International Journal of Computer Processing of Oriental Languages
Vol. 16, No. 1 (2003) 41–62

World Scientific
www.worldscientific.com

A Statistical Model for Automatic Extraction of Korean Transliterated Foreign Words

JONG-HOON OH* AND KEY-SUN CHOI†

Computer Science Division, Department of EECS,
Korea Terminology Research Center for Language and
Knowledge Engineering (KORTERM)
Korea Advanced Institute of Science and Technology (KAIST),
373-1, Kusong-Dong, Yusong-Gu, Taejon, 305-701, Korea
*rovellia@world.kaist.ac.kr
†kschoi@world.kaist.ac.kr

In this paper, we will describe a Korean transliterated foreign word extraction algorithm. In the proposed method, we reformulate the foreign word extraction problem as a syllable-tagging problem such that each syllable is tagged with a foreign syllable tag or a pure Korean syllable tag. Syllable sequences of Korean strings are modelled by Hidden Markov Model whose state represents a character with binary marking to indicate whether the syllable is part of a transliterated foreign word or not. The proposed method extracts a transliterated foreign word with high recall rate and precision rate. Moreover, our method shows good performance even with small-sized training corpora.

Keywords: Transliteration; Hidden Markov Model; Syllable Tagging; Korean Transliterated Foreign Word; Cross Lingual Information Retrieval.

1. Introduction

Recently, the use of words of foreign origin in Korean texts is growing at a high speed. They are mostly from English words and are usually used as transliterated forms. In Korean, the transliterated foreign words[1] are written in various forms though there is a standard form [10, 11]. For example, an English word 'data' can be transliterated into Korean words like '*deiteo*', '*deita*' and so on. This causes errors such as the well-known unknown word problem in a morphological analyser.

[1]In this paper, we use a term 'foreign word' and 'transliterated word' as the meaning of a transliterated foreign word.

One spacing unit in Korean is called a word phrase. A typical word phrase consists of a sequence of content words (like noun or verb stem) and functional words (like postposition or verbal ending). A sentence '*na+neun hag-gyo+e geu+rang ga+n-da*[2]' ("I go to school with him.") can be analysed as Table 1: (the underlined morphemes '*neun*', '*e*', '*rang*' and '*n-da*' are topical marker, dative marker, co-participant marker, and verbal ending respectively.)

Table 1. The analysed result of '*na+neun hag-gyo+e geu+rang ga+n-da*'.

Meaning	I+topic	school+dative	he+with	Go+verbal ending
Korean	*na+neun*	*hag-gyo+e*	*geu+rang*	*ga+n-da*

Since the set of functional words is closed and they are usually positioned after content words, content words (like '*na*', '*hag-gyo*', '*geu*' and '*ga*' in the above example) and their part-of-speech may be identified just by deleting the closed set of functional words, even if some content words are not registered in the dictionary [7]. In Korean, this is a widely used heuristic to handle the unknown word problem. However, this simple heuristic can cause an error when it comes to a non-Korean transliterated word such as '*o-pe-ra-neun*' (whose right morphological structure is '*o-pe-ra+neun*'; '*opera*'+topical marker). We have two ambiguities here because of two possible candidate postposition markers: '*ra-neun*' and '*neun*'. Unfortunately, the simple heuristic produces a wrong result, *'o-pe+ra-neun'* ('*o-pe*'+maker for quotation). Because the pure Korean words show very little confusion to differentiate the combination of content and functional words according to their surface alphabetic peculiarities, almost all errors come from the transliterated spelling of non-Korean transliterated foreign words[3]. For example, former French president 'Mitterand' is transliterated into Korean spelling '*mi-te-rang*'. Here, '*rang*' can be a co-participant marker like '*rang*' in Table 1, and then '*mi-te-rang*' is wrongly analysed into *'mi-te+rang*' ('*mi-te*'+dative marker) that means "and under" or "with *mi-te*", where "*mi-te*" may be a proper noun in Korean.

[2]Korean romanized transcription will be written in italic script. In the transcription, '+' indicates a word boundary and '-' indicates a syllable boundary.

[3]We analyzed unknown words which produced this type of error in KT and KRIST test collection, when the simple heuristic[7] was applied. In the analysis, 93% was caused by transliterated foreign words, such as '*bol-cheu-**man***'(Boltzman), '*mo-di-pa-**i***' (modify), '*o-veo-le-**i***' (overlay) and '*ha-i-deu-**lo***' (hydro). Here, the syllables in bold, '*man*', '*i*', and '*lo*', are used as functional words in Korean.

To avoid the problem, we need a secure extraction algorithm to recognize transliterated words in the word stream. This paper tries to find an effective method for automatic detection[4] and extraction[5] of transliterated foreign words.

This paper is organized as follows: Section 2 describes the related works. Section 3 deals with details of our method and Section 4 shows several experiments. Section 5 deals with discussion. Conclusion and future works are drawn in Sections 6.

2. Related Works

Recent works on extracting transliterated foreign words from Korean corpora [8, 15] used statistical information regarding the use of different sequence of syllables between transliterated words and pure Korean words. They identified extraction of transliterated foreign words in two-step procedures — detection step and extraction step. In the first step (detection step), statistical information was used to determine whether a given word phrase was likely to contain a transliterated foreign word or not. Such statistical information was acquired by unigram and bi-gram statistics of syllable sequences and they used Equation (1) for the decision.

$$D(W) = \frac{P(Transliterated \mid W)}{P(pure\ \ Korean \mid W)} = \frac{P(W \mid Transliterated) \times P(Transliterated)}{P(W \mid pure\ \ Korean) \times P(pure\ \ Korean)} \quad (1)$$

where

$P(Transliterated \mid W)$: the conditional probability, which a word phrase W is a transliterated foreign word

$P(pure\ Korean \mid W)$: the conditional probability, which a word phrase W is a pure Korean word

In Equation (1), if $D(W) > 1$, the algorithm determined that W contains a transliterated foreign word and passed W to the second step. They estimated $P(Transliterated \mid W)$ and $P(pure\ Korean \mid W)$ from the frequency of transliterated foreign words and pure Korean words in training corpora. For estimating $P(W \mid Transliterated)$ and $P(W \mid pure\ Korean)$, they used bigram and unigram of syllables in a word phrase as Equation (2). Note that the condition $D(W) > 1$ is the prerequisite to make a word phrase W fed into the second step.

[4] 'Detection of transliterated word' can be defined as a binary decision whether a given word phrase contains transliterated words or not.

[5] 'Extraction of transliterated words' can be defined as segmenting a given word phrase into words and classfying them into transliterated words and pure Koean words.

In the second step (extraction step), a transliterated foreign word was extracted by eliminating pure Korean words and particles in the word phrases where D(W) is higher than 1 in the previous step — it means that word phrases, in which D(W) is lower than 1 or equal to 1, are not handled in this step [14]. The method produced a relative good result — precision rates about 78.98% and recall rates about 63.54%.

$$
\begin{aligned}
&P(W \mid Transliterated) \approx \\
&\lambda \times p(s_1 \mid Transliterated) \times p(s_2 \mid Transliterated) \times \ldots \times p(s_n \mid Transliterated) + \\
&(1-\lambda_1) \times p(\phi s_1 \mid Transliterated) \times p(s_1 s_2 \mid Transliterated) \times \ldots \times p(s_n \phi \mid Transliterated) \\
&p(s_i \mid Transliterated) = \frac{frequency(transliterated\ \ words\ \ containing\ \ s_i)}{frequency(transliterated\ \ words)} \\
&p(s_i s_{i+1} \mid Transliterated) = \frac{frequency(transliterated\ \ words\ \ containing\ \ s_i s_{i+1})}{frequency(transliterated\ \ words)}
\end{aligned}
\tag{2}
$$

where

s_i is the *ith* syllable of a given word phrase W

ϕ is the symbol indicating start and end points of word phrases

However, this method has some limitations in extracting a transliterated foreign word. First, since this method is composed of two steps, the errors in the first step can be propagated to the second step. It means that the method does not attempt to extract a transliterated foreign word when the given word phrase is determined as consisting of only pure Korean words in the detection step, even if it contains a transliterated foreign word. Second, the detection of a transliterated foreign word depends on the number of syllables written in a transliterated foreign word of the given word phrase (Equation (2)). The method works very well, when a transliterated foreign word appears alone. However, in Korean, transliterated foreign words can be used to compose compound nouns with pure Korean words and are attached with functional words[6]. This makes it difficult for the method to detect transliterated foreign words under the condition that there are more syllables, which are part of pure Korean words, than those, which are part of transliterated foreign words. For example, the method cannot detect a transliterated word in the given word phrase, *'gaek-che+ji-hyang+si-seu-tem+e-seo'* ("in the object-oriented system"), since there are six syllables of pure Korean words — '*gaek*', '*che*', '*ji*', '*hyang*', '*e*', '*seo*' — and three syllables

[6]Word phrases which are composed of pure Korean words and transliterated foreign words will be called 'word phrases with combination forms' in this paper.

of a transliterated foreign word — '*si*', '*seu*', '*tem*' that represents "system" in English. To avoid these limitations, we use Hidden Markov Model for detecting and extracting transliterated words from word phrases.

3. Detecting and Extracting Transliterated Words

3.1. Preliminaries

The main idea of extracting a transliterated foreign word is that the composition of transliterated foreign words would be different from that of pure Korean words, since the phonetic system for the Korean language is different from that for the foreign language. Especially, several English consonants that occur frequently in English words, such as 'p', 't', 'c', and 'f', are transliterated into Korean consonants '*p*', '*t*', '*k*', and '*p*' respectively. These consonants do not occur frequently in pure Korean words. This can be an important clue for extracting transliterated foreign words from Korean texts. For example, in a word phrase '*o-pe-ra'* (opera), the syllable '*pe*' has high probability to be a syllable of the transliterated foreign word, since the consonant, '*p*', in the syllable '*pe*' is usually not used in a pure Korean word. However, solely consonant information may not offer enough information to distinguish the difference between a syllable in a pure Korean word and that in a transliterated foreign word, since a Korean syllable is composed of a consonant in the beginning position, a vowel in the middle position, and a consonant in the last position. We call them '*cho-seong*', '*jung-seong*', and '*jong-seong*', respectively. For example, '*tem*' is composed of *cho-seong* '*t*', *jung-seong* '*e*', and *jong-seong* '*m*'. Since, consonants are usually not used alone in Korean texts, we extract transliterated words in two conditions — 1) with only syllable information, 2) with syllable information and consonant information.

In our method, we reformulate the foreign word extraction problem as a **'syllable-tagging'** problem such that each syllable is tagged with a foreign syllable tag or a pure Korean syllable tag. Syllable sequences of Korean strings are modelled by Hidden Markov Model whose state represents a character with binary marking to indicate whether the character is part of a pure Korean word or not. We use transition probability, consonant probability, and syllable probability acquired from a syllable-tagged corpus to determine whether a syllable in a given word phrase is likely to be a part of a transliterated foreign word or not. For a given word phrase, each syllable in the word phrase is tagged with 'F' or 'K' (a syllable with tag 'F' means that the syllable is part of a transliterated foreign word, and a syllable with tag 'K' means that the syllable is part of a pure Korean word). For example, word phrases '*o-pe-ra* + *neun* (opera + topical

marker)' and '*mi-te-rang* (Mitterand)' that contain transliterated foreign words - '*o-pe-ra* (opera in English)' and '*mi-te-rang* (Mitterand in English)' — can be tagged as Table 2.

Table 2. Syllable tagged results of '*o-pe-ra* + *neun*' and '*mi-te-rang*'.

Word Phrase	Syllable tagged result
o-pe-ra+*neun* (opera+neun)	o/F + pe/F + ra/F + neun/K
mi-te-rang (Mitterand)	mi/F + te/F + rang/F

A series of 'F' tags makes it possible to detect and extract transliterated foreign words in the tagged results. If there is a series of 'F' tags in the result, we can determine that a given word phrase contains transliterated words and the words corresponding to the series of 'F' tags can be extracted as transliterated words. The whole procedures for identifying and extracting a transliterated word are depicted in Figure 1.

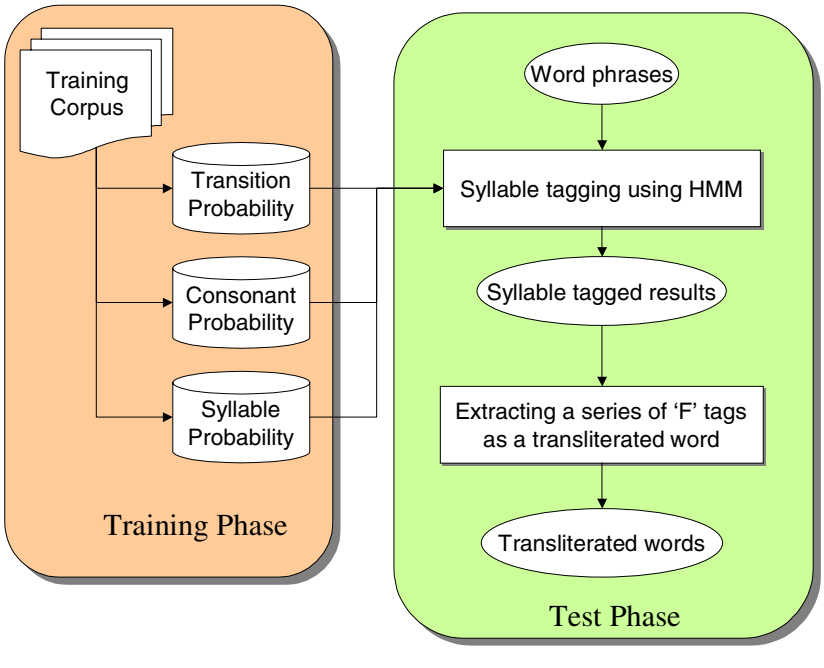

Figure 1. Procedures for identifying a transliterated foreign word.

3.2. Syllable tagging using Hidden Markov Model

The problem of extracting a transliterated foreign word can be defined as syllable-tagging. For a given word phrase S, which is composed of syllables $s_{1...n}$, the task is to find the sequence of tags T which is composed of tags $t_{1...n}$ that maximize the probability $p(t_{1...n} | s_{1...n})$. This can be described as Equation (3).

$P(S|T)P(T)$ in Equation (3) can be rewritten into Equation (4), since they can be represented as a sequence of conditional probability [20]. By the Markov assumption [3], Equation (4) can be transformed to Equation (5).

$$\phi(S) \overset{def}{=} \underset{T}{argmax}\, P(T \mid S) = \underset{T}{argmax}\, P(S \mid T)P(T) \tag{3}$$

$$P(S \mid T)P(T) = \prod_{i=1}^{n} p(s_i \mid t_n,..,t_1, s_{i-1},..., s_1)p(t_i \mid t_{i-1}, t_{i-2},...t_1) \tag{4}$$

$$P(S \mid T)P(T) = \left[\prod_{i=1}^{n} p(s_i \mid t_i, t_{i-1}, s_{i-1})\right] \times p(t_1 \mid t_0) \times \left[\prod_{i=2}^{n} p(t_i \mid t_{i-1}, t_{i-2})\right] \tag{5}$$

where

s_i: the *ith* syllable in the given word phrase S
t_0: a tag for start of a given word phrase S
t_i: the *ith* tag ('F' or 'K') of the syllable in the given word phrase S

Next, $p(t_i|t_{i-1},t_{i-2})$ is estimated as in Equation (6) and $p(s_i|s_{i-1},t_i,t_{i-1})$ is estimated as in Equation (7) and Equation (10) respectively using syllable-tagged corpora, which are manually tagged with 'F' and 'K' tags [2]. There are two versions of syllable probability, $p(s_i|s_{i-1},t_i,t_{i-1})$, whether it uses linear interpolation using consonant information or not. Equation (7) represents syllable probability without linear interpolation using consonant information and Equation (10) represents syllable probability with linear interpolation using consonant information [16, 18]. Note that in Equation (7), each probability is estimated only with syllable (s_i) and its tag (t_i). But in Equations (8) and (9), each probability is estimated with syllable (s_i), its consonants (c_i) and its tag (t_i). In Equation (10), the probabilities estimated in Equations (8) and (9) are linearly interpolated with parameter λ_1. We will use Equation (7) and Equation (10) for extracting transliterated words in two conditions as described in Section 3.1.

$$p(t_i \mid t_{i-1}, t_{i-2}) = \frac{C(t_i, t_{i-1}, t_{i-2})}{C(t_{i-1}, t_{i-2})}, \qquad p(t_i \mid t_{i-1}) = \frac{C(t_i, t_{i-1})}{C(t_{i-1})} \tag{6}$$

$$p'(s_i \mid t_i) = \frac{C(s_i, t_i)}{C(t_i)}, \qquad p'(s_i \mid s_{i-1}, s_i, t_{i-1}) = \frac{C(s_i, s_{i-1}, t_{i-1}, t_i)}{C(s_{i-1}, t_i, t_{i-1})}$$
$$p(s_i \mid s_{i-1}, t_i, t_{i-1}) = \lambda_1 \times p'(s_i \mid t_i) + (1-\lambda_1) \times p'(s_i \mid s_{i-1}, t_i, t_{i-1}) \tag{7}$$

$$p''(s_i \mid t_i) = \lambda_2 \times \frac{C(s_i, t_i)}{C(t_i)} + (1-\lambda_2) \times \frac{C(c_i, t_i)}{C(t_i)} \tag{8}$$

$$p''(s_i \mid s_{i-1}, t_i, t_{i-1}) = \lambda_2 \times \frac{C(s_i, s_{i-1}, t_{i-1}, t_i)}{C(s_{i-1}, t_i, t_{i-1})} + (1-\lambda_2) \times \frac{C(c_i, c_{i-1}, t_{i-1}, t_i)}{C(c_{i-1}, t_{i-1}, t_i)} \tag{9}$$

$$p(s_i \mid s_{i-1}, t_i, t_{i-1}) = \lambda_1 \times p''(s_i \mid t_i) + (1-\lambda_1) \times p''(s_i \mid s_{i-1}, t_i, t_{i-1}) \tag{10}$$

where,

c_i[7]: consonants in s_i, c_i represents two consonants in s_i
$C(T)$: represents the occurrences of T in the training corpora

We can extract a transliterated foreign word in one step. If there is a series of 'F' tags in the syllable tagged result, the series of the 'F' tagged syllables can be extracted as a transliterated foreign word. For example, '*gaek-che+ji-hyang+si-seu-tem+e-seo-neun*' ("in the object oriented system") may be tagged as follows:

Syllable	*geak*	*che*	*ji*	*hyang*	*si*	*seu*	*tem*	*e*	*seo*	*neun*
tag	K	K	K	K	F	F	F	K	K	K

Although, there are more pure Korean syllables than transliterated syllables in the given word phrase, our model can capture whether the word phrase contains a transliterated foreign word or not — [8, 15] cannot do it as described in Section

[7]A Korean syllable is composed of a consonant in the beginning position, a vowel in the middle position, and a consonant in the last position — note that a consonant in the last position is optional. c_i represents the set of consonants in s_i such as c_i = {a consonant in the beginning position in s_i , a consonant in the last position in s_i}. In Korean, the number of possible consonants in the beginning position is 19 and that in the last position is 31 — the number of possible combination of the two consonants is 589 ($19 \times 31 = 589$). In our test collection (KT collection and KRIST collection), only 189 combinations appeared. This may be the fact that complex consonants in the last position such as 'lt', 'lb' and so on are not frequently used in our collection. The frequently used combination is '{g , null, 24793}', '{s, null, 21868}', '{j, null,14533}', '{l, null, 12196}' and so on. Here, a format of the combination is '{a consonant in the beginning position, a consonant in the last position, frequency}.

2.1. The substring '*si-seu-tem* ('system' in English)' with continuous 'F' tags in the syllable tagged result will be recognized as a transliterated foreign word. Figure 2 shows graphical representation of the syllable tagging procedure.

t_{i-2}	t_{i-1}	t_i	$p(t_i \mid t_{i-1},t_{i-2})$	t_{i-2}	t_{i-1}	t_i	$p(t_i \mid t_{i-1},t_{i-2})$
	ϕ	F	0.19110	F	F	K	0.09287
	ϕ	K	0.80889	F	K	ϕ	0.70731
ϕ	F	ϕ	0.03960	F	K	F	0.00238
ϕ	F	F	0.95606	F	K	K	0.29030
ϕ	F	K	0.00433	K	F	ϕ	0.02021
ϕ	K	ϕ	0.02147	K	F	F	0.97288
ϕ	K	f	0.00286	K	F	K	0.00690
ϕ	K	K	0.97566	K	K	ϕ	0.46870
F	F	ϕ	0.32571	K	K	F	0.00944
F	F	F	0.58141	K	K	K	0.52185

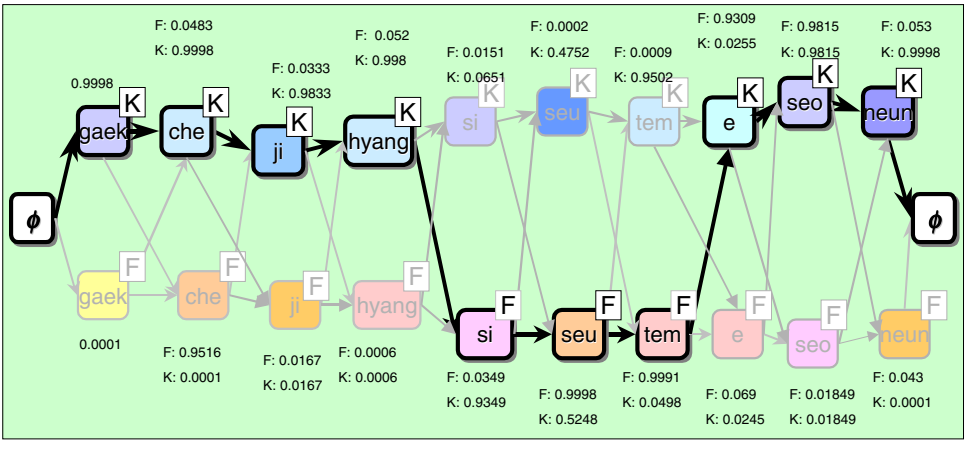

Figure 2. Graphical representation of the syllable-tagging procedure.

In Figure 2, the probability for each node is syllable probability and probability in the table is transition probability. Since, there is one path from 'ϕ' to '*gaek*[K][8]', we can acquire probability $p(\phi \rightarrow gaek[K])=p(K|\$) \times p(gaek|\phi,K)$ Via Equation (5). As the same manner, we can acquire $p(\phi \rightarrow gaek[F])= p(F|\phi) \times p(gaek|\phi,F)$. For the second syllable '*che*', there are two paths for each case — $gaek[K] \rightarrow che[K]$, and $gaek[F] \rightarrow che[K]$, for '*che*[K]' and $gaek[K] \rightarrow che[F]$, and $gaek[F] \rightarrow che[F]$ for '*che*[F]'. For these ambiguous cases, a path, which has higher probability, is selected. For example, $p(gaek[K] \rightarrow che[K])$ has higher probability than $p(gaek[F] \rightarrow che[K])$, and $gaek[K] \rightarrow che[K]$ is selected as the relevant path for '*che*[K]' in Figure 2. For each syllable and each tag, we

[8]'syllable [syllable tag]'.

can select a path using probability and finally the path with bold arrows in Figure 2 can be acquired using the Viterbi algorithm [3].

4. Evaluation

4.1. Experimental setup

The proposed method was tested on two kinds of corpus. The first one was called KT collection [19] that contained 4,414 documents in a computer science field. The second one was called KRIST collection [9] that contained 13,515 documents in a scientific field including biology, physics and so on. We tagged manually both collections with 'F' and 'K' for experiments.

Table 3 shows the number of word phrases, which contain transliterated foreign words, and pure Korean word phrases in each collection. KW means the number of word phrases, which are composed of only pure Korean words and TFW means the number of word phrases, which contain transliterated foreign words. In Table 3, the KT collection contains more word phrases containing transliterated foreign words (about 27.9% of a total word phrase in the collection) than the KRIST collection does (about 12.4% of a total word phrase in the collection). We will describe the effect of the characteristics on performance in analysing evaluation results.

Table 3. Characteristics of each collection.

	KW	TFW	Total
KRIST collection	52,598 (87.58%)	7,456 (12.42%)	60,054
KT collection	29,762 (72.24%)	11,495 (27.86%)	41,257

In this paper, we perform six kinds of experiment to evaluate performance. One of them is on transliterated word detection and the others are on transliterated word extraction.

a) An experiment on transliterated word detection.
b) To evaluate the robustness of our method, we examined the performance in two conditions — homogeneous test and cross test. In the homogeneous test, the same kinds of collection were used as a training set and a test set — for example, a training and a test set can be extracted from the same collection, say, KT collection. In the cross test, different kinds of collection

are used — for example, a training set is from KT collection and a test set is from KRIST collection.

c) The previous works [8, 15] were compared with the proposed method to measure the performance improvement.

d) Performance using Equation (7) is compared with that using Equation (10). This experiment will show the effect of consonant information on performance.

e) An experiment according to word phrase types — pure Korean word phrase, pure transliterated word phrase, and combination of pure Korean words and transliterated words.

f) An experiment according to the training data size.

The results were evaluated by precision rates and recall rates [21]. The precision rate is defined as the proportion of the correct answers to the extracted results, and the recall rate is defined as the proportion of the correct answers to the result that should be extracted from the given corpus.

$$precision = \frac{correctly\ extracted\ trasliterated\ words}{extracted\ transliterated\ words}$$
$$recall = \frac{correctly\ extracted\ trasliterated\ words}{transliterated\ words\ in\ the\ text} \tag{11}$$

4.2. Transliterated words detection

Table 4 shows the result of a foreign word detection experiment. The table shows that the unigram and bigram model produces high precision but relatively low recall rate, especially for KRIST collection. Since a transliterated foreign word detection procedure of the unigram and bigram model depends on the number of transliterated foreign words in the collection, some transliterated foreign words cannot be detected. This is the reason why recall rate is lower than precision rate in the unigram and bigram model. Because of the nature of the unigram and bigram model — a pipelined method, foreign word extraction after foreign word detection — these results can have a negative effect on foreign word extraction.

Our model detects transliterated foreign words well in both collections. Recall rate is about 97% for KRIST collection and about 99% for KT collection. Precision rate is about 98% for KRIST collection and about 99% for KT collection. Our model produces better performance in the foreign word detection experiment — about 23% recall improvement on the average, especially 38.5% recall improvement for KRIST collection for unigram and bigram model.

Table 4. Results of transliterated word detection[9].

	Collection	Recall	Precision
Unigram and Bigram Model [8, 15]	KRIST	69.27%	98.08%
	KT	89.12%	95.09%
HMM model (The proposed method)	KRIST	97.82%	98.53%
	KT	98.77%	99.14%

4.3. Transliterated words extraction

4.3.1. Homogeneous test and cross test

We used all word phrases in the KT collection and randomly selected 40,000 word phrases from the KRIST collection to balance the size of each collection. Then, we randomly selected 90% of word phrases in each collection as training set and select the remaining 10% as test set.

The result, depicted in Table 5, can be interpreted as follows. If training and test set are the KRIST collection, recall and precision are 94.32% and 95.01% respectively — a result of homogeneous test. If a training set is KRIST collection and a test set is KT collection, recall rates and precision rates are 84.17% and 85.32% respectively — it is a result of cross test. In the result, we find that our method shows higher performance both in homogeneous test and in cross test, when training set is KT collection than when training set is KRIST collection. This may be caused by the fact that there are more word phrases containing a transliterated foreign word in the KT collection than those in the KRIST collection. Thus, relevant probability for a transliterated foreign word may be more easily acquired from KT collection.

Table 5. Precision rate and Recall rate of homogeneous test and cross test.

Train set/ Test set	Cross Test		Homogeneous Test	
	KRIST / KT	KT / KRIST	KRIST / KRIST	KT / KT
Recall	84.17%	91.31%	94.32%	97.05%
Precision	85.32%	94.92%	95.01%	97.42%

[9]In the experiment, we used the same test collection in training and test. This means that we perform the homogeneous test in transliterated word detection.

These results show that our model extracts a transliterated foreign word accurately both in the homogeneous test and in the cross test.

4.3.2. Comparison with previous work

We randomly selected 90% of word phrases in each collection as training set and selected the rest as test set to compare the performance of our model with that of the unigram and bigram model.

Table 6 shows the performance of the previous method [8, 15] and that of the proposed method (HMM model). In the table, the unigram and bigram model shows relatively low recall rate in both conditions (in overall performance and in only extraction performance). The main reason why the recall rate of the previous method in overall performance is lower than ours is that the unigram and bi-gram model can not detect a transliterated foreign word effectively, when there is a word phrase that contains more syllables in pure Korean word than those in transliterated foreign words. The low performance of the foreign word extraction step in the previous method, also affect the overall performance. This means that there are some errors when functional words and Korean words are eliminated in the step.

The result can be summarized as follows:

- Precision and recall rates of our proposed method are higher at any collection than those of the previous works.
- The proposed method shows significant improvement — the recall rate and the precision rate about 42.8% and 17% respectively — when it is compared with unigram and bigram model.
- Our proposed method outperforms the previous works

Table 6. Results of foreign word extraction.

	Collection	Recall	Precision
Unigram and Bi-gram model [8, 15] (when both detection and extraction steps are evaluated-overall performance)	KRIST KT	64.10% 66.78%	82.04% 77.85%
Unigram and Bigram Model [8, 15] (when only extraction step is evaluated)	KRIST KT	86.7% 73.8%	86.6% 80.2%
HMM model (The proposed method)	KRIST KT	94.32% 97.05%	95.01% 97.42%

4.3.3. Performance comparison between information with and without consonant

To examine the effect of consonant information, randomly select 90% of word phrases in each collection as training set and the rest as test set are used. Equation (7) (without consonant information) and Equation (10) (with consonant information) are used for this experiment. Table 7 shows the result in each case. In table 7, Equation (10) — with consonant information — produces better results in both collections than Equation (7) — without consonant information. As a result of this experiment, we found that consonant information is helpful to extract transliterated foreign words. We believe that consonant information may support probability of syllables, which did not appear in training data. For example, if '*tem*' did not appear in training data, we can estimate its probability by consonant information of '*t*' and '*m*', which may consist of syllables in training data, such as '*tam*', '*tim*', '*teom*' and so on.

Table 7. Performance using Equation (7) and using Equation (10).

	Test collection	Precision	Recall
Without consonant information (Eq. (7))	KT collection	97.26%	96.07%
	KRIST collection	92.05%	92.33%
With consonant information (Eq. (10))	KT collection	97.05%	97.42%
	KRIST collection	94.32%	95.01%

4.3.4. Performance according to word phrase types

To examine the performance according to word phrase types, randomly select 90% of word phrases in each collection as training set and the rest as test set are used. Word phrases are classified into three types — pure Korean, pure transliterated, combination. The pure Korean type contains word phrases, which are composed of only pure Korean words and the pure transliterated type contains word phrases, which are composed of only transliterated words. 'Word phrases with combination forms' are classified into the combination type. Tables 8 and 9 show the number of classified word phrases in each collection and each set. In the tables, the pure Korean type and the pure transliterated type more frequently appears in each collection than the combination type. This distribution affects the performance of each word phrase type as shown in Table 10. In Table 10, the pure Korean and the pure transliterated types show relatively higher

performance than that of the combination type. Especially the wrong boundaries between pure Korean words and transliterated words are the main reason of errors in this type. The main reason for the result can be analysed as the data sparseness problem of the combination type. As shown in Tables 8 and 9, the number of word phrases in combination type is smaller than others. This leads to make the transition probability, on condition that previous syllable tag is different from current syllable tag such as '$t_{i-1} = F$', '$t_i = K$' and '$t_{i-1} = F$', '$t_i = K$', smaller than that in other conditions.

Table 8. Distribution of word phrases for each word phrase type in training data.

Type	KT collection	KRIST collection
Pure Korean	24,941 (69.28%)	47261 (87.52%)
Pure Transliterated	9,202 (25.56%)	3391 (6.28%)
Combination	1,857 (5.16%)	3348 (6.20%)
Total	36,000	54,000

Table 9. Distribution of word phrases for each word phrase type in test data.

Type	KT collection	KT collection
Pure Korean	2,772 (69.3%)	5,292 (88.20%)
Pure Transliterated	1,007 (25.18%)	366 (6.10%)
Combination	221 (5.52%)	342 (5.70%)
Total	4,000	6,000

Table 10. Experimental results for each word phrase type.

Type	Collection	Recall	Precision
Pure Korean	KT	99.61%	99.61%
	KRIST	99.80%	99.80%
Pure Transliterated	KT	98.82%	99.87%
	KRIST	95.52%	97.71%
Combination	KT	89.09%	90.96%
	KRIST	93.03%	95.05%

4.3.5. Performance according to training data size

Figure 3 shows recall rates and precision rates according to the size of training set. For the experiment, we combined the two collections into one. The total number of word phrases in the combined collection was about 100,000. We randomly selected 10,000 word phrases from combined collection and fixed it as test set. In Figure 3, the x coordinate indicates the proportion of training set size to the combined collection size and the y coordinate indicates the percentage of recall rates and precision rates. In Figure 3, both recall rates and precision rates converge from the 30% point of x coordinate (when the size of training set is about 30,000 word phrases). Moreover, the method produces a relative good result, even if the size of training set is about 3,000 word phrases (3% in the x coordinate). This implies that our method works very well even with the small size of training data — training data size is about 3,000 word phrases and test set size is 10,000.

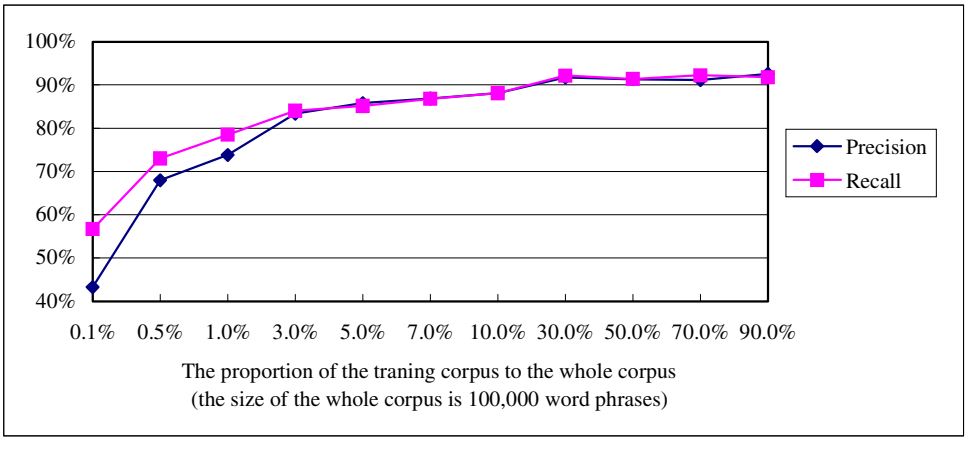

Figure 3. Performance according to training data size.

4.4. Analysing errors

The errors are classified into two major causes. The first one is word phrases, which are composed of one syllable and the second one is a syllable, which '*cho-seong*' — the beginning consonant of a syllable — is not a phoneme. '*i-eung*' in '*cho-seong*' is used as only grapheme which has no phonetic value in Korean.

Word phrases composed of one-syllable cause errors by lack of information. Since there is little information such as in 'a start symbol of a word phrase (ϕ)', 'syllable (s_i)', and 'an end symbol of a word phrase (ϕ)', it is difficult to produce

correct answers for this type of a word phrase. Table 11 shows examples of errors in this type. In the table, the word phrase '*bool*', '*jon*', and '*bi*' are wrongly tagged as 'K'. In the result, we find that errors wrongly tagged as 'K' are dominant. The main reason for this phenomenon may be caused by the fact that there are more word phrases of pure Korean type than that of pure transliterated type.

Table 11. Examples of errors caused by one-syllable word phrases.

Meaning	Result	Correct answer
Bool	'*bool*' [K]	'*bool*' [F]
Zone	'*jon*' [K]	'*jon*' [F]
Alphabet 'B'	'*bi*' [K]	'*bi*' [F]

A syllable containing '*cho-seong*'[10], which is not a phoneme, has more ambiguity in syllable tagging than that containing '*cho-seong*', which is a phoneme, because 'zero consonant' is frequently used in not only syllables which consist of a pure Korean word but syllables which consist of a transliterated word. Table 12 shows examples of errors of this type. In the example '*eo-sem-beul-ri-eo*'[FFFFK], the beginning consonant in the syllable '*eo*' is the 'zero consonant' — it is used as a Korean word '*eo*', which means 'language' and used as a part of a transliterated word '*eo-sem-beul-ri*' which means 'assembly'. The zero consonants, '*o*', '*in*', '*ok*' and '*e*', are also used as components of pure Korean words and transliterated words — '*o*' and '*in*' in '*o-in-sik*', '*ok*' in '*ok-tet*' and '*e*' in '*e-ji*'. It will need to make rules to reduce these errors.

Table 12. Examples of errors caused by zero consonants.

Meaning	Result	Correct answer
Octet	'*ok-tet*' [KK]	'*ok-tet*' [FF]
Assembly language	'*eo-sem-beul-ri-eo*' [FFFFF]	'*eo-sem-beul-ri-eo*' [FFFFK]
Edge	'*e-ji*'[KK]	'*e-ji*'[FF]
Misunderstanding	'*o-in-sik*' [FKK]	'*o-in-sik*' [KKK]

[10]We call it 'zero consonant' in this paper. The transciribed syllable containing zero consonant usually start with a vowel — for example, syllable '*eo*', '*ok*', '*o*', '*e*' etc. in this paper.

The results of experiments in Section 4 can be summarized as follows:

- In both detecting and extracting transliterated foreign words, our method produced good results.
- In both homogeneous and cross test, our method extracted transliterated words well.
- There was significant improvement on the previous work in both detection and extraction of transliterated foreign words.
- Consonant information was useful.
- Our method showed a good result for three word phrase types — pure Korean type, pure transliterated type and combination type.
- Our method captured transliterated foreign words even with small-sized training data.

5. Discussion

There was a partially supervised algorithm to extract transliterated foreign words [6]. The algorithm aims at finding transliterated foreign words without large amounts of syllable-tagged corpora. In the work, a syllable-tagging method that is similar to our method and a word segmentation algorithm are mainly used. The mechanism of the syllable-tagging is the same as ours except parameter estimation, syllable probability[11] and newly introduced syllable tag 'J' for a syllable, which is part of functional words. The model parameters are estimated not from syllable-tagged corpora but from entries of dictionaries — Korean noun dictionary, transliterated foreign word dictionary and functional word dictionary. The Korean noun dictionary is used as a source of pure Korean words and the transliterated foreign word dictionary is used as a source of pure transliterated words. However the dictionaries can not give data for word phrase, where transliterated foreign words are attached with functional words or compose compound nouns with pure Korean words. This makes it difficult to estimate parameters directly from the dictionaries. Thus, some parameters are indirectly estimated using the first syllable of entries in each dictionary. Transition probability is also indirectly estimated with equally distributed probabilities among possible transitions.

Then, the syllable-tagged word phrases are segmented in the word segmentation phase. In this stage, functional words are detached and compound nouns are segmented using the dictionaries and syllable-tagging information.

[11]Instead of $p(s_i|s_{i-1},t_i,t_{i-1})$ in our method, the method use $p(s_i|\ t_i,t_{i-1})$ as syllable probability.

Finally foreign words are extracted by performing syllable-tagging again on the word segments from functional word detachment and compound noun segmentation. For newly syllable-tagged segments, the algorithm extracts transliterated foreign words when the segments contain more than 50% foreign syllables.

Even though the direct comparison between our method and the partially supervised algorithm [6] may not be reasonable, because they use different language resources and aims at different learning methods — Kang's method used a partially supervised algorithm to find foreign words without large amounts of syllable tagged corpora but our method is a supervised algorithm to find foreign words using a fully statistical method —, we compare the result of our model with that from the partially supervised algorithm using the same test set. The algorithm [6] showed relatively high performance for KT collection — about 83% precision and 94% recall — but relatively low performance for KRIST collection — about 40% precision and 86% recall. We find that the errors are mainly caused by the indirectly estimated model parameters of the syllable tagging. This means that the syllable-tagging method of the algorithm mainly caused errors when word phrases were a combination of pure Korean word, transliterated foreign words and functional words. For example, '*e-seu-te-reu-wa*' (with ester) that is a combination of a transliterated foreign word '*e-seu-te-reu*' (ester) and a functional word '*wa*' (with), was tagged as 'FFFFF' — correct one is 'FFFFJ'. The performance for each test set is also related to 'word phrases with combination forms'. The proportion of 'word phrases with combination forms' to total word phrases in the KT test set was only 14% but that in the KRIST test set was 43%. This is the reason why the performance for KRIST collection is lower than that for KT collection.

6. Conclusion

In this paper, we proposed a method to handle the problems associated with transliterated foreign words in Korean texts. In this method, we reformulated the foreign word detection and extraction problem as a syllable-tagging problem such that each syllable was tagged with a foreign syllable tag or a pure Korean syllable tag. Our method showed the higher performance in transliterated word detection and extraction than that of the previous model (unigram and bi-gram model) in both precision rates and recall rate. The proposed method based on HMM showed a relatively good result in the homogeneous test and cross test and for three word phrase types. We also showed that consonant information was helpful and our method captured transliterated foreign words even with small-sized training data.

In future works, we will devise rules for reducing errors which statistical measure did not handle. Although we showed that consonant information is helpful, the performance improvement did not come up to our expectations. The reason may be the fact that various types of information may not be integrated efficiently and we will integrate them with an efficient model. In training data and test data, there are less word phrases containing transliterated words than those containing pure Korean words. Using our method we may acquire large amounts of training data from Korean texts to relieve the data sparseness problem..

Our work can be applied to some applications such as automatic term recognition, namely entity recognition, automatic transliteration, and construction of bi-lingual dictionary. Since many terms and many proper nouns — person name, location name, and organization name — are of foreign origin in Korean texts, our method will be useful for automatic term recognition (ATR) [17] and named entity (NE) recognition [13]. For automatic transliteration, our work can be used as a pre-processor to find target words [4, 5, 10, 11]. Our method can be also used as a pre-processor to find target words for constructing bi-lingual dictionary, in which entries are transliterated words [12, 22]. Due to the statistical nature of HMM, our approach may be extensible to different domains such as automatic word spacing [23] and word segmentation — with binary marking to indicate whether the syllable is an end point of a word (a word boundary) or not.

Acknowledgements

KORTERM is sponsored by the Ministry of Culture and Tourism under the program of King Sejong Project. This work was partially supported by the Ministry of Science and Technology through the “Knowledge base prototype construction and its application for human knowledge processing modelling” (M1-0107-00-0018) project at the BSRC. This work was also supported by the Korea Science and Engineering Foundation (KOSEF) through the “Basic Research on Mobile Communication Systems Based upon Natural Language Processing” (GH1424V) project.

References

[1] Berger, V. Della Pietra and S. Della Pietra, “A Maximum Entropy Approach to Natural Language Processing”, *Computational Linguistics* 22(1), 1995, 39–71.

[2] X.D. Huang, Y. Ariki and M.A. Jack, *Hidden Markov Models for Speech Recognition*, Edinburgh University Press, 1990.

[3] A. James, *Natural Language Understanding*, Benjamin/Cummings, 1995.
[4] B.J. Kang, J.S. Lee and K.S. Choi, "The Phonetic Similarity Measure for the Korean Transliterations of Foreign Words", *Journal of Korea Information Science Society* 26(10), 1999, 1237–1246.
[5] B.J. Kang and K.S. Choi, "Two Approaches for the Resolution of Word Mismatch Problem Caused by English Word and Various Korean Transliterations in Korean Information Retrieval", *in Proceedings of the International Workshop on Information Retrieval with Asian Languages (IRAL2000)*, 2000, pp. 133–140.
[6] B.J. Kang and K.S. Choi, "Effective Foreign Word Extraction for Korean Information Retrieval", *Journal of Information Processing and Management* 38(1), 2002, 91–109.
[7] S.S. Kang, H.I. Kwon and D.R. Kim, "The Role of Morphological Analyses for Korean Automatic Indexing", *in Proceedings of the 22nd KISS Spring Conference,* 1995, pp. 929–932 (in Korean).
[8] Y.H. Kwon, K.S. Jeong and S.H. Myaeng, "Foreign Word Identification Using Statistical Method for Information Retrieval", *in Proceedings of the 17th International Conference of Computer Processing of Oriental Languages*, 1997.
[9] J.H. Lee, K.N. Choi, H.S. Han, J.W. Kim and S.W. Nam, "Development of the KRIST Test Collection for Researchers in Information Retrieval", *Journal of Korea Society for Information Management* 12(2), 1995 (in Korean).
[10] J.S. Lee and K.S. Choi, "English to Korean Statistical Transliteration for Information Retrieval", *Int. J. Computer Processing of Oriental Languages* 12(1), 1998, 17–37.
[11] J.S. Lee, *An English-Korean Transliteration and Re-transliteration Model for Cross-lingual Information Retrieval*. Ph.D. dissertation, Department of Computer Science, Korea Advanced Institute of Science and Technology, 1999 (in Korean).
[12] J.S. Lee, "Automatic Construction of a Transliteration Dictionary from Bilingual Corpus", *in Proceedings of the 11th Conference on Hangul and Korean Language Information Processing*, 1999, pp. 142–149 (in Korean).
[13] K.H. Lee, *Study on Named Entity Recognition in Korean Text*, M.S. thesis, Department of Computer Science, Korea Advanced Institute of Science and Technology, 2000 (in Korean).
[14] S.H. Myaeng and D.H. Jang, "On Language-Dependency in Indexing", *in Proceedings of the Workshop on Information Retrieval with Oriental Languages*, Taejon, Korea, June 1996.

[15] S.H. Myaeng, K.S. Jeong and Y.H, Kwon, "The Effect of a Proper Handling of Foreign and English Words in Retrieving Korean Test", *in Proceedings of the 2nd International Workshop on Information Retrieval with Asian Languages*, Tsukuba, Japan, Oct. 8-9, 1997, pp. 114–122.

[16] J.H. Oh and K.S. Choi, "Automatic Extraction of Transliterated Foreign Words in the Domain Specific Text", *in Proceedings of the 11th Conference on Hangul and Korean Language Information Processing*, 1999, pp. 137–141 (in Korean).

[17] J.H. Oh, J.H. Lee, K.S. Lee and K.S. Choi, "Japanese Term Extraction Using Dictionary Hierarchy and Machine Translation System", *Journal of Terminology* 6(2), 2000, pp. 287–311.

[18] J.H. Oh and K.S. Choi, "Automatic Extraction of Transliterated Foreign Words Using HMM", in *Proceedings of the 19th International Conference on Computer Processing of Oriental Languages (ICCPOL 2001)*, 2001, pp. 433–438.

[19] Y.C. Park, K.S. Choi, J.K. Kim and Y.H. Kim, "Development of the KT Test Collection for Researchers in Information Retrieval", *in Proceedings of the 23rd KISS Spring Conference*, 1996 (in Korean).

[20] L. Rabiner, "Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *in Proceedings of the IEEE*, Vol. 77, No 2, 1989.

[21] G. Salton and M.J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, 1983.

[22] K. Tsuji, "Automatic Extraction of Translational Japanese-KATAKANA and English Word Pairs from Bilingual Corpora", *in Proceedings of the 19th International Conference on Computer Processing of Oriental Languages (ICCPOL 2001)*, 2001, pp. 245–250.

[23] Do-Gil Lee, Sang-Zoo Lee, Hae-Chang Rim, and Heui-Seok Lim, "Automatic Word Spacing Using Hidden Markov Model for Refining Korean Text Corpora", *in Proceedings of the 3rd Workshop on Asian Language Resources and International Standardization*, 2002.