Korean-to-Chinese Word Translation using Chinese Character Knowledge

Yuanmei Lu[*], Toshiaki Nakazawa[*,] and Sadao Kurohashi[*]**

[*]Graduate School of Informatics, Kyoto University

[**]Japan Science and Technology Agency

{*lu, kuro*}*@nlp.ist.i.kyoto-u.ac.jp, nakazawa@pa.jst.jp*

1 Introduction

The quality of the statistical machine translation highly relies on the amount of parallel corpora available, and improving the lexical coverage of the parallel corpora seems to play an important role in reducing the number of out-of-vocabulary (OOV) words. However, the number of vocabularies of languages keeps growing, especially for technical terms. It is impossible to cover all the newly appeared words by augmenting the parallel corpora; 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. Although there may exist enough resources between English and the other language, there is less resources between two non-English languages, such as Korean, Chinese and Japanese.

As in the same linguistic area, Korean, Chinese and Japanese have much in common in their languages. One of the aspects is that these languages use Chinese characters. In Japan, they use Kanji, which is originated from Chinese, and Korean use Sino-Korean vocabularies, in which characters (Hangul) can be converted to corresponding Chinese characters (Hanzi). Even though the forms are different, most of the vocabularies in these three languages have one-to-one correspondence in character. In this paper, we propose a method of translating Korean words into Chinese using the Chinese character knowledge. We use the Hangul-to-Hanzi mapping table to generate translation candidates and rank the candidates considering the possibility of the character combination and contextual similarity.

2 Sino-Korean Words

2.1 Chinese Words in Korean and Japanese

In Korean, a significant portion of the words are composed of Sino-Korean. The *Standardized Korean Language Dictionary* (표준국어대사전, 標準国語大辞典) published by National Institute of the Korean Language (NIKL) (국립국어원, 國立國語院) in 2004 contains near 57% of Sino-Korean words (漢字語); the *Survey of Korean Vocabulary frequency*, which is conducted in 1956, has shown that about 70% of the frequently used words are Sino-Korean. Nowadays, this kind of words are much frequently used in writings, such as newspapers and dissertations. Most of the Sino-Korean words are not written in Hanzi, but in Hangul. However, we can convert them into Hanzi because there is a correspondence between them. Actually, some papers are published with combinations of Hangul and Hanzi in order to specify definitions of vocabularies or emphasize them.

The situation is similar in Japanese. According to *Studies on the Vocabulary of Modern Newspapers* Ⅲ published by National Institute for Japanese Language and Linguistics (国立国語研究所), the percentage of Kanji (Chinese characters used in Japanese) words are over 70% in newspapers. 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.

2.2 Related Work

Since Korean characters are phonogram, we can find a corresponding Hangul for a given Hanzi. Actually, almost all of the Hanzi can be converted

to one (or scarcely several) Korean characters. (Huang,et,al. 2000) constructed a Chinese-Korean Character Transfer Table (CKCT Table) to reflect the correspondence between Hanzi and Hangul. The table contains 436 Hangul with corresponding 6763 Hanzi [1]. The number of daily-used Hanzi in Korea is known as only 1800[*1], and 3500 Hanzi are required to learn for practical Chinese character level test[*2]. Obviously, many of the Hanzi in their table cannot be considered as practical ones.

In the following sections, we introduce the proposed method of our system in detail, and further demonstrate the result of conducted experiment .

3 Proposed Method

Figure 1 shows the overview of our Korean-to-Chinese word translation system. In this study, we only focus on the translation of Korean nouns because the large number of technical terms are nouns.

Given a Korean sentence, we first apply morphological analyses to extract Korean nouns. Then we look up the Chinese translations of the Korean words in a Korean-Chinese parallel dictionary. The words not included in the parallel dictionary are passed to the next step: generating possible Chinese character combinations as the translation candidates using the Hangul-Hanzi mapping table. The candidates are ranked by the *combination score* and *context similarity score*. The combination score represents the possibility of the sequence of the chinese characters calculated on the large Chinese web corpus. The context similarity score considers the context of the input sentence and that of the sentences in the large Chinese web corpus.

3.1 Extracting Nouns

Korean words are separated from each other with spaces, however, each words splitted by a space may contain one or more morphological elements. For example, Korean word 학교에서 is composed of 학교 (noun) and 에서 (particle). Since most of the Sino-Korean words are nouns, to obtain Sino-Korean words, we need a Korean morphological analyzer to extract nouns.

For the given Korean sentences that contain Sino-Korean vocabularies, apart from splitting words with spaces, we extracted nouns from these sentences with the help of a Java-library based morphological analyzer[*3]. The precision of the analyzer is announced to be higher than 95%.

Table 1: A portion of the mapping table

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

3.2 Translation by Dictionary Matching

Some of the Korean nouns are translated into Chinese with a parallel dictionary as the initial step. As is well known, Wikipedia offers a wide range of parallel data for many languages, among them is aligned Wikipedia titles. In our method, we use the aligned Wikipedia title pairs of Chinese and Korean as a parallel dictionary. In addition, 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 of the definition part to augment the parallel dictionary.
- Convert Chinese characters of traditional Chinese into simplified Chinese.

The Korean nouns which cannot be translated with the parallel dictionary are passed to the next process. In addition, the Korean nouns which have multiple translation candidates (such as homonyms or ambiguous words) are also passed to the next process.

3.3 Generating Translation Candidates

The aim 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 1, we can generate the translation candidates for "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 the combined words may have no practical meaning. So we need to select the most appropriate combination.

3.4 Rank the Translation Candidates

Now we have a large amount of combinations of Hanzi characters. In order to select the most appropriate one, we utilize combination score and context

[*1] Wikipedia:상용한자

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

[*3] KOMORAN ver 2.3 (Java Korean morphological analyzer)

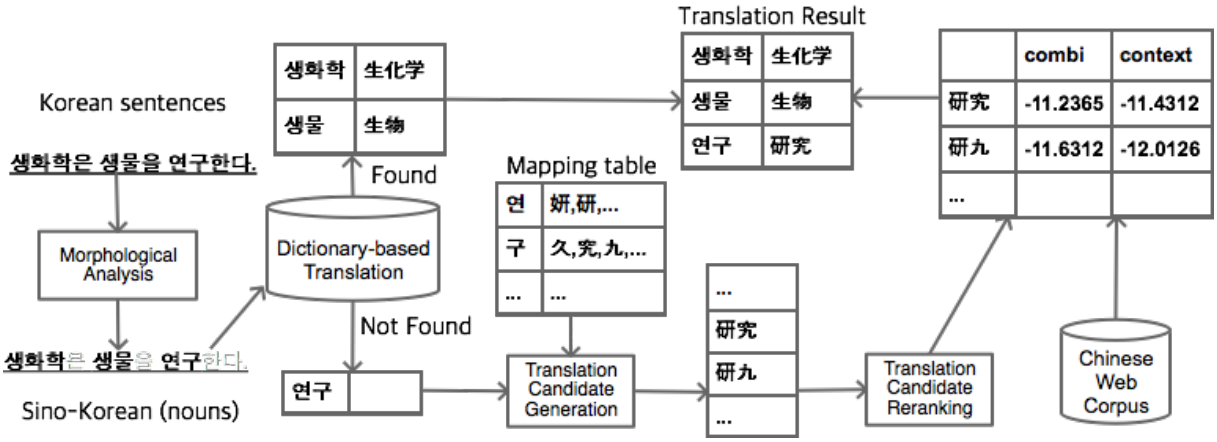

Figure 1: Lexicon Construction System

similarity score calculated using a large Chinese web corpus.

3.4.1 Combination Score

Combination score S_{combi} measures the strength of the link between the characters. For example, the combination score for "汉字" is calculated as

$$S_{combi}(汉字) = P(汉 \mid 字) \times P(字 \mid 汉)$$

3.4.2 Context Similarity Score

For each combination, character-based context vector is constructed using the web corpus. We use sentences which contains the combination as the context window, and each element of the vector is the co-occurrence count of Chinese characters. We ignore stop characters such as 的 and 了 [*4] and characters with less than 100 times of occurrence.

We also construct another context vector of the input Korean sentence using the formerly translated Korean words. The context similarity score $S_{context}$ can be calculated as the cosine similarity of the two context vectors.

3.4.3 Interpolation

The combination score is useful to examine if the combination is appropriate or not, and the context similarity score is useful to select the appropriate one according to the context where two or more combinations have practical meanings. Therefore, we interpolate the two scores and calculate the score of the translation candidate $S(cand)$ as follows:

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

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

Table 2: The Korean-Chinese aligned resource

Hanzi	Meaning in Korean	Hangul
强	강할	강
京	서울	경
界	지경	계
計	셀	계

The specified value of α ([0,..1]) is determined with 5-fold cross validation. We divide the set into 5 parts and recursively select one of them to test the precision for each α. The final translation result is get with the best performed α. The character combination with the highest score is regarded as the final translation result.

4 Experiment

4.1 Settings

For the Hanja-Hanzi mapping table, [Chu, et, al. 2012] have produced a *Chinese character mapping table for Japanese (Kanji), Traditional Chinese (TC) and Simplified Chinese (SC)* [2], thus for constructing a table further for Japanese, Chinese and Korean, we need to construct rather Japanese-Korean or Chinese-Korean table and merge these two tables. We used the 3500 Chinese characters, which is mentioned in the Section 2.2, the Table 2 briefly shows data sample of the resource.

We merged the table with the [Chu, et, al. 2012]'s and checked the compatibility with web-engined Hanja dictionary[*5][*6]. Finally, we achieved the map-

[*5] http://hanja.naver.com (네이버 한자사전)

[*6] http://small.dic.daum.net/index.do?dic=hanja (Daum 한자사전)

Hangul	Hanja	Kanji	Hanzi
가 :	珂	珂	珂
	呵	呵	呵 ,诃
	柯	柯	柯
	价	價	价
	可	可	可
	歌	歌	歌
	暇	暇	暇

Figure 2: Hangul-Hanja-Kanji-Hanzi mapping table

Table 3: The α -Precision relation for each testset

testset	1	2	3	4	5
α	0.71	0.71	0.71	0.71	0.71
Precision(%)	72.92	76.52	74.73	76.96	78.87

ping table, and its data sample is as shown below: We prepared 100 Korean sentences with 3281 words for test. After morphological analyzing, 1014 words among them returned analyzing result as nouns. 458 words of them were found data from the Wikipedia. For the left words, we obtained the possible combination using the mapping table shown in Figure 2. For querying the web corpus, we used the KenLM[3] model, which utilizes a character based process.

4.2 Result

We conducted the experiment with 5-fold cross validation, and obtained the best-performed (highest precision) α for each testset, as shown in Table 3.

The result of the translation with $\alpha = 0.71$ is as shown below (Table 4). Table 5 shows some good and bad examples of the translation result. The first lines of each table shows the original Korean test sentences and words being discussed is underlined. The column "By Context" shows translation result by only considering context features, and "By Context+Combi" shows result by considering both context features and combination scores.

5 Conclusion and Future Work

In this paper, we conducted an automatic character-based Korean-to-Chinese translation. The aim of our system is to construct useful resources in MT between Japanese, Korean and Chinese. A Java library based morphology analyzer was induced to extract Sino-Korean like words (considered only nouns), with which the precision is announced be higher than 95%. In the translation step, we both considered the context vectors and probabilities of each Hanzi combination. In the future, we plan to use the model to conduct a Korean-Chinese-Japanese machine translation.

Table 4: The translate result table

	Wikipedia	+Context	+Context + Probability
Correct	395	389	426
InCorrect	63	150	113
NoTranstion	556	17	17
Precision(%)	38.95	69.96	76.62

Table 5: Examples of good and bad translations.

Good example:

양성자, 전자, 광자 모두 입자성과 파동성을 가지고 있다.

Korean	Correct	By Context	By Context+Combi
전자	电子	电磁	电子

Bad example:

명제 논리는 논리식로 명제를 기술 하는 형식 체계이다.

Korean	Correct	By Context	By Context+Combi
기술	记述	记述	技术

References

[1] Jin-Xia Huang and Key-Sun Choi. Chinese-Korean Word Alignment Based on Linguistic Comparison. In *ACL*, 2000.

[2] Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. 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, May 2012.

[3] Kenneth Heafield. 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, July 2011.