Term Extraction from Korean Corpora via Japanese

Atsushi Fujii, Tetsuya Ishikawa
Graduate School of Library,
Information and Media Studies
University of Tsukuba
1-2 Kasuga, Tsukuba
305-8550, Japan
{fujii,ishikawa}@slis.tsukuba.ac.jp

Jong-Hyeok Lee
Division of Electrical and
Computer Engineering,
Pohang University of Science and Technology,
Advanced Information Technology Research Center
San 31 Hyoja-dong Nam-gu,
Pohang 790-784, Republic of Korea
jhlee@postech.ac.kr

Abstract

This paper proposes a method to extract foreign words, such as technical terms and proper nouns, from Korean corpora and produce a Japanese-Korean bilingual dictionary. Specific words have been imported into multiple countries simultaneously, if they are influential across cultures. The pronunciation of a source word is similar in different languages. Our method extracts words in Korean corpora that are phonetically similar to Katakana words, which can easily be identified in Japanese corpora. We also show the effectiveness of our method by means of experiments.

1 Introduction

Reflecting the rapid growth in science and technology, new words have progressively been created. However, due to the limitation of manual compilation, new words are often out-of-dictionary words and decrease the quality of human language technology, such as natural language processing, information retrieval, machine translation, and speech recognition. To resolve this problem, a number of automatic methods to extract monolingual and bilingual lexicons from corpora have been proposed for various languages.

In this paper, we focus on extracting foreign words (or loanwords) in Korean. Technical terms and proper nouns are often imported from foreign languages and are spelled out (or transliterated) by the Korean alphabet system called *Hangul*. The similar trend can be observable in Japanese and Chinese. In Japanese, foreign words are spelled out by its special phonetic alphabet (or phonogram) called *Katakana*. Thus, foreign words can be extracted from Japanese corpora with a high accuracy, because the Katakana characters are seldom used to describe the conventional Japanese words, excepting proper nouns.

However, extracting foreign words from Korean corpora is more difficult, because in Korean both the conventional and foreign words are written with Hangul characters. This problem remains a challenging issue in computational linguistic research.

It is often the case that specific words have been imported into multiple countries simultaneously, because the source words (or concepts) are usually influential across cultures. Thus, it is feasible that a large number of foreign words in Korean can also be foreign words in Japanese.

In addition, the foreign words in Korean and Japanese corresponding to the same source word are phonetically similar. For example, the English word "system" has been imported into both Japanese and Korean. The romanized words are /`sisutemu`/ and /`siseutem`/ in both countries, respectively.

Motivated by these assumptions, we propose a method to extract foreign words in Korean corpora by means of Japanese. In brief, our method performs as follows. First, foreign words in Japanese are collected, for which Katakana words in corpora and existing lexicons can be used. Second, from Korean corpora the words that are phonetically similar to Katakana words are extracted. Finally, extracted Korean words are compiled in a lexicon with the corresponding Japanese words.

In summary, our method can extract foreign words in Korean and produce a Japanese-Korean bilingual lexicon in a single framework.

2 Methodology

2.1 Overview

Figure 1 exemplifies our extraction method, which produces a Japanese-Korean bilingual lexicon using a Korean corpus and Japanese corpus and/or lexicon. The Japanese and Korean corpora do not have to be parallel or comparable. However, it is desirable that both corpora are associated with the same domain. For the Japanese resource, the corpus and lexicon can alternatively be used or can be used together. Note that compiling Japanese monolingual lexicon is less expensive than that for a bilingual lexicon. In addition, new Katakana words can easily be extracted from a number of on-line resources, such as the World Wide Web. Thus, the use of Japanese lexicons does not decrease the utility of our method.

First, we collect Katakana words from Japanese resources. This can systematically be performed by means of a Japanese character code, such as EUC-JP and SJIS.

Second, we represent the Korean corpus and Japanese Katakana words by the Roman alphabet (i.e., romanization), so that the phonetic similarity can easily be computed. However, we use different romanization methods for Japanese and Korean.

Third, we extract candidates of foreign words from the romanized Korean corpus. An alternative method is to first perform morphological analysis on the corpus, extract candidate words based on morphemes and parts-of-speech, and romanize the extracted words. Our general model does not constrain as to which method should be used in the third step. However, because the accuracy of analysis often decreases for new words to be extracted, we experimentally adopt the former method.

Finally, we compute the phonetic similarity between each combination of the romanized Hangul and Katakana words, and select the combinations whose score is above a predefined threshold. As a result, we can obtain a Japanese-Korean bilingual lexicon consisting of foreign words.

It may be argued that English lexicons or corpora can be used as source information, instead of Japanese resources. However, because not all English words have been imported into Korean, the extraction accuracy will decrease due to extraneous words.

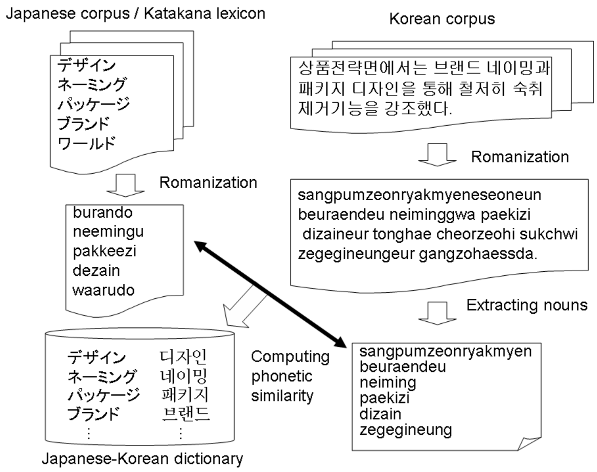


Figure 1: Overview of our extraction method.

2.2 Romanizing Japanese

Because the number of phones consisting of Japanese Katakana characters is limited, we manually produced the correspondence between each phone and its Roman representation. The numbers of Katakana characters and combined phones are 73 and 109, respectively. We also defined a symbol to represent a long vowel. In Japanese, the Hepbern and *Kunrei* systems are commonly used for romanization purposes. We use the Hepburn system, because its representation is similar to that in Korean, compared with the *Kunrei* system.

However, specific Japanese phones, such as /ti/, do not exist in Korean. Thus, to adapt the Hepburn system to Korean, /ti/ and /tu/ are converted to /chi/ and /chu/, respectively.

2.3 Romanizing Korean

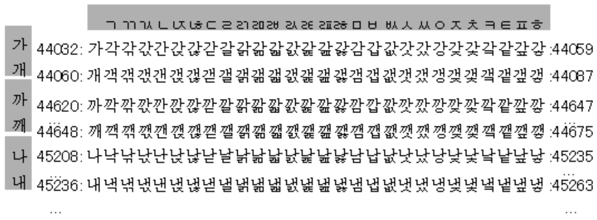
The number of Korean Hangul characters is much greater than that of Japanese Katakana characters. Each Hangul character is a combination of more than one consonant. The pronunciation of each character is determined by its component consonants.

In Korean, there are types of consonant, i.e., the first consonant, vowel, and last consonant. The numbers of these consonants are 19, 21, and 27, respectively. The last consonant is optional. Thus, the number of combined characters is 11,172. However, to transliterate imported words, the official guideline suggests that only seven consonants be used as the last consonant. In EUC-KR, which is a standard coding system for Korean text, 2,350 common characters are coded independent of the pronunciation. Therefore, if we target corpora represented by EUC-KR, each of the 2,350 characters has to be corresponded to its Roman representation.

We use Unicode, in which Hangul characters are sorted according to the pronunciation. Figure 2 depicts a fragment of the Unicode table for Korean, in which each line corresponds to a combination of the first consonant and vowel and each column corresponds to the last consonant. The number of columns is 28, i.e., the number of the last consonants and the case in which the last consonant is not used. From this figure, the following rules can be found:

- the first consonant changes every 21 lines, which corresponds to the number of vowels,
- the vowel changes every line (i.e., 28 characters) and repeats every 21 lines,
- the last consonant changes every column.

Based on these rules, each character and its pronunciation can be identified by the three consonant types. Thus, we manually corresponded only the 68 consonants to Roman alphabets.



```
            ㄱ ㄲ ㄳ ㄴ ㄵ ㄶ ㄷ ㄹ ㄺ ㄻ ㄼ ㄽ ㄾ ㄿ ㅀ ㅁ ㅂ ㅄ ㅅ ㅆ ㅇ ㅈ ㅊ ㅋ ㅌ ㅍ ㅎ
가 44032: 가각갂갃간갅갆갇갈갉갊갋갌갍갎갏감갑값갓갔강갖갗갘같갚갛 :44059
개 44060: 개객갞갟갠갡갢갣갤갥갦갧갨갩갪갫갬갭갮갯갰갱갲갳갴갵갶갷 :44087
까 44620: 까깍깎깏깐깑깒깓깔깕깖깗깘깙깚깛깜깝깞깟깠깡깢깣깤깥깦깧 :44647
깨 44648: 깨깩깪깫깬깭깮깯깰깱깲깳깴깵깶깷깸깹깺깻깼깽깾깿꺀꺁꺂꺃 :44675
나 45208: 나낙낚낛난낝낞낟날낡낢낣낤낥낦낧남납낪낫났낭낮낯낰낱낲낳 :45235
내 45236: 내낵낶낷낸낹낺낻낼낽낾낿냀냁냂냃냄냅냆냇냈냉냊냋냌냍냎냏 :45263
     ...                                                         ...
```

Figure 2: A fragment of the Unicode table for Korean Hangul characters.

We use the official romanization system for Korean, but specific Korean phones are adapted to Japanese. For example, /j/ and /l/ are converted to /z/ and /r/, respectively.

It should be noted that the adaptation is not invertible and thus is needed for both J-to-K and K-to-J directions.

For example, the English word "cheese", which has been imported to both Korean and Japanese as a foreign word, is romanized as /`chiseu`/ in Korean and /`ti:zu`/ in Japanese. Here, /`:`/ is the symbol representing a Japanese long vowel. Using the adaptation, these expressions are converted to /`chizu`/ and /`chi:zu`/, respectively, which look more similar to each other, compared with the original strings.

2.4 Extracting term candidates from Korean corpora

To extract candidates of foreign words from a Korean corpus, we first extract phrases. This can be performed systematically, because Korean sentences are segmented on a phrase-by-phrase basis.

Second, because foreign words are usually nouns, we use hand-crafted rules to remove post-position suffixes (e.g., *Josa*) and extract nouns from phrases.

Third, we discard nouns including the last consonants that are not recommended for transliteration purposes in the official guideline. Although the guideline suggests other rules for transliteration, existing foreign words in Korean are not necessarily regulated by these rules.

Finally, we consult a dictionary to discard existing Korean words, because our purpose is to extract new words. For this purpose, we experimentally use the dictionary for SuperMorph-K morphological analyzer[1], which includes approximately 50,000 Korean words.

2.5 Computing Similarity

Given romanized Japanese and Korean words, we compute the similarity between the two strings and select the pairs associated with the score above a threshold as translations. We use a DP (dynamic programming) matching method to identify the number of differences (i.e., insertion, deletion, and substitution) between two strings, on a alphabet-by-alphabet basis.

In principle, if two strings are associated with a smaller number of differences, the similarity between them becomes greater. For this purpose, a Dice-style coefficient can be used.

However, while the use of consonants in transliteration is usually the same across languages, the use of vowels can vary significantly depending on the language. For example, the English word "system" is romanized as /`sisutemu`/ and /`siseutem`/ in Japanese and Korean, respectively. Thus, the differences in consonants between two strings should be penalized more than the differences in vowels.

In view of the above discussion, we compute the similarity between two romanized words by Equation (1).

$$1 - \frac{2 \cdot (\alpha \cdot dc + dv)}{\alpha \cdot c + v} \qquad (1)$$

Here, dc and dv denote the numbers of differences in consonants and vowels, respectively, and α is a parametric constant used to control the importance of the consonants. We experimentally set $\alpha = 2$. In addition, c and v denote the numbers of all consonants and vowels in the two strings. The similarity ranges from 0 to 1.

3 Experimentation

3.1 Evaluating Extraction Accuracy

We collected 111,166 Katakana words (word types) from multiple Japanese lexicons, most of which were technical term dictionaries.

We used the Korean document set in the NTCIR-3 Cross-lingual Information Retrieval test collection[2]. This document set consists of 66,146 newspaper articles of Korean Economic Daily published in 1994. We randomly selected 50 newspaper articles and used them for our experiment. We asked a graduate student excluding the authors of this paper to identify foreign words in the target text. As a result, 124 foreign word types (205 word tokens) were identified, which were less than we had expected. This was partially due to the fact that newspaper articles generally do not contain a large number of foreign words, compared with technical publications.

We manually classified the extracted words and used only the words that were imported to both Japan and Korea from other languages. We discarded foreign words in Korea imported from Japan, because these words were often spelled out by non-Katakana characters, such as Kanji (Chinese character). A sample of these words includes "*Tokyo* (the capital of Japan)", "*Heisei* (the current Japanese era name)", and "*enko* (personal connection)". In addition, we discarded the foreign proper nouns for which the human subject was not able to identify the source word. As a result, we obtained 67 target word types. Examples of original English words for these words are as follows:

> digital, group, dollar, re-engineering, line, polyester, Asia, service, class, card, computer, brand, liter, hotel.

Thus, our method can potentially be applied to roughly a half of the foreign words in Korean text.

We used the Japanese words to extract plausible foreign words from the target Korean corpus. We first romanized the corpus and extracted nouns by removing post-position suffixes. As a result, we obtained 3,106 words including all the 67 target words. By discarding the words in the dictionary for SuperMorph-K, 958 words including 59 target words were remained.

For each of the remaining 958 words, we computed the similarity between each of the 111,166 Japanese words. For evaluation purposes, we varied a threshold for the similarity and investigated the relation between precision and recall. Recall is the ratio of the number of target foreign words extracted by our method and the total number of target foreign

[1] http://www.omronsoft.com/

[2] http://research.nii.ac.jp/ntcir/index-en.html

words. Precision is the ratio of the number of target foreign words extracted by our method and the total number of words obtained by our method.

Table 1 shows the precision and recall for different methods. While we varied a threshold of a similarity, we also varied the number of Korean words corresponded to a single Katakana word (N). By decreasing the value of the threshold and increasing the number of words extracted, the recall can be improved but the precision decreases. In Table 1, the precision and recall are in an extreme trade-off relation. For example, when the recall was 69.5%, the precision was only 1.2%.

We manually analyzed the words that were not extracted by our method. Out of the 59 target words, 12 compound words consisting of both conventional and foreign words were not extracted. However, our method extracted compound words consisting of only foreign words. In addition, the three words that did not have counterparts in the input Japanese words were not extracted.

Table 1: Precision/Recall for term extraction.

	Threshold for similarity		
	>0.9	>0.7	>0.5
N=1	50.0/8.5	12.7/40.7	4.1/47.5
N=10	50.0/8.5	7.4/47.5	1.2/69.5

3.2 Application-Oriented Evaluation

During the first experiment, we determined a specific threshold value for the similarity between Katakana and Hangul words and selected the pairs whose similarity was above the threshold. As a result, we obtained 667 Korean words, which were used to enhance the dictionary for the SuperMorph-K morphological analyzer.

We performed morphological analysis on the 50 articles used in the first experiment, which included 1,213 sentences and 9,557 word tokens. We also investigated the degree to which the analytical accuracy is improved by means of the additional dictionary. Here, accuracy is the ratio of the number of correct word segmentations and the total segmentations generated by SuperMorph-K. The same human subject as in the first experiment identified the correct word segmentations for the input articles.

First, we focused on the accuracy of segmenting foreign words. The accuracy was improved from 75.8% to 79.8% by means of the additional dictionary. The accuracy for all words was changed from 94.6% to 94.8% by the additional dictionary.

In summary, the additional dictionary was effective for analyzing foreign words and was not associated with side effect for the overall accuracy. At the same time, we concede that we need larger-scale experiments to draw firmer conclusions.

4 Related Work

A number of corpus-based methods to extract bilingual lexicons have been proposed (Smadja et al., 1996). In general, these methods use statistics obtained from a parallel or comparable bilingual corpus and extract word or phrase pairs that are strongly associated with each other. However, our method uses a monolingual Korean corpus and a Japanese lexicon independent of the corpus, which can easily be obtained, compared with parallel or comparable bilingual corpora.

Jeong et al. (1999) and Oh and Choi (2001) independently explored a statistical approach to detect foreign words in Korean text. Although the detection accuracy is reasonably high, these methods require a training corpus in which conventional and foreign words are annotated. Our approach does not require annotated corpora, but the detection accuracy is not high enough as shown in Section 3.1. A combination of both approaches is expected to compensate the drawbacks of each approach.

5 Conclusion

We proposed a method to extract foreign words, such as technical terms and proper nouns, from Korean corpora and produce a Japanese-Korean bilingual dictionary. Specific words, which have been imported into multiple countries, are usually spelled out by special phonetic alphabets, such as Katakana in Japanese and Hangul in Korean.

Because extracting foreign words spelled out by Katakana in Japanese lexicons and corpora can be performed with a high accuracy, our method extracts words in Korean corpora that are phonetically similar to Japanese Katakana words. Our method does not require parallel or comparable bilingual corpora and human annotation for these corpora.

We also performed experiments in which we extracted foreign words from Korean newspaper articles and used the resultant dictionary for morphological analysis. We found that our method did not correctly extract compound Korean words consisting of both conventional and foreign words. Future work includes larger-scale experiments to further investigate the effectiveness of our method.

References

Kil Soon Jeong, Sung Hyon Myaeng, Jae Sung Lee, and Key-Sun Choi. 1999. Automatic identification and back-transliteration of foreign words for information retrieval. *Information Processing & Management*, 35:523–540.

Jong-Hoon Oh and Key sun Choi. 2001. Automatic extraction of transliterated foreign words using hidden markov model. In *Proceedings of ICCPOL-2001*, pages 433–438.

Frank Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1–38.