Journal of the Korean Society of Marine Engineering, Vol. 37, No. 3, pp. 300~306, 2013 (ISSN 1226-9549)
http://dx.doi.org/10.5916/jkosme.2013.37.3.300

Bilingual lexicon induction through a pivot language

Jae-Hoon Kim† · Hyeong-Won Seo[1] · Hong-Seok Kwon[2]

(Received May 1, 2013 ; Revised May 7, 2013 ; Accepted May 13, 2013)

Abstract: This paper presents a new method for constructing bilingual lexicons through a pivot language. The proposed method is adapted from the context-based approach, called the standard approach, which is well-known for building bilingual lexicons using comparable corpora. The main difference between the standard approach and the proposed method is how to represent context vectors. The former is to represent context vectors in a target language, while the latter in a pivot language. The proposed method is very simplified from the standard approach thereby. Furthermore, the proposed method is more accurate than the standard approach because it uses parallel corpora instead of comparable corpora. The experiments are conducted on a language pair, Korean and Spanish. Our experimental results have shown that the proposed method is quite attractive where a parallel corpus directly between source and target languages are unavailable, but both source-pivot and pivot-target parallel corpora are available.

Keywords: Bilingual lexicon induction, Parallel corpus, Comparable corpora, Pivot language

1. Introduction

Bilingual lexicons are an important language resource in many domains, for example, machine translation, cross-language information retrieval, and so on. Automatic construction of bilingual lexicons has received great interest since the beginning of 1990 **[1]**. The relatively easy method for automatically constructing bilingual lexicons is to align source words with the corresponding target words using a parallel corpus **[2]**, which contains source texts and their translations. The parallel corpus for all language pairs, however, is not always publicly available and is also difficult to collect some language pairs. For these reasons, many researchers in bilingual lexicon extraction have focused on comparable corpora **[3]-[5]**. These corpora are also hard to build on less-known language pairs, for instances, Korean and Spanish. Therefore, some researchers have studied the use of a pivot language as an intermediary language to extract bilingual lexicons **[6]-[8]**. Tanaka and Ummemura (1994) **[6]** used a pivot language to construct a bilingual lexicon by using the structure of dictionaries and morphemes. Wu and Wang (2007) **[7]** utilized a pivot language to build a bilingual lexicon through phrase tables for statistical machine translation. Tsunakawa et al. (2008) **[8]** made use of a pivot language to increase the number of translation pairs obtained from two other bilingual lexicons.

Unlike the previous works, we use a pivot language to represent context vectors, which is constructed by using two parallel corpora: One is be-

† Corresponding Author: Division of Information Technology, Korea Maritime University, Youngdo-gu, Busan 606-791, Republic of Korea, E-mail: jhoon@hhu.ac.kr, Tel: 051-410-4574
1 Department of Computer Engineering, Korea Maritime University, Youngdo-gu, Busan 606-791, Republic of Korea, E-mail: wonn24@gmail.com, Tel: 051-410-4896
2 Department of Computer Engineering, Korea Maritime University, Youngdo-gu, Busan 606-791, Republic of Korea, E-mail: hong8c@naver.com, Tel: 051-410-4896

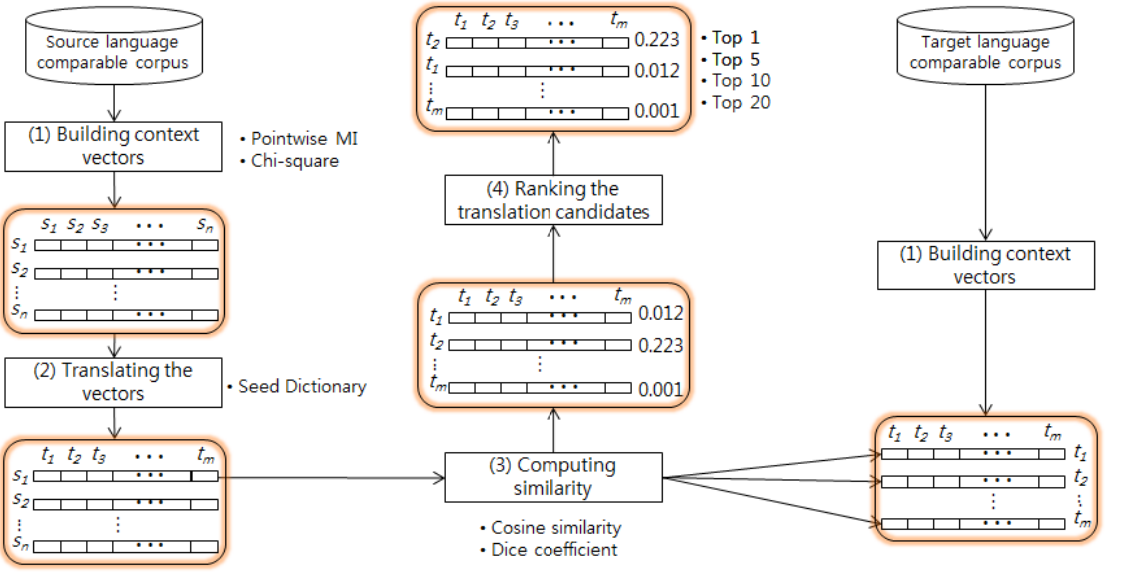

Figure 1: The concept of the standard approach for constructing a bilingual lexicon with a comparable corpus

tween the source language and the pivot language. The other is between the pivot language and the target language. By doing this, the proposed method is a much simpler than the standard approach (see Section 2). 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 works to get a large scale seed dictionary.

The remainder of this paper is organized as follows: Section 2 describes the standard approach as the related work. Section 3 represents the proposed method in detail and Section 4 presents the experimental result and some discussions. Finally, Section 5 draws conclusions and suggests directions for the future works.

2. Related work

Most of previous works **[1][9]-[11]** for extracting bilingual lexicons are based on lexical context analysis and depends on the assumption that a word and its translation are going to appear in the same lexical contexts **[12]**. This approach is called the context-based approach or the standard approach. The works are closely related to comparable corpora and large scale seed dictionaries. **Figure 1** shows the concept of the standard approach, which can be carried out by applying the following four steps **[9][13]**:

(1) **Context characterization**: To build context vectors for each word in source and target texts of comparable corpora, respectively. All words in the context vectors are weighted as word association measure like mutual information and the log-likelihood.

(2) **Context vector translation**: To translate the context vector represented by source words into that represented by target words by using a seed bilingual dictionary. If the bilingual dictionary provides several translations for a source word, we consider all of them but weighted the difference translations according to their frequency in the target language. The dimension of the translated vector is the same as that of the target vector.

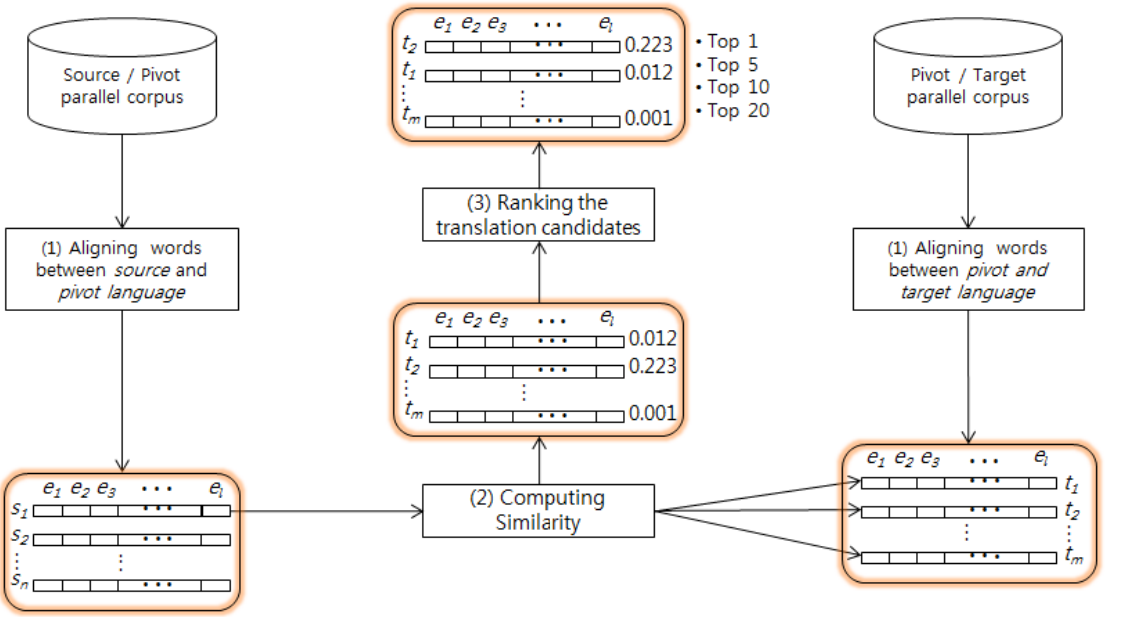

Figure 2: Overall structure of the proposed method.

(3) **Similarity calculation**: To calculate the similarity between each word represented in the source context vector and all words represented in the target context vectors through the vector distance measures such as cosine similarity and weighted Jaccard coefficient.

(4) **Candidate translation selection**: To select the candidate translation from the top *k* word pairs ranked following the similarity.

However, the accuracy of bilingual lexicon extraction via comparable corpora is quite poor **[4]**. Besides, large scale seed dictionaries are needed in order to improve the accuracy **[1][3]**. Alternatives to overcome these disadvantages are previously proposed by researchers **[8][14]**. The methods, nevertheless, have a flaw as known as the ambiguity problem in pivot words since it is to combine two independent bilingual lexicons into one using a pivot language.

3. Proposed method

3.1 Motivation

As mentioned before, the standard approach builds context vectors from two comparable corpora. Then, the source context vectors are translated into target language words using a seed dictionary. In such point of view, the seed dictionary is essentially required. For some language pairs, it is not easy to obtain the seed dictionary from public domains. Furthermore, it is difficult to build a parallel corpus and/or a comparable corpus between them. Thus, the standard approach can not be directly applied for less-known language pairs like Korean (KR) and Spanish (ES).

To overcome this problem, we present a new method for extracting a bilingual lexicon from two parallel corpora (KR-EN and EN-ES) sharing with a common pivot language (EN) instead of the seed dictionary.

3.2 Methodology

In this paper, we propose a new method for building bilingual lexicons between less-known language pairs by using a pivot language. The pivot language

is used for representing both of context vectors $\vec{s_i}$ and $\vec{t_j}$ of a source language and a target language. Unlike the previous studies using comparable corpora, we use two parallel corpora sharing the pivot language like KR-EN and EN-ES. The overall structure of the proposed method is depicted in **Figure** 2. The proposed method can be summarized in the following three steps:

(1) **Context characterization:** To build a source context vector $\vec{s_i}$ (resp. target vector $\vec{t_j}$) for each source word s_i (resp. target word t_j) in source texts (resp. targrt texts) of a source-pivot parallel corpus (resp. a pivot-target parallel corpus). Both of the context vectors $\vec{s_i}$ and $\vec{t_j}$ are represented as $(w_1, w_2, ..., w_m)^T$, where w_k is the association score between the source word s_i (resp. the target word t_j) and the k-th pivot word e_k and can be obtained from the two parallel corpora. As a result, the two vectors are comparable because the dimensions of $\vec{s_i}$ and $\vec{t_j}$ are same.

(2) **Similarity calculation**: This step is the same as the third step of the standard approach (see Section 2 and Figure 1).

(3) **Candidate translation selection**: This step is also the same as the fourth step of the standard approach (see Section 2 and Figure 1).

The proposed method in **Figure 2** is simplified by comparison with the standard approach in **Figure 1**. The proposed method does not require the step of the context vector translation any more and thus we do not use any seed dictionary as seen in **Figure 2**. If source and target context vectors are built once through each parallel corpus, the two context vectors can be compared with each other to get its similarity between them. You can read off the difference through **Figures** 1 and 2 at a glance. The main differences between the standard approach and the proposed method is that context vectors in the proposed method are represented by pivot words, while in the standard approach they are represented by target words translated by using the seed dictionary. The proposed method does not use any linguistic resources such as seed dictionaries except parallel corpora sharing a pivot language, which is a resource-rich language like English. We can obtain more accurate alignment information by using parallel corpora instead of comparable corpora.

4. Experiments and results

As discussed in Section 3.2, for each source word s_i (resp. target word t_j), we build a source context vector $\vec{s_i}$ (resp. target context vector $\vec{t_j}$) represented by pivot words e_k. To define the vector elements, we use two association measures: One is Chi-square (denoted hereafter as CHI-SQUARE) and the other one is the word translation probability (denoted hereafter as ANYMALIGN) estimated by Anymalign [15] which is a freely available word aligner. We conduct experiments on a language pair, Korean (KR) and Spanish (ES) and the pivot language is English (EN).

4.1 Experimental setting

4.1.1 Parallel corpora

We use two parallel corpora KR-EN and EN-ES. The KR-EN parallel corpus (433,151 sentence pairs) was compiled by Seo et al. (2006) [16] and the EN-ES parallel corpus is a sub-corpus (500,000 sentence pairs) that are randomly selected from the ES-EN parallel corpus in the Europarl parallel corpus [17]. The average number of words per sentence is described in **Table** 1. The number of words in ES-EN parallel corpus is nearly similar, but the number of KR words (called eojeol in Korean) in KR-EN parallel corpus is smaller 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.

Table 1: The average number of words per sentence

KR-EN		EN-ES	
KR	EN	EN	ES
19.2	31.0	25.4	26.4

4.1.2 Data preprocessing

All words in the parallel corpora are tokenized by the following tools: Hannanum[1] **[18]** for Korean, TreeTagger[2] **[19]** for English and Spanish. All words in English and Spanish are converted to lowercase and those in Korean are morphologically analyzed into morphemes and pos-tagged by Hannanum.

4.1.3 Building evaluation dictionary

To evaluate the performance of the proposed method, we build bilingual lexicons, KR-ES and ES-KR, 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. The frequent words are randomly selected from 50% in high 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.

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

Evaluation dictionary	The number of words
KR-ES	7.36
ES-KR	10.31

1) http://kldp.net/projects/hannanum
2) http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
3) http://dic.naver.com/

4.2 Results

For evaluating the proposed method, we have used the accuracy, which is the percentage of the test cases where the correct translation is found among the top k candidate words extracted by the proposed method. **Figure 3** and **4** show the accuracy of the

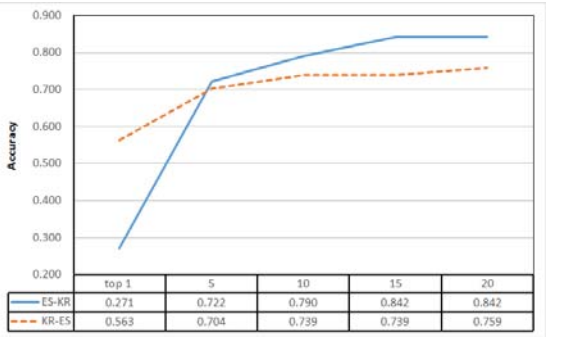

Figure 3: The accuracy of translation candidates extracted by CHI-SQUARE.

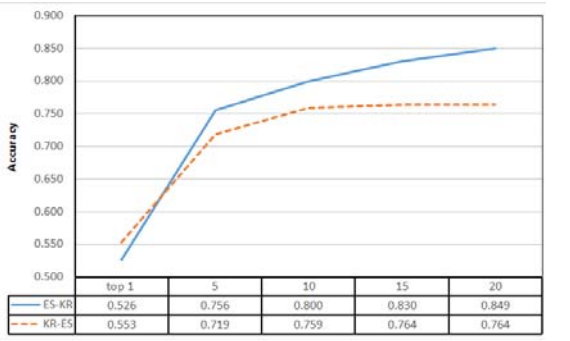

Figure 4: The accuracy of translation candidates extracted by ANYMALIGN.

translation candidates extracted by CHI-SQUARE and ANYMALIGN, respectively. we have observed that the shapes of the two graphs are very similar. At the top 5 below the slopes are steep and at the top 5 above the slopes are gentle. It indicates that most of correct translations lie below the top 5. We can consider translation candidates at the top 5 above to be quite rare. On the whole, we have shown that ANYMALIGN outperforms just a little bit

CHI-SQUARE, but there is no big difference between CHI-SQUARE and ANYMALIGN. At the top 1, CHI-SQUARE simply outperforms ANYMALIGN with KR-ES. As shown in figures, the ES-KR outperforms the KR-ES at the top 5 above. We have presumed that this is related to the degree of ambiguity, that is, the more the degree of ambiguity, the higher the accuracy of translation candidates at the top 5 above.

Furthermore, our experiment results surpassed Fung's result **[11]** (30% to 76% when top 1 to 20 translation candidates are considered). Surely, the experimental results in this paper come from different circumstances with the Fung's study. Nonetheless, the results are quite encouraging because Fung's study used linguistic resources such as a seed dictionary. Our experimental results have shown that the proposed method is quite attractive where a parallel corpus directly between source and target languages are unavailable, but both source-pivot and pivot-target parallel corpora are available.

5. Conclusions and future works

This paper proposed a new method for extracting a bilingual lexicon from two parallel corpora sharing an intermediary language as a pivot language. The proposed method is adapted from the context-based approach (called the standard approach) which uses context vectors. The main differences between the standard approach and the proposed method is that context vectors in the proposed method are represented by pivot words, while in the standard approach, they are represented by target words translated using the seed dictionary. The proposed method does not use any linguistic resources such as seed dictionaries except parallel corpora sharing with a pivot language, which is a resource-rich language like English.

Our experimental results show that the proposed method is quite attractive where public bilingual corpora between two languages are directly unavailable but public bilingual parallel corpora based on specific language such as English is available.

For the future works, multi-word expression should be handled and words with similar meaning should be clustered to improve the performance. Furthermore, bilingual lexicons built manually need to be fixed to have regular translation numbers for a fair evaluation (more translation candidates, more coverage).

Acknowledgements

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

References

[1] R. Rapp, "Automatic identification of word translations from unrelated English and German corpora", Proceedings of the Association for Computational Linguistics, pp. 519-526, 1999.

[2] D. Wu and X. Xia, "Learning an English-Chinese lexicon from a parallel corpus", Proceedings of the First Conference of the Association for Machine Translation in the Americas, pp. 206-213, 1994.

[3] P. Fung, "Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus", Proceedings of the Third Workshop on Very Large Corpora, pp. 173-183, 1995.

[4] K. Yu and J. Tsujii, "Bilingual dictionary extraction from Wikipedia", Proceedings of the 12th Machine Translation Summit, pp. 379-386, 2009.

[5] A. Ismail and S. Manandhar, "Bilingual lexicon extraction from comparable corpora using in-domain terms", Proceedings of the International Conference on Computational Linguistics, pp. 481-489, 2010.

[6] K. Tanaka and K. Umemura, "Construction of a bilingual dictionary intermediated by a third language", Proceedings of the 15th International

Conference on Computational Linguistics, pp. 297-303, 1994.

[7] H. Wu and H. Wang, "Pivot language approach for phrase-based statistical machine translation", Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, pp. 856-863, 2007.

[8] T. Tsunakawa, N. Okazaki, and J. Tsujii, "Building bilingual lexicons using lexical translation probabilities via pivot languages", Proceedings of the International Conference on Computational Linguistics, pp. 18-22, 2008.

[9] R. Rapp, "Identifying word translations in non-parallel texts", Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics, pp. 320-322, 1995.

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

[11] A. Hazem and E. Morin, "Adaptive dictionary for bilingual lexicon extraction from comparable corpora", Proceedings of the 8th International Conference on Language Resources and Evaluation pp. 288-292, 2012.

[12] G. Grefenstette, "Corpus-derived, first, second and third-order affinities, Proceedings of EuroLex, pp. 279-290, 1994.

[13] P. Fung and K. R. McKeown, "Finding terminology translations from non-parallel corpora", Proceedings of the 5th International Workshop of Very Large Corpora, pp. 192-202, 1997.

[14] F. Bond, R. Sulong, T. Yamazaki, and K. Ogura., "Design and construction of a machine-tractable Japanese-Malay dictionary", Proceedings of Machine Translation Summit VIII, pages 53-58.

[15] A. Lardilleux, Y. Lepage, and F. Yvon, "The contribution of low frequencies to multilingual sub-sentential alignment: a differential associative approach", International Journal of Advanced Intelligence, vol. 3, no. 2, pp. 189-217, 2011.

[16] H.-W. Seo, H.-C. Kim, H.-Y. Cho, J.-H. Kim and S.-W. Yang, "Automatically constructing English-Korean parallel corpus from web documents", Proceedings of the 26th KIPS Fall Conference, vol. 13, no. 2, pp. 161-164, 2006.

[17] P. Koehn, "Europarl: a parallel corpus for statistical machine translation", Proceedings of the conference on the 10th Machine Translation Summit, pp. 79-86, 2005.

[18] W. Lee, S. Kim, G. Kim and K. Choi, "Implementation of modularized morphological analyzer", Proceedings of the 11th Annual Conference on Human and Cognitive Language Technology, pp. 123-136, 1999.

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