SELECTION OF KOREAN PROPER TRANSLATION WORDS USING BI-GRAM-BASED HISTOGRAMS

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

[1] Information System Division, KISTI, Korea
Email: {jhm, wksung, dipark}@kisti.re.kr
[2] Practical Information Science, UST, Korea
Email:hkkoo@kisti.re.kr

ABSTRACT

This paper describes a proper translation-selecting and translation-clustering algorithm for Korean translation of words automatically extracted from newspapers. As about 80% of the English words in Korean newspapers appear in abbreviated form, it is necessary to make clusters of translation words to construct easily bilingual knowledge bases such as dictionaries and translation patterns. As a seed to acquiring a translation cluster, we selected a proper translation word from a given translation set using bi-gram-based histograms. Translation words that share bi-grams with the chosen proper translation word are assigned to the cluster for the proper word. The given translation set then picks out the translation words of the cluster. These processes continue until the translation set becomes empty. Experimental results show that our algorithms are superior to bi-gram-based binary vectors including Dice coefficient and Jaccard coefficient in selecting the proper translation word for each translation cluster.

Keywords: Translation word, Bi-gram-based histogram, Clustering, Terminology, Term life cycle, Transliteration

1 INTRODUCTION

As information technology develops, many terminologies evolve and are then discarded. Newspapers are excellent resources for acquiring newly coined terms and inspecting their life cycles (Jung et al., 2005). About 90% of such terms in Korean newspapers, in particular, originate in foreign languages such as English and Chinese[1] (Choi & Chae, 2000). Some of them are accompanied by original words in English for readers to grasp the meaning easily, for example, “세계무역기구 (WTO).” However, many English words (about 82% in our test set) are 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 ambiguities, for example, “Internet Service Provider,” “Information Strategic Planning,” and “Image Signal Processor” for “ISP.” Newspapers also usually use parentheses to represent the pair of possible translations, but the entire meaning may not be limited to the pair. Many extraction errors are caused by the free use of parentheses such as “모델명S3C2410 (CPU)” and “경제한파 (IMF).”[2] Korean transliteration is another consideration for the design of a translation-clustering model because the transliteration does not contain any translation meaning. These phenomena require making translation clusters and selecting proper translation words, which are crucial for the building of translation knowledge bases.

However, previous studies failed to notice the need for translation clusters (Jung et al., 2000) (Lee, 2000). They focused only on automatic transliteration and unabbreviated word translation. We believe this work does not collect and analyze the real status of a huge newspaper corpus. In this paper, we introduce the sub-

[1] For example, “아펙” is a Korean translated word for English word “APEC,” and “경제” is for Chinese word “經濟.”
[2] “모델명 S3C2410” = “모델명 (Model No.)” + “S3C2410”
“경제한파” = “경제 (Economic)” + “한파 (Cold wave)”

sequent methods of managing translations in a real newspaper corpus of about 30 million Korean words: proper translation word finding and translation clustering.
Automatic transliteration can be implemented by direct and pivot-based translation (Oh & Choi, 2002). 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 (Jung et al., 2000) (Lee, 2000) (Oh & Choi, 2002). They considered only English unabbreviated words that generate many possible transliteration candidates. As a result, they introduced statistical methods to rank the candidates. However, comparison between English and Korean words is much easier than generating the best transliteration candidate for a given English word. In addition, the ratio of abbreviated English words in the Korean newspaper corpus is over 80%, which indicates that complex pronunciations (e.g. "er" and "eo") appear less than unabbreviated words.

Example-based translation systems (Izuha, 2005) generally use linguistic information and statistical information. The number of element words in each language becomes a basic feature in acquiring linguistic information. However, the number of element words in abbreviated forms cannot be directly calculated. Statistical information for corresponding probability is also meaningless because extracted bilingual words are from translation patterns not from the bilingual corpus.

Important issues in our research scope are the generation of translation clusters and the recommendation of proper translation words for the clusters from the monolingual corpus. These are the points in which our research differs from the alignment and the extraction of translation patterns from the bilingual corpus (Ohara et al., 2003) (Tufis et al., 2003). Unfortunately, there is no study of the issues for the Korean newspaper corpus. No one has previously 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 is another reason to ignore these issues.

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

2 SYSTEM OVERVIEW

To generate translation clusters for an English word, we introduce three functions: ***SelectProperTranslation*** (*see Section 3.1*), ***FindTranslationCluster*** (*see Section 3.2*), and ***VerifyTranslationCluster*** (*see Section 3.3*). The algorithm for transaction-clustering is given in Figure 1. A proper translation word is automatically obtained before generating a cluster for it.

```
TranslationClustering ($T_e$) {
        i = 1;
        Repeat while $T_e$ is not NULL {
                $C_i$.proper_translation = SelectProperTranslation ($T_e$);

                $C_i$ = FindTranslationCluster ($C_i$.proper_translation, $T_e$);
                $T_e$ = $T_e$ – $C_i$;

                Increase i;
        }

        Return C;
}
```

Figure 1. Translation-clustering process (T_e is the translation set of an English term e, and C_i.proper_translation is a translation word of translation cluster C_i.)

In the algorithm T_e is the Korean translation set of an English term “e.” It is divided into 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*). The loop to find translation clusters continues until the translation set T_e becomes NULL. Whenever the iteration ends, we find a translation cluster including a proper translation word in it. ***SelectProperTranslation*** gives us the proper translation word, that is, the word with the most shared bi-grams in translation set T_e. ***FindTranslationCluster*** generates a translation cluster in the manner of matching the proper word (C_i.proper_translation) with the translation words in set T_e. In the case that a Korean word shares one or more bi-grams with the proper word, we consider the two translations are in the same translation cluster. Finally, we acquire a set of translation clusters (C = {C_1, C_2 …}).

Table 1. An example of acquiring translation clusters for the abbreviated English term “WTO” (∅ indicates that it failed to find one or more unabbreviated forms corresponding to the abbreviated form.)

Initial State	
T_e	{국제무역기구, 세계무역기구, 반도체시장개방세계무역기구, 외견상세계무역기구, 더블유티오}
1st Iteration	
*After **SelectProperTranslation***	
C_1.proper_translation	세계무역기구
*After **FindTranslationCluster***	
C_1	{국제무역기구, 세계무역기구, 세계관광기구, 반도체시장개방세계무역기구, 외견상세계무역기구}
T_e	{ 더블유티오}
2nd Iteration	
*After **SelectProperTranslation***	
C_2.proper_translation	더블유티오[3]
*After **FindTranslationCluster***	
C_2	{더블유티오}
T_e	{}
Translation Clusters	
C	{C_1, C_2}

3 TRANSLATION-CLUSTERING

3.1 Selection of a proper translation word using bi-gram-based histograms

First, we define a modified bi-gram-based method to select a proper translation word from a given translation cluster. Let the proper translation word of a translation cluster C_i be C_i.proper_translation whose proper-value (PV) is the maximum in translation set T_e. PV(k) is for translation word k in set T_e, and is defined by the subsequent equation (1). The number of translation words in set T_e is n. X_k is the bi-gram set of translation word k. PV(k) increases as k shares bi-grams with more and more words. This method naturally favors preference of a shorter word to a longer word when a tie occurs.

$$\mathrm{PV}(k) = \frac{\sum_{j=1}^{n} | X_k \cap X_j |}{| X_k | + 1} \qquad \text{where } k \neq j \tag{1}$$

However, in most cases in which the size of a given translation cluster is rather small, and thus shared bi-grams are not enough to select a proper one, a shorter translation word would be the winner. To compensate

[3] Korean transliteration of “WTO”

for this problem, we introduce a bi-gram-based histogram analysis, whose main idea is that more consecutively matched bi-grams should receive high matching scores. If the difference of two bi-gram frequencies is less than 2, then the system recognizes the two as consecutively matched bi-grams and multiplies their frequencies. Otherwise, it adds their frequencies. Equation (2) explains this algorithm as given in Figure 2 using stack operations. B_i is the frequency of the *i*th bi-gram.

$$\mathrm{PV}(k) = \frac{ContinuityCheck(X_k)}{|X_k| + 1} \tag{2}$$

```
ContinuityCheck($X_k$) {
        $Score = B_0$;
For (i=1; i<SizeOf($X_k$); i++) {
                $Score$ *= $B_i$;                  if ($B_i$<($B_{i-1}$*2) or $B_{i-1}$<($B_i$*2))
                PUSH($Score$); PUSH(+);  $Score = B_i$;        otherwise
        }
        PUSH($Score$);
Binary calculation with POP() until stack underflows;
        Return calculation result;
}
```

Figure 2. Algorithm for bi-gram-based histogram analysis

Let us show an example to select a proper translation from a translation set {“봉지재,” “반도체봉지제,” “반도체봉지재,” “반도체용봉지제”} for “EMC.” Table 2 shows bi-gram frequencies of the above translation set. The system automatically finds the boundary lines in the bi-gram-based histograms of the translation words (*see scissors notation in fig. 3*). The proper values of the translation words are then calculated by equation (2) as follows, and “반도체봉지제” or “반도체봉지재” is selected as a proper translation (C_i.proper_translation).

Table 2. Bi-gram frequency of “EMC”

Bi-gram	Frequency	Bi-gram	Frequency
봉지	4	지재	2
반도	3	도체	3
체봉	2	지제	2
체용	1	용봉	1

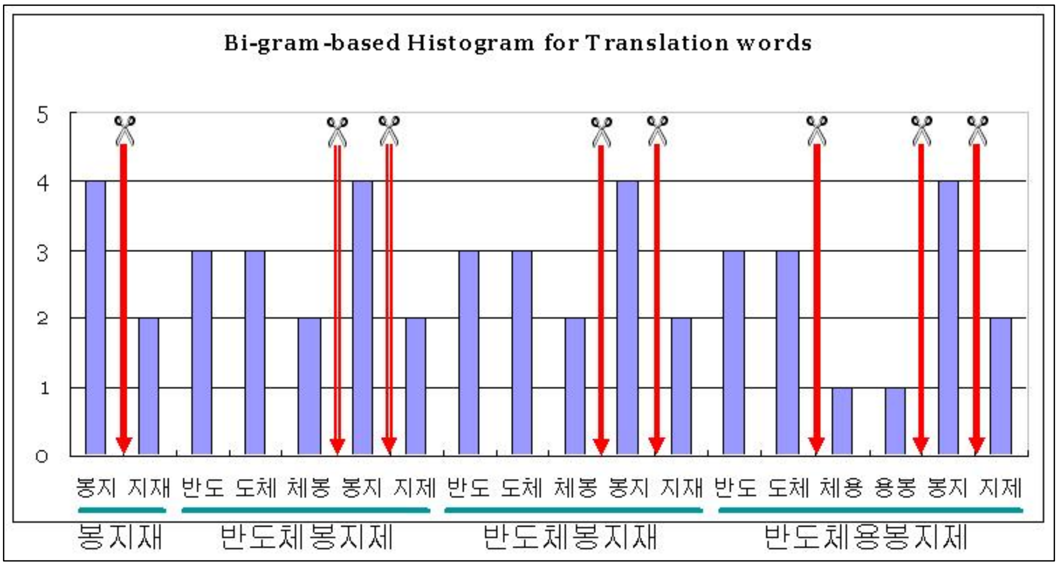

Figure 3. Bi-gram-based histogram for a translation set {“봉지재,” “반도체봉지제,” “반도체봉지재,” “반도체용봉지제”} for “EMC” (Scissors notation means boundary lines of the histogram.)

PV (“봉지재”) = (4 + 2) / (2 + 1) = 2.00

PV (“반도체봉지제”) = (3 * 3 * 2 + 4 + 2) / (5 + 1) = 4.00

PV (“반도체봉지재”) = (3 * 3 * 2 + 4 + 2) / (5 + 1) = 4.00

PV (“반도체용봉지제”) = (3 * 3 + 1 * 1 + 4 + 2) / (6 + 1) = 2.29

3.2 Finding a translation cluster for a proper translation word

After obtaining a proper translation word from a current Korean translation set, we find a translation cluster for it. Let a proper word be X_{Ci} and the Korean translation set be T_e. A translation word X_j with the value of greater than zero is assigned to cluster C_i.

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

In the above example, “한국과학기술기술원” and “연기한국과학기술원” become members of C_i because they share bi-grams with “한국과학기술원” which is C_i.proper_translation.

4 EXPERIMENTAL RESULTS

From a Korean IT newspaper corpus[4], we extracted Korean-English pairs combined with parentheses such as “북미자유무역협정 (NAFTA)” and “나프타 (NAFTA).” A total of 1,806 Korean translation sets were acquired after re-arranging on English words, and 200[5] of them were used to measure the subsequent performance.

We chose the Dice coefficient as a criterion with which to compare our method and modified it, so PV(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*).

The Jaccard coefficient (PV(k) = $\sum_{j=1}^{n} \frac{| X_k \cap X_j |}{| X_k \cup X_j |}$ where $k \neq j$) is another criterion. The Korean word X_k would be selected when it has the highest PV value. According to our clustering algorithms, different proper translation words have different clusters.

To measure the performance, we manually attached cluster tags to the above-mentioned 200 translation sets, which consist of one or more semantically separate clusters without consideration of surface forms. An example is given below of the tagging results for our answer set. C1, C2, and C3 are cluster tags, and A is answer proper translation word.). Proper words can be a multiple within 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

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

[5] 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.

초고속교환기/C3 초고속국가정보통신/C3[6]

Table 3 shows the comparison between our three methods and the Dice/Jaccard Coefficients. Precision for selecting the proper translation word is the ratio of the number of correct words to the number of chosen proper words. The system automatically checks whether proper words and clusters are correct by comparison with the answer set. In the case that one or more translation words in an acquired cluster have different cluster tags from 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 if and only if all the translation words in it are exactly matched with those of a cluster in the answer set. We experimented on our methods with two ways: (1) without bi-gram-based histogram analysis and (2) with bi-gram-based histogram analysis.

Table 3. Comparison among our two methods, Dice coefficient, and Jaccard coefficient[7] (BH: Bi-gram-based Histograms)

	Our Method (without BH, equation (1))	Our Method (with BH, equation (2))	Dice/Jaccard Coefficient Method
The Number of Translation Clusters	613	611	620
Average Size of Translation Clusters	3.675	3.736	3.633
Recall for Selecting Proper Translation Word	86.370% (602/697)	**87.088% (607/697)**	75.323% (525/697)
Precision for Selecting Proper Translation Word	98.206% (602/613)	**99.345% (607/611)**	84.677% (525/620)
F-measure for Selecting Proper Translation Word	91.909%	**92.814%**	79.727%
Recall for Clustering Translations	64.419% (449/697)	64.132% (447/697)	62.123% (433/697)
Precision for Clustering Translations	73.246% (449/613)	73.159% (447/611)	69.839% (433/620)
F-measure for Clustering Translations	68.550%	68.349%	65.755%

The reason why our method shows higher performance than Dice coefficients and Jaccard coefficients is that we discriminatively apply length information to eliminate superfluous words attached to the proper translation word because of automatic extraction from the corpus, for example, "로열티한국전자통신연구원 (ETRI)" and "셀러론중앙처리장치 (CPU)"[8]. These redundancies can be easily eliminated as they appear rarely.

We found two factors that decrease the performance of our methods; (1) "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 a synonym set (e.g. "공개=공모" and "미=미국"[9]) would be helpful in enhancing the clustering performance. (2) As previously mentioned, newspapers widely use parentheses to expatiate translation words. The pairs can accidentally share bi-grams with other translation pairs, for example, "삼성전자 (ETRI)" and "한국전자통신연구원 (ETRI)."[10] Such analysis might be a way to reduce wrongly extracted translation pairs, for example, by referring to English unabbreviated words corresponding to English abbreviated forms.

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

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

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

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

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

5 CONCLUSIONS

We introduced a practical proper translation-selecting and translation-clustering algorithm for translation pairs automatically extracted from the Korean newspaper corpus using bi-gram-based histograms. It has an important meaning in that previous studies could not consider a great portion of abbreviated forms that appeared in a real newspaper corpus. Knowledge builders have only to confirm proper translations and their clusters generated from a given translation set.

6 REFERENCES

Choi, K. & Chae, Y. (2000) Terminology in KOREA: KORTERM. *Proc. LREC-2000*.

Izuha, T. (2005) Machine Translation Using Bilingual Term Entries Extracted from Parallel Texts. *Journal Systems and Computers 36*(8).

Jung, H., Koo, H., Lee, B., & Sung, W. (2005) Toward Managing the Life Cycle of Terms Using Term Dominance Trend. *Proc. the Pacific Association for Computational Linguistics*.

Jung, S., Hong, H., & Baek, E. (2000) An English to Korean Transliteration Model of Extended Markov Window. *Proc. the 18th conference on Computational linguistics*.

Lee, J. (1999) *An English-Korean Transliteration and Retransliteration Model for Cross-lingual Information Retrieval*. PhD Thesis. Korea Advanced Institute of Science and Technology.

Oh, J. & Choi, K. (2002) An English-Korean Transliteration Model Using Pronunciation and Contextual Rules. *Proc. the International Conference on Computational Linguistics*.

Ohara, M., Matsubara, S., & Inagaki, Y. (2003) Automatic Extraction of Translation Patterns from Bilingual Legal Corpus. *Proc. the International Conference on Natural Language Processing and Knowledge Engineering*.

Tufis, D., Barbu, A., & Ion, R. (2004) Extracting Multilingual Lexicons from Parallel Corpora. *Journal Computers and Humanities 38*(2).