Bilingual Lexicon Extraction via Pivot Language and Word Alignment Tool

Hong-Seok Kwon Hyeong-Won Seo Jae-Hoon Kim
Korea Maritime University,
Dongsam-Dong, Yeongdo-Gu, Busan, South Korea
hong8c@naver.com, wonn24@gmail.com, jhoon@hhu.ac.kr

Abstract

This paper presents a simple and effective method for automatic bilingual lexicon extraction from less-known language pairs. To do this, we bring in a bridge language named the pivot language and adopt information retrieval techniques combined with natural language processing techniques. Moreover, we use a freely available word aligner: Anymalign (Lardilleux et al., 2011) for constructing context vectors. Unlike the previous works, we obtain context vectors via a pivot language. Therefore, we do not require to translate context vectors by using a seed dictionary and improve the accuracy of low frequency word alignments that is weakness of statistical model by using Anymalign. In this paper, experiments have been conducted on two different language pairs that are bi-directional Korean-Spanish and Korean-French, respectively. The experimental results have demonstrated that our method for high-frequency words shows at least 76.3 and up to 87.2% and for the low-frequency words at least 43.3% and up to 48.9% within the top 20 ranking candidates, respectively.

1 Introduction

Bilingual lexicons are an important resource in many domains, for example, machine translation, cross-language information retrieval, and so on. The direct way of bilingual lexicon extraction is to align words from a parallel corpus (Wu and Xia, 1994), which contains source texts and their translations. For some language pairs, however, collecting the parallel corpus is not easy and are restricted to specific domains. For these reasons, many researchers in bilingual lexicon extraction have focused on comparable corpora (Fung, 1995; Yu and Tsujii, 2009; Ismail and Manandhar, 2010). These corpora are also hard to build on less-known language pairs, for instances, Korean and Spanish, Korean and French, and so on. Therefore, some researchers have studied the use of pivot languages as an intermediary language to extract bilingual lexicons (Tanaka and Ummemura, 1994; Wu and Wang, 2007; Tsunakawa et al., 2008).

On the other hand, some researchers adopt information retrieval (IR) techniques to extract bilingual lexicons (Fung, 1998; Gaussier et al., 2004; Hazem et al., 2012). The techniques are collecting all the lexical units from each of two languages, L_1 and L_2, respectively, and then are generating context vectors S and T for the collected lexical units in $L_1$and L_2, respectively. The context vector, S and T are translated using seed dictionaries, which are manually constructed by hand and of which the size is huge for accurate translation. Finally, the context vectors, S and T are compared with each other in order to get their translation candidates.

In this paper, we propose a simple and effective method for bilingual lexicons between two less-known language pairs using a pivot language and IR techniques. The pivot language is used for representing both of context vectors of a source language and a target language and IR techniques for calculating the similarity between the source context vector and the target context vector represented by the pivot language. Unlike the previous studies, therefore, we use two parallel corpora, Korean (KR)-English (EN) and English (EN) and English (EN)-Spanish (ES). Here English is the pivot language. We also use a free available word aligner, called Anymalign to generate the context vectors easily.

The proposed method has many advantages such as easy adaptation to less-known language pairs through a pivot language like English, easy extension to multi-word expression, and dramatic reduction in labor-intensive words to get a large scale seed dictionary.

The remainder of this paper is organized as follows: we describe the proposed approach in Section 2. The experimental results are presented in Section 3. Finally Section 4 draws conclusions and discusses the future works.

Proceedings of the 6th Workshop on Building and Using Comparable Corpora, pages 11–15,
Sofia, Bulgaria, August 8, 2013.

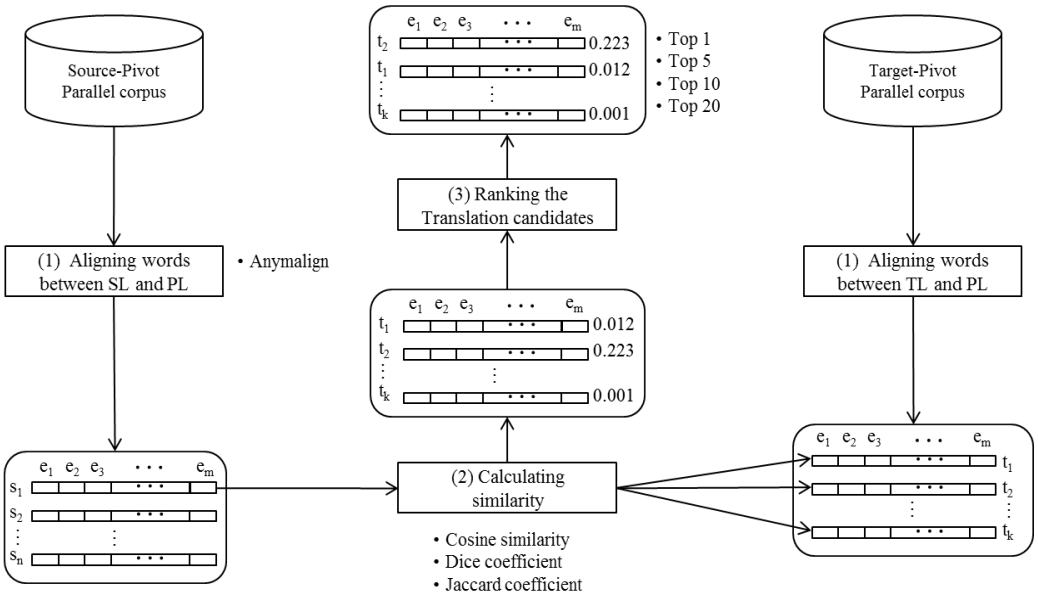


Figure 1. Overall structure of the proposed method.

2 Proposed Approach

In this paper, a simple and effective method for bilingual lexicons between two less-known language pairs using a pivot language and IR techniques. We use parallel corpora with more accurate alignment information instead of comparable corpora. It, however, is difficult to obtain parallel corpora for less-known language pairs. For such reasons, we use a pivot language which is well-known like English.

The pivot language is used for representing both of context vectors of a source language and a target language. Unlike the previous studies using comparable corpora, therefore, we use two parallel corpora through the pivot language like Korean (KR)-English (EN) and English (EN)-Spanish (ES) and IR techniques for calculating the similarity between the source context vector and the target context vector represented by the pivot language.

In the previous works, translating context-vectors is required using a seed dictionary, but in this paper, translating them is not needed anymore. Therefore, any bilingual dictionaries are not expected. Besides, we use a free available word aligner, called Anymalign, to construct context-vectors. Anymalign shows high accuracy for low-frequency words to extract translation candidates (Lardilleux et al., 2011). Overall structure of the proposed method is depicted in Figure 1. The proposed method can be summarized in the following three steps:

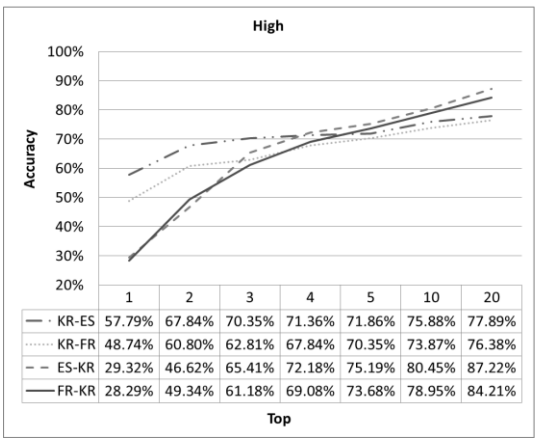
i. To build source context vectors and target source context vectors for each word in the source language (eg. KR) and the target language (eg. ES) using two sets of independent parallel corpora that are KR-EN and EN-ES, respectively. All words in context vectors are weighted by Anymalign.
ii. To calculate the similarity between each word in source context vector and all words in the target context vectors on the basis of the cosine measure
iii. To sort the top *k* word pairs based on their similarity scores

Two parallel corpora share a pivot language, English, in our case, and are used to build context vectors because Korean-Spanish bilingual corpora are publicly unavailable. Anymalign is used to weight all words in the context vectors.

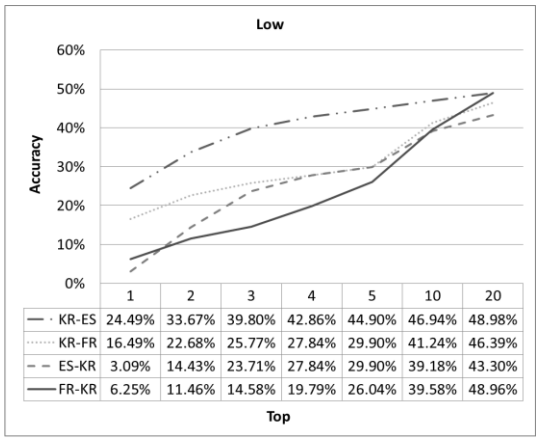
As mentioned before, in the previous work, a seed dictionary is required to translate context vectors at this time, but we do not carry out them. After context vectors are built once, all source and target context vectors are compared each other to get its similarity between them by using the cosine measure. Finally, top *k* word pairs are extracted as a result.

3 Experiments and Results

In this paper, we extract translation candidates from two different language pairs that are bi-directional KR-ES and KR-FR.



	1	2	3	4	5	10	20
KR-ES	57.79%	67.84%	70.35%	71.36%	71.86%	75.88%	77.89%
KR-FR	48.74%	60.80%	62.81%	67.84%	70.35%	73.87%	76.38%
ES-KR	29.32%	46.62%	65.41%	72.18%	75.19%	80.45%	87.22%
FR-KR	28.29%	49.34%	61.18%	69.08%	73.68%	78.95%	84.21%



	1	2	3	4	5	10	20
KR-ES	24.49%	33.67%	39.80%	42.86%	44.90%	46.94%	48.98%
KR-FR	16.49%	22.68%	25.77%	27.84%	29.90%	41.24%	46.39%
ES-KR	3.09%	14.43%	23.71%	27.84%	29.90%	39.18%	43.30%
FR-KR	6.25%	11.46%	14.58%	19.79%	26.04%	39.58%	48.96%

Figure 2. Accuracies of the proposed method for HIGH and LOW words.

3.1 Experimental setting

3.1.1 Parallel corpora

We used the KR-EN parallel corpora compiled by Seo et al. (2006) (433,151 sentence pairs), and two sets of sub-corpora (500,000 sentence pairs each) that are randomly selected from ES-EN and FR-EN in the Europarl parallel corpus (Koehn, 2005). The average number of words per sentence is described in Table 1 below. The number of words in ES-EN and FR-EN parallel corpora is nearly similar, but the number of KR words (called eojeol in Korean) in KR-EN parallel corpus is lower than that of EN words. In fact, KR words are a little bit different from EN words and others. Korean words consist of one morpheme or more. Therefore, the number of KR words can be similar to that of EN words if morphemes instead of words are counted.

KR-EN		ES-EN		FR-EN	
KR	**EN**	**ES**	**EN**	**FR**	**EN**
19.2	**31**	26.4	**25.4**	29.7	**27.1**

Table 1. The average number of words per sentence.

3.1.2 Data preprocessing

All words are tokenized by the following tools: Hannanum[1] (Lee et al., 1999) for Korean, TreeTagger[2] (Schmid, 1994) for English, Spanish and French. All words in English, Spanish, and French are converted to lower case, and those in Korean are morphologically analyzed into morphemes and pos-tagged by Hannanum.

[1] http://kldp.net/projects/hannanum

[2] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

3.1.3 Building evaluation dictionary

To evaluate the performance of the proposed method, we build two sets of bilingual lexicons (KR-ES and KR-FR) manually using the Web dictionary[3]. Each lexicon is unidirectional, meaning that they list the meanings of words of one language in another, and contains 100 high frequent words (denoted by HIGH hereafter) and 100 low rare words (denoted by LOW hereafter), respectively. The frequent words are randomly selected from 50% in high rank and the rare words from 20% in low rank. Table 2 shows the average number of the translations per source word in each lexicon. The number means the degree of ambiguity and is same as the number of polysemous words.

Evaluation dictionary	HIGH	LOW
KR-FR	5.79	2.26
KR-ES	7.36	3.12
ES-KR	10.31	5.49
FR-KR	10.42	6.32

Table 2. The average number of the translations per source word in the evaluation dictionaries.

3.1.4 Evaluation metrics

We evaluate the quality of translation candidates extracted by the proposed systems. Similar to the evaluation in information retrieval, the accuracy, the recall, and the mean reciprocal rank (MRR) (Voorhees, 1999) are used as evaluation metrics. The accuracy is the fraction of its translation candidates that are correct. The recall is the ratio of the suggested translation candidates that agree with the marked answer to the total number of translations in the evaluation words. The MRR is

[3] http://dic.naver.com/

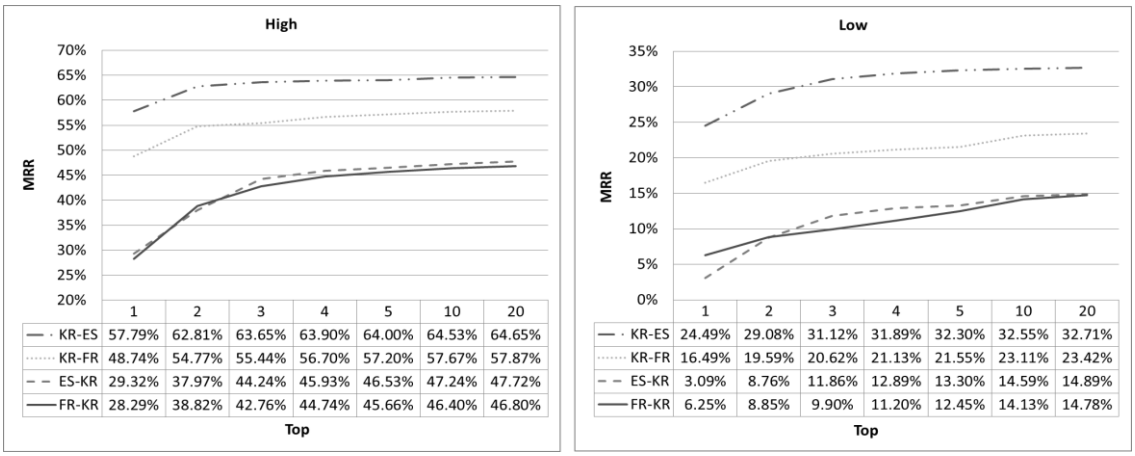


Figure 3. MRR of the proposed method for HIGH and LOW words.

the average of the reciprocal ranks of translation candidates that are correct translations for a sample of evaluation words.

3.2 Results

The accuracies of the HIGH and LOW words are shown in Figure 2. As seen in the figure, at the top 4 below, the accuracies of ES-KR and FR-KR are lower than the others. The difference can be attributed to stopwords such cardinal, ordinal, etc. The stopwords is normalized by Tree-Tagger for ES and FR, but not normalized by Korean POS-tagger (Hannanum). KR stopwords can badly affect the accuracies of ES-KR and FR-KR. In Table 3 below, '300' and '4' are stopwords and examples of the mistranslation of atención (attention)' in Spanish. Accordingly, '주목 (attention)' can be extracted as the first translation candidate if '300' and '4' are removed as stopwords.

Rank	Source language	Target language	Similarity score
1	atención	300	0.999
2	atención	주목 (attention)	0.993
3	atención	4	0.894
4	atención	눈(eye)	0.838
5	atención	모으(gather)	0.802

Table 3. Top 5 translation candidates of 'atención (attention)'.

The MRR results of the proposed method are shown in Figure 3. As shown in Figure 3, the MRR of the HIGH words is rapidly increased until the top 5, after then the MRR is steadily increased. This means that correct translation candidates tend to appear within the top 5. In the same experiments, the correct translation candidates for the LOW words tend to appear within top 10.

Lastly, the recalls of HIGH and LOW words are calculated in Table 4 below. As seen in the figure, the best recall is 32.7% on the KR-FR for HIGH words. One of reasons can be why words usually have one sense per corpus in parallel corpus (Fung, 1998). Another reason can be why words do not belong to various domains and our data sets only come from European Parliament proceedings and news article.

	Top20 Recall	
Language pairs	**High 100**	**Low 100**
KR-FR	32.73%	24.20%
KR-ES	27.49%	26.20%
ES-KR	29.55%	20.64%
FR-KR	27.30%	20.52%

Table 4. Recalls for HIGH and LOW words.

Our experimental results show that the proposed method is encouraging results because we do not use any linguistic resources such as a seed dictionary, and that the proposed method is sufficiently valuable where parallel corpus is unavailable between source and target languages.

4 Conclusion

We have presented an IR based approach for extracting bilingual lexicons from parallel corpus via pivot languages. We showed that the proposed method overcomes some of the problems of previous works that need a seed dictionary and use comparable corpora instead of parallel corpora in terms of lack of linguistic resources.

In future work, we will remove stopwords, and some words that have similar meaning could be clustered to improve the performance. Furthermore, we will handle multi word expression. Lastly, we plan to resolve a domain-constraint.

Acknowledgments

This work was supported by the Korea Ministry of Knowledge Economy (MKE) under Grant No.10041807

References

P. Fung. 1995. Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. In *Proceedings of the Third Workshop on Very Large Corpora (VLC'95)*, pages 173-183.

P. Fung. 1998. A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora. In *Proceedings of the Parallel Text Processing*, pages 1-16.

E. Gaussier, J.-M. Renders, I. Matveeva, C. Goutte and H. Dejean. 2004. A geometric view on bilingual lexcion extraction from comparable corpora. In *Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics*, Barcelona, Spain, pages 527-534.

A. Hazem and E. Morin. 2012. Adaptive dictionary for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, pages 288-292.

A. Ismail and S. Manandhar. 2010. Bilingual lexicon extraction from comparable corpora using in-domain terms. In *Proceedings of the International Conference on Computational Linguistics*, pages 481-489.

P. Koehn. 2005. EuroParl: A parallel corpus for statistical machine translation. In *proceedings of the Conference on the* 10^{th} *Machine Translation Summit*, page 79-86.

W. Lee, S. Kim, G. Kim and K. Choi. 1999. Implementation of modularized morphological analyzer. In *Proceedings of The 11th Annual Conference on Human and Cognitive Language Technology*, pages 123-136.

A. Lardilleux, Y. Lepage, and F. Yvon. 2011. The contribution of low frequencies to multilingual sub-sentential alignment: a differential associative approach. *International Journal of Advanced Intelligence*, 3(2):189-217.

H Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK, pages 44-49.

H. Seo, H. Kim, H. Cho, J. Kim and S. Yang, 2006. Automatically constructing English-Korean parallel corpus from web documents. *Korea Information Proceedings Society*, 13(2):161-164.

K. Tanaka and K. Umemura. 1994. Construction of a Bilingual Dictionary Intermediated by a Third Language. In *Proceedings of the 15th International Conference on Computational Linguistics (Coling' 94)*, Kyoto, Japan, August, pages 297-303.

T. Tsunakawa, N. Okazaki, and J. Tsujii. 2008. Building Bilingual Lexicons Using Lexical Translation Probabilities via Pivot Languages. In *Proceedings of the* 22^{nd} *International Conference on Computational Linguistics*, Posters Proceedings, pages 18-22.

E. Voorhees. 1999. The TREC-8 Question Answering Track Report. In *8th Text Retrieval Conference (TREC-8)*, pages 77-82.

D. Wu and X. Xia. 1994. Learning an English-Chinese lexicon from a parallel corpus. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA 1994, Columbia, Maryland, USA, October)*, pages 206-213.

H. Wu and H. Wang. 2007. Pivot Language Approach for Phrase-Based Statistical Machine Translation. In *Proceedings of 45th Annual Meeting of the Association for Computational Linguistics*, pages 856-863.

K. Yu and J. Tsujii. 2009. Bilingual dictionary extraction from Wikipedia. In *Proceedings of the 12th Machine Translation Summit (MTS 2009)*, Ottawa, Ontario, Canada.