Clustering of English-Korean Translation Word Pairs Using Bi-grams

Hanmin Jung[1], Hee-Kwan Koo[2], Won-Kyung Sung[1] and Dong-In Park[1]

[1] NTIS Division, KISTI, Korea
jhm@kisti.re.kr
[2] Practical Information Science, UST, Korea

Abstract. This paper describes a clustering algorithm for Korean translation words automatically extracted from Korean newspapers. Since above 80% of English words appear with abbreviated forms in Korean newspapers, it is necessary to make the clusters of their Korean translation words to easily construct bi-lingual knowledge bases such as dictionaries and translation patterns. As a seed to acquire each translation cluster, we repeat to choose an adequate translation word from a remaining translation set using an extended bi-gram-based binary vector matching until the set becomes empty. We also deal with several phenomena such as transliterations and acronyms during the clustering. Experimental results showed that our algorithm is superior to Dice coefficient and Jaccard coefficient in both determining adequate translation words and clustering translations.

1 Introduction

As information technology develops in recent years, many terminologies are rapidly created and discarded. Newspapers are excellent resources to acquire new-coined terms and to inspect their life cycle [4]. About 90% of terms in Korean newspapers, in particular, are originated from foreign languages such as English and Chinese[1] [1]. Some of them are accompanied by original words in English for readers to easily grasp the meaning, for example, "세계무역기구 (WTO)." However, many English words (about 82% in our test set) appear with abbreviated forms, and translations differ like "아시아태평양경제협력기구," "아시아태평양경제협력체," "아태경제협력체," and "아태경제협력회의" for "APEC; Asia-Pacific Economic Cooperation." Such English abbreviated forms tend to cause word sense ambiguities, for example, "Internet Service Provider," "Information Strategic Planning," and "Image Signal Processor" for "ISP." Newspapers also usually use parentheses to represent a pair of translation pairs, but they are not limited to the pairs. Many extraction errors are caused by the free uses of parentheses such as "모델명 S3C2410 (CPU)"[2] and

[1] E.g., "아펙" is a Korean transliterated word for English "APEC," and "경제" is for Chinese "經濟."
[2] "모델명 S3C2410" = "모델명 (Model No.)" + "S3C2410."

© *A. Gelbukh (Editor)*
Advances in Natural Language Processing
Research in Computing Science 18, 2006, pp. 15-24

"경제한파 (IMF)."[3] Korean transliteration is another consideration for the design of a translation clustering model since it does not contain any translation meaning but imply pronunciation rules. These phenomena should be resolved to make translation clusters and to determine adequate translation words, which are crucial for the building of translation knowledge bases.

However, previous studies failed to notice the need for the clustering [4, 5]. They focused only on automatic transliteration and unabbreviated word translation. We think they might not collect and analyze the real status of a huge newspaper corpus. In this paper, we will introduce the subsequent methods to manage translations in a real newspaper corpus with the amount of about 30 million Korean words: transliteration clustering, translation clustering including acronyms, and adequate translation determination.

Automatic transliteration can be implemented by direct and pivot-based translation systems [6]. Previous studies tried to generate several possible candidate words based on pronunciation derived by dictionaries and statistical approaches such as Markov window and decision tree [4, 5, 6]. However, they considered only English unabbreviated words that generate many possible transliteration candidates. It is the reason why they introduced statistical methods to rank the candidates. Comparison of an English word with a Korean word is much easier than generating the best transliteration candidate for the given English word. In addition, the ratio of English abbreviated words in Korean newspaper corpus is above 80%, which indicates that complex pronunciations (e.g. "er" and "eo") appear less than unabbreviated words.

Example-based translation systems like [2] usually use linguistic information and statistical information. The number of element words in each language becomes a basic feature to acquire linguistic information. However, the number of element words in abbreviated forms cannot be directly calculated. Statistical information for corresponding probability is also meaningless because we extract bilingual words from translation patterns not bilingual corpus.

Important issues in our research scope are to make translation clusters and to determine an adequate translation word for each cluster from monolingual corpus. These are the points that our research scope differs from the alignment and the extraction of translation patterns from bilingual corpus [7, 8]. Unfortunately, there is no study of the issues for Korean newspaper corpus. Nobody tried to extract a set of Korean translations for an English word in a real newspaper. Ignoring English abbreviated forms that frequently appear in the corpus would be another reason to skip the issues.

We found that clustering method using similarity between surface forms is more efficient than using dictionaries and partial translation word matching since translation words appear with various forms and parentheses are widely used to clarify the meaning of the words. For example, Korean translations for "EC" are, at least, morphologically classified into three groups: "Electronic Commerce," "European Commission," and "Electrolytic Condensers." The whole process including an extended bi-gram-based binary vector matching to measure semantic distance between two translation words and to determine an adequate one for a cluster will be introduced in Section 2 and 3.

[3] "경제한파" = "경제 (Economic)" + "한파 (Cold wave)." "국제금융기구" is a right translation of "IMF."

2 System Overview

To generate translation clusters for an English word, we introduce four functions: *FindTransliterationCluster* (see Section 3.1), *DetermineAdequateTranslation* (see Section 3.2), *FindTranslationCluster* (see Section 3.3), and *FindAcronymCluster* (see Section 3.4). An adequate translation word is automatically obtained before generating a cluster for it.

```
TranslationClustering (T) {
        $C_1$ = FindTransliterationCluster (T);
        T = T – $C_1$;
        i = 2;
        Repeat while T is not NULL {
                $C_i$.adequate_translation = DetermineAdequateTranslation (T);
                $C_i$ = FindTranslationCluster ($C_i$.adequate_translation, T);
                T = T – $C_i$;
                $C_i$ = $C_i$ + FindAcronymCluster ($C_i$.adequate_translation, T);
                T = T – $C_i$;
                Increase i;
        }
        Return C;
}
```

Fig. 1. Translation clustering process including transliterations and acronyms (Both T and C_i are the sets of translations, and C_i.adequate_translation is a translation word.)

T is the Korean translation set of an English word. It includes one or more translation clusters that will be found as the above process goes ahead. A translation cluster consists of Korean translation words with the same meaning. *TranslationClustering* finds these clusters C_1, C_2, and so on (see Section 3). *FindTransliterationCluster* generates a translation cluster whose components are transliterations, for example, "시스템온칩" and "플랫폼[4]." Since they have no meaning in Korean, we separate them as a distinct cluster (C_1) from translation set T (see Section 3.3). The loop to find translation clusters continues until the translation set T becomes NULL. Whenever iteration ends, we find a translation cluster including an adequate translation word in it. *DetermineAdequateTranslation* gives us the adequate translation word, that is, the word with the most shared bi-grams in translation set T (see Section 3.2). *FindTranslationCluster* generates a translation cluster in the manner of matching the adequate (C_i.adequate_translation) with the translation words in set T. In the case that a Korean word shares one or more bi-grams with the adequate, we consider the two translations are in the same translation cluster. *FindAcronymCluster* discovers acronyms for the adequate (C_i.adequate_translation). Finally, we acquire a set of translation clusters (C = {C_1, C_2 ...}).

[4] "시스템온칩" = "시스템 (System)" + "온 (On)" + "칩 (Chip)"
"플랫폼" = "플랫폼 (Platform)"

Table 1. An example to acquire translation clusters for English word "KAIST; Korea Advanced Institute of Science and Technology" (Underlined are newly added terms.)

Initial State	
T	{한국과학기술원, 한국과학기술기술원, 연기한국과학기술원, 카이스트, 과기원, 나노종합팩, 석사·박사[5]}
After FindTransliterationCluster	
C_1	{카이스트}
1st Iteration	
After DetermineAdequateTranslation	
C_2.adequate_translation	한국과학기술원[6]
After FindTranslationCluster	
C_2	{한국과학기술원, 한국과학기술기술원, 연기한국과학기술원}
After FindAcronymCluster	
C_2	{한국과학기술원, 한국과학기술기술원, 연기한국과학기술원, 과기원[7]}
2nd Iteration	
After DetermineAdequateTranslation	
C_3.adequate_translation	나노종합팩
After FindTranslationCluster	
C_3	{나노종합팩}
3rd Iteration	
After DetermineAdequateTranslation	
C_4.adequate_translation	석사·박사
After FindTranslationCluster	
C_4	{석사·박사}
Translation Clusters	
C	{C_1, C_2, C_3, C_4}

3 Translation-Clustering

3.1 Finding a Transliteration Cluster

In Korean, there are two ways to make translation words; one is transliteration (e.g. "이미지시그널프로세싱[8]") and the other is liberal translation using Chinese characters (e.g. "영상신호처리[9]"). Transliterations should not be assigned into other trans-

[5] The translation set T includes an extraction error "연기한국과학기술원" and a typing error "한국과학기술기술원." "나노종합팩" and "석사·박사" are a little irrelevant terms with "KAIST."

[6] 한국과학기술원 (韓國科學技術院). It is the best translation in Korean.

[7] 과기원 (科技院). It is an acronym of "과학기술원."

[8] "Image Signal Processing" ↔ "이미지시그널프로세싱" = "이미지 (Image)" + "시그널 (Signal)" + "프로세싱 (Processing)"

[9] "Image Signal Processing" ↔ "영상신호처리" = "영상[映像] (Image)" + "신호[信號] (Signal)" + "처리[處理] (Processing)"

lation clusters since they have no meaning in Korean. Thus, we generate a separate translation cluster for these by applying *FindTransliterationCluster*.

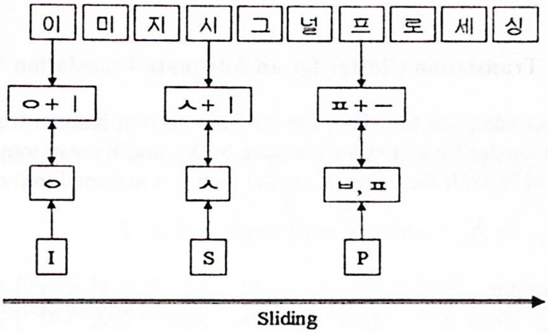

Fig. 2. An example to compare an English word "ISP" with a Korean transliteration candidate "이미지시그널프로세싱"

Unlike [4] and [5] which tried to automatically generate the best transliteration for an English word, it is easier to determine a translation word whether it is a transliteration or not. We make a set of simple mapping rules that each contains a possible Korean alphabet[10] list corresponding to an English character, for example, {'ㅂ,' 'ㅍ'} for 'p.' We convert each character of an English word into a sequence of Korean alphabet lists. Figure 2 shows an example of this mapping. It is very similar to the acronym-finding process of Section 3.3.

3.2 Choosing an Adequate Translation Word

Let the adequate translation word of a translation cluster C_i be C_i.adequate_translation whose adequate-value (AV) is the maximum in translation set T. AV_k is for a translation word *k* in set T, and is defined by the subsequent equation (1). The number of translation words in set T is *n*. AV_k increases as *k* shares bi-grams with the other words more and more. We prefer a shorter word to a longer word when tie occurs.

$$AV_k = \frac{\sum_{j=1}^{n} |X_k \cap X_j|}{|X_k|} \quad \text{where } k \neq j \tag{1}$$

Let us show an example to determine an adequate one for a translation set {"한국과학기술원," "한국과학기술기술원," "연기한국과학기술원," "과기원"}. "한국과학기술원" shares 6 bi-grams with "한국과학기술기술원," [11] and 6 with "연기한국과학기술원." [12] Since its bi-gram length is 6, AV becomes 2. The ade-

[10] A Korean alphabet is a phonetic unit that can be a consonant or a vowel.

[11] The shared bi-grams are "한국," "국과," "과학," "학기," "기술," and "술원."

[12] The shared bi-grams are the same as the above.

quate-value of "한국과학기술원" is the maximum among the others, thus it is chosen as the adequate translation word from the example set. A translation cluster then is generated from the set by matching with the adequate word as follows.

3.3 Finding a Translation Cluster for an Adequate Translation Word

After obtaining an adequate translation word from current Korean translation set, we find a translation cluster for it. Let an adequate be X_{C_i} and Korean translation set be T. A translation word X_j with the value of greater than 0 is assigned into cluster C_i.

$$| X_{C_i} \cap X_j | \text{ where } X_j \text{ is an element of set T} \tag{2}$$

In the above example, "한국과학기술기술원" and "연기한국과학기술원" become members of C_i, since they share bi-grams with "한국과학기술원" which is C_i.adequate_translation.

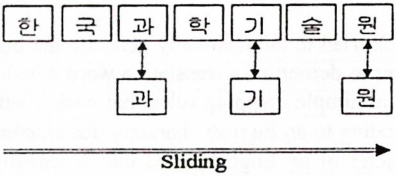

Fig. 3. An example to find an acronym ("과기원") using a Korean unabbreviated word ("한국과학기술 원") which is an adequate translation word.

3.4 Finding an Acronym Cluster

Some Korean unabbreviated words, in particular, organization names, written in Chinese characters[13] have acronyms such as "한국과학기술원 (韓國科學技術院) → 과기원 (科技院)" and "정보통신부 (情報通信部) → 정통부 (情通部)." As a Korean unabbreviated word and its acronym have the same character sequences, we can easily match the two words in the manner of left-to-right scanning. Matched acronyms then are assigned into the previously acquired translation cluster (see C_2 in Table 1).

4 Experimental Results

From Korean IT newspaper corpus,[14] we extracted Korean-English pairs combined with parentheses such as "북미자유무역협정 (NAFTA)" and "나프타 (NAFTA)."

[13] Most of the Korean words are originated from Chinese.

[14] Electronic Times (http://www.etnews.co.kr/)

Total 1,806 Korean translation sets were acquired after re-arranging on English words, and 200[15] of them were used to measure the subsequent performance.

We choose Dice coefficient as a criterion to compare with our method, and modify it like

$$\mathrm{AV_k} = \sum_{j=1}^{n} \frac{2|X_k \cap X_j|}{|X_k| + |X_j|}$$

where $k \neq j$ (*see Section 3.2 to refer the notations*). Jaccard coefficient

$$\mathrm{AV_k} = \sum_{j=1}^{n} \frac{|X_k \cap X_j|}{|X_k \cup X_j|}$$

where $k \neq j$, is another criterion. As these two algorithms are bi-gram approaches to easily calculate the similarity between strings, they were chosen. Korean word $\mathrm{X_k}$ would be selected when it has the highest AV value. According to our clustering algorithms, different adequate translation words have different clusters.

To measure the performance, we manually attached cluster tags to the above-mentioned 200 translation sets. A translation set consists of one or more semantically separate clusters without consideration of surface forms. The subsequent shows an example of the tagging results for our answer set (C1, C2, and C3 are cluster tags, and A is answer adequate translation word.). Adequate words can be multiple in a cluster.

[ATM] 현금자동입출금기/C1/A 현금입출금기/C1/A
초대형현금자동입출금기/C1 자동화기기/C1
비동기전송모드/C2/A 비동기전송방식/C2/A 장비-비동기전송모드/C2
초고속정보통신망/C3/A 초고속국가망/C3
초고속교환기/C3 초고속국가정보통신/C3[16]

Table 2 shows the comparison between our methods and the Dice/Jaccard Coefficients. Precision for choosing adequate translation word is the ratio of the number of correct words to the number of chosen adequate words. The system automatically checks whether adequate words and clusters are correct or not by comparing with the answer set. In the case that one or more translation words in an acquired cluster have different cluster tags with the other words in the cluster, we consider the cluster is wrong. Recall for clustering translations is the ratio of the number of correct clusters to the number of answer clusters. A cluster is recognized as correct one if and only if all the translation words of it are exactly matched with those of a cluster in the answer set.

[15] We selected translation sets with more than five translation words. The number of total words is 2,253 and the average number of translation words for a translation set is 11.265.

[16] The words with C3 are not translation words of "ATM." However, we always attached answer tags because our research topic does not concern about the determination of whether a word really is translation word or not.

Table 2. Comparison among our method, Dice coefficient, and Jaccard coefficient[17]

	Our Method	Dice Coefficient	Jaccard Coefficient
The Number of Translation Clusters	617	623	623
Average Size of Translation Clusters	3.651	3.617	3.617
Recall for Choosing Adequate Translation Word	82.496% (575/697)	75.036% (523/697)	75.036% (523/697)
Precision for Choosing Adequate Translation Word	93.193% (575/617)	83.949% (523/623)	83.949% (523/623)
F-measure for Choosing Adequate Translation Word	87.519%	79.243%	79.243%
Recall for Clustering Translations[18]	64.275% (448/697)	61.549% (429/697)	61.549% (429/697)
Precision for Clustering Translations	72.609% (448/617)	68.104% (429/623)	68.104% (429/623)
F-measure for Clustering Translations	68.188%	64.661%	64.661%

The reason why our method shows higher performance than the other two is that we discriminatively apply length information to eliminate superfluous words attached adequate translation word due to automatic extraction from corpus, for example, "로열티한국전자통신연구원 (ETRI)" and "셀러론중앙처리장치 (CPU)"[19]. These redundancies can be easily eliminated since they appear rarely.

We found three factors that decrease the performance: word sense ambiguities of English abbreviated words, synonyms without sharing bi-gram, and fake translation pairs with parentheses. (1) "WTO" has two meanings: "World Trade Organization" and "World Tourism Organization." Their representative translation words are "세계무역기구" and "세계관광기구." They share "세계" and "기구," thus the two translations are recognized as the same members of a cluster by our algorithms. Some English abbreviated words including widely used components such as "system" and "technology" tend to have many word sense ambiguities. Using stop words for them would be helpful to reduce the ambiguities. (2) "IPO" as "Initial Public Offering" has several translation words with the same meaning such as "기업공개" and "주식공모," even though they do not share any bi-gram. Introducing morphological analysis and synonym set (e.g. "공개=공모" and "미=미국"[20]) would be helpful to

[17] It is interesting that the two coefficients show the same performance even though adequate values are different.

[18] Each translation set has one or more translation clusters, and even garbage clusters including extraction errors, because of translation ambiguity. This is the reason why the number of correct translation clusters is 697 not 200.

[19] "로열티" is for "Royalty" and "셀러론" is for "Celeron."

[20] "미 (美)" is a Korean abbreviated form of "미국 (美國)." Both words are represented for "USA."

enhance the clustering performance. (3) As previously mentioned, newspapers widely use parentheses to expatiate translation words. The pairs can accidentally share bigrams with other translation pairs, for example, "삼성전자 (ETRI)" and "한국전자통신연구원 (ETRI)."[21] It will be a way to reduce wrongly extracted translation pairs by referring English unabbreviated words corresponding to English abbreviated forms.

5 Conclusions

We introduced a practical translation-clustering algorithm for translation pairs automatically extracted from newspaper corpus by using an extended bi-gram-based binary vector matching. To increase clustering coverage, our research scope included transliterations and acronyms. It has an important meaning in that previous studies could not consider a great portion of abbreviated forms appeared in a real newspaper corpus. Knowledge builders can easily confirm the clustering results because the system shows both adequate translation words and translation clusters. We now consider introducing cluster verification by Web search and histogram-based adequate translation determination as future works.

References

1. Choi. K. and Chae, Y.: Terminology in KOREA: KORTERM. Proceedings of LREC-2000.
2. Izuha, T.: Machine Translation Using Bilingual Term Entries Extracted from Parallel Texts. Journal of Systems and Computers. Vol.36 No.8 (2005)
3. Jung, H., Koo, H., Lee, B., and Sung, W.: Toward Managing the Life Cycle of Terms Using Term Dominance Trend. Proceedings of the Pacific Association for Computational Linguistics (2005)
4. Jung, S., Hong, H., and Baek, E.: An English to Korean Transliteration Model of Extended Markov Window. Proceedings of the 18th conference on Computational linguistics (2000)
5. Lee, J.: An English-Korean Transliteration and Retransliteration Model for Cross-lingual Information Retrieval. Ph.D. Thesis. Korea Advanced Institute of Science and Technology (1999)
6. Oh, J. And Choi, K.: An English-Korean Transliteration Model Using Pronunciation and Contextual Rules. Proceedings of the International Conference on Computational Linguistics (2002)
7. Ohara, M., Matsubara, S., and Inagaki, Y.: Automatic Extraction of Translation Patterns from Bilingual Legal Corpus. Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering (2003)
8. Tufis, D., Barbu, A., and Ion, R.: Extracting Multilingual Lexicons from Parallel Corpora. Journal of Computers and Humanities. Vol.38 No.2 (2004)

[21] "삼성전자" is for "Samsung Electronics Inc." and "한국전자통신연구원" is for "Electronics and Telecommunications Research Institute."