Korean-to-Chinese Word Translation using Chinese Character Knowledge

Yuanmei Lu lu@nlp.ist.i.kyoto-u.ac.jp
Toshiaki Nakazawa nakazawa@pa.jst.jp
Japan Science and Technology Agency, Tokyo, 102-8666, Japan

Sadao Kurohashi kuro@nlp.ist.i.kyoto-u.ac.jp
Graduate School of Informatics, Kyoto University, Kyoto, 606-8501, Japan

Abstract

In this paper, we present a way of translating Korean words to Chinese using Chinese character information. A mapping table of Korean and Chinese characters is constructed and used to obtain possible combinations as translation candidates. The candidates are ranked by the *combination score* which accounts for the possibility of the character combination and *context similarity score*, which indicates contextual information among words. Parallel resources like Wikipedia aligned data or Wiktionary data are used for preliminary translation and also used during ranking candidates.

1 Introduction

The quality of the statistical machine translation heavily correlates with the amount of parallel corpora used, and improving the lexical coverage of the parallel corpora plays an important role in reducing the number of out-of-vocabulary (OOV) words. However, the vocabularies of languages keep growing over time, especially in the case of technical terms. It is impossible to cover all the newly appearing words by augmenting the parallel corpora and therefore we need to prepare bilingual dictionaries for the new words, or translate them separately, for example, using the transliteration technique.

There are some parallel dictionaries available for limited language pairs and limited domains. In addition, we can extract parallel resources from Wikipedia. It offers hyper-linked pages of the same topic in different languages, and the title pairs of the linked pages can be used as a parallel dictionary. However, the coverage is not sufficient for both cases especially for technical terms. In some cases, titles aligned by hyper-linked pages can not be seen as exact translations of each other since many of them may just be related to similar concepts. Another limitation in this domain is that although there may exist enough resources between English and other languages, there are very few resources between the non-English languages, such as Korean, Chinese and Japanese.

Korean, Chinese and Japanese use Chinese characters. In Japan, they use Kanji, which originated from China, while Korean uses Sino-Korean vocabularies, in which characters (Hangul) can be converted to corresponding Chinese characters (Hanzi). Just as Japanese writing includes both Kana and Kanji, Korean contains two kinds of characters–Hangul and Hanja. Hangul are Korean characters that are most commonly used and seen, and many Sino-Korean words have their Hanja writings–Hanja. For the sake of clarity refer to Table 1 which contains a few examples of Hangul and Hanja.

Hangul (Korean)	애정	노동
Hanja (Korean)	愛情	勞動
Kanji (Japanese)	愛情	労働
Hanzi (Chinese)	爱情	劳动

Table 1: Difference of characters written in Hangul, Hanja, Kanji and Hanzi

In reality, Hanja writing is seldom used in modern Korean language. It is used in several technical writings to emphasize the definition of certain words. It is not hard to see that in many cases, the same Sino-Korean word may written in different Hanja that represent various meanings according to the context. Thus, the Hanja play important roles in word sense disambiguation (WSD), especially when it comes to terminology words.

Even though the forms are different, most of the vocabularies in these three languages (Korean, Chinese and Japanese) have one-to-one correspondence with respect to the characters. Using this characteristic, we can perform word translation and use the result to construct a terminological or scientific dictionary, which can further be used in machine translation systems for the above mentioned languages. However, there is neither a publicly available Hangul-Hanzi-Kanji mapping table, nor is there an official parallel dictionary between Sino-Korean and Chinese/Japanese (there exists an online dictionary for common words, but not for terminological words) on the web.

In this paper, we propose a method of translating Korean words (mainly terminological words) into Chinese and Japanese using the Chinese character knowledge. The input is Korean sentences, and our objective is the translation of terminological words in these sentences. We treat nouns as potential terminological words. A morphological analyzer is employed for extracting nouns, as the pre-processing of our model. We also construct a Hangul-to-Hanzi mapping table and use it to generate translation candidates, since Hangul and Hanzi have a one-to-one correspondence. Then we rank the candidates using the character combination and contextual similarity scores. We also apply a machine learning method for interpolating the two aforementioned scores of candidates.

2 Sino-Korean Words

2.1 Chinese Words in Korean and Japanese

In Asian languages like Korean and Japanese, majority of words are borrowed or adopted from other languages, especially when considering terminological words (Matsuda et al. (2008)). Apart from borrowing/adopting words phonetically like transliteration, a majority of words in these languages are originally borrowed from ancient Chinese. An example is the adoption of Kanji words (漢字語) of Japanese, or Sino-Korean words of Korean.

According to *Studies on the Vocabulary of Modern Newspapers* Ⅲ published by *National Institute for Japanese Language and Linguistics*, Kanji (Chinese characters used in Japanese) accounts for over 70% of the readable content in newspapers (National Institute for Japanese Language and Linguistics (1972)). The Kanji words such as "使用 (use) " are preferably used than the native Japanese words such as "使う (to use) " in Japanese formal writings.

The situation is similar in Korean wherein a significant portion of the words are composed of Sino-Korean. The *Standardized Korean Language Dictionary* by *National Institute of the Korean Language (NIKL)* published in 2004 had 57% of its content as Sino-Korean words (鄭虎聲 (2000)); the *Survey of Korean Vocabulary frequency*, which was conducted in 1956, has shown that about 70% of the frequently used words are Sino-Korean (文教部 (1956)). Nowadays, the percentage of Sino-Korean words of spoken language being generally used is grad-

ually decreasing but these kinds of words are still frequently used in formal writings, such as newspapers and dissertations. Most of the Sino-Korean words are not written in Hanja directly, but in Hangul. However, we can convert them into Hanja and further to Hanzi because there is a correspondence among them. Actually, some papers are published with combinations of Hangul and Hanzi in order to specify definitions of vocabularies or emphasize them.

2.2 Related Work

There are some previous work on character conversion, both within language or between two languages (Chen and Lee (2000), Huang et al. (2004)). During character conversion, a bilingual dictionary is needed for candidate selection. For a low resource language like Korean, a bilingual dictionary containing enough data, including polysemous words, is often hard to obtain. Moreover, unlike sentence translation, character translation often ignores the context information of the input source sentence.

Chinese character knowledge is widely used in cross-language information retrieval (Hasan and Matsumoto (2000)), or translation of names of people (Wang et al. (2007, 2008, 2009)). During translation, they select named entities by removing the postpositions or the endings, by applying the maximum matching algorithm. For Sino-Korean words that are written using same Hangul word but expressing different meanings according to various context environments (ambiguous words), they adopt some mutual information score to evaluate the co-relation between the query term and the candidates. In languages such as Japanese and Korean, there is more ambiguity about where word boundaries should be, wherein some particles may also be a part of a noun, or a verb, their method of extracting target words are often not efficient.

Moreover, unlike information retrieval, the machine translation method also has to ensure the meaning of the sentence in order to be fluent. In other words, we should also consider context features of the sentences.

Since Korean characters are phonograms, we can find corresponding Hangul characters for given Hanzi characters. Actually, almost all of the Hanzi can be converted to one (or in some rare cases, many) Korean characters. Huang et al. constructed a Chinese-Korean Character Transfer Table (CKCT Table) to reflect the correspondence between Hanzi and Hangul (Huang and Choi (2000)). It is reported that the table contains 436 Hangul with corresponding 6763 Hanzi. Practically, there are 4888 common Hanja used in Korean (KATS (Korean Agency for Technology and Standards) (1997), Hanyang Systems (1992)). Moreover, the number of daily-used Hanzi in Korea numbers around 1800[1], while 3500 Hanzi are required to learn for practical Chinese character level test[2]. Obviously, most of the Hanzi in their table cannot be considered as practical ones. After all, the table is non-public to ordinary users.

3 Proposed Method

Figure 1 gives an overview of our Korean-to-Chinese terminological word translation methodology using Chinese character knowledge. The determined translation result is marked in a darker color. Since our objective is to perform terminological word translation, and in scientific documents, terminological words are often Sino-Korean words, mostly nouns, we focus on the translation of Korean nouns to Chinese.

Given a Korean sentence, we first need to extract the nouns for which we use morphological analyzers to extract Korean nouns (Section 3.1). In the example in Figure 1, after morphological analysis, Korean words, 화합물, 구조, 결정 are extracted from the input sentence. We

[1] Wikipedia:상용한자

[2] Korea Foreign Language Evaluation Institute
http://www.pelt.or.kr/cs/10/main/main.aspx

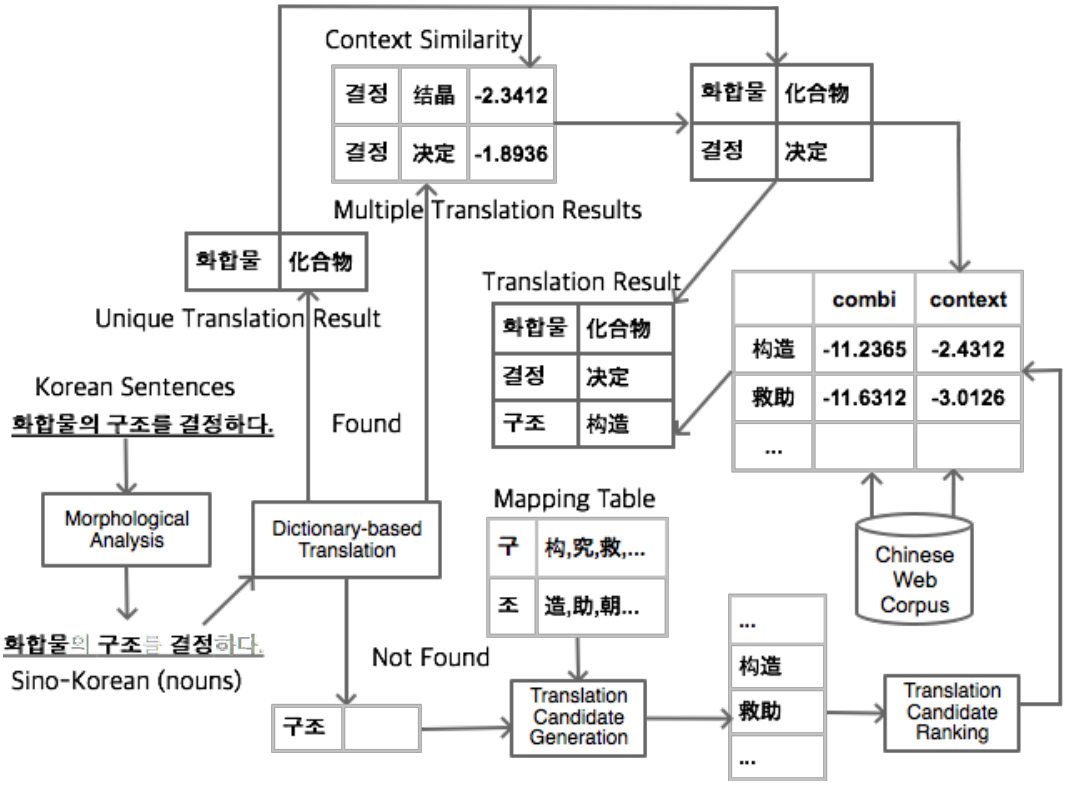

Figure 1: Korean-to-Chinese Word Translation System

then look up the Chinese translations of the Korean words in a Korean-Chinese parallel dictionary (Section 3.2). There are three different cases for a given word: unique translation (only 1 candidate per word), multiple translations (more than 1 candidate per word) and untranslatable. The words cannot be translated by the parallel dictionary are further processed wherein possible Chinese character combinations are generated as the translation candidates, using the Hangul-Hanzi mapping table constructed in Section 3.3. The candidates are ranked by two scores: *combination score* and *context similarity score* (Section 3.4). The *combination score* accounts for the possibility of the Chinese character sequences and is calculated using the Chinese web corpus, whereas the *context similarity score* considers the context feature of the input sentences and that of the sentences in the Chinese web corpus. Finally, we interpolate these two scores to handle the score of each translation candidate. The candidate with the highest score is considered as the final translation result. During the whole translation procedure, some data selection methods are involved.

3.1 Extracting Nouns

In Korean sentences, words are separated by spaces among them wherein each word may contain one or more morphological elements. On the other hand, a functional word may have different morphologies under different conditions Li et al. (2013). For example, in Korean sentences:

(1) 학교 에 가다
school DAT go
'Go to school.'

(2) 비 가 내리다
rain NOM fall
'It rains.'

In the first sentence, the word "학교에" is composed of "학교 (noun)" and "에 (particle)", while in the two sentences, "가" may act as either a verb (as in 가다) or a particle (as in 비가). To get terminologies from a sentence, some methods of extraction, which involves processes like tokenization and POS tagging should be implemented. However, there is no freely available tool to accomplish this task of extracting Korean terminological words. Instead, since most of the terminological words are nouns, we can use a Korean morphological analyzer to extract nouns from the given Korean sentences and treat them as terminological words.

For the given Korean sentences that contain Sino-Korean vocabularies, we extract nouns from these sentences with the help of a morphological analyzer, during which also performs word tokenization. After comparing several POS tag sets (KKMA (2011)) that mainly used for Korean language, we treat NNP and NNG as POS type for terminological words.

3.2 Translation by Dictionary Matching

Some of the Korean words are translated into Chinese with a parallel dictionary as an initial step. We use a dictionary of Wikipedia and Wiktionary aligned data to achieve this.

As a multilingual online encyclopedia, Wikipedia titles can be used as parallel data for many languages. In our model, we use the aligned Wikipedia title pairs of Chinese and Korean and Wiktionary data as a parallel dictionary. In addition, for aligned Wikipedia titles, we apply the following processes to improve the quality and coverage of the parallel dictionary.

- Make full use of redirect pages of each page, and validate the correctness using the first sentence (definition sentence) of each page to augment the parallel dictionary. Definition sentences of some pages containing more than one key words are analyzed to determine whether these key words are synonyms, by checking whether the definition sentence contains the word "또는" and "혹은", which means "or" between the key words. These synonyms have the same Chinese translation result.

- Convert Chinese characters of traditional Chinese into simplified Chinese for normalization. This step is done because translation result in dictionary matching step will be used as context feature of each of the character combination candidates (simplified Chinese characters).

On the other hand, we collect Sino-Korean words from Korean Wiktionary. Some of the Wiktionary pages have marked Chinese information which serves to indicate that the Chinese content is a translation of the title (語源: 漢字).

During the preliminary translation, some words may have more than one translation result (Chinese alignment). Some Wikipedia titles may contain brackets within them, the brackets often contain information for disambiguation. For example, consider a Korean word 전자. The word is contained in two titles 전자 and 전자 (언어학), having aligned Chinese "电子" (electricity) and "转字" (transliteration), respectively. We combine titles of shared Korean words (without brackets) as translation candidates. The example above, 전자 has two aligned Chinese, 电子 and 转字 translations. Korean words that have more than one Chinese translations

한	闲, 韩, 恨, 限, 汉
자	姐, 字, 磁, 子, 仔, 姿, 刺, 自, 资, 瓷

Table 2: A portion of the mapping table

are considered ambiguous. In the Korean sentences we work with, we have Korean words belonging to both the above mentioned types. In the case of Korean words with unique Chinese translations, we do not need any further processing. However, in the case of multiple candidates, we compare the context vectors of the candidates calculated with the formerly translated results and determine the most appropriate candidate as the translation result. The Korean nouns which cannot be translated with the parallel dictionary are subjected to further processing.

3.3 Generating Translation Candidates

The purpose of this step is to generate possible translation candidates by combining Hanzi characters converted from the Hangul characters, using the Hangul-Hanzi mapping table. For instance, using the mapping table in Table 2, we can generate the translation candidates for the Korean word 한자 (*Chinese character*, 汉字) and the possible character combinations could be:

한자 (*Chinese character*, 汉字): 闲姐, 韩姐, ...汉字, 汉子....

Whether these combinations have the proper meaning or not is still unknown. Most of them may not have practical meanings. So we need to select the most appropriate combination.

3.4 Rank the Translation Candidates

In this step, we utilize the segmented Chinese web corpus as a filter. First of all, we check the existence with the corpus and unify the POS type of input Korean words and their Hanzi combinations. According to Xia (2000), in Chinese, there are three POS tags for nouns: NR (Proper Noun), NT (Temporal Noun) and NN (Other Noun). Thus, according to these definitions, we extract all NN and NR type nouns from the segmented Chinese web corpus that contain POS information (see Shen et al. (2013)) as a Chinese noun corpus. After that, we generate all possible combinations and check whether they are included in this corpus. This step plays a significant role for reducing the cost of our model. In order to determine the most appropriate translations among the selected combinations, we utilize *combination score* and *context similarity score*.

3.4.1 Combination Score

Combination score S_{combi} measures the strength of the link between the characters. Here, we utilize a language model to get the score. The language model returns the possibility (log score) of each combination according to the Chinese web corpus (see as equation (1)). We acquire the score of the original character sequence of the candidate and also the score for the reverse sequence. Equation (2) indicates the way the language model computes the score for the reverse sequence. Suppose that the combination consists of n characters, $c_{1..n}$

$$S(c_{1..n}) = log(\prod_{i=1}^{n-1} P(c_i|c_{1..i-1})) \tag{1}$$

$$S(c_{n..1}) = log(\prod_{i=1}^{n-1} P(c_i|c_{i+1..n})) \tag{2}$$

For example, the scores for “汉字语” and “语字汉”may be calculated as,

$$S(\text{汉字语}) = log(P(\text{汉}) \times P(\text{字}\ |\text{汉}) \times P(\text{语}|\text{汉字})),$$

$$S(\text{语字汉}) = log(P(\text{语}) \times P(\text{字}\ |\text{语}) \times P(\text{汉}|\text{语字}))$$

We combined these two scores by simply adding them. As the length of the word increases, the score would decrease drastically. We divide the sum with *length-1* for normalization. The following formula defines the combination score. For words with many characters, the score will be much smaller after each multiplication and thus we divide the score with the length of the word. *n* here indicates the number of characters the word contains.

$$S_{combi}(c_{1 \cdot\cdot n}) = \frac{S(c_{1..n}) + S(c_{n..1})}{n-1}, \quad (n \geq 2)$$

When n=1 (words with single character), we utilize unigram score as their combination scores. For example, the combination score for "汉字语" may be calculated as

$$S_{combi}(\text{汉字语}) = \frac{\begin{array}{r} log(P(\text{汉}) \times P(\text{字}\ |\text{汉}) \times P(\text{语}|\text{汉字})) \\ +\ \ log(P(\text{语}) \times P(\text{字}\ |\text{语}) \times P(\text{汉}|\text{语字})) \end{array}}{3-1}$$

Since the potential number of combinations may run into tens of thousands we conduct a selection again before moving on to the next step. We sort combinations according to their scores, and keep only the combinations whose scores are not lower than the highest *score*-2. For example, suppose S_{combi}(汉字)= -5.7126 is the highest one among the combinations, we remove combinations whose combination score is lower than -7.7126. If we still have many candidates, we keep top 10 candidates among them.

3.4.2 Context Similarity Score

Now that we have obtained candidates using the most possible combinations. To acquire the most proper ones for the given sentences, we consider context features. For each combination, context vector is constructed using the Chinese corpus. We use sentences which contain each combination as the context window, so the element of the vector is the co-occurrence Chinese characters (for character-based context vector) and frequency of them. We ignore 125 stop words (characters) such as 的 and 了 [3]. Some combinations that have higher frequency often contain bigger and wider range of count values in their context vectors. For words with this situation, co-occurrence characters with lower occurrence frequency are less effective as context information. Moreover, considering all of the co-occurrence characters is impractical since it will only increases complexity of time and space. We set the threshold as 100, that is to say, we ignore combinations that occur less than 100 times. Of course, there may also be some combinations such that all of their co-occurrence characters appear less than 100 times. For these candidates, we keep all these low-frequency characters as context information.

We also construct another context vector with the information of the input Korean sentence. We use the formerly dictionary translated Korean words (Section 3.2).

The context similarity score $S_{context}$ is defined as the *cosine similarity* of the two context vectors–one created using the Chinese corpus that contains Chinese sentences and the other using the translation results using the dictionary.

3.4.3 Interpolation

The combination score is useful to examine whether the combination is appropriate or not, and the context similarity score is helpful for selecting the the most appropriate one according to the context features where two or more combinations have practical meanings. Therefore, we

[3] https://code.google.com/p/verymatch/downloads/detail?name=stopwords.txt

Hanja	Meaning in Korean	Hangul
强	강할	강
界	지경	계
計	셀	계

Table 3: The Korean-Chinese aligned resource

interpolate two scores obtained in the former two sub chapters and calculate the score of each translation candidate $S(cand)$ as follows:

$$S(cand) = \alpha S_{combi} + (1 - \alpha) S_{context}$$

The specified value of α ($0 \leq \alpha \leq 1$) is determined using a method described in next chapter. The character combination with the highest score is regarded as the final translation result. If there are more than 2 words that cannot be translated by Wikipedia or Wiktionary, we process the procedure in sequence (from beginning to end of the sentence) and use the translation result as context feature to the remaining unknown words.

4 Experiment

4.1 Settings

Chu et al. had produced a *Chinese character mapping table for Japanese (Kanji), Traditional Chinese (TC) and Simplified Chinese (SC)* Chu et al. (2012). Thus, for constructing a table that contains mapping relationship between Korean Hangul and Chinese Hanzi, we need to construct rather Kanji-Hangul or Hanzi-Hangul tables and merge them. We collected Hangul-Hanja mapping information from the web.

- We acquired 1365 Hanja characters with their aligned Hangul from a freely accessible webpage[4]. These characters are contained by words whose frequency is higher than 5965 times in some Sino-Korean corpus (there is no specific information about the mentioned Sino-Korean corpus).
- There are also some materials that contain Hanja characters that are used in practice. We collected 3500 Chinese characters that are required for Hanja level tests PELT (2005), as mentioned in Section 2.2, to get most generally used Hanja characters with their aligned Hangul.
- As we mentioned previously, more data is needed to guarantee the coverage of the generally used Hanja characters. In other to collect as much data as possible about Hanja and their Hangul alignment, we crawled additional content from some Wikia pages[5], which contain more than 20000 Hangul-Hanja correspondence. Of course most of them are not used in general.

We merged these data to get a Hangul-to-Hanja alignment, and combined this table with the one created by Chu et al. (2012). This combination is possible because most Hanja are traditional Chinese characters, and have one-to-one correspondences with Hangul. Hanja-Hangul tables mentioned above may contain a large amount of unused characters, which can affect time efficiency and the correctness of candidates. We checked the compatibility of the table with some Hanja dictionaries obtained from the web[6,7]. There are also Hanja characters that cannot

[4] http://korean.nomaki.jp/site_j/kanji16.html
[5] 사용자:Masoris/hani converter.js - 한자위키- Wikia
[6] http://hanja.naver.com (네이버한자사전)
[7] http://small.dic.daum.net/index.do?dic=hanja (Daum한자사전)

testset	1	2	3	4	5
α	0.44	0.44	0.20	0.44	0.44
Precision(%)	93.75 (15 / 16)	100.00 (5 / 5)	72.22 (13 / 18)	90.00 (9 / 10)	92.31 (12 / 13)

Table 4: The α-Precision relation for each test set

be converted to Hanzi and so we ignore them. This way, we can create a Hangul-Hanja-Kanji-Hanzi mapping table.

4.1.1 Experimental data

For our experiments, we obtained 100 Korean sentences from several technical documents (on natural science, such as biology, chemistry, physics, earth science) from the web, which contain 3281 words in total[8] (1014 nouns, as translation object).

Totally, we have 5368 Hanzi characters map to 481 Hangul characters in our Hangul-Hanzi mapping table. The size of the aligned title dictionary is around 7.1MB (Wikipedia 6.6MB, Wiktionary 495KB) and that of the Chinese web corpus that contains ordinary Chinese sentences for calculating the context score and combination score of each translation candidate, is 47GB, among them are 184MB of NN, NR words. For querying the web corpus, we used the KenLM Heafield (2011) language model with smoothing technique included, ignoring start and end symbols and utilizes a character based process.

We used Google Translate as one of the baselines and compared the result with reference. We checked the precision with exact-match mechanism and obtained a precision of 38.17% [9].

4.2 Result

We manually set the reference translation results for the test data, and calculated precision for evaluation. To obtain the best value for α during interpolation, we tried 5-fold cross validation. We divided the test set into 5 parts and recursively selected four of them to get the α that gives the best score and used it on the remaining part to test the performance of the translation (see Table 4). We calculated precision of each test set for evaluation, and the best-performing α is the one with which we can get the highest precision. Figure 2 demonstrates the variation of precision with the change of α, for a randomly selected test set, testset 4.

We obtained an α for each test set, took the mean value and used it on the test sets again. The translation results are as shown in Table 5. (a) shows when we do not consider ambiguity of words in the dictionary, while (b) shows the results obtained when we determine ambiguous words with currently determined translation results (Chinese). The last columns of the two tables shows the result obtained with mean value of α, (α=0.18 in (a) and α=0.39 in (b)). Using the mean value of α, the precision dropped. (c) demonstrates the results when considering the scores separately as well as together. The column "+Combi" and "+Context" shows the translation result by only considering combination scores and context features, separately, and "+Combi +Context" shows result by considering both combination scores and context features.

In another experiment, we utilized a machine learning algorithm that utilizes these features in order to get a better translation result.

We employed SVM rank, Support Vector Machine for Ranking. Taking a Support Vector approach, the resulting training problem is tractable even for large numbers of queries and large numbers of features Joachims (2002, 2006).

[8] words are separated by space in the sentence

[9] https://docs.google.com/spreadsheets/d/1hFiIjaZ7SiSzIRp8Nyk8WRW68D3yRGipD4zeh2D4G4k/edit#gid=0

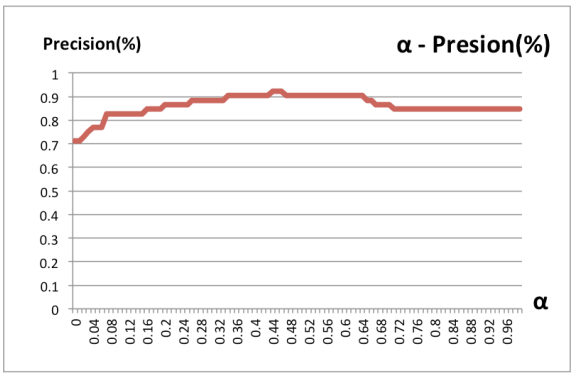

Figure 2: Relationship between α and precision of testset 4

We used combination score of normal sequence and reversed one separately as features, with the third one as context similarity score. For the two combination scores, in order to pick candidates to process, we implemented a selection mechanism. We first added two scores, then sorted the sum and got the candidates among top 10 scores. We queried combinations scores of these candidates respectively and utilized as value of these two features.

On the other hand, the value of the third feature–context score, was calculated with formerly determined translation results in the given sentence. We again divided the sentences into 5 parts, and performed a 5 fold cross validation. As a result, the precision is 93.55%. This method outperforms the former one, is fast and has a low resource requirement.

4.3 Discussions

Among the 1014 words that is analyzed as nouns by the morphological analyzer, 59 are original Korean words, which are neither adopted words nor terminological words, 46 are transliterated words. Analysis of mistranslations is presented in Table 6.

When analyzing the experimental results, we found that among the 13 words that could not be translated, 11 of them are contain phonetic parts within them, such as 메카니즘 (mechanism), 시스템 (system). These words are untranslatable since they could neither be looked up in the dictionary (Wikipedia aligned data and Wiktionary data) nor did they have Hanzi mapping for the phonetic characters.

There is a compound Korean word "서양음악사학 (西洋音乐史学)", consisting of three words, "서양 (西洋)", "음악 (音乐)", "사학 (史学)". During the generation of Hanzi combination, the generated combination could not be found in the web corpus. We did use a corpus that contains compound words (nouns), and found that it contains "西洋音乐" and "音乐史学", but no 西洋音乐史学 in it. For compound words that are space separated and do not have translation candidates in the dictionary, we checked the existence of their components in the dictionary and obtained candidates by combining the translations of the individual components. However, for compound words that contain more than 4 characters and do not have a space among them, the number of character combination increases drastically as n (length of the word) increases, which leads to an increase in space and time complexity. Thus, for words in this case, preprocessing during the morphological analysis phase gives better results. For example, for long words (containing more than 4 characters and do not have space among them), we can implement a split-and-analyze step before moving on to the translation step. We attempt to do further separation by adding spaces among characters and run the morphological analyzer again, until two divided parts are both analyzed as nouns. In this method, we successfully segmented "서양음악사학" into "서양" and "음악사학", and confirmed that they can be translated by the

	Dictionary	+Combination +Context (α =0.18)
Correct	646	43
InCorrect	234	19
NoTranstion	75	13
Precision(%)	73.41 (646/880)	69.35 (43/62)

(a) Experimental results for not having consider ambiguous words in dictionary

	Dictionary		+Combination +Context (α =0.39)
	Sole	Multi	
Correct	549	190	50
InCorrect	35	106	12
NoTranstion	371	75	13
Precision(%)	94.01 (549/584)	64.19 (190/296)	80.65 (50/62)

(b) Experimental results when considering ambiguous words in dictionary

	+ Combi	+Context	+Context + Combi
Correct	44	47	50
InCorrect	18	15	12
NoTranstion	13	13	13
Precision(%)	70.97 (44/62)	75.80 (47/62)	80.65 (50/62)

(c) Experimental results when considering ambiguous words in dictionary (for each score)

Table 5: Experimental results

Wiki-dictionary (Wikipedia and Wiktionary).

In our experiments, instead of considering all combination candidates (obtained using the web corpus), we used several selection rules like: sort the candidates according to combination scores in the web corpus and select candidates to whose score differs by two points from the candidate with the highest combination score. From the candidates, we selected up to 10 candidates that have higher scores. We also utilize the segmented Chinese web corpus as a filter. We check the existence with the corpus and unify the POS type of input Korean words and their Hanzi combinations. However, this kind of filtering sometimes excludes "useful" candidates. For example, 유사점 is another word that could not be translated. All of the character combinations of it, including correct translation 类似点 is not contained in the web corpus, thus the word did not have any translation candidate. There are also other words, such as 영문자 and 상점가, which have candidates in the web corpus but correct translation is filtered during the procedure. It is mainly caused by the segmentation error of the web corpus.

Table 7 shows some good and bad example of translation results when using combination and context scores. The words being discussed are under-lined and the correct translations are marked with "*". The table contains combination scores, context scores and interpolated candidate scores. Determined candidates by each score are marked in bold.

InCorrect trans	24 failed to find correct translation result in wik data	
	correct candidate has lower score	82 words in Wik dictionary
		12 words consider combi & context sim

No trans result	2 have no combination as nouns in web corpus (NN,NR)
	11 are transliterated words

Table 6: Error analysis of mistranslations

Good example:

용액으로부터 이온이 석출되는 경우에도 비슷한 현상이 일어나며, 이때는 결함 자리에 도착한 이온이 그 주위의 여러이온들과 정전기적 상호작용을하게된다.

Korean	Candi	Combi	Context	Context+Combi
현상	现象*	-10.5459	**-2.2506**	**-5.4858**
	悬赏	**-10.3712**	-3.4574	-6.1538

Bad example:

이 분광법은 분해력과 감도가 높아서 가벼운 원소에 대해서도 높은 감도를 나타낸다.

Korean	Candi	Combi	Context	Context+Combi
감도	感度*	-12.1544	**-2.0072**	-5.9646
	感到	**-10.9311**	-2.0130	**-5.4910**

Table 7: Examples of good and bad translations.
(* indicates correct translation result)

5 Conclusions

In this paper, we present an automatic Korean-to-Chinese terminological translation mechanism that uses Chinese character knowledge. The ultimate goal of this work is to use the translation result for constructing useful resources in machine translation between Korean and Chinese. A morphology analyzer was used to extract nouns as Sino-Korean words. We used aligned Wikipedia title data and Hangul-Hanja information in Wiktionary to obtain reference translations. Some polysemous words may have more than one Chinese translation in the dictionary. To select the most proper one, we compared their context features within the sentence. In order to rank candidates, we both considered context information and probabilities of occurrence in the web corpus. We carried out character-based translation, for which segmented Chinese web corpus is used to create a character-based and word-based context vector for each translation candidate. Some incorrect reference translation results (insufficiency in alignment data or failed to be selected correctly with context similarity) caused some incorrect translation results. Moreover, some candidates which are correct translations were excluded from consideration since have too lower combination score among the candidate set.

In the future, we intend to determine candidates that are less ambiguous and easy to determine during candidate selection, to guarantee more contextual information. Moreover, we want to evaluate the proposed approaches on larger test data.

References

Chen, Z. and Lee, K.-F. (2000). A New Statistical Approach to Chinese Pinyin Input. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, pages 241–247, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chu, C., Nakazawa, T., and Kurohashi, S. (2012). Chinese Characters Mapping Table of Japanese, Traditional Chinese and Simplified Chinese. In *Proceedings of the Eighth Conference on International Language Resources and Evaluation (LREC 2012)*, pages 2149–2152, Istanbul, Turkey.

Hanyang Systems (1992). KS X 1001:1992 (CJKV Information Processing, Appendix L). `http://examples.oreilly.com/cjkvinfo/AppL/ksx1001.pdf`.

Hasan, M. M. and Matsumoto, Y. (2000). Chinese-Japanese cross language information retrieval: a Han character based approach. In *Proceedings of the ACL-2000 workshop on Word senses and multi-linguality*, pages 19–26. Association for Computational Linguistics.

Heafield, K. (2011). KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom.

Huang, J.-X., Bae, S.-M., and Choi, K.-S. (2004). A Statistical Model for Hangeul-Hanja Conversion in Terminology Domain. Association for Computational Linguistics.

Huang, J.-X. and Choi, K.-S. (2000). Chinese-Korean Word Alignment Based on Linguistic Comparison. In *ACL*.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM.

Joachims, T. (2006). Training linear SVMs in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 217–226. ACM.

KATS (Korean Agency for Technology and Standards) (1997). KS X 1001 Hanja. `http://jinjucci.korcham.net/cms/board/Download.jsp?fileId=IUAjJDU3MDI2LS0kJA==`.

KKMA (2011). Korean Part Of Speech tagging table. `http://kkma.snu.ac.kr/documents/?doc=postag`.

Li, S., Wong, D. F., and Chao, L. S. (2013). Experiments with POS-based restructuring and alignment-based re-ordering for statistical machine translation. *ACL 2013*, page 82.

Matsuda, M., Takahashi, T., Goto, H., Hayase, Y., Nagano, R. L., and Mikami, Y. (2008). Technical Terminology in Asian Languages: Different Approaches to Adopting Engineering Terms. In *IJCNLP*, pages 25–32.

National Institute for Japanese Language and Linguistics (1972). Studies on the Vocabulary of Modern Newspapers Vol.III.

PELT (2005). Korea Foreign Language Evaluation Institute. `http://www.hanja.com/content/test01.php`.

Shen, M., Kawahara, D., and Kurohashi, S. (2013). Chinese word segmentation by mining maximized substrings. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 171–179.

Wang, Y.-C., Lee, Y.-H., Lin, C.-C., Tsai, R. T.-H., and Hsu, W.-L. (2007). Korean-Chinese Person Name Translation for Cross Language Information Retrieval.

Wang, Y.-C., Tsai, R. T.-H., and Hsu, W.-L. (2008). Learning Patterns from the Web to Translate Named Entities for Cross Language Information Retrieval. In *IJCNLP*, pages 281–288.

Wang, Y.-C., Tsai, R. T.-H., and Hsu, W.-L. (2009). Web-based pattern learning for named entity translation in Korean–Chinese cross-language information retrieval. *Expert Systems With Applications*, 2(36):3990–3995.

Xia, F. (2000). The part-of-speech tagging guidelines for the Penn Chinese Treebank (3.0).

文教部 (1956). Survey of Korean Vocabulary frequency.

鄭虎聲 (2000). Statistical analysis of vocabularies in Standardized Korean Language Dictionary.