frequency of the *i*th *n*-gram word, and m is the number of the $(n+1)$-gram words that contain the *i*th *n*-gram word.

In Eq. (2), if the *i*th *n*-gram word does not contain in the subset of the *j*th $(n+1)$-gram word, then the difference of the duplicate combination frequency of the *i*th *n*-gram word with the *j*th $(n+1)$-gram word is equal to the original duplicate combination frequency of the *i*th *n*-gram word. Table 1 gives an example to show how to compute the difference of the duplicate combination frequency. Suppose the duplicate combination frequency of the bi-gram word "速公 (sung kung)"is equal to 10 in this example. The difference of the duplicate combination frequency of the bi-gram word "速公" with the tri-gram words "高速公 (gao su gong)" and "速公路 (su gong lu)" which simultaneously contains the bi-gram word "速公" can be computed as,

$$\begin{aligned} \mathrm{DDCF}_{2(\text{速公})} = \mathrm{DCF}_{2(\text{速公})} - (\mathrm{DCF}_{3(\text{高速公})} \\ + \mathrm{DCF}_{3(\text{速公路})}) = 10 - (6+6) = -2 \end{aligned} \tag{3}$$

Step 3. Judging whether the *i*th *n*-gram word should be preserved as a potential new word or filtered out from the set of all potential new words extracted from news titles based on a predetermined DDCF_{ni} threshold.

In this step, suppose the threshold value of DDCF_{ni} is set to R_n for determining whether the *n*-gram word should be preserved as a likely candidate new word or filtered out from the set of all likely new words. The decision rule can be described as

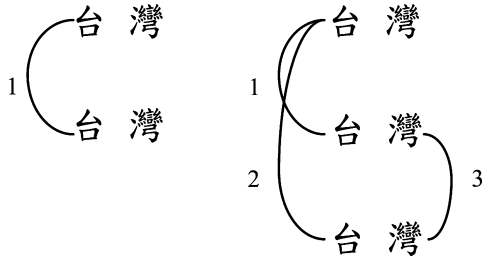

Fig. 1. An example of calculating the duplicate combination frequency.

Table 1
An example for computing the difference of the duplicate combination frequency

W_{ni}	DCF_{ni}	DDCF_{ni}
高速 (high speed)	15	9
速公 (su gong)	10	-2
公路 (highway)	21	15
高速公 (gao su gong)	6	0
速公路 (su gong lu)	6	0
高速公路 (super highway)	6	6

$$\begin{aligned}&\text{IF } \mathrm{DDCF}_{ni} \geqslant R_n \text{ THEN } W_{ni} \text{ is correct}\\&\text{ELSE } W_{ni} \text{ is wrong}\\&n = 2,3,4,\ldots,k\end{aligned} \tag{4}$$

Actually, the threshold value of DDCF_{ni} plays a key role and it is difficult to be appropriately determined because the assigned threshold value will be trade-off between the amount of likely new words and the quality of new words. If the R_n value is set to 5 in the given example, then the likely new words are illustrated as Table 2.

Based on the statistics analysis, this study found that bi-gram words detected by *Step 3* for extracting potential new words have a very high accuracy rate. This result provides benefit in terms of eliminating incorrect over bi-gram new words, and detailed in *Step 4*. This study also found that potential over bi-gram new words detected by *Step 3* are frequently incorrect. This is because different word combinations could occur in potential over bi-gram new words, thus generating ambiguous word segmentation results. For example, tri-gram words may be combined by a 1-gram word and a bi-gram word, and quad-gram words may be combined by two bi-gram words, two 1-gram words and a bi-gram word or a 1-gram word and a tri-gram word. To solve this problem, eliminating some possibly unreasonable tri-gram or quad-gram words based on the correct bi-gram words detected by *Step 3* is a practical method. The next step details the procedure of eliminating incorrect tri-gram and quad-gram words.

Step 4. Filtering out incorrect tri-gram and quad-gram words based on the bi-gram words obtained in the Step 3.

This step aims at checking the potential new words extracted by Step 3 once again in order to filter out some incorrect tri-gram and quad-gram words based on the correct bi-gram words. When a bi-gram word is included in a tri-gram or quad-gram word, the proposed decision rule is employed to judge whether the tri-gram or quad-gram word is correct by measuring whether the calculated DDCF_{2i} is smaller and equal to $M \times \mathrm{DDCF}_{3i}$ or $M \times \mathrm{DDCF}_{4i}$, where M is a constant. First, the proposed decision rule for judging the correct tri-gram words can be described as

$$\begin{aligned}&\text{IF } W_{ni} \subset W_{(n+a)j} \text{ and } \mathrm{DDCF}_{ni} \leqslant M \times \mathrm{DDCF}_{(n+a)j}\\&\text{THEN } W_{(n+a)j} \text{ is correct ELSE } W_{(n+a)j} \text{ is wrong}\\&\text{where } n = 2,\ a = 1,\ \text{and } M \text{ is a constant}\end{aligned} \tag{5}$$

Similarly, the likely incorrect quad-gram words can be filtered out by the correct tri-gram words, and so on. The proposed decision rule for judging the correct quad-gram words can be described as,

$$\begin{aligned}&\text{IF } W_{ni} \subset W_{(n+a)j} \text{ and } \mathrm{DDCF}_{ni} \leqslant M \times \mathrm{DDCF}_{(n+a)j}\\&\text{THEN } W_{(n+a)j} \text{ is correct ELSE } W_{(n+a)j} \text{ is wrong}\\&\text{where } n = 3,\ a = 1,\ \text{and } M \text{ is a constant}\end{aligned} \tag{6}$$

Finally, Table 3 summarizes the correct new words determined by *Step 3* under M is set to 3.

Step 5. Merging potential new words into the lexicon by comparing whether the extracted new words have been included in the original lexicon.

This step aims at avoiding duplicate words contained in the lexicon, thus reducing the performance of word identification. So far, a lexicon with new word extension can be successfully implemented to enhance word identification capability of lexicon-based Chinese word segmentation system ECScanner mentioned later. Moreover, the excellent word identification capability of ECScanner has provided benefits in terms of developing an intelligent news agent for tracking user-interested news events in our another study (Chen & Liu, 2006), such that a user can monitor a specific on-line news event more accurately.

Table 2
The list of all potential new words under the R_n value is set to 5 for the given example listed in Table 1

W_{ni}	DCF_{ni}	DDCF_{ni}
高速 (high speed)	15	9
公路 (highway)	21	15
高速公路 (super highway)	6	6

Table 3
The correct and reasonable potential new words determined by *Step 3* under M is set to 3

W_{ni}	DDCF_{ni}	Final decision
高速 (high speed)	9	Correct
公路 (highway)	15	Correct
高速公路 (super highway)	6	Correct

2.3. Chinese word segmentation system with automatic lexicon extension

At present, most Chinese word segmentation systems, such as CKIP (CKIP Chinese Parser, 2007) and CScanner (A Chinese Lexicon Scanner, 2000), were designed to perform Chinese word segmentation process for general domain Chinese texts based on a lexicon manually maintained by manpower. In general, it cannot handle the Chinese word segmentation well for specific domains, such as Chinese financial news. For example, the CKIP performed Chinese word segmentation process for the financial news title "聯電股價走勢強勁盤中完成填權 (The UMC stock price is rising and price recovery was finished during trading)" as "聯電 (UMC)/ 股價 (Stock Price)/走勢 (Trend)/強勁 (Rising)/盤中 (Trading)/完成 (Finish)". The result indicates that the word "填權" (Price Recovery) is a non-meaningful word after performing Chinese word segmentation process. This is because the current word lexicon in CKIP has not contained the word "填權" (Price Recovery) yet. Although manually maintaining new words for enhancing Chinese word segmentation systems is a major approach, it is a time-consuming and inefficient job. In particular, the contents of Web or news articles are highly variable with time, so that manually maintaining lexicon becomes very difficult.

To solve this problem, developing a Chinese word segmentation system with lexicon extension of new words is urgently needed specially for highly changeable texts. Therefore, the proposed scheme of new word extraction was employed to enhance lexicon-based Chinese word segmentation systems herein. Currently, the proposed scheme of new word extraction has been applied to automatically expand the lexicon of the Chinese word segmentation system CScanner (A Chinese Lexicon Scanner) in order to speed up the word identification capability. Additionally, to achieve new word extension automatically, an intelligent news crawler was also developed to extract Google News according to the assigned schedule time. The CScanner is an open source Chinese lexical scanner based on two variants of the maximum matching heuristic of word identification to perform Chinese word segmentation process. In the study, the Chinese word segmentation system CScanner with new word extension mechanism is renamed as ECScanner (ECScanner (A Chinese Lexicon Scanner with Lexicon Extension), 2007). Nowadays, the ECScanner has been published at the web site http://dlll.nccu.edu.tw/~rank/ecscanner/ to provide excellent Chinese word segmentation service by SOAP web service mechanism.

Fig. 2 shows the implemented ECScanner with the proposed mechanism of new word discovery, which provides the Chinese word segmentation service by the SOAP web service. The experimental results discussed later prove that the performance of Chinese word segmentation of the proposed ECScanner with the mechanism of new word discovery is superior to CKIP with the new word guessing mechanism.

Moreover, any new word discovery schemes have difficulty to produce completely correct new words. In other words, when the lexicon used the Chinese word segmentation system ECScanner completely accepts all new words discovered from the proposed scheme of new word extraction, the precision rate of the Chinese word segmentation will be declined. A relatively better strategy is to develop a new word management system that provides a friendly interface for linguistic experts to assess whether the discovered new words should be accepted formally as new words or rejected as un-meaningful words through an administrator interface. Fig. 3 presents the user interface of the implemented new word management system with three operation modes—accept, reject or recommend—to assist linguistic experts in determining whether candidate words discovered by the proposed scheme of new word discovery should be accepted as new words or rejected as un-meaningful words. After candidate new words are confirmed as new words by linguistic experts, the Chinese word lexicon used in the ECScanner is immediately updated based on a planning execution time managed by the administrator. To promote the efficiency of maintaining new words, all rejected new words confirmed by linguistic experts will not be recommended as likely new words again in the future. This mechanism will gradually reduce the working load of linguistic experts with time.

2.4. Inferring life cycles of extracted new words

A lexicon-based Chinese word segmentation system with new word extension could suffer from too large lexicon problem so that the performance of word identification is gradually descended with the growth of lexicon size. To conquer this problem, this study presents the concept of inferring life cycle for the extracted new words. The basic idea is to eliminate some out-of-date words from the expanded lexicon based on the extracted new words with

Fig. 2. The Chinese word segmentation system ECScanner.

Fig. 3. The news words management interface for determining whether the candidate words should be accepted as formal new words or rejected as unmeaningful words by linguistic experts.

the corresponding time stamp of occurrence time. The time stamp can be easily obtained from Google News site because a piece of news published on Google News site always contains the announcing time. Fig. 4 illustrates an example to display the distribution of the occurrence time of a new word recorded by the developed lexicon in the Chinese word segmentation system ECScanner (A Chinese Lexicon Scanner with Lexicon Extension, 2007). In Fig. 4, suppose the current time is t, and t_j represents the time that some new word appears in the Google News site in the jth time. The study presents two definitions related to the occurrence time of a new word to infer whether a new word should be eliminated from the lexicon by fuzzy inference (Lin & George Lee, 1996) based on a pre-designed fuzzy rule base.

$$T_1 = t - t_j$$
$$T_2 = \frac{t_j - t_1}{j} \tag{7}$$

where T_1 represents how long a new word has not appeared, T_2 stands for the average appearance period of a new word.

To obtain a simple fuzzy rule base for inferring the confidence degree of a new word, this study sets the numbers of input and output linguistic variables as three and five, respectively. The simplified representation notations of the input and output linguistic variables used for designing a fuzzy rule base to eliminate out-of-date words from the original lexicon in a Chinese word segmentation system

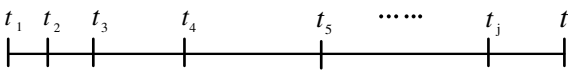

Fig. 4. The distribution of occurrence time of a new word.

Table 4
The linguistic variable for the input variables T_1 and T_2

Linguistic variable	Representation
Short	S
Moderate	M
Long	L

Table 5
The linguistic variable for the output variable "confidence degree of a new word"

Linguistic variable	Representation
Very lowly confidence degree	VLCD
Lowly confidence degree	LCD
Moderately confidence degree	MCD
Highly confidence degree	HCD
Very highly confidence degree	VHCD

are listed in Tables 4 and 5. Moreover, the triangle membership functions are applied to describe the linguistic variables of input and output for the fuzzy inference mechanism. To construct a reasonable fuzzy rule base, the membership functions used in both the input and output linguistic variables must be logically determined in advance. In this work, the K-means clustering algorithm (Rui & Wunsch, 2005) was applied to automatically determine the centers of the triangle fuzzy membership functions of input linguistic variables according to the data distribution of the linguistic variables T_1 and T_2 herein. Fig. 5 shows how to determine the fuzzy membership functions of input linguistic variables T_1 and T_2 based on three clustering centers determined by the K-means clustering algorithm. Moreover, the determined triangle fuzzy membership functions for the output linguistic variable are shown as Fig. 6.

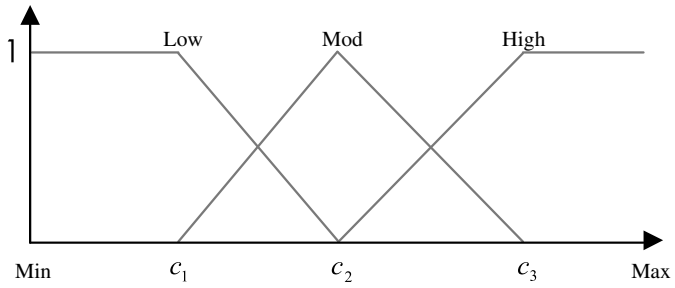

Fig. 5. The fuzzy membership functions automatically determined by the *K*-means clustering algorithm for the input linguistic variables T_1 and T_2.

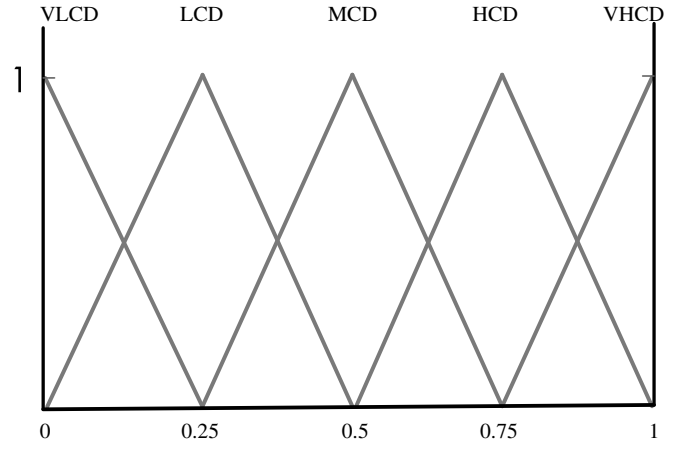

Fig. 6. The determined membership functions for the confidence degree of a new word.

Moreover, according to analyzing the appearance behavior of a new word, nine basic fuzzy rules were summarized to infer the confidence degree of a new word for determining the life cycle of an extracted new word. Table 6 illustrates the designed fuzzy rule base for inferring the confidence degree of a new word. To infer the confidence degree of a new word, the reasoning process of Mandani's minimum fuzzy implication (Lin & George Lee, 1996) was employed to integrate the triggered fuzzy rules. Moreover, the defuzzification method of center of gravity (Lin & George Lee, 1996) was used to obtain the crisp value of the confidence degree of a new word in order to serve as an index for determining whether the word is an out-of-date word. In this work, the range of inferred confidence degree for each new word is ranged between 0 and 1. Hence, a pre-assigned threshold of confidence degree must be determined to eliminate the new words that their corresponding confidence degrees are smaller than a pre-assigned threshold.

Table 6
The designed fuzzy rule base for inferring the confidence degree of a new word

The confidence degree of a new word	How long a new word does not appear		
	S	M	L
The average appearance period of a new word			
S	VHCD	HCD	MCD
M	HCD	MCD	LCD
L	MCD	LCD	VLCD

3. Experimental results

To show the excellent performance of the proposed scheme of new word extraction, Section 3.1 first reveals the performance evaluation results by extracting new words from Google News articles, and Section 3.2 presents the performance of the Chinese word segmentation system ECScanner with the proposed scheme of new word discovery. Finally, Section 3.3 assesses the performance of eliminating out-of-date words by inferring life cycles of extracted new words.

3.1. Performance evaluation of the proposed scheme of new word extraction

To demonstrate the performance of the proposed new word extraction, the proposed method was compared with Lin and Yu's method (Lin & Yu, 2001) in terms of the amount of extracting new words and the accuracy rate for 10 randomly selected Google News stories/categories which contain various numbers of news events. In the experiment, news events classified in the same story/category are respectively used to extract new words by the proposed method with Lin and Yu's method (Lin & Yu, 2001). Since the news titles always convey the most simplified information which is helpful to discover meaningful new words, the study used the news titles to evaluate the performance of the proposed scheme of new word extraction. Lin and Yu's method (Lin & Yu, 2001) adopted the proposed net frequency to determine the correct new words and the decision rule for new word extraction is to judge whether the net frequency is larger than 1 (i.e. net frequency >1). To give a fair comparison as possible, the threshold R_n in the proposed scheme of new word extraction is set to 1 and another parameter M is set to 2. Ten randomly selected Google News stories/categories which contain various numbers of news titles were adopted to assess new word extraction capability for the proposed scheme with Lin and Yu's method. Additionally, three Chinese language experts were invited to evaluate the same news titles used in this experiment for new word extraction. Table 7 shows the experimental results of new word extraction for a news story containing thirty news titles and the notation "*" stands for the incorrect new words identified by Chinese language experts. Moreover, the accuracy rate of new word extraction is defined and computed as follows:

$$\text{Accuracy rate} = \frac{\text{The number of correctly detected new words}}{\text{The total number of detected new words}} \times 100\% \quad (8)$$

Table 7
The experimental results of new word extraction for a news story (Lin & Yu, 2001)

Method	The potential new words	The number of extracted new words	Accuracy rate (%)
Lin and Yu's method	總統 (president)/中國 (china)/台灣 (Taiwan)/兩岸 (cross-Strait)/阿扁 (a bian)/武器 (weapon)/*扁倡 (bian chang)/美國 (America)/發表 (publish)/善意 (grace)/誠意 (sincerity)/*籲兩岸 (yu cross-Strait)/*議兩岸 (yi cross-Strait)/*十裁示 (ten decision)/邱義仁 (chiu yi ren)/國民黨 (KMT)/國台辦 (guo tai ban)/陳水扁 (chen shuei bian)/*欺天下 (ci tian sia)/*十點裁示 (ten decision)/*十項裁示 (ten decision)	21	66.67
The proposed method	總統 (president)/一中 (one china)/*十項 (ten items)/中國 (china)/台灣 (Taiwan)/台辦 (tai ban)/兩岸 (cross-Strait)/拒認/阿扁 (a bian)/武器 (weapon)/*扁倡 (bian chang)/*扁提 (bian ti)/美國 (America)/裁示 (decision)/發表 (publish)/善意 (grace)/誠意 (sincerity)/談話 (conversations)/*可復談 (ke fu tan)/邱義仁 (chiu yi ren)/國民黨 (KMT)/陳水扁 (chen shuei bian)/*欺天下 (ci tian sia)/九二共識 (common view in 1992)	24	79.17

The experimental results listed in Table 7 indicate that the proposed method surpasses the Lin and Yu's method in terms of the number of new word extraction and accuracy rate for a randomly selected news story. In addition, Table 8 summarizes the comparison results of the proposed method with Lin and Yu's method for 10 randomly selected Google News stories/categories which contain various numbers of news titles. The experimental results also confirm that the proposed method is superior to the Lin and Yu's method for almost all testing news stories. The average accuracy rates of the Lin and Yu's method and the proposed method for 10 randomly selected Google News stories/categories are 45.39% and 78.397%, respectively. To analyze the major reasons of performance promotion of the proposed method, this study summarized that the proposed duplicate combination frequency and the strategy of filtering out incorrect tri-gram and quad-gram words based on the correct bi-gram words mainly contribute the accuracy promotion of new word extraction.

Furthermore, to avoid affecting the accuracy rate of the proposed scheme of new word extraction, only a piece of

Table 8
The comparison results of the proposed method with Lin and Yu's method

Extracted Google News	The number of news titles	Lin and Yu's method (net frequency >1) (Lin & Yu, 2001)		The proposed method ($R_n = 1$, $m = 2$)	
		The number of extracted new words	Accuracy rate (%)	The number of extracted new words	Accuracy rate (%)
News story (1)	30	21	66.67	24	79.17
News story (2)	28	9	77.78	13	100
News story (3)	33	16	56.25	25	84
News story (4)	44	15	53.33	17	94.12
News story (5)	18	5	60	5	100
News story (6)	4	1	0	2	100
News story (7)	40	13	61.54	17	76.47
News story (8)	18	4	25	11	45.45
News story (9)	5	3	33.33	6	33.33
News story (10)	14	5	20	7	71.43
Average performance	23.4	9.2	45.39	12.7	78.397

news title is preserved and the others must be filtered out if news articles appear the completed same news title in a news story. Moreover, the experimental results also indicated that the news stories containing few pieces of news titles are not appropriately selected as corpora for new word extraction. This is logical because more news articles contained in a news story will increase the confidence of the used words.

3.2. Evaluating performance for the Chinese word segmentation system ECScanner with the lexicon extension

In this section, the precision rates of Chinese word segmentation schemes, which include identifying Chinese news events as several distinct words by linguistic experts, CKIP with unknown term guessing scheme and the ECScanner with the proposed scheme of new word discovery, are compared, respectively. Table 9 displays a randomly selected Chinese news event for identifying entire sentence into several distinct words by the linguistic expert, CKIP with unknown term guessing scheme and ECscanner with the proposed discovery scheme of new words. In this results, the notation "**" stands for the new words that have been identified by Chinese language experts, but the CKIP with unknown term guessing scheme cannot identify them well. The notation "*" stands for the incorrect new words identified by both the CKIP with unknown term guessing scheme and ECscanner with the proposed scheme of new word discovery. Moreover, Table 10 illustrates the statistic results of Chinese word segmentation by CKIP with unknown term guessing scheme and the ECScanner with the proposed discovery scheme of new words for the news event listed in Table 9. The results show that the ECscanner with the proposed scheme of new word discovery is superior to the CKIP with unknown words guessing scheme in terms of the number of identifying correct words and the number of unknown words. Furthermore, one hundred news events are randomly sampled from the Google News site for evaluating the performances of various Chinese word segmentation systems. Table 11 illustrates the comparison results for various Chinese word segmentation systems. The experimental results demonstrate that the proposed Chinese word segmentation system ECScanner with new word extension mechanism is indeed helpful to promote the accuracy and recall rates of Chinese word segmentation as well as reduce the unknown rate of Chinese word segmentation.

Table 9
A randomly selected news title for identifying entire sentence into several separated words by linguistic experts, CKIP with unknown term guessing scheme and the ECscanner with the proposed scheme of new word discovery

A new title randomly selected from Google News site	綠營：泛藍提罷免或倒閣政治鬥爭意圖奪權 (Pan-green alliance said that pan-blue alliance push the recall vote or topple the cabinet, and the purpose of politics strife is to take over power)
The Chinese word segmentation results by linguistic experts	綠營(Pan-Green Alliance)/泛藍(Pan-Blue Alliance)/罷免(Recall Vote)/倒閣(Topple the Cabinet)/政治(Politics)/鬥爭(Strife)/意圖(Purpose)/** *奪權(Take Over Power)*
The Chinese word segmentation results by CKIP with unknown word guessing scheme	綠營(Pan-Green Alliance)/* **泛藍提(Pan-Blue Alliance Push)**/罷免(Recall Vote)/倒閣(Topple the Cabinet)/政治(Politics)/鬥爭(Strife)/意圖(Purpose)
The Chinese word segmentation results by ECScanner with the proposed scheme of new word discovery	綠營(Pan-Green Alliance)/* **泛藍提(Pan-Blue Alliance Push)**/罷免(Recall Vote)/倒閣(Topple the Cabinet)/政治(Politics)/鬥爭(Strife)/意圖(Purpose)/** *奪權(Take Over Power)*

Table 10
The statistic results of Chinese word segmentation by CKIP with unknown term guessing scheme and the ECScanner with the proposed scheme of new word discovery for the news event listed in Table 9

Compared scheme	The number of correct words	The number of incorrect words	The number of unknown words
The Chinese word segmentation results by CKIP with unknown term guessing scheme	6	1	1
The Chinese word segmentation results by ECScanner with the proposed scheme of new word discovery	7	1	0

Table 11
The performance comparison of Chinese word segmentation for CKIP with unknown term guessing scheme and ECScanner with the proposed scheme of new word discovery for one hundred news events randomly selected from the Google News site

Compared scheme	Recall rate (%)	Precision rate (%)	Unknown rate (%)
The Chinese word segmentation results by CKIP with unknown word guessing scheme	88.96	95.39	6.75
The Chinese word segmentation results by ECscanner with the proposed scheme of new word discovery	93.35	96.86	3.62

Table 12
The summarization information of the gathered Google news stories

Item	Description
The dates of collection news	From 2004-11-9 to 2006-6-16; about 584 days
The number of news stories	1,191,065 stories
The number of news categories	60,269 categories
The average number of news stories per day	2039.49 stories/day
The average number of news stories per category	19.76 stories/category
The news category with maximum number of news stories	3007 stories
The news category with minimum number of news stories	2 stories

Table 13
Some likely out-of-date words with low confidence degree detected by the proposed scheme of eliminating out-of-date words

Likely out-of-date words	Property of out-of-date word	Likely out-of-date words	Property of out-of-date word
克林頓 (Clinton)	Out-of-date people name	炒金(Chao Jin)	Error word in the lexicon
割喉 (Cut off throat)	Infrequent word	元兌 (Yuan Dui)	Error word in the lexicon
撤並 (Che Bing)	Error word in the lexicon	林明成 (Lin Ming Cheng)	Out-of-date people name
世界第三 (The third in the word)	Infrequent word	西氣東輸 (West area gas transports to east area)	Infrequent word

3.3. Eliminating out-of-date words by inferring confidence degree of new word

In this experiment, a large number of news events collected from Google News site was used to evaluate out-of-date words by the proposed fuzzy inference scheme of life cycles of new words. Table 12 illustrates the summarization information of the gathered Google news. The number of news stories per category is from 2 to 20 in most parts of news categories. The period of collecting news is totally 584 days. Table 13 lists some likely out-of-date words with low confidence degree detected by the proposed scheme of eliminating out-of-date words. First of all, the results show that people names, such as 克林頓 (Clinton), and 林明成 (Lin Ming Cheng), are easily viewed as out-of-date words by the proposed scheme. This is because the name of Clinton president has infrequently appeared in Taiwan daily news since he has left from his president position for a long time. Moreover, infrequent words, such as 割喉 (Cut off throat), 世界第三 (The third in the word), and 西氣東輸 (West area gas transports to east area), are also easily viewed as out-of-date words because news media rarely report daily news using these words. Finally, this study also found that some error words contained in the lexicon, such as 炒金 (Chao Jin), and 元兌 (Yuan Dui), are easily detected as out-of-date words. This result is very helpful to promote the quality of lexicon. To analyze the main reason, the most of error words result from unintentionally incorrect decisions while linguistic experts confirmed new words by the new words management interface.

4. Discussion

This paper has proposed an excellent scheme of new word extraction for supporting lexicon-based Chinese word segmentation systems. However, two critical issues need to be further investigated.

4.1. How to appropriately determine the used parameters for the proposed scheme of new word extraction

First, the parameters R_n and M used in the proposed scheme of new word extraction will affect the amount of extracted new words and the accuracy rate of new words. If we set large values for parameters R_n and M, then the accuracy rate will be obviously promoted, but the amount of new words will be reduced. Conversely, if we set small values for parameters R_n and M, then the accuracy rate will be descended, but the amount of new words will be increased. Namely, how to determine parameters R_n and M is a trade-off issue between the amount of new words and accuracy rate of new words. It is very encouraging that the proposed scheme of new word extraction obtains a high accuracy rate even though the parameter R_n is set as the

lowest threshold, i.e. 1, in our experiment. Actually, the parameters R_n and M should be adaptively tuned according to the number of new titles. This problem has been considered as a research issue in our future study.

4.2. How to design reasonable fuzzy rule knowledge base for eliminating out-of-date words from the original lexicon

Although this study proposes a fuzzy rule base to infer the confidences of new words for eliminating likely out-of-date words from the original lexicon based on the time stamp of occurrence time of each word, the critical problem is how to design a reasonable fuzzy rule base according to appearance behavior of new words. To solve this problem, a practical method is to design an intelligent agent which can automatically monitor and collect the appearance behavior of new words during a long time.

5. Conclusion

In this paper, a novel statistics-based scheme for extracting new words based on the categorized corpora of Google News titles automatically retrieved from the Google News site is presented to promote the word identification capability for the lexicon-based Chinese word segmentation system ECScanner. In addition, to avoid reduce the performance of word identification due to over large lexicon derived from new word extension, this study also proposes a fuzzy rule knowledge base to eliminate out-of-date new words based on the inferred confidence degrees of new words. The proposed scheme of new word extraction was compared with Lin and Yu's method in terms of the amount of extracted new words and accuracy rate of new words based on the same texts of Google new titles. The experimental results show the proposed scheme of new word extraction is obviously superior to Lin and Yu's method as well as is helpful to promote the word identification capability of lexicon-based Chinese word segmentation system ECScanner. Currently, the developed ECScanner with automatic lexicon extension supported by the proposed scheme of new word extension has been successfully implemented and published on the Web to provide excellent Chinese word segmentation service by SOAP web service mechanism.

References

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, C.-M., & Liu, C.-Y. (2006). Personalized E-news monitoring agent system for tracking user-interested news events. *IEEE International Conference on Systems, Man, and Cybernetics*, 1062–1067.

Chen, K. J., & Liu, S. H. (1992). Word identification for Mandarin Chinese sentences. *Proceedings of COLING*, 101–107.

Chen, K. J., & Ma, W.-Y. (2002). Unknown word extraction for Chinese documents. *Proceedings of COLING*, 169–175.

Chowdhury, G. G. (2004). *Introduction to modern information retrieval*. London: Facet.

CKIP Chinese Parser. (2007). <http://140.109.19.112/>.

Cscanner (A Chinese Lexicon Scanner). (2000). <http://technology.chtsai.org/cscanner/>.

ECScanner (A Chinese Lexicon Scanner with Lexicon Extension). (2007). <http://dlll.nccu.edu.tw/~rank/ecscanner/>.

Foo, S., & Li, H. (2004). Chinese word segmentation and its effect on information retrieval. *Information Processing & Management, 40*(1), 161–190.

Google News. (2007). <http://news.google.com.tw/>.

Lin, C.-T., & George Lee, C. S. (1996). *Neural fuzzy systems: A neuro-fuzzy synergism to intelligent systems*. Prentice-Hall Inc.

Lin, S. H., Shih, C. S., Chen, M. C., Ho, J. M., Ko, M. T., & Huang, Y. M. (1998). Extracting classification knowledge of internet documents with mining term associations: A semantic approach. In *Proceedings of the annual international ACM SIGIR conference on research and development in information retrieval* (pp. 241–249).

Lin, Y. J., & Yu, M. S. (2001). Extracting Chinese frequent strings without a dictionary from a Chinese corpus and its applications. *Journal of Information Science and Engineering, 17*(5), 805–824.

Lu, W. H., Chien, L. F., & Lee, H. J. (2002). Mining anchor texts for translation of web queries. *ACM Transactions on Asian Language Information Processing, 1*(2), 159–172.

Ma, W. Y., & Chen, K. J. (2003). A bottom-up merging algorithm for Chinese unknown word extraction. *Proceedings of ACL Workshop on Chinese Language Processing*, 31–38.

Rui, X., & Wunsch, D. II, (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks, 16*(3), 645–678.

Shan, H., & Jie, Z. (2001). A bootstrap method for Chinese new words extraction. In *Proceedings of ICASSP-2001*, 1. Speech-L12: Acoustic & Lexical Modeling (p. I-581).

Wai, L., Cheung, P.-S., & Huang, R. (2004). Mining events and new name translations from online daily news. In *Proceedings of the 2004 joint ACM/IEEE conference on digital libraries* (pp. 287–295).

Yeh, C. L., & Lee, H. J. (1991). Rule-based word identification for Mandarin Chinese sentences – A unification approach. *Computer Processing of Chinese & Oriental Languages, 5*, 97–118.

Zhang, M.-Y., Lu, Z.-D., & Zou, C.-Y. (2004). A Chinese word segmentation based on language situation in processing ambiguous words. *Information Sciences, 162*(3–4), 275–285.