J Inf Process Syst, Vol.12, No.3, pp.455~467, September 2016
http://dx.doi.org/10.3745/JIPS.02.0044

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Bilingual Multiword Expression Alignment by Constituent-Based Similarity Score

Hyeong-Won Seo*, Hongseok Kwon**, Min-Ah Cheon***, and Jae-Hoon Kim***

Abstract

This paper presents the constituent-based approach for aligning bilingual multiword expressions, such as noun phrases, by considering the relationship not only between source expressions and their target translation equivalents but also between the expressions and constituents of the target equivalents. We only considered the compositional preferences of multiword expressions and not their idiomatic usages because our multiword identification method focuses on their collocational or compositional preferences. In our experimental results, the constituent-based approach showed much better performances than the general method for extracting bilingual multiword expressions. For our future work, we will examine the scoring method of the constituent-based approach in regards to having the best performance. Moreover, we will extend target entries in the evaluation dictionaries by considering their synonyms.

Keywords

Bilingual Lexicon, Compositionality, Context Vector, Multiword Expression, MWE Alignment, Pivot Language

1. Introduction

A lexicon, including multiword expressions (MWEs), is a very important resource for many natural language processing (NLP) applications, such as building ontologies [1], information retrieval (IR) [2], text alignment [3], machine translation [4,5], and so on [6]. In general, extracting bilingual lexicons requires many linguistic resources, such as bilingual corpora (i.e., parallel or comparable), bilingual seed dictionaries, and some heuristics, for mapping identical word pairs between two languages. However, extracting bilingual pairs of MWEs from bilingual corpora not only requires them, but is also much harder than extracting single words.

In this situation, there are several studies [7-10] for extracting bilingual MWEs from parallel corpora. Most of them extract MWEs in resource-rich language pairs, such as English–French or English–Chinese. Bilingual corpora from resource-rich language pairs, such as English–*(any language), are readily available online, while those for resource-poor language pairs, such as Korean–French are hard to get. This is one of the disadvantages for extracting bilingual MWEs from parallel corpora. Under the circumstance, Seo et al. [11] have proposed a novel method, denoted as the pivot context-based

Manuscript received June 19, 2015; first revision December 21, 2015; accepted March 25, 2016.

Corresponding Author: Jae-Hoon Kim (jhoon@kmou.ac.kr)
* International Vaccine Institute, Seoul, Korea (HyeongWon.Seo@ivi.int)
** Dept. of Computer Science and Engineering, POSTECH, Pohang, Korea (hong8c@naver.com)
*** Dept. of Engineering, Korea Maritime and Ocean University, Busan, Korea (dkahffk0218@naver.com, jhoon@kmou.ac.kr)

approach for multiwords (PCAM) in the rest of the paper, for extracting bilingual MWEs by using parallel corpora in a resource-poor language pair. However, PCAM has a weak point that targets constituents instead of complete translation equivalents, which might have high scores of similarity if their contexts are not enough. In order to compensate for this shortcoming, this paper presents a novel method to compute constrained similarity scores not only between source words and translation equivalents, but also between source words and constituents of the translation equivalents. We call the reinforced approach a constituent-based approach (CTA), which significantly outperforms PCAM in terms of accuracy, as confirmed through several types of experiments.

The rest of the paper is organized as follows: Section 2 presents several works related to the bilingual translation extraction of multiword, including PCAM. In Section 3, we provide the motivation behind the method we are proposing and address the resources used to implement it. Section 4 gives the experimental results and Section 5 draws conclusions and describes further studies.

2. Related Work

2.1 MWE Identification

Early approaches to MWE identification focused on collocational behaviors between MWEs [12]. Other research compares 55 association measures, such as point-wise mutual information (PMI) to score German adjective-noun and preposition-verb collocation candidates [13]. Mixing different kinds of association measures is more effective than using one stand-alone measure. Other studies based on association measures have been proposed [14,15] to find the measure that shows the highest efficiency for identifying several types of MWEs in several languages. However, computing co-occurrence measures is probably not effective for obtaining the best performance. Piao et al. [16] combined word statistics with linguistic information to extract idiomatic MWEs because about 68% of MWEs occur only once or twice in their corpus.

Several researchers have used latent semantic analysis (LSA) to distinguish between compositional and non-compositional preferences of expressions [17,18]. They show that compositional MWEs are generally more likely similar to their constituents than other non-compositional MWEs. In other words, non-compositional MWEs probably have a different sense with the combination of their constituents. For example, the score from LSA between the expression “under the weather” and the word “sick” is much higher than the score between the expression and its constituents. The important thing is that we need a lot of idiomatic information for MWEs to distinguish whether the usage of MWEs is idiomatic or non-idiomatic. Unfortunately, however, such information is usually not available.

In this paper, we only focus on compositional MWEs (i.e., identifying idiomatic expressions are ignored in limited resource-poor circumstances). Thus, any type of external language resources, such as a bilingual dictionary or a parser, is not needed. Furthermore, we assumed that pivot single-words are enough to play the role of the bridge that connects two languages, source and target. Therefore we only extracted MWE candidates of both source and target languages, and those for a pivot language were ignored. Based on these circumstances, we identified potential MWE candidates for source and target languages by using both linguistic and statistical information [19,20]. The identification method (see

[11] for more details) can be described as follows:

1. To extract all possible n-grams ($2 \leq n \leq 3$) from monolingual corpora.
2. To select good monolingual candidates by statistical scores, such as PMI.
3. To extract specific POS sequences to select potential MWE candidates.

2.2 MWE Alignment

We used the pivot-based context approach [21] to align MWE candidates. It is for extracting bilingual single-words from resource-poor language pairs. The key idea is to use a pivot language, such as English, because parallel corpora, such as English–*, are readily available online.

The PCAM is the extended version for MWEs. It first identifies MWE candidates from monolingual corpora, and then concatenates them into single tokens to directly use the pivot-based context approach for MWE alignments. The experimental results of the PCAM showed quite meaningful performances while no external linguistic resources (e.g., bilingual dictionaries, syntactic parser, and so forth) were used. However, the method can give errors for low-frequency words. For example, there are several incorrect translations that occurred due to a lack of contexts. If a target constituent has a richer context than its parent expression (i.e., MWEs containing the constituent), the constituent can be selected instead of the parent expression. This is because of the identification method of the PCAM, which focuses on compositional MWE candidates. All extracted MWE candidates are based on a collocational measure, such as PMI. For this reason, we reinforced the method by considering the similarity scores of MWE candidates and constituents.

3. Constituent-Based Approach

In this section, we address the method for extracting bilingual MWEs in a resource-poor language pair, such as Korean–French/Spanish. For this purpose, this paper enhances the performance of PCAM as a baseline, which has difficulties when the contexts of multiword expressions are insufficient. In order to complement this shortcoming, this paper presents a novel method to compute constrained similarity scores not only between source words and target translation equivalents, but also between source words and constituents of the translation equivalents. In this paper, we call this a constituent-based context approach. In the following subsections, we describe motivation, an augmented similarity measure, and bilingual MWE lexicon extraction using the similarity measure.

3.1 Motivation

The task we address in this paper is aligning MWEs, such as noun phrases, in comparable corpora. Although the alignment method is not language-specific, it focuses on resource-poor language pairs, such as Korean–French. For several resource-rich language pairs (e.g., English–*), various types of bilingual resources such as parallel corpora or bilingual dictionaries are readily available. However, for some other (i.e., resource-poor) language pairs (e.g., Korean–*), such resources are usually unavailable. The problem is that building these resources manually requires a huge amount of effort and cost.

Although collecting monolingual resources is much easier than bilingual resources, collecting any kind of monolingual or bilingual resources for all language pairs is realistically impossible.

The alignment method (i.e., PCAM) is encouraged when such resource-poor language pairs are considered. It builds context vectors from two parallel corpora of resource-rich language pairs, such as Korean–English and English–French. These pairs share one pivot language (e.g., English) to bridge a resource-poor language pair, such as Korean–French. And then, the method computes similarity scores between two different vectors (i.e., for source and target languages) to select the top *k* target equivalents that are the closest to the source word.

A major limitation of the method is that most errors occur due to a lack of contexts. The shortage can be shown if domains of two parallel corpora are different from each other or the words are low-frequency in their corpus to build context vectors. Several types of errors derived from the shortage can be arranged as follows: First, target translation equivalents that are similar to a correct answer or that have the same topic but are incorrect. For example, where the French multiword *point de vue* ("point of view" in English) is given as a source word, the Korean word 세계관 (*vue métaphysique*, *vision du monde* in French; "world view" in English) as the incorrect target equivalent is extracted. Second, the target constituents of multiwords are extracted as the top *k* target equivalents. For example, where the Korean word 언어학과 ("department of linguistics" in English) is given as a source word, the French word *département* (or *linguistique*) as the target equivalent is extracted instead of the French multiword *département de linguistique*. This phenomenon can be seen when the context of the complete multiword (e.g., *département de linguistique*) is insufficient for being represented as a vector, but that of its constituents (i.e., *département* or *linguistique*) is so much richer. However, most of these multiwords are low-frequency (or rare) words. In fact, the frequency of multiwords should be lower than that of their constituents in a corpus although the multiword is a high- frequent word. In this paper, we focus on the latter type of errors (i.e., extracting constituents as target equivalents). For this, we used the PCAM that aligns bilingual MWE candidates, but the PCAM is not naturally able to handle this type of situation.

3.2 Augmented Similarity Measure

The contribution of the CTA is that it considers the relationship not only between source expressions and target equivalents, but also between the expressions and constituents of the target equivalents. This is mostly for low-frequency words where their context vectors are poor in regards to having high similarity scores. Note that we define a source expression as a landmark case, so constituents of source expressions are ignored in this work. We modified the cosine similarity formula to compute the similarity score between two vectors, $\vec{\boldsymbol{s}}$ of the source term s, and $\vec{\boldsymbol{t}}$ of the source term t, and the formula is given in Eq. (1):

$$sim(s,t) = \alpha\left(\frac{\vec{s}\cdot\vec{t}}{|\vec{s}||\vec{t}|}\right) + \beta\left(\frac{1}{|t|}\sum_{k=1}^{|t|}\frac{\vec{s}\cdot\overrightarrow{t_k}}{|\vec{s}||\overrightarrow{t_k}|}\right) \qquad (1)$$

where, $|t|$ denotes the number of constituents of the target equivalent. The parameters α and β are constants and can be estimated under various experiments.

3.3 Extracting a Bilingual MWE Lexicon Using the Augmented Measure

The steps to extract a bilingual MWE lexicon from two parallel corpora (i.e., the source-pivot parallel corpus and the pivot-target parallel corpus as PCAM) are as listed below.

(1) **MWE identification**: As mentioned in Section 2.1, all possible n-grams ($2 \leq n \leq 3$) are extracted from monolingual corpora (i.e., source and target language) independently. And then, reasonable collocations are extracted from the n-grams by an association measure (PMI is empirically determined in this work). Some of them, which have lower scores than a specific threshold, are eliminated. After that, several POS sequence patterns (see Section 4.1) are given to remove irrelevant MWE candidates. This identification method requires morphological analyzers and noun phrase patterns for each language. Performing this process is relatively easy because this information is already available for most languages. In this paper, we assumed that extracted MWE candidates are accepted as actual MWEs.

(2) **Building context vectors**: After MWE candidates are extracted, context vectors from two parallel corpora are separately built. The MWE candidates are first converted into single tokens by concatenating them with a specific symbol like '_' (for example, "department_of_linguistics"). These converted MWEs are treated as other single-words. As mentioned in Section 2.2, MWEs in the pivot languages are unnecessary and single pivot words are sufficient to connect both source and target languages.

(3) **Computing similarity scores**: After context vectors are built, similarity scores between one source word and all target words are computed. The thing that is the most different between this approach and the PCAM is whether each of the constituents of translation equivalents is taken into account or not. This approach considers all constituents when similarity scores are measured. This method is not measure-specific, so any similarity measure can be adapted here. In this thesis, only cosine similarity was considered. The modified measurement is described in Equation (2). For example, the similarity score sim (언어학과, *département de linguistique*) between the Korean word 언어학과 ("department of linguistics" in English) and the French phrase *département de linguistique* can be scored as follows:

$$sim(\text{언어학과}, département\ de\ linguistique)$$
$$= 0.6 \times \left(\frac{\overrightarrow{\text{언어학과}} \cdot \overrightarrow{département\ de\ linguistique}}{\left|\overrightarrow{\text{언어학과}}\right| \left|\overrightarrow{département\ de\ linguistique}\right|} \right)$$
$$+0.4 \times \frac{1}{2}\left(\frac{\overrightarrow{\text{언어학과}} \cdot \overrightarrow{département}}{\left|\overrightarrow{\text{언어학과}}\right| \left|\overrightarrow{département}\right|} + \frac{\overrightarrow{\text{언어학과}} \cdot \overrightarrow{linguistique}}{\left|\overrightarrow{\text{언어학과}}\right| \left|\overrightarrow{linguistique}\right|} \right)$$

We assumed that two parameters α and β are 0.6 and 0.4, respectively. As can be seen, only content words (i.e., nouns, verbs, adjectives, or adverbs) are included in the translation equivalents. The measurement that considers constituents to augment the score of MWEs is the key feature.

(4) **Selecting similar context vectors**: After all similarity scores for each source word are computed, the top × candidates are picked and then added to the bilingual lexicon.

4. Experimental Results

The main objective of this paper is to show that constituents consisting of a multiword are a form of useful information for improving the performance of a bilingual multiword lexicon extraction system, especially based on the PCAM. In this section, we bi-directionally evaluate our approach for two different language pairs: Korean-Spanish (KR-ES) and Korean-French (KR-FR).

4.1 Experimental Environments

4.1.1 Parallel corpora

To experiment on our approach, we needed several linguistic resources, such as parallel corpora (for source–pivot and pivot–target language pairs), morphological analyzers or lemmatizers, and part-of-speech taggers. For Korean–English (KR-EN) parallel corpus, we used the KMU parallel corpus (https://sites.google.com/site/nlpatkmu/Resources/Corpora) [22], which contains several bilingual news articles. For French/Spanish–English (FR/ES-EN) parallel corpus, we used the Europarl parallel corpus [23], which were extracted from the Proceedings of the European Parliament. In this paper, we only used the part of the corpus to maintain balance with the KR-EN parallel corpus in regards to size (see Table 1).

Table 1. Statistics of the parallel corpora

	Parallel corpora					
	Korean	**English**	**French**	**English**	**Spanish**	**English**
Sentences	433,151		500,000		500,000	
Words	8,283,222	13,381,739	13,292,137	12,750,062	13,196,180	12,713,067
Types	1,110,499	374,175	185,815	144,457	210,485	145,531
Avg. words*	19.1	30.9	26.6	25.5	26.4	25.4

* The Avg. words indicates the average number of words per sentence.

4.1.2 Preprocessing

In general, a Korean word usually contains one or more morphemes (2.3 morphemes per word in our experiment; but it depends on the domain or corpus). Therefore, a Korean word should be separated into several simplified/minimized units, since each word has one or more senses. For the other languages (i.e., English, French, and Spanish), we needed to access the lemmas of word tokens to reduce the size of context vectors and for richer statistics. To do this, some preprocessing steps should be performed. For Korean, the U-tagger (http://nlplab.ulsan.ac.kr) [24] was used to tokenize sentences and to induce POS tags of morpheme tokens. For the other languages (English, French, and Spanish), the TreeTagger (http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger) [25] was used to lemmatize word tokens and to induce their POS tags. All word/morpheme tokens were annotated and then transformed to lowercase letters.

4.1.3 Extracting MWE candidates

After preprocessing, we extracted MWE candidates. For this task, all stop-words, numeric strings, or punctuation marks were excluded. And then, word/morpheme n-grams, $2 \leq n \leq 3$, that occurred more than or equal to three times in each monolingual corpus were extracted by computing an association measure (see more details in Section 2.1) and applying them to light POS filters. The POS filter to extract French/Spanish noun phrases comes from the approach by Bouamor et al. [6], and for Korean are based on the French filter. The Korean POS filter (N-N, N-N-N, V-E-N, J-E-N, N-G-N) contains five patterns (resp. eight patterns for French) where N is a noun, G is a genitive case marker, V is a verb, J is an adjective, E is an adnominal ending, and P is a preposition. Most French patterns (N-N, N-N-N, N-J, J-N, J-N-J, N-N-J, N-J-J, N-P-N) consist of a noun and an adjective. To maintain balance with the French filter, the Korean filter should include POS sequences that are as similar as possible to the French patterns. Finally, single content word/morpheme tokens (i.e., nouns, verbs, adjectives, or adverbs) and extracted MWE candidates remained as input texts.

4.1.4 Building an evaluation dictionary

For evaluation, we needed evaluation dictionaries, which consist of source MWEs and target translations. Four evaluation dictionaries of KR→FR, FR→KR, KR→ES, and ES→KR were required and were manually built from web pages (http://dic.naver.com). All words in the source part were multiwords and in the target part were single-words or multiwords. The number of source MWEs for evaluations and the average number of their translations are listed in Table 2. Note that the MWEs or their translations are neither domain-specific nor over-fitted (i.e., they are considered as general terms) because the source MWEs came from web dictionaries. This means that the MWEs can be frequent in their corpora or not, although both MWEs and one of their translations must occur at least once in their corpora.

Table 2. Statistics of source MWEs in evaluation dictionaries

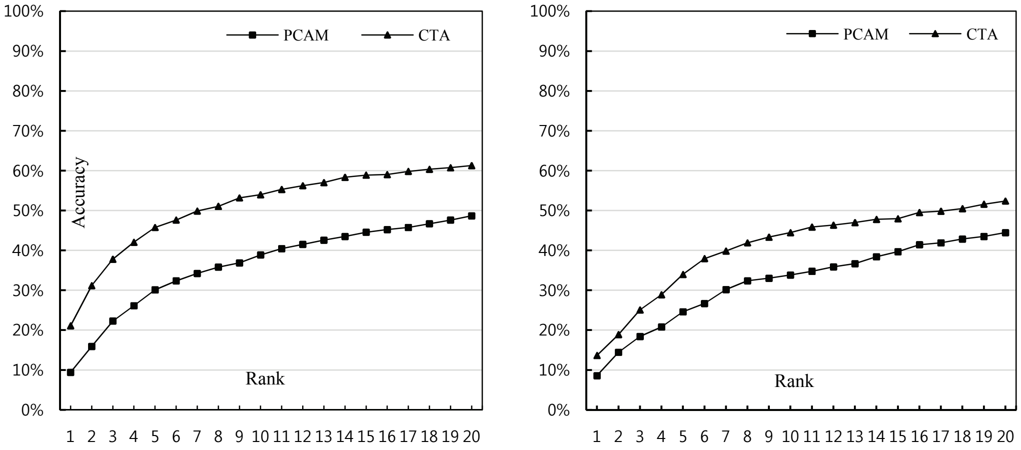
	Korean–French		Korean–Spanish	
Collected	15,287	28,961	8,489	15,540
Selected	754	630	426	529
Avg. translations	1.6	1.2	1.4	1.2

4.2 Performance Evaluation

In this section, we describe the experiments that we performed with the source MWEs in the evaluation dictionaries, as described in Table 2. We used the PCAM as the baseline. The PCAM used a general cosine similarity score between two context vectors (i.e., similarity scores for constituents are ignored) while the CTA used Eq. (1) for the relationship between one source word and constituents of the translation equivalent. Fig. 1 (resp. Fig. 2) indicates the accuracy from the top 1 to 20 “for Korean–French pairs (resp. Korean–Spanish pairs) (i.e., the percentage of source words that have at least one exact translation in top *x* candidate translations).

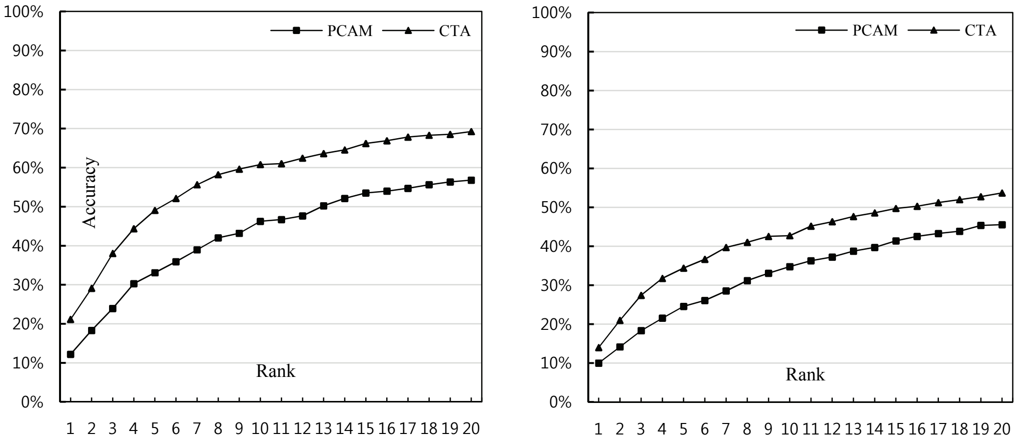
As seen in Figs. 1 and 2, the CTA significantly outperforms the PCAM. As for Korean to French translations, the best accuracy of 61.3% (455 out of 754 Korean source MWEs) was obtained for the top 20 by the CTA, while 48.7% (367 out of 754 Korean source MWEs) was obtained by the PCAM. As for the opposite translations (i.e., French to Korean), the best accuracy 52.4% (330 out of 630 French source MWEs) was obtained for the top 20 by the CTA, while 44.4% (280 out of 630 French source MWEs) was obtained by the PCAM. These results are very meaningful because they prove that considering constituents is clearly helpful in improving the performance of the PCAM.

Regarding Korean to Spanish translations, the best accuracy of 69.3% (295 out of 426 Korean source MWEs) was obtained for the top 20 by the CTA, while 56.8% (242 out of 426 Korean source MWEs) was obtained by the PCAM. In the case of Spanish to Korean translations, the best accuracy of 53.7% (284 out of 529 Spanish source MWEs) was obtained for the top 20 by the CTA, while 45.6% (241 out of 529 Spanish source MWEs) was obtained by the PCAM.



(a) (b)

Fig. 1. Accuracy on the Korean–French parallel corpora. (a) Korean to French and (b) French to Korean.



(a) (b)

Fig. 2. Accuracy on the Korean–Spanish parallel corpora. (a) Korean to Spanish and (b) Spanish to Korean.

Considering all of this information, we observed that the CTA generally outperforms the PCAM. Also note that these results are meaningful in terms of considering that evaluated words are not high-frequency or are general terms not fitted to specific domains.

4.3 Error Analysis

Table 3 shows the statistics of overall errors observed for the top 20. The CTA has reduced errors by an average of 10.4%. This means that considering constituents is helpful in improving the performance of the MWE alignment for resource-poor language pairs, even if the approach generates other types of errors.

Table 3. Statistics of errors

	Korean–French		**Korean–Spanish**	
Number of source MWEs	754	630	426	529
Number of source MWEs with no translation				
From the PCAM	387 (51.3)	354 (56.2)	184 (43.2)	288 (54.4)
From the CTA	292 (38.7)	300 (47.6)	131 (30.8)	245 (46.3)

Values are presented as number (%).

Even though the CTA outperformed the PCAM in our experiments, the performance of our proposed approach still needs improvement to reach much further. In particular, generating Korean translations (i.e., the case of French/Spanish to Korean translations) comparatively has more difficulties (error rate of 47.6% for French to Korean, 46.3% for Spanish to Korean translations; see Table 3) than the opposite cases (error rate of 38.7% for Korean to French, 30.8% for Korean to Spanish translations).

All errors have been classified into three types: (I) the case of "the translation equivalent is a reference translation, which is not included in an evaluation dictionary," (II) the case of "there is no correct translation, but a translation equivalent and a translation in an evaluation dictionary come from the same domain," and (III) the case of "the translation equivalent is one of constituents of a correct MWE translation." These error types will be discussed with several examples, and the statistics for error types are presented in Table 4.

Table 4. Statistics of error types

Language pair	**Method**	**Type I**	**Type II**	**Type III**
Korean → French	PCAM (387)	9 (2.3)	80 (20.7)	144 (37.2)
	CTA (292)	13 (4.5)	110 (37.7)	187 (64.0)
French → Korean	PCAM (354)	54 (15.3)	156 (44.1)	96 (27.1)
	CTA (300)	35 (11.7)	150 (50.0)	69 (23.0)
Korean → Spanish	PCAM (184)	11 (6.0)	59 (32.1)	64 (34.8)
	CTA (131)	13 (9.9)	58 (44.3)	89 (67.9)
Spanish → Korean	PCAM (288)	19 (6.6)	89 (30.9)	48 (16.7)
	CTA (245)	37 (15.1)	149 (60.8)	47 (19.2)

Values are presented as number (%).

The results (except for French to Korean translations) corresponding to Type I obtained by the CTA show higher percentages (i.e., 2.3% to 4.5% for Korean to French translations, 6.0% to 9.9% for Korean to Spanish translations, and 6.6% to 15.1% for Spanish to Korean translations). Here, two examples are presented for these results. First, the Korean to French translation pair 비상 사태 ("state of emergency") → *état d'urgence* already exists in the evaluation dictionary. The French phrase *état d'urgence* also has the meanings in terms of synonyms (or reference translations of 비상 사태), such as *situation d'urgence* ("emergency situation"), *situation critique* ("plight"), and *situation de danger* ("dangerous situation"). However, the acceptable translation equivalents, *situation d'urgence*, *situation critique*, and *situation de danger* will be marked as incorrect because the evaluation dictionary contains neither the synonyms for translations nor the reference translations of source MWEs. If the evaluation dictionaries are extended either manually or automatically, the performance of the approach can be much improved. Alternatively, *col blanc* ("white collar") has the Korean translation in terms of the literal meaning of 하얀색 깃 as well as the idiomatic meaning of 사무 직원 ("clerical worker," "office worker"). As mentioned before, such idiomatic expressions are ignored in this work so the latter example cannot be solved with the CTA.

The error corresponding to Type II indicates that extracted translation equivalents are incorrect but they treat the same topic as being a correct translation. For example, when the Korean to Spanish evaluation dictionary includes the pair 민간 항공 ("civil aviation") → *aviación civil*, the Spanish translation equivalents, such as *avión* ("plane"), *aeronave* ("aircraft"), and *línea internacional* ("international line") are extracted as the top *x* equivalents. All of these equivalents, the target translation *aviación civil*, and the source MWE 민간 항공 are related to the same topic of "flight." In other words, these words share common context words in the pivot language (i.e., in English, "flight," "airplane," "international," "domestic," and so on). However, the exact target translation *aviación civil* does not exist in the target monolingual corpus, or it has a very poor context even though it exists. This could be due to a misalignment of parallel sentences or a mismatching of domains between the source and target corpora. Each parallel corpus shares the same domain, respectively, but source/target monolingual (i.e., Korean/French or /Spanish, and vice versa) corpora do not. The fortunate thing is that the CTA extracts some words that share the same topic much more than the baseline does. Concerning just the number itself, the listed numbers obtained by the CTA slightly decreased (e.g., 156 to 150, and 59 to 58) or greatly increased (e.g., 80 to 110, and 89 to 149), but the percentages of them prove that the claim is true. Considering all of this information, this approach gathers more and more equivalents that share a context that is as similar as possible to each other.

Lastly, the error corresponding to type III is motivated to improve this kind of the error in this paper. As mentioned before, type III indicates that an equivalent as a whole is not extracted, but a part of its constituents are extracted as the top x equivalents. Note that this phenomenon especially occurs when the contexts of a translation equivalent as a whole are very poor, while the contexts of a part of constituents are rich. Usually, low-frequency words correspond to this type. As can be seen from Type III in Table 4, the results (except for French to Korean translations) obtained by the CTA show higher percentages than the baseline (i.e., 37.2% to 64.0% for Korean to French translations, 34.8% to 67.9% for Korean to Spanish translations, and 16.7% to 19.2% for Spanish to Korean translations). Taken as a part of the type of errors, these performances seem somewhat poor. Taken as a whole, however, the error rates are decreased by the CTA.

5. Conclusions

In this paper, we presented the constituent-based context approach, which measures augmented cosine similarity scores between source and target vectors. The measure computes a general cosine similarity score between two vectors and also computes the average score for each of the constituents of the target equivalent. Finally, it adds together two cosine similarity scores to augment the original score. As can be seen in our experimental results, this approach is quite meaningful and helpful in improving the performance of the alignment of MWEs.

For our future work, we will examine the parameters, α and β (mentioned in Section 3.2) in order to maximize the performance of this approach. We will also extend the evaluation dictionaries by extracting the synonyms of target translations.

Acknowledgement

This work was supported by the ICT R&D program of MSIP/IITP (R7119-16-1001, Core technology development of the real-time simultaneous speech translation based on knowledge enhancement). This paper has also been extended and updated from Chapter 6 of the Ph.D. dissertation of the first author.

References

[1] S. Venkatsubramanyan and J. Perez-Carballo, “Multiword expression filtering for building knowledge,” in *Proceedings of the 2nd ACL Workshop on Multiword Expressions: Integrating Processing*, Barcelona, Spain, 2004, pp. 40-47.

[2] Doucet and H. Ahonen-Myka, “Non-contiguous word sequences for information retrieval,” in *Proceedings of the 2nd ACL Workshop on Multiword Expressions: Integrating Processing*, Barcelona, Spain, 2004, pp 88-95.

[3] S. Venkatapathy and A. Joshi, “Using information about multiword expressions for the word-alignment task,” in *Proceedings of the COLING/ACL Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, Sydney, Australia, 2006, pp. 20-27.

[4] T. Baldwin and T. Tanaka, “Translation by machine of complex nominals: Getting it right,” in *Proceedings of the 2nd ACL Workshop on Multiword Expressions: Integrating Processing*, Barcelona, Spain, 2004, pp. 24-31.

[5] K. Uchiyama, T. Baldwin, and S. Ishizaki, “Disambiguating Japanese compound verbs,” *Computer Speech & Language*, vol. 19, no. 4, pp. 497-512, 2005.

[6] D. Bouamor, N. Semmar, and P. Zweigenbeaum, “Automatic construction of a multiword expressions bilingual lexicon: a statistical machine translation evaluation perspective,” in *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon (CogALex-III)*, Mumbai, India, 2012, pp. 95-108.

[7] F. Smadja, K. McKeown, and V. Hatzivassiloglou, “Translating collocations for bilingual lexicons: a statistical approach,” *Computational Linguistics*, vol. 22, no. 1, pp. 1-38, 1996.

[8] B. Daille, S. Dufour-Kowalski, and E. Morin, “French-English multi-word terms alignment based on lexical content analysis,” in *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC2004)*, Lisbon, Portugal, 2004, pp. 919-922.

[9] D. Wu and X. Xia, “Learning an English-Chinese lexicon from a parallel corpus,” in *Proceedings of the 1st Conference on Association for Machine Translation in the Americas*, Columbia, MD, 1994, pp. 206-213.

[10] B. Lu and B. K. Tsou, “Towards bilingual term extraction in comparable patents,” in *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC2009)*, Hong Kong, 2009, pp. 755-762.

[11] H. W. Seo, H. S. Kwon, M. A. Cheon, and J. H. Kim, “Bilingual multiword lexicon construction via a pivot language,” *Journal of Contemporary Engineering Sciences*, vol. 7, no. 23, pp. 1225-1233, 2014.

[12] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Computational Linguistics*, vol. 16, no. 1, pp. 22-29, 1990.

[13] P. Pecina, "A machine learning approach to multiword expression extraction," in *Proceedings of the LREC Workshop towards a Shared Task for Multiword Expressions (MWE2008)*, Marrakech, Morocco, 2008, pp. 54-57.

[14] Villavicencio, V. Kordoni, Y. Zhang, M. Idiart, and C. Ramisch, "Validation and evaluation of automatically acquired multiword expressions for grammar engineering," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language*, Prague, Czech Republic, 2007, pp. 1034-1043.

[15] G. Bouma, "Collocation extraction beyond the independence assumption," in *Proceedings of the ACL 2010 Conference Short Papers*, Uppsala, Sweden, 2010, pp. 109-114.

[16] S. S. Piao, P. Rayson, D. Archer, and T. McEnery, "Comparing and combining a semantic tagger and a statistical tool for MWE extraction," *Computer Speech and Language*, vol. 19, no. 4, pp. 378-397, 2005.

[17] G. Katz and E. Giesbrecht, "Automatic identification of non-compositional multiword expressions using latent semantic analysis," in *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, Sydney, Australia, 2006, pp. 12-19.

[18] T. Baldwin, C. Bannard, T. Tanaka, and D. Widdows, "An empirical model of multiword expression decomposability," in *Proceedings of the ACL 2003 Workshop on Multiword Expressions*, Singapore, 2003, pp. 89-96.

[19] B. Daille, E. Gaussier, and J. M. Lange, "Towards automatic extraction of monolingual and bilingual terminology," in *Proceeding of the 15th Conference on Computational Linguistics*, Kyoto, Japan, 1994, pp. 515-521.

[20] Kunchukuttan, "Multiword expression recognition," Ph.D. dissertation, Indian Institute of Technology, Bombay, India, 2007.

[21] H. W. Seo, H. S. Kwon, and J. H. Kim, "Context-based lexicon extraction via a pivot language," in *Proceeding of the 13th Conference on Pacific Association for Computational Linguistics (PACLING 2013)*, Tokyo, Japan, 2013.

[22] H. W. Seo, H. C. Kim, H. Y. Cho, J. H. Kim, and S. I. Yang, "Automatically constructing English-Korean parallel corpus from web documents," *Journal of KIISE: Software and Applications*, vol. 13, no. 2, pp. 161-164, 2006.

[23] P. Koehn, "Europarl: a parallel corpus for statistical machine translation," in *Proceeding of the 10th Conference on Machine Translation Summit*, Phuket, Thailand, 2005, pp. 79-86.

[24] J. C. Shin and C. Y. Ock. "A stage transition model for Korean part-of-speech and homograph tagging," *Journal of KIISE: Software and Applications*, vol. 39, no. 11, pp. 889-901, 2012.

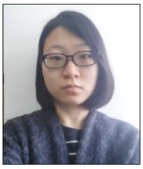
[25] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *Proceedings of the Conference on New Methods in Language Processing*, Manchester, UK, 1994, pp. 44-49.

Hyeong-Won Seo http://orcid.org/0000-0002-4256-1928

He received M.S. degree and Ph.D. in Department of Computer Engineering from Korea Maritime and Ocean University in 2015. Currently, he works at Development and Delivery Unit in International Vaccine Institute (IVI).

Hongseok Kwon http://orcid.org/0000-0002-4680-8705

He received B.S and M.S. degrees in School of Computer Engineering from Korea Maritime and Ocean University in 2012 and 2014, respectively. Since March 2015, he is with the School of Computer Science and Engineering from Pohang University of Science and Technology as a PhD candidate. His current research interests include machine translation and deep learning.



Min-Ah Cheon http://orcid.org/0000-0003-2925-7013

She received B.S. and M.S. degrees in School of Computer Engineering from Korea Maritime and Ocean University in 2014 and 2016, respectively. Since March 2016, she is with the School of Computer Engineering from Korea Maritime and Ocean University as a PhD candidate.

Jae-Hoon Kim http://orcid.org/0000-0001-8655-2591

He has been a professor at Department of Computer Engineering in Korea Maritime and Ocean University since September 1997. He was a visiting researcher at Beckman Institute in UIUC from August 2007 to August 2008 and at Information Sciences Institute in USC from February 2001 to March 2002. He was also a senior member of research staff in ETRI, Korea from February 1988 to August 1997. He received a B.A. degree in computer science from Keimyung University in 1986 and M.S. and Ph.D. in computer science from KAIST in 1988 and 1996, respectively. His current research interests include natural language processing and automated scoring.