PERGAMON

Information Processing and Management 38 (2002) 91–109

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Effective foreign word extraction for Korean information retrieval

Byung-Ju Kang *, Key-Sun Choi

Division of Computer Science, Department of Electrical Engineering & Computer Science, Advanced Information Technology Research Center (AITrc), Korea Terminology Research Center for Language and Knowledge Engineering (KORTERM), Korea Advanced Institute of Science and Technology, 373-1 Kusong-dong, Yusong-gu, Taejon 305-701, South Korea

Received 4 April 2000; accepted 27 October 2000

Abstract

In Korean text, foreign words, which are mostly transliterations of English words, are frequently used. Foreign words are usually very important index terms in Korean information retrieval since most of them are technical terms or names. So accurate foreign word extraction is crucial for high performance of information retrieval. However, accurate foreign word extraction is not easy because it inevitably accompanies word segmentation and most of the foreign words are unknown. In this paper, we present an effective foreign word recognition and extraction method. In order to accurately extract foreign words, we developed an effective method of word segmentation that involves unknown foreign words. Our word segmentation method effectively utilizes both unknown word information acquired through the automatic dictionary compilation and foreign word recognition information. Our HMM-based foreign word recognition method does not require large labeled examples for the model training unlike the previously proposed method.

Keywords: Information retrieval; Foreign word recognition; Word segmentation

1. Introduction

In Korean text, the use of foreign words, which are mostly transliterations of English words, is growing at high speed. This is mainly due to the World Wide Web and the Internet that enable the instant access of new information at the global scale. The newly introduced foreign words usually

* Tel: +82-2-042-869-4365, Fax: +82-2-042-869-8700.
E-mail addresses: bjkang@world.kaist.ac.kr (B.-J. Kang), kschoi@world.kaist.ac.kr (K.-S. Choi).

0306-4573/01/$ - see front matter
PII: S0306-4573(00)00065-0

remain unregistered in any dictionary during the significant amount of time. This causes the well-known unknown word problem in natural language processing. Even after foreign words are registered, however, the problem is not completely solved. There may be more than one transliteration for the same English word. For example, for the English word 'digital', three different foreign words, '디지털 (*ti-ci-thel*)', '디지탈 (*ti-ci-thal*)', and '디지틀 (*ti-ci-thul*)', are actually used in Korean text. In the dictionary, however, only '디지털 (*ti-ci-thel*)' is present as a standard form. Consequently, the other two transliterations continuously remain unknown. [1]

In Korean information retrieval, usually only nouns are indexed, so accurate noun extraction is the most paramount. However, the unknown word problem caused by foreign words significantly hinders the noun extraction task since noun extraction in Korean inevitably accompanies word segmentation problem. It is well known that word segmentation involving unknown words is a very difficult problem. Most of the foreign words are nouns and also important content carriers so that they are usually first-class index terms. Hence accurate foreign word extraction is a very important issue in Korean information retrieval.

In Chinese, Japanese, Thai, etc., word segmentation must be preceded before any further non-trivial natural language processing since a sentence is written as a single long string without natural delimiters such as spaces between words. Unfortunately, the word segmentation problem especially involving unknown words has yet remained more or less as an unsolved problem. This kind of word segmentation problem is also very seve7re in Korean, even though it is not of the same degree, since in Korean some of the words in a sentence are written without space. We call the text segments, which are delimited by spaces, as *eojeol*. Eojeol needs segmentation.

Korean words can be broadly classified into content words (noun, verb, adjective, etc.) and functional words (particle, ending, etc.). Functional words cannot appear by themselves in a sentence but are always used as attachments at the end of content words and usually determine the grammatical role of content words in a sentence. More than one functional word can be attached to a content word. In addition, nouns are relatively freely joined together to form compound noun. So the configuration of an eojeol containing noun is mostly noun sequence (or compound noun) followed by functional word sequence. Therefore, noun extraction task in Korean consists of detaching functional word sequence from noun or compound noun and segmenting compound noun into simple nouns. For example, an eojeol '분산데이터베이스시스템은 (*pwun-san-te-yi-the-pe-yi-su-si-su-theym-un*)', which is composed of three nouns and a functional word, can be segmented as follows:

분산 (*pwun-san*, distribution)/ [2] 데이터베이스 (*te-yi-the-pe-yi-su*, database)/ 시스템 (*si-su-theym*, system) + 은 (*un*, func. word, subject marker),

where '데이터베이스 (*te-yi-the-pe-yi-su*, database)' and '시스템 (*si-su-theym*, system)' are foreign words.

Unknown words make very ambiguous both functional word detachment and compound noun segmentation. In order to solve the word segmentation problem involving unknown words, we

[1] Although all the various transliterations are identified, if they are not recognized as the same English word, recall may not be improved. The resolution method of the word mismatch problem caused by foreign words and English words is described in our other paper (Kang & Choi, 2000; Jeong, Myaeng, Lee, & Choi, 1999).

[2] We use '/' to denote the boundaries between nouns and '+' to denote the boundaries between noun and functional word.

need a way of guessing unknown words with high probability. One of the simplest methods is to use a variation of a maximum matching algorithm (Chena & Liu, 1992). One possible variation of the maximum matching algorithm extracts the longest word in the input string. Then the algorithm is recursively applied on the substrings to the left and right of the string. The remained substrings that are unmatched to the last are considered as unknown words. The most significant shortcoming of this simple approach is that the substring may constitute a valid word purely by chance. This mishap drastically increases as the size of noun dictionary increases. Moreover, if the compound noun is composed of only unknown words, this method is not applicable at all. The recent more advanced method is to automatically compile a dictionary of unknown words from target corpus (Fung & Wu, 1994). Basically, repeatedly occurring strings or character sequences are extracted as words. This approach has limitation when unknown words have low frequency. To remedy this low frequency problem, more linguistically oriented approaches were tried. Special recognizers depending on the type of unknown words – personal names, transliterated foreign names, and morphologically derived words – were used to help word segmentation (Sproat, Shih, Gale, & Chang, 1996; Jin, 1995). But the recognition accuracy is not so high due to the inherent difficulty of the problem itself.

In Korean, fortunately, it is possible to relatively accurately recognize foreign words since the syllable sequences in transliterated foreign words are usually very rare in pure Korean words. The differences in syllable pattern stem from the drastically different phonetic systems of Korean and English. So statistical methods utilizing the differences in syllable unigram or bigram patterns between pure Korean word and foreign word have been developed (Jeong, Kwon, & Myaeng, 1997; Oh & Choi, 1999). However, one of the difficulties in the purely statistical methods is that it is very difficult to distinguish foreign word syllables and functional word syllables. This is because most of the syllables used in functional words are also frequently used in foreign words. Moreover, overall recognition accuracy is not still high enough for the reliable foreign word extraction. Therefore, for the more accurate foreign word extraction, word segmentation is needed to supplement the foreign word recognition information (Jeong et al., 1999).

In this paper, we present a new effective method of foreign word extraction through word segmentation. Our word segmentation method effectively utilizes both unknown word information acquired through automatic dictionary compilation and foreign word recognition information. Our main strategy for the word segmentation involving unknown words (foreign words) is to apply different segmentation methods depending on the frequency of unknown words so that both high and low frequency unknown word information and foreign word recognition information are effectively utilized. High frequency unknown word dictionary alone is used with maximum matching segmentation to decrease the risk of accidental word formation problem. On the other hand, low frequency unknown word dictionary, together with Korean noun dictionary and foreign word dictionary, is used with complete matching segmentation that is more conservative in segmentation by disallowing unmatched segments to compensate the lower word guessing accuracy. However, there may still remain many undetected foreign words. So lastly, foreign word recognition information is used to detect possible low frequency unknown foreign words. Once segmentation is done, only foreign words are detected by our HMM-based foreign word recognition method. Our method does not require large labeled examples, which are very difficult to obtain, unlike the previous method of Oh and Choi (1999) and needs only easily accessible resources while not degrading accuracy much.

2. Foreign word recognition

Kwon, Jeong, and Myaeng (1997) proposed a simple foreign word detection method that estimates whether a given word is a foreign word or a pure Korean word. They used the following formula for the decision:

$$D(W) = \frac{P(\text{Foreign} \mid W)}{P(\text{Korean} \mid W)} \tag{1}$$

$$= \frac{P(W \mid \text{Foreign})P(\text{Foreign})}{P(W \mid \text{Korean})P(\text{Korean})} \text{ by Bayes' rule,} \tag{2}$$

where $P(\text{Foreign} \mid W)$ and $P(\text{Korean} \mid W)$ represent the conditional probability that the word W is a foreign word and a pure Korean word, respectively. If $D(W) > 1$, W is decided as a foreign word. For the real probability computation, they used the formula rewritten by Bayes' rule. $P(\text{Foreign})$ and $P(\text{Korean})$ are prior probabilities and may be estimated by computing the ratio of the foreign word and Korean words in the training corpus. For the estimation of $P(W \mid \text{Foreign})$, the syllables s_is constituting W are assumed to occur independently from each other. So the following formula is obtained:

$$P(W \mid \text{Foreign}) = \lambda_1 \cdot \prod_{i=1}^{n} P(s_i \mid \text{Foreign}) + \lambda_2 \cdot \prod_{i=1}^{n+1} P(s_{i-1}s_i \mid \text{Foreign}), \tag{3}$$

where s_0 and s_{n+1} are interpreted as special syllables indicating the start of a word and the end of a word, respectively, and λ_1 and λ_2 represent the weights of unigram and bigram probabilities, respectively, so that $\lambda_1 + \lambda_2 = 1$. $P(W \mid \text{Korean})$ is similarly estimated.

This method works very well if a foreign word appears alone. However, in many cases, foreign words compose compound nouns with pure Korean words and are attached with functional words. So in the case that more Korean syllables than foreign syllable exist in a word it is difficult to detect the existence of foreign words. What is worse than this limitation is that it cannot tell the specific boundaries between the Korean word and the foreign word.

To remedy these handicaps, Oh and Choi (1999) developed more effective foreign word detection method. They reformulated the foreign word recognition problem as a *syllable-tagging* problem such that each syllable is tagged with the foreign syllable tag (*F*) or the pure Korean syllable tag (*K*), then the syllable sequences tagged with contiguous *F* tags are recognized as foreign words. For example, a Korean eojeol '분산데이터베이스시스템은 (*pwun-san-te-yi-the-pe-yi-su-si-su-theym-un*)' may be tagged as follows:

분	산	데	이	터	베	이	스	시	스	템	은
K	K	F	F	F	F	F	F	F	F	F	K

The substring "데이터베이스시스템 (database system)" corresponding to *F*'s is determined as a foreign word. The syllable-tagging problem may be effectively modeled using the hidden Markov model (HMM) like POS-tagging problem. Therefore, the problem is to determine the most probable tag sequence T given an eojeol S:

$$T^* = \arg\max_T (S \mid T)P(T). \tag{4}$$

This HMM-based method had turned out to be much more effective in detecting even short foreign syllable sequences surrounded with many Korean syllables than the previous method. However, their method is still far from perfection. Especially, it is not good at distinguishing foreign syllables and functional word syllables. This is because most of the functional word syllables are also frequently used syllables in foreign words.

In order to alleviate the problem, we introduce one more tag *J* for the functional word syllable. Better distinction between foreign syllables and functional word syllables may significantly reduce mistaggings of foreign syllables. The previous tagging example may be tagged as follows:

분	산	데	이	터	베	이	스	시	스	템	은
K	K	F	F	F	F	F	F	F	F	F	J

Note that "은 (*un*)" is a functional word and is tagged with *J*.

Due to the introduction of tag *J*, we need to device a new HMM model that better reflects the internal structure of Korean eojeol (Fig. 1). An unrealistic assumption in our model is not to allow one-syllable word in both Korean and foreign words. One reason about the restriction is that in Korean one-syllable word is relatively few and in the case of the allowance the tagging error rate drastically increases.

Applying the chain rule and assuming that the current tag depends only on the previous two tags and the current syllable observation depends only on the current tag and the immediate previous tag, the term $P(S \mid T)P(T)$ in the Eq. (4) may be simplified as follows:

$$P(S \mid T)P(T) = \prod_{i=1}^{n} [p(s_i \mid t_i t_{i-1}) \times p(t_i \mid t_{i-1} t_{i-2})], \tag{5}$$

where we define $t_0 = B$ (begin-of-word symbol) and $p(t_1 \mid t_0 t_{-1}) = p(t_1 \mid t_0)$ to simplify our notation. In the above formula, we use $p(s_i \mid t_i t_{i-1})$ instead of $p(s_i \mid t_i)$ differently with Oh and Choi (1999). Since more context is considered, we believe that our model will perform better than theirs.

Next, we need to estimate the model parameters, i.e., transition and observation probabilities. For the ideal parameter estimation huge syllable-tagged corpus is required but in general this kind

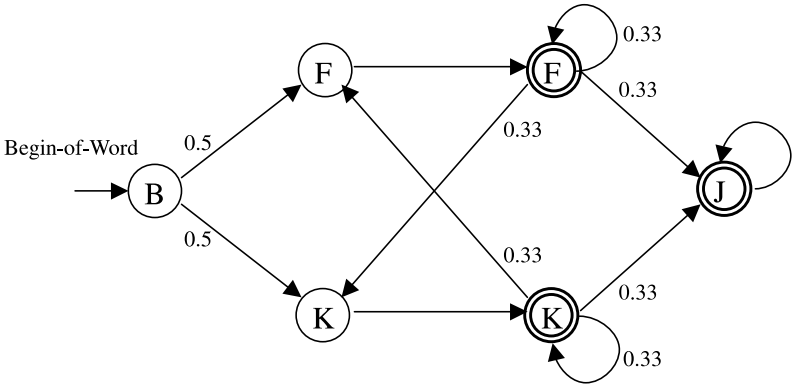

Fig. 1. The HMM model for Korean eojeol.

of resource is not available. Therefore, we propose a method of indirectly estimating all the parameters by only using pure Korean noun dictionary, foreign word dictionary and functional word sequence list that are all easily obtainable. In order to make a list of the functional word sequences, we need to compile all the possible functional word combinations from corpus. This can be easily done by using POS-tagged corpus and collecting functional word sequences that immediately follow nouns.

Probabilities $p(s \mid FF)$, $p(s \mid KK)$, $p(s \mid JJ)$ may be directly estimated by simply counting syllables, except the first syllable in a word, respectively, in foreign word and pure Korean word dictionary and functional word sequences list. However, $p(s \mid FK)$, $p(s \mid KF)$, $p(s \mid FB)$, $p(s \mid KB)$, $p(s \mid JF)$, $p(s \mid JK)$, etc. cannot be directly estimated, so $p(s \mid F_{\text{first}})$, $p(s \mid K_{\text{first}})$, $p(s \mid F_{\text{first}})$, $p(s \mid K_{\text{first}})$, $p(s \mid J_{\text{first}})$, $p(s \mid J_{\text{first}})$ are used instead, respectively. F_{first}, K_{first}, and J_{first} indicate the first syllable of a foreign word, a Korean noun and a functional word sequence, respectively. The actual computation of the observation probabilities is as follows:

$$p(s \mid FF) = \frac{\text{Number of non-first-syllable } s \text{ in foreign word dictionary}}{\text{Total number of non-first-syllable } s \text{ in foreign word dictionary}},$$

$$\left.\begin{matrix} p(s \mid FB) \\ p(s \mid FK) \end{matrix}\right\} = p(s \mid F_{\text{first}}) = \frac{\text{Number of first-syllable } s \text{ in foreign word dictionary}}{\text{Total number of first-syllable } s \text{ in foreign word dictionary}}.$$

Other probabilities may be similarly computed.

The trigram transition probabilities are impossible to be estimated since there does not exist syllable-tagged corpus. Note that we have no direct statistical information about transitions between different tags. So we just equally distributed probabilities among possible transitions (Fig. 1). However, we do not deny the possibility of a more clever estimation. One such an estimation is to consider frequency ratio between the word types. Generally foreign word occurs much less frequently than Korean words even though it depends on the domain under consideration. Therefore, it should be $p(K \mid B) \gg p(F \mid B)$ and so on. But in this paper, we want our recognizer to be more sensitive to foreign words. So our equal distribution of probability makes sense.

3. Foreign word extraction

Our overall foreign word extraction process consists of the following six phases (Fig. 2):

1. morphological analysis,
2. syllable tagging,
3. unknown word dictionary construction,
4. functional word sequence detachment,
5. compound noun segmentation,
6. foreign word detection.

In the phase of morphological analysis, eojeols containing unknown words are extracted. This can be simply achieved by collecting the eojeols that fail for morphological analysis. Actually, the large portion of morphological analysis failures comes from misspelling and misspacing.

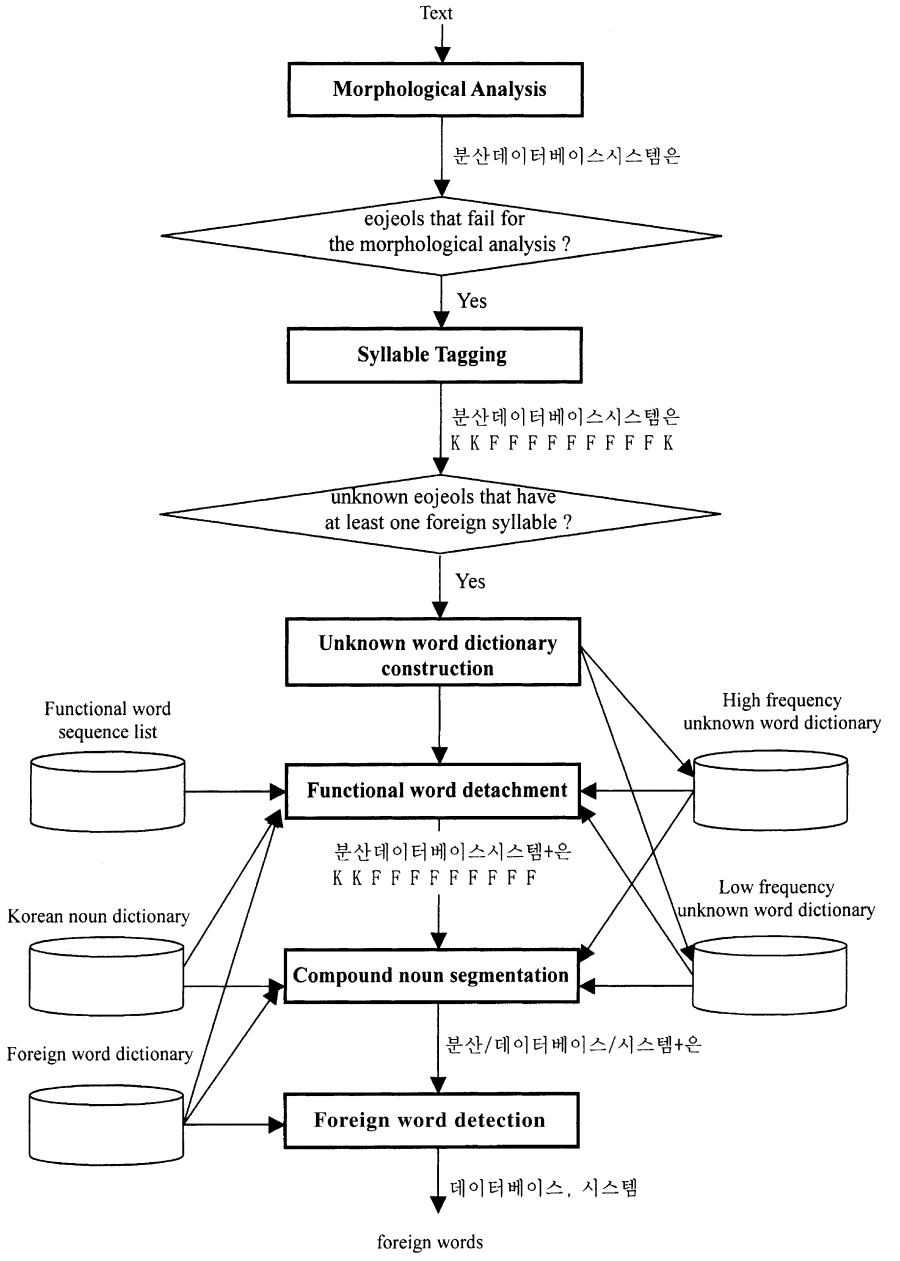

Fig. 2. Foreign word extraction process.

However, we assume that all the failures are due to unknown words since it is difficult to distinguish the sources of the failures. All the nouns are also extracted. This is because the noun dictionary, which is used by the morphological analyzer, may contain foreign words. So we need to pick out foreign words from the nouns. This may be done by looking up a foreign word dictionary.

In the second phase, syllable tagging is performed for all the unknown eojeols that are collected in the previous phase. The syllable-tagging result is a sequence of *K*, *F*, and *J* tags. But we convert *J* tags into *K* tags in this phase. This is because we introduced *J* tag only for a better distinction of foreign syllables and functional word syllables. Moreover, after syllable tagging, *J* tag information is not used anymore in the following segmentation phase since *J*-tagging accuracy is not high enough to be directly used for the functional word detachment. Eojeols that have at least one foreign syllable are passed to the next phase, i.e., only potential foreign-word-containing eojeols are gathered.

In the third phase, unknown word dictionary is automatically constructed from the unknown eojeols, i.e., eojeols that fail for morphological analysis, that have at least one foreign syllable. Unknown word dictionary may be constructed by extracting unknown words that can be guessed with high confidence. Repeatedly occurring strings are good candidates of word. We collected all the unknown eojeols of which frequency is above some fixed threshold.

We use different segmentation strategies depending on the frequency of unknown words so that both high and low frequencies unknown word information are effectively utilized. So we constructed two kinds of unknown word dictionary: a high frequency unknown word dictionary and a low frequency unknown word dictionary. The high frequency unknown word dictionary means the unknown word dictionary constructed with high threshold word frequency. On the other hand, the low frequency unknown word dictionary means the unknown word dictionary constructed with low threshold word frequency. We set the high and the low threshold frequencies, respectively, as 10 and 2. We simply chose 10 since we believed that 10 times of occurrence are high enough to be confident for the unknown words. On the other hand, two times are the minimum number of occurrences to admit any possibility of being word.

Our simple approach of collecting repeatedly occurring strings as words has a problem in handling compound nouns, which may be composed of both foreign words and Korean words or only of foreign words or only of Korean words, since they also appear repeatedly. These compound nouns must be filtered out. Fortunately, we are able to relatively accurately guess whether foreign syllables are contained. So if both Korean and foreign syllables are detected through syllable tagging the candidate can be filtered out. In the case of compound words that consist of only foreign words, the other method should be sought. However, we provide no solution for this case.

Next, the syllable-tagged eojeols are fed into the fourth phase of functional word detachment. First, unambiguous functional word sequences are detached and then ambiguous functional word sequences are detached utilizing eojeol patterns and unknown word dictionaries that were automatically compiled from the target corpus in the previous phase.

In the fifth phase, compound noun segmentation is performed using all the available information such as the unknown word dictionaries, syllable-tagging information, Korean noun dictionary, foreign word dictionary, and the functional word sequence list.

Finally foreign words are actually extracted by performing syllable tagging again on the word segments from compound noun segmentation. The segments that contain more than 50% foreign syllables are decided as foreign words.

In the following, we explain in more detail the functional word sequence detachment phase, the compound noun segmentation phase, and the foreign word detection phase.

3.1. Functional word detachment

Our functional word detachment procedure consists of the following four steps:
1. unambiguous functional word sequence detachment,
2. ambiguous functional word sequence detachment using eojeol pattern,
3. maximum matching segmentation using high frequency unknown word dictionary,
4. complete matching segmentation using low frequency unknown word dictionary.

In the first step, unambiguous functional word sequences are detached. Detaching functional word sequence from an unknown word is inherently ambiguous. But there exist some functional word syllables that are highly improbable to occur in foreign words such as '의 (*uy*)', '으' (the first syllable of functional word '으로 (*u-lo*)'), '같' (the first syllable of functional word '같이 (kath-i)'), and '밖' (the first syllable of functional word '밖이 (pakk-i')). So it is possible to relatively safely detach the functional word sequences that start with one of these syllables. Especially '의 (*uy*, genitive marker)' is the most frequently used functional word so that a large portion of the functional word sequences may be detached at this step.

The unambiguous functional word sequences we identified are the ones that start with one of the above four functional words. Table 1 lists some of the unambiguous functional word sequences. We took a simple approach for a criterion whether a functional word is ambiguous or not. We investigated about 40,000 entries in a foreign word dictionary and found that these four first syllables of functional words are never used as foreign word syllables. Of course, we may not completely exclude the possibility of using these syllables in unseen or new foreign words. However, it is highly improbable because their phonetic characteristics are very specific to Korean. On the other hand, if a more sophisticated method, for example, probabilistic approach, is used, we may identify more unambiguous functional word sequences.

In the second step, we detach ambiguous functional word sequences using eojeol patterns. We sort the unknown eojeols in Korean alphabetic order. There is a very effective cue to identify functional word sequence in the sorted list. Let us see the following consecutive eojeols in the list:

인터프리터로	인터프리터(*in-the-phu-li-the*, interpreter)+로 (*lo*, func. word)
인터프리터에	인터프리터(*in-the-phu-li-the*, interpreter)+에 (*ey*, func. word)
인터프리터와	인터프리터(*in-the-phu-li-the*, interpreter)+와 (*wa*, func. word)
인터프리터하거나	인터프리터(*in-the-phu-li-the*, interpreter)+하거나 (*hakena*, func. word)

This kind of pattern, repeatedly occurring substring of fixed length plus various possible functional word sequences, is a strong indicator of a common stem. We detect this kind of pattern, determine a common stem and detach the functional word part. The minimum number of eojeols that is capable of composing the pattern is two in this paper.

In the third step, we detach ambiguous functional word sequences using the high frequency unknown word dictionary. After detaching possible functional word sequence, if the string remained exists in the unknown word dictionary, the detachment is accepted. For example, if

Table 1
Some of the unambiguous functional word sequences

의
으로
으로까지
으로까지는
으로부터까지는커녕
으로부터
으로서는
으로만은
으로도
으로서조차도
으로서처럼
으로써
으로야
같이
같이는
같이만
같이만은
같이만이라도
같이도
같이나
밖에
밖에도
밖에는
밖엔

'로보트 (*lo-po-thu*, robot)' were extracted as an unknown word in the previous phase, '로보트가 (*lo-po-thu-ka*)' can be segmented as '로보트 (*lo-po-thu*, robot) + 가 (*ka*, subject marker)' since '가 (*ka*)' composes a valid functional word. The validity check of a functional word sequence is done by looking up a list of foreign word sequences that is automatically extracted from a POS-tagged corpus. Only functional word sequences trailing after nouns were collected. However, this simple method will miss many possible detachments. For example, '이동로보트가 (*i-tong-lo-po-thu-ka*)' would be missed since '이동로보트 (*i-tong-lo-po-thu*, mobile robot)' does not exist in the unknown dictionary. So we do maximum matching segmentation using the high frequency unknown word dictionary and if the last segment composes a valid functional word sequence and the segment right before the last segment exists in the unknown word dictionary the last segment is detached as a valid functional word sequence. The maximum matching segmentation algorithm segments [3] '이동로보트가 (*i-tong-lo-po-thu-ka*)' into '이동 (*i-tong*, mobile)/로보트 (*lo-po-thu*, robot)/가 (*ka*, subject marker)'. The last segment '가 (*ka*)' is a valid functional word and '로보트 (*lo-po-thu*)' exist in the unknown dictionary, so '가 (*ka*)' can be detached.

In the fourth step, we detach ambiguous functional word sequences using the low frequency unknown word dictionary. The procedure is similar to the previous third step but this time we

[3] The maximum matching segmentation algorithm is explained in detail in Section 3.2.

employ complete matching segmentation algorithm. [4] The last segment must compose a valid functional word sequence and the immediately preceding segment must exist in the low frequency dictionary or Korean noun dictionary or foreign word dictionary.

3.2. Compound noun segmentation

Our compound noun segmentation phase consists of the following three steps (Fig. 3):
1. maximum matching segmentation using high frequency unknown word dictionary,
2. complete matching segmentation using low frequency unknown word dictionary,
3. segmentation using syllable-tagging information.

Actually, the first and the second step of compound noun segmentation phase and the third and the fourth step of functional word detachment phase are overlapped. In another words, the maximum matching and the complete matching segmentation algorithm both detach functional word sequence and segment compound noun simultaneously in a single pass. This will be clear after reading this section. For just ease of explanation they were separately described.

In the first step, maximum matching segmentation is performed using the high frequency unknown word dictionary. The algorithm first tries to find the longest match with the dictionary entry. If the longest match is found, the identified left and right word boundaries are marked and the algorithm is recursively applied on the left and right substrings of the input string. We may use a Korean noun dictionary extended with the high frequency unknown word dictionary. However, the simple maximum matching method with the extended noun dictionary may produce many wrong words. This is mainly because some substring may compose a valid word purely by chance. In order to alleviate this accidental word formation problem, we perform the maximum matching segmentation only with the high frequency unknown word dictionary. If the maximum matching segmentation is successful, we directly go to the third step. But if it fails, we go to the second step.

In the second step, complete matching segmentation is performed using the Korean noun dictionary, foreign word dictionary, and the low frequency unknown word dictionary. Here we define the complete matching segmentation slightly differently with the maximum matching segmentation. The complete matching segmentation does not allow unmatched substring unlike the maximum matching segmentation. The complete matching segmentation takes a greedy approach. It tries to find the longest match with the dictionary entry. If the search fails, the algorithm exits with failure. If the longest match is found, the identified left and right word boundaries are marked and the algorithm is recursively applied on the left and right substrings of the input string.

In the third step, for the eojeols that failed for both the maximum matching and the complete matching segmentation and also for the each segment resulted from the maximum matching or the complete matching segmentation, the word boundaries between foreign words and Korean words are detected using syllable-tagging information and then segmentation is performed at the boundaries between Korean word and foreign word. It is obvious that segmentations should happen at the boundaries between Korean words and foreign words. Therefore, if the following condition is satisfied, segmentation occurs at the position:

[4] The complete matching segmentation algorithm is explained in detail in Section 3.2.

MAXIMUM_MATCHING_SEGMENTATION(S)

- Given the input string S, find the longest substring match, S_{match}, with the dictionary entry. (The default dictionary is the high frequency unknown word dictionary)
- If the search fails or S_{match} is same with S, return;
- Else do the following
  - $S \leftarrow S_{left}/S_{match}/S_{right}$; (mark the left and right word boundary with '/')
  - MAXIMUM_MATCHING_SEGMENTATION(S_{left});
  - MAXIMUM_MATCHING_SEGMENTATION(S_{right});

COMPLETE_MATCHING_SEGMENTATION(S)

- Given the input string S, find the longest substring match, S_{match}, with the dictionary entry. (The default dictionary is the low frequency unknown word dictionary + Korean noun dictionary + foreign word dictionary)
- If the search fails, exit with failure;
- If S_{match} is same with S, return;
- Else do the following
  - $S \leftarrow S_{left}/S_{match}/S_{right}$; (mark the left and right word boundary with '/')
  - COMPLETE_MATCHING_SEGMENTATION(S_{left});
  - COMPLETE_MATCHING_SEGMENTATION(S_{right});

SYLLABLE_TAGGING_SEGMENTATION(w)

- For each syllable boundary of w
  - If (*Condition 1*) and (*Condition 2*) are satisfied
    - mark the word boundary with '/' at the position;

COMPOUND_NOUN_SEGMENTATION(S)

- MAXIMUM_MATCHING_SEGMENTATION(S);
- If the maximum matching segmentation fails on S
  - COMPLETE_MATCHING_SEGMENTATION(S);
- For each word segments w of S
  - SYLLABLE_TAGGING_SEGMENTATION(w);

Fig. 3. Compound noun segmentation algorithm.

Condition 1. Tag changes at the position from *F* to *K* or *K* to *F*.

Condition 2. The string corresponding to the contiguous *F*s exist in the low frequency unknown word dictionary or foreign word dictionary or the string corresponding to the contiguous *K*s composes a valid Korean word.

For example, '확장트리 (*hwak-cang-thy-li*, extended tree)' is tagged as ''KKFF'' and the substring '확장 (*hwak-cang*, extension)' corresponding to K's is a valid Korean word. So it is correctly segmented as '확장 (*hwak-cang*, extension)/트리 (*hwak-cang-thy-li*, tree)'. On the other hand, '화일시스템 (*hwa-il-si-su-theym*, file system)' is tagged as ''KKFFF'' and substring '화일 (*hwa-il*, file)' is not in the Korean noun dictionary but '시스템 (*si-su-theym*, system)' exists in the foreign word dictionary or the unknown word dictionary. So it is also correctly segmented as '화일 (*hwa-il*, file)/시스템 (*si-su-theym*, system)'.

3.3. Foreign word detection

Now, foreign word extraction is very straightforward. First, foreign word dictionary is looked up. Those word segments that are found in the foreign word dictionary are extracted as foreign words. Next, all the remained word segments are syllable tagged and the segments that contain more than 50% foreign syllables are decided as foreign words. One thing that we have to be cautious at this point is oversegmentation. For example, '마이크로프로세스 (*ma-i-khu-lo-phu-lo-sey-su*, microprocessor)' is segmented as '마이크로 (*ma-i-khu-lo*, micro)' and '프로세스 (*phu-lo-sey-su*, processor)' but this is over-segmentation. To alleviate this over-segmentation problem, we also extract compound nouns before segmentation as index terms. So, in the above example, '마이크로프로세스 (*ma-i-khu-lo-phu-lo-sey-su*, microprocessor)' is also extracted as an index term in addition to '마이크로 (*ma-i-khu-lo*, micro)' and '프로세스 (*phu-lo-sey-su*, processor)'.

4. Experiments

4.1. Experiment data and evaluation metrics

As experiment data, we used KTSET 1.0 (Kim et al., 1994) that is a standard Korean IR test collection. Some important characteristics of the test suit are given in Table 2. The documents in KTSET 1.0 contain relatively high proportion of foreign words so that the effectiveness of foreign word extraction to information retrieval would be more easily observed. Actually foreign words take about 16% of the index terms (noun equivalents) (Table 3).

Table 2
Some characteristics of KTSET 1.0 Korean standard IR test collection

No. of documents	1000
Avg. length of a document	110 words (eojeols)
No. of queries	30
Avg. length of a query	116/30 = 3.86 terms
Source	Abstract of technical papers
Subject	Computer science and information science

Table 3
Frequency of foreign words and English words in KTSET 1.0

Index terms	No. of terms
Pure Korean noun	74,000 (72%)
Foreign word	15,500 (16%)
English word	12,500 (12%)
Total	102,000 (100%)

This is a significant amount of percentage that is capable of greatly influencing information retrieval performance.

In Korean information retrieval, only nouns are usually extracted as index terms and morphological analyzer is used to remove non-nouns (verb, adjective, adverb, etc.) and to detach the trailing functional words from noun stems or even to segment nominal compounds. From the 1000 documents of KTSET 1.0, all nouns were extracted and the unknown eojeols, whose morphological analysis failed, were also extracted. In total 1,01,954 terms were extracted. On these terms we performed syllable tagging. If at least one foreign syllable is detected, the term was extracted as potential foreign word containment. Total 15,989 foreign word candidates were extracted and there were 799 unique terms in the candidates. Now what we have to do is to extract genuine foreign words from the 799 terms.

We use three important external language resources for our foreign word extraction (Fig. 2). They are the foreign word dictionary, Korean noun dictionary, and functional word sequence list. The foreign word dictionary of about 40,000 entries was selected from Nam's foreign word dictionary (Nam, 1997). The Korean dictionary of about 90,000 entries was built by simply eliminating non-nouns from the general Korean dictionary. The functional word sequence list of about unique 2000 entries was automatically extracted from the POS-tagged corpus. [5]

We use the following metrics for the evaluation of foreign word extraction performance:

$$\text{Precision} = \frac{\text{Number of foreign words correctly recognized}}{\text{Number of foreign words recognized}},$$

$$\text{Recall} = \frac{\text{Number of foreign words correctly recognized}}{\text{Total number of foreign words in a test set}}.$$

4.2. Syllable tagging

For training of our HMM-based syllable-tagging model, we used the foreign word dictionary, the functional word sequence list, and a pure Korean noun dictionary. The pure Korean noun dictionary was built by simply eliminating foreign words found in our foreign word dictionary from the Korean noun dictionary.

[5] KAIST POS-tagged corpus (Chae & Choi, 1999) was used.

Table 4
The performance of foreign word extraction by syllable tagging

		One-syllable word is not allowed	One-syllable word is allowed
Two tags	Precision	70.10%	62.89%
	Recall	81.23%	70.02%
Three tags	Precision	73.59%	64.36%
	Recall	82.35%	71.84%

We experimented how accurately foreign words can be extracted only using syllable-tagging information. Table 4 shows our experiment results. We conducted the experiments using both two tags (*K*, *F*) and three tags (*K*, *F*, *J*) and simultaneously prohibiting or allowing one-syllable word. The experiment result shows that prohibiting one-syllable word results in much higher precision and recalls than allowing one-syllable word. This means that prohibiting one-syllable word significantly increases the number of foreign words correctly recognized. On the other hand, using three tags results in higher precision but only trivial changes in recall when compared with using two tags. This result may mean that using three tags is more conservative in tagging a syllable with foreign word tag. Therefore, we may conclude that increasing the number of tags is a precision enhancing scheme and prohibiting one-syllable word is both a precision and recall enhancing scheme.

In all the following experiments in this paper, syllable tagging with 3 tags and one-syllable word prohibition were used.

4.3. Foreign word extraction

We may directly use syllable-tagging information for the extraction of foreign words but the syllable-tagging accuracy is not high enough for practical use. However, the foreign word extraction accuracy may be significantly improved when it is supported by functional word detachment and compound noun segmentation. Table 5 shows the experiment results of our foreign word extraction on KTSET 1.0. The functional word detachment brought non-trivial increases in both precision and recall. Furthermore, after compound noun segmentation, both precision and recall were significantly increased when compared with those obtained only using syllable tagging.

The experimental results in Table 5 are the evaluation of the mere extraction of foreign word parts from surrounding Korean syllables, not considering the segmentation of foreign compound nouns that are composed of only foreign words. Our compound noun segmentation method is capable of decomposing the foreign word compounds. If the decomposition of the foreign word compounds is considered, the performance improvement would be more significant.

Table 5
The performance of foreign word extraction by word segmentation

	Precision (% change)	Recall (% change)
Syllable tagging	73.59%	82.35%
+ Functional word detachment	77.01% (+4.64)	85.85% (+4.25)
+ Compound noun segmentation	84.33% (+14.59)	92.01% (+11.73)

4.4. Information retrieval experiment

We conducted an information retrieval experiment to see the effectiveness of foreign word extraction to the information retrieval. The retrieval system we used is SMART system developed at the Cornell university. SMART system is based on the vector space model and ranks documents based on term vector similarity (Salton, Wong, & Yang, 1975). We chose the standard *tf*idf* scheme for both index term and query term weighting (Salton & Buckley, 1988). The test suit we used was KTSET 1.0. (Table 3).

The experiment result is shown in Table 6. The performance is measured in terms of 11-point average precision. The average precision was increased by 8.0% when we extract foreign words by detaching functional words and segmenting compound nouns. The baseline is the performance before foreign word extraction, i.e., when eojeols contain unknown words, the eojeols themselves without segmentation are indexed. This is a quite impressive improvement when considering the errors made in the foreign word extraction.

4.5. Discussions

In the 799 foreign word candidates, 717 terms actually contained foreign words. This is a very high proportion of foreign words among unknown words. This high ratio is partly due to the technical content of the documents. And in the 717 terms, 558 unknown foreign words, which were not registered in the foreign word dictionary, were found. Most of the newly found foreign words were non-standard transliterations, technical terms, and proper nouns. Table 7 lists some of the foreign words that are newly found.

We may use bigger IR test collection such as KTSET 2.0 (Park, Choi, Kim, & Kim, 1996) and KRIST (Lee, Choi, Han, Kim, & Nam, 1995). The main reason for using smaller KTSET 1.0 is for the purpose of evaluating the foreign word extraction. The 799 terms are small enough to manually segment and extract foreign words. The evaluations were possible by this manually constructed answer set.

Table 6
Information retrieval performance before and after foreign word extraction

Recall	Baseline	After foreign word extraction
0.0	0.6376	0.6293
0.1	0.5765	0.5756
0.2	0.5130	0.5267
0.3	0.4149	0.4773
0.4	0.3969	0.4547
0.5	0.3541	0.4152
0.6	0.3160	0.3721
0.7	0.2851	0.3133
0.8	0.2450	0.2671
0.9	0.2180	0.2360
1.0	0.1771	0.1985
11-pt Avg. precision	0.3758	0.4060
% Change		+8.0%

Table 7
Some of the newly found foreign words

그래피칼 (graphical)
그래픽스 (graphics)
로보트 (robot)
스레드 (thread)
다이아그래밍 (diagramming)
데이터베이스 (database)
디멀티플렉서 (demultiplexer)
디소러스 (thesaurus)
디소어러스 (thesaurus)
레이블링 (labeling)
리드뮬러 (Read-Muller)
릴레이셔날 (relational)
마르코프 (markov)
매니퓰레이터 (manipulator)
멀티플렉싱 (multiplexing)
메트릭스 (matrix)
벡타 (vector)
스케듈링 (scheduling)
스케줄링 (scheduling)
스케쥴링 (scheduling)
인스탄스 (instance)
트랜스퓨터 (transputer)
퍼셉트론 (perceptron)
프로시쥬어 (procedure)

Our method is of course applicable to the extraction of unknown words that does not contain a foreign word. Our segmentation method using high and low frequency unknown word dictionaries is capable of identifying any kind of unknown words. In this paper, we focused on extracting only foreign words and evaluating how much impact it gives to information retrieval performance. To extend our method to any kind of unknown words we need a just minor modification to our process. In Fig. 2, the second phase of syllable tagging and the final phase of foreign word extraction are not needed in this case.

5. Conclusion

We have presented a new effective statistical foreign word recognition method based on the observation that the syllable sequence patterns are very different between pure Korean words and foreign words. Our HMM-based method does not require any labeled example for training and more effectively distinguish foreign word syllable from functional word syllable. However, the recognition accuracy is not high enough for the direct use of foreign word extraction. Moreover, it is not capable of decomposing foreign word compounds. So we developed a new foreign word extraction method that is based on word segmentation. Our word segmentation method utilizes both unknown word information that is automatically compiled from the target corpus and the

foreign syllable recognition information. We used different word segmentation methods depending on the frequency of unknown words so that the high frequency unknown word information, the low frequency unknown word information, and the foreign word recognition information were all effectively utilized. We have succeeded in obtaining significant improvement in foreign word extraction accuracy by utilizing both the unknown word information and the foreign word recognition information when compared to using only foreign word recognition information. We also showed that the impact of accurate foreign word extraction on Korean information retrieval performance is great.

Acknowledgements

This work was supported by the Korea Science and Engineering Foundation (KOSEF) through the Advanced Information Technology Research Center (AITrc).

References

Chae, Y. S., & Choi, K. S. (1999). The KIBS project: KAIST corpus and part-of-speech and syntactic tagset. In *Proceedings of workshop on multi-lingual information processing and Asian language processing (MAL'99)* (pp. 19–22). Beijing, China.

Chen, K. J., & Liu, S. H. (1992). Word identification for Mandarin Chinese sentences. In *Proceedings of the fifteenth international conference on computational linguistics (COLING-92)* (pp. 101–107). Nantes, France.

Fung, P., & Wu, D. (1994). Statistical augmentation of a Chinese machine-readable dictionary. In *Proceedings of the second annual workshop on very large corpora (WVLC-2)* (pp. 69–86). Kyoto, Japan.

Jeong, K. S., Kwon, Y. H., & Myaeng, S. H. (1997). The effect of a proper handling of foreign and English words in retrieving Korean test. In *Proceedings of the second international workshop on information retrieval with Asian languages* (pp. 114–122). Tsukuba, Japan.

Jeong, K. S., Myaeng, S. H., Lee, J. S., & Choi, K. S. (1999). Automatic identification and back-transliteration of foreign words for information retrieval. *Information Processing and Management*, *35*(4), 523–540.

Jin, W., & Chen, L. (1995). Identifying unknown words in Chinese corpora. In *Proceedings of natural language processing Pacific Rim symposium (NLPRS95)* (pp. 234–239). Seoul, Korea.

Kang, B. J., & Choi, K. S. (2000). Two approaches for the resolution of word mismatch problem caused by English words and foreign words in Korean information retrieval. In *Proceedings of the fifth International workshop on information retrieval with Asian languages*. (pp. 133–140), Hong Kong.

Kim, J. G., Kim, Y. W., & Kim S. H. (1994). Development of the data collection (KTSET) for Korean information retrieval studies. In *Proceedings of the sixth conference on Hangul and Korean information processing* (pp. 378–385). Taejon, Korea (in Korean).

Kwon, Y. H., Jeong, K. S., & Myaeng, S. H. (1997). Foreign word identification using a statistical method for information retrieval. In *Proceedings of International conference on computer processing of oriental languages* (pp. 675–680). Hong Kong, China.

Lee, J. H., Choi, K. N., Han, H. S., Kim, J. W., & Nam, S. W. (1995). Developing the KRIST test collection for researches in information retrieval. *Journal of the Korean Information Management Society*, *12*(2), 225–232 (in Korean).

Nam, Y. S. (1997). The latest foreign word dictionary. *Sung-An-Dang Press* (in Korean).

Oh, J. H., & Choi, K. S. (1999). Automatic extraction of technical terminologies from scientific text based on hidden Markov model. *The Eleventh Hangul and Korean Information Processing* (in Korean).

Park, Y. C., Choi, K. S., Kim, J. K., & Kim, Y. W. (1996). KT test collection version 2.0 for Korean information retrieval research. In *Proceedings of 23rd Korean information science society, Spring Conference* (pp. 59–65). Seoul, Korea (in Korean).

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of ACM*, *18*(5), 613–620.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, *24*(5), 513–523.

Sproat, R., Shih, C., Gale, W., & Chang, N. (1996). A stochastic finite-state word-segmentation algorithm for Chinese. *Computational Linguistics*, *22*(3), 377–404.