Hindawi Publishing Corporation
Computational Intelligence and Neuroscience
Volume 2015, Article ID 434153, 13 pages
http://dx.doi.org/10.1155/2015/434153

Research Article

Evaluating a Pivot-Based Approach for Bilingual Lexicon Extraction

Jae-Hoon Kim, Hong-Seok Kwon, and Hyeong-Won Seo

Department of Computer Engineering, Korea Maritime and Ocean University, 727 Taejong-ro, Yeongdo-gu, Busan 606-791, Republic of Korea

Correspondence should be addressed to Jae-Hoon Kim; jhoon@kmou.ac.kr

Received 29 December 2014; Revised 26 March 2015; Accepted 9 April 2015

Academic Editor: Francesco Camastra

A pivot-based approach for bilingual lexicon extraction is based on the similarity of context vectors represented by words in a pivot language like English. In this paper, in order to show validity and usability of the pivot-based approach, we evaluate the approach in company with two different methods for estimating context vectors: one estimates them from two parallel corpora based on word association between source words (resp., target words) and pivot words and the other estimates them from two parallel corpora based on word alignment tools for statistical machine translation. Empirical results on two language pairs (e.g., Korean-Spanish and Korean-French) have shown that the pivot-based approach is very promising for resource-poor languages and this approach observes its validity and usability. Furthermore, for words with low frequency, our method is also well performed.

1. Introduction

Bilingual lexica are very important resources for bilingual works like machine translation, cross-language information retrieval, foreign language learning, terminology-related studies, and so on. Therefore, many researchers [1–4] have attracted bilingual lexicon extraction (BLE) and have proposed several approaches under various circumstances. The approaches can be broadly classified into three categories based on the different usage of language resources like corpora and dictionaries.

The first category is the simplest approach using a machine readable dictionary (MRD) [5]. This approach directly extracts them from MRDs or Web-based dictionaries like Wikitionary (http://www.wiktionary.org/) and Wikipedia (http://www.wikipedia.org/). It is very simple, but there are no MRDs in some language pairs (in particular, resource-poor language pairs like Korean-Bengali). Besides, many MRDs are not well formed like XML, although MDRs might be, and it is difficult to extract them from MDRs.

The second category is an approach using parallel corpora [6]. This approach extracts translation equivalents from parallel corpora based on word alignment models, which are well established [7, 8]. This approach is relatively high in the quality of extracted bilingual lexica if the size of parallel corpora is large enough [6]. Also this approach is very practical and easily applicable since there are several word alignment tools publicly available (http://web.eecs.umich.edu/~mihalcea/wa/). However, the effectiveness of the methods seems to be highly dependent on the availability of sizeable and high-quality parallel corpora. Also most parallel corpora are only provided for well-known language pairs like English and Spanish and the existing parallel corpora do not cover most domains [9].

The third category is an approach using comparable corpora and seed dictionaries [2, 3]. This approach extracts translation equivalents by comparing source context vectors and target context vectors which are made from comparable corpora. In fact dimensions of the two vectors are different from each other and source context vectors have to be translated in target languages using seed dictionaries in order that the dimensions agree. Recently this approach has been most widely used. Comparable corpora cause the accuracy of BLE to be lower than the ones obtained from parallel corpora of similar sizes [2] and the applicability of this approach depends on the size of the initial seed dictionary which

is needed for translating target context vectors [2, 10]. The size is important to achieve good applicability because the larger the size is, the higher the applicability is. Hence, some researchers [11] have studied extending the initial seed dictionary automatically. Since the quality of the extracted bilingual lexica is relatively sensitive to the type of corpora used in the extraction process, one would be naturally skeptical with the ideas of building a bilingual lexicon automatically using comparable corpora [9].

One common thing among those studies is that corpora like comparable and/or parallel corpora are essentials for building bilingual lexica. Unfortunately, for some language pairs like Korean and Bengali, it is not easy to get such corpora in the public domain. What was worse, it is difficult to build such corpora because it is very expensive and tedious to build for new language pairs parallel and/or comparable corpora needed for extracting the bilingual lexica. Just like in the corpora, there is the same difficulty for the initial seed dictionaries of such language pairs. In order to alleviate these problems, some researchers proposed pivot-based approach [12–17]. There are three different phases in pivot-based approaches. The first is to combine two bilingual dictionaries sharing with one common language as a pivot language, for example, Japanese-English and English-Chinese dictionaries (to build a Japanese-Chinese dictionary) [12, 17]. The second is to merge two phrase tables sharing with one common language as a pivot language [15]. The two phrase tables are extracted from two parallel corpora on the basis of phrase based statistical machine translation (SMT). The third is to build bilingual lexica based on the similarity of context vectors in a pivot language [13, 14]. The context vectors (called source and target context vectors) in this approach are represented by words in a pivot language and are produced using two parallel corpora sharing the pivot language, that is, source-pivot and pivot-target parallel corpora. We call this approach a pivot-based standard approach (PBSA). This paper is related to this approach, which will be described further in detail.

In this paper, we will evaluate the PBSA in company with two different methods for estimating context vectors represented by words in a pivot language (called pivot-based context vectors). The purpose of this paper is twofold: one is to show the validity of pivot-based context vectors, which are produced by the two estimation methods. The other is to show the usability of the PBSA in between resource-poor language pairs to extract bilingual lexica.

The paper is organized as follows: Section 2 describes the related works of the standard approach and the PBSA, from which our works are derived. Then, Section 3 presents two estimation methods of context vectors and Section 4 describes some experiments and discussions on Korean-French and Korean-Spanish lexicon extraction. Finally, Section 5 draws conclusions and discusses some perspectives on our future works.

2. Related Works

There are many works focused on extracting bilingual lexica from comparable corpora [3, 18–20]. Most of them are based on the context-based approach, so called the standard approach, proposed by Rapp [10]. In this section, first of all, we describe the standard approach (SA) in short and then the pivot-based standard approach (PBSA) proposed by Kim et al. [13, 14].

2.1. Standard Approach. The basic assumption of the standard approach is distributional hypothesis [21], which states that words with a similar meaning are likely to appear in similar context across languages. Therefore, a word can be represented as a context vector of similar words. Under BLE from comparable corpora, there are two context vectors, source and target context vectors for source and target words, respectively. Each element in the context vectors represents its association with a word which occurs within a window of words. The two vectors, however, are incomparable because the dimensions of them are totally different from each other. That is, one is represented by words in a source language and the other in a target language. In order to enable the comparison of source and target context vectors, words in the source context vectors (or target context vectors) are translated into the target language (or source language) using a seed bilingual dictionary between the source language and the target language. A vector similarity like cosine similarity and Jarcard coefficient is used and target words with the highest vector similarity are treated as translation candidates. In summary, the standard approach can be carried out by applying the following four steps:

(1) context characterization: to build source and target context vectors for each word of source and target languages, respectively. All words in the context vectors can be weighted with an association measure like χ^2 score,

(2) context vector translation: to translate words in a source context vector in a target language using a seed dictionary,

(3) similarity calculation: to compute similarity between the translated context vector and all target context vectors through vector distance measures such as cosine similarity,

(4) candidate translation selection: to rank translation candidates according to the similarity score.

The advantage of the standard approach is that it is a fast and affordable way to construct bilingual lexica in resource-poor language pairs and new domains [22]. However, it also presupposes the availability of a seed bilingual dictionary to translate context vectors, which is not the case for many language pairs or domains. Hence, some researchers had extended the seed dictionary automatically. Comparable corpora are also essentials. Unfortunately, for some language pairs like Korean and Bengali, it is not easy to get such corpora in the public domain.

2.2. Pivot-Based Standard Approach. As mentioned in the previous section, there are many advantages in the standard approach, but the standard approach still has some problems

on resource-poor language pairs like Korean and Bengali. To overcome these problems, the PBSA was proposed by Kim et al. [13, 14], with adapting from the standard approach. Figure 1 is the overall structure of the PBSA, where s_i, t_i, and e_i are the ith word in a source language, a target language, and a pivot language, respectively, and n, m, and l are also the number of words in a source language, a target language, and a pivot language, respectively. As shown in Figure 1, unlike the standard approach mentioned before, there are three steps: (1) *context characterization*, (2) *similarity calculation*, and (3) *candidate translation selection*. That is, there is no step of context vector translation using a seed dictionary. The second step of similarity calculation and the third step of candidate translation selection are the same as those of the standard approach, but the first step of context characterization is totally different from that of the standard approach. In the PBSA, context vectors are made up of words in a pivot language as a bridge language, which should be very popular like English. We call the words "pivot words" hereafter. That is, both of source and target context vectors are represented by pivot words instead of their own words. As a result, they have the same dimensions and can be compared with each other to get similarity between them. Now the remaining problem is how to get the context vectors. To find out probable pivot words for a source word (resp., target word), in this paper, we use two parallel corpora, a source-pivot parallel corpus (SL-PL parallel corpus) and a pivot-target parallel corpus (PL-TL parallel corpus). If the parallel corpora could be publicly available, pivot-based context vectors should be easily estimated through word alignment [9, 23, 24]. In this paper, we evaluate two estimation methods of the context vectors, which are described in the next section.

Table 1 shows the difference between the standard approach and the PBSA. The PBSA has much strength in terms of a seed dictionary, context vector translation, and computational efficiency, but some weaknesses as to domain adaptation and corpus preparation. The PBSA does not use any linguistic resources such as seed dictionaries except parallel corpora sharing a pivot language. We can obtain more accurate alignment information by using parallel corpora instead of comparable corpora.

3. Estimation Methods of Context Vectors

In the PBSA using two parallel corpora, SL-PL and PL-TL, estimating a context vector for a given source word (resp., target word) is the same as aligning source words (resp., target word) into pivot words in the SMT. If a bilingual dictionary between a source language (resp., target language) and a pivot language is publicly available, it could be used for estimating the context vector. Generally it is not simple to get such a bilingual dictionary and also to build it from an MRD because the MRD is not well formatted in a markup language like XML. Accordingly, most of works on word alignments are based on cooccurrence [25, 26]. We evaluate two methods for estimating context vectors based on word cooccurrence. One is based on a word alignment tool, which is publicly available and can produce translation probabilities. The other is based on word association between source words (resp., target words) and pivot words in an SL-PL parallel corpus (resp., PL-TL parallel corpus) through a contingency table.

3.1. Word Alignment Tools. One of the word alignment tools like Moses (http://www.statmt.org/moses/) and Anymalign (https://sites.google.com/site/adrienlardilleux/) can be used for estimating a source context vector (resp., target context vector) for each word s_i (resp., t_i) in a source language (resp., target language) using an SL- PL (resp., PL-TL) parallel corpus. In this paper, we use Anymalign [23] as a word alignment tool. For a parallel corpus, Anymalign can produce bidirectional translation probabilities, that is, $\Pr(e_k \mid s_i)$ and $\Pr(s_i \mid e_k)$, where s_i and e_k are the ith word and the kth word in the source language and the pivot language, respectively. The size of context vectors estimated in such a way is very large because of some noises like improper translations. In order to reduce the noises and improve the reliability of the context vectors, consider Table 2 partial examples of translation probabilities between Korean words as source words and English words as pivot words.

In Table 2, the first row is a good example, but the rest of rows are not. In case of the second row, it is much less possible that s_i and e_k may be translation equivalents because the difference between two probabilities is too big. In case of the third row, it is also improper translation because the two probabilities are too low. In summary, an association between s_i and e_k, $a(s_i, e_k)$ can be estimated as follows:

$$a(s_i, e_k) = \begin{cases} \dfrac{\Pr(e_k \mid s_i) + \Pr(s_i \mid e_k)}{2} & \text{if } \Pr(e_k \mid s_i) > \theta_1,\ \Pr(s_i \mid e_k) > \theta_1,\ \left|\Pr(e_k \mid s_i) - \Pr(s_i \mid e_k)\right| < \theta_2 \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where θ_1 and θ_2 are constants as thresholds.

3.2. Contingency Table. Measures of bilingual word association are used to determine if a word in one language and another word in a second language are dependent on one another or not in a parallel corpus [27, 28]. In our case, the two words can be a source word s_i (resp., a target word t_j) and a pivot word e_k. To measure the dependence of these two words, we determined their frequency counts using a simple statistical model. We considered the two words to be represented by binary random variables that simply indicate if a word occurred or not in their corresponding pieces. A word pair can fall into one of the four possible categories, and we can represent the associated count data using a two

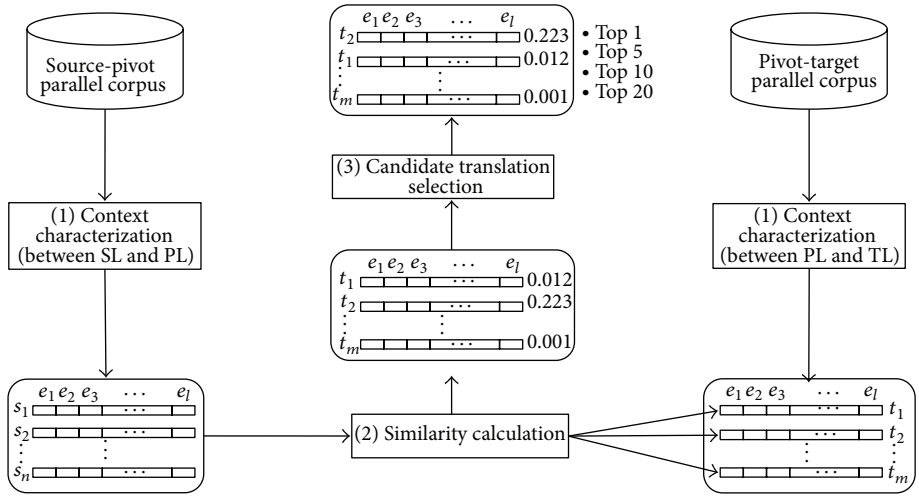


Figure 1: Overall structure of the PBSA for BLE (adapted from the paper [13]).

Table 1: Comparisons between the standard approach and the PBSA.

Category	Standard	Pivot-based
Corpora for estimating models	Two comparable	Two parallel
Representation of source context vectors	Source words	Pivot words
Representation of target context vectors	Target words	Pivot words
Context vector translation	Required	Not required
Seed dictionary	Required	Not required
Domain adaptation	Relatively easy	Not easy

by two contingency table. Table 3 shows an example of a 2 by 2 contingency table for a Korean word s_i = *kyeong-chal* and an English word e_k = *police*. Measure of association in finding word translations was proposed by Gale and Church [29], who employed the φ-coefficient and pointwise mutual information as their measures of association. Since then, several word association measures have been proposed in the literature [26]. In this paper, we make use of χ^2 score defined as (2) through the 2 by 2 contingency table:

$$\chi^2 = \frac{\left[(n_{11} \times n_{22}) - (n_{12} \times n_{21})\right]^2 \times n_{++}}{n_{1+} \times n_{2+} \times n_{+1} \times n_{+2}}. \quad (2)$$

The χ^2 value is 24.78 for Table 3 and the critical value is $\chi^2 = 3.841$ (you can find the χ^2 distribution table from http://easycalculation.com/statistics/chisquare-table.php). So we reject the null hypothesis that *kyeong-chal* and *police* occur independent of each other. That is, the Korean word *kyeong-chal* and the English *police* can be a good translation candidate.

4. Experiments and Evaluation

In order to evaluate the performance of the PBSA with the two estimation methods of context vectors, we perform experiments over two language pairs, Korean-Spanish (KR-ES) and Korean-French (KR-FR). In this section, we first give experimental setups such as parallel corpora and evaluation dictionaries, and then we present the results of experiments conducted on the two language pairs. Finally we analyze some errors and discuss some issues on BLE through these results.

4.1. Experimental Environments

4.1.1. Parallel Corpora. We need three parallel corpora, KR-EN, EN-ES, and EN-FR, since we use English (EN) as a pivot language. In the case of the KR-EN parallel corpus, we use the KMU Parallel Corpus [30], which is freely available on the Web (https://sites.google.com/site/nlpatkmu/Resources/Corpora). The parallel corpus was built in 2006 and consists of 433,151 sentence pairs of KR-EN. In the case of the EN-ES and EN-FR parallel corpora, we use the Europarl Parallel Corpus [31], which is also freely available on the Web (http://www.statmt.org/europarl/). Actually we use only subcorpora to keep balance with the KR-EN parallel corpus in size. The two sets of subcorpora of EN-ES and EN-FR consist of 500,000 sentence pairs each that are randomly selected from the original corpus of the Europarl Parallel Corpus. To tell the truth, the KMU Parallel Corpus and the Europarl Parallel Corpus are built from different domains (news article and European Parliament proceedings). The average number of words per sentence in the three parallel corpora is shown in Table 4.

As you can see in Table 1, word distributions are similar except those of KR-EN. The number of Korean words

Table 2: The partial output of Anymalign for Korean as a source language and English as a pivot language.

Number	Source word, s_i	Pivot word, e_k	$\Pr(e_k \mid s_i)$	$\Pr(s_i \mid e_k)$	Remarks
(1)	*kyeong-chal (police)*[a]	*police*	0.890421	0.877882	Good
(2)	*kyeong-chal (police)*	*scenario*	1.000000	0.000167	Bad
(3)	*jeong-bu (government)*	*balance*	0.000004	0.000506	Bad

[a]In the second column, Korean words are romanized and are composed of syllables, which are separated by hyphens (-). Words in the parenthesis are their meanings. For example, the meaning of Korean word *kyeong-chal* is *police*.

Table 3: An example of a 2 by 2 contingency table for a bilingual word pair, the Korean word s_i = *kyeong-chal* and the English word e_k = *police*.

	e_k = *police*	$e_k \neq$ *police*	Total
s_i = *kyeong-chal*	n_{11} = 80	n_{12} = 4,667	n_{1+} = 4,747
$s_i \neq$ *kyeong-chal*	n_{21} = 4,510	n_{22} = 460,123	n_{2+} = 464,633
Total	n_{+1} = 4,590	n_{+2} = 464,790	n_{++} = 469,930

n_{11} is the number of times *kyeong-chal* and *police* appears in the parallel corpus.
n_{12} is the number of times *kyeong-chal* appears in the Korean text and *police* does not occur in the English text.
n_{21} is the number of times *kyeong-chal* does not appear in the Korean text and *police* occurs in the English text.
n_{22} is the number of times *kyeong-chal* does not appear in the Korean text and *police* does not occur in the English text.
n_{1+} is the number of times *kyeong-chal* appears in the Korean text.
n_{2+} is the number of times *kyeong-chal* does not appear in the Korean text.
n_{+1} is the number of times *police* appears in the English text.
n_{+2} is the number of times *police* does not appear in the English text.

(called Eojeol in Korean) in KR-EN corpus is lower than others. This is caused by the Korean characteristic that Korean words usually have one morpheme or more (average number of morphemes per word = 2.1, but varies with a corpus). Considering Korean morphemes, the number could be similar to that of EN.

4.1.2. Preprocessing. All words are tokenized and tagged in POS (part-of-speech) by the following tools: Hannanum (http://kldp.net/projects/hannanum) [32] for Korean and TreeTagger (http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger) [28] for English, Spanish, and French. Moreover, all words except content words (nouns, main verbs, adjectives, and adverbs) are removed as stop words and then are converted to lower case.

4.1.3. Evaluation Dictionary. For the evaluation, we use the KMU Evaluation Dictionary for BLE, which is freely available on the Web (https://sites.google.com/site/nlpatkmu/Resources/Lexicons). The dictionary was manually built using the Web dictionary (http://dic.naver.com/) and contains four bilingual lexica, KR-ES, KR-FR, ES-KR, and FR-KR. Each lexicon is unidirectional and contains 200 frequent words (denoted as HIGH) and 200 rare words (denoted as LOW). The frequent words and the rare words in the lexica are randomly sampled from three parallel corpora (mentioned in the previous section) based on the frequency of source words, respectively. Their translation equivalence distribution is described in Table 5. Those are considered as to the degree of ambiguity. Overall, the degree of ambiguity for Korean as a source language is lower than that as a target language without respect to the frequency of HIGH or LOW.

Table 4: The average number of words per sentence in the three parallel corpora.

KR-EN		EN-ES		EN-FR	
KR	EN	EN	ES	EN	FR
19.2	31.0	25.4	26.4	27.1	29.7

Table 5: The average number of the translations per source word in the evaluation dictionaries.

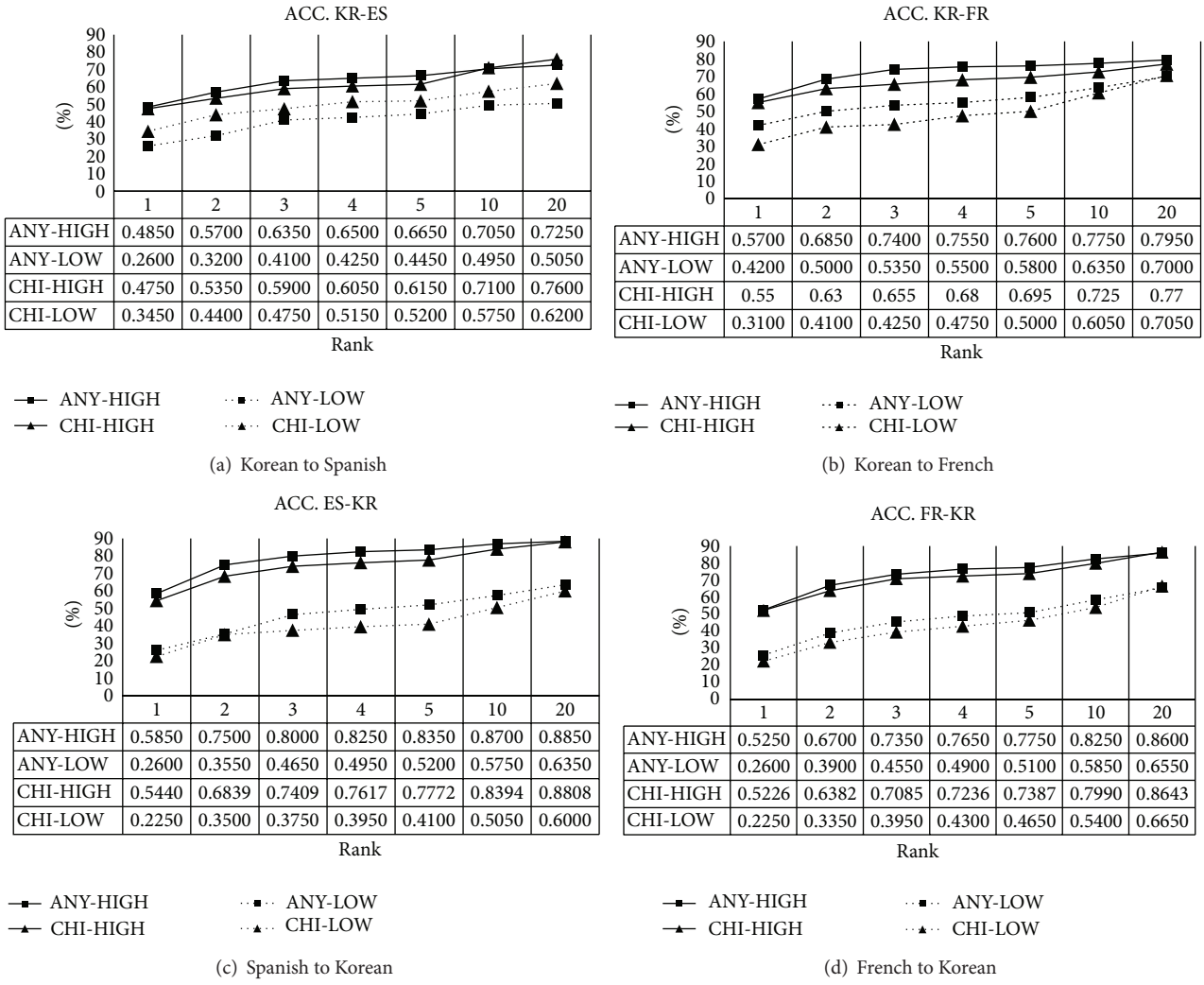
Evaluation dictionary	HIGH	LOW
KR-ES	5.79	2.26
KR-FR	7.36	3.12
ES-KR	10.31	5.46
FR-KR	10.42	6.32

4.2. Performance Evaluation. In this section, we discuss the performance of the two estimation methods, a word alignment tool of Anymalign (ANY) and a contingency table for word association of χ^2 (CHI), which were explained in Section 3.

Each estimation method has model parameters. For ANY, there are two model parameters, θ_1 and θ_2. θ_1 is fixed to 0.5 for both HIGH and LOW and θ_2 is set to 0.005 for HIGH and 0.003 for LOW through several experiments. On the other hand, for CHI, a model parameter is the critical value of χ^2 as 3.841 (the α level of significance is 0.05 and the degree of freedom is 1 on the the χ^2 distribution table.) as mentioned in Section 3.2.

To evaluate the quality of translation candidates extracted by systems, we use accuracy (ACC) [33], mean reciprocal rank (MRR) [34], precision (PRE) [35], recall (REC) [35], and rated recall (RRC) [36] as evaluation metrics. ACC and MRR show the performance for at least one correct answer, while PRE, REC, and RRC present the performance for multiple correct answers. The formal definition of each evaluation metrics will be described in subsequent sections.

4.2.1. Performance of Accuracy. Generally, accuracy (ACC) is the fraction of their translation candidates that are correct and we use top-k ACC, that is, the proportion of source words which have at least one translation equivalent among top k of



ACC. KR-ES	1	2	3	4	5	10	20
ANY-HIGH	0.4850	0.5700	0.6350	0.6500	0.6650	0.7050	0.7250
ANY-LOW	0.2600	0.3200	0.4100	0.4250	0.4450	0.4950	0.5050
CHI-HIGH	0.4750	0.5350	0.5900	0.6050	0.6150	0.7100	0.7600
CHI-LOW	0.3450	0.4400	0.4750	0.5150	0.5200	0.5750	0.6200

(a) Korean to Spanish

ACC. KR-FR	1	2	3	4	5	10	20
ANY-HIGH	0.5700	0.6850	0.7400	0.7550	0.7600	0.7750	0.7950
ANY-LOW	0.4200	0.5000	0.5350	0.5500	0.5800	0.6350	0.7000
CHI-HIGH	0.55	0.63	0.655	0.68	0.695	0.725	0.77
CHI-LOW	0.3100	0.4100	0.4250	0.4750	0.5000	0.6050	0.7050

(b) Korean to French

ACC. ES-KR	1	2	3	4	5	10	20
ANY-HIGH	0.5850	0.7500	0.8000	0.8250	0.8350	0.8700	0.8850
ANY-LOW	0.2600	0.3550	0.4650	0.4950	0.5200	0.5750	0.6350
CHI-HIGH	0.5440	0.6839	0.7409	0.7617	0.7772	0.8394	0.8808
CHI-LOW	0.2250	0.3500	0.3750	0.3950	0.4100	0.5050	0.6000

(c) Spanish to Korean

ACC. FR-KR	1	2	3	4	5	10	20
ANY-HIGH	0.5250	0.6700	0.7350	0.7650	0.7750	0.8250	0.8600
ANY-LOW	0.2600	0.3900	0.4550	0.4900	0.5100	0.5850	0.6550
CHI-HIGH	0.5226	0.6382	0.7085	0.7236	0.7387	0.7990	0.8643
CHI-LOW	0.2250	0.3350	0.3950	0.4300	0.4650	0.5400	0.6650

(d) French to Korean

Figure 2: Accuracy of k best candidates with different ranks.

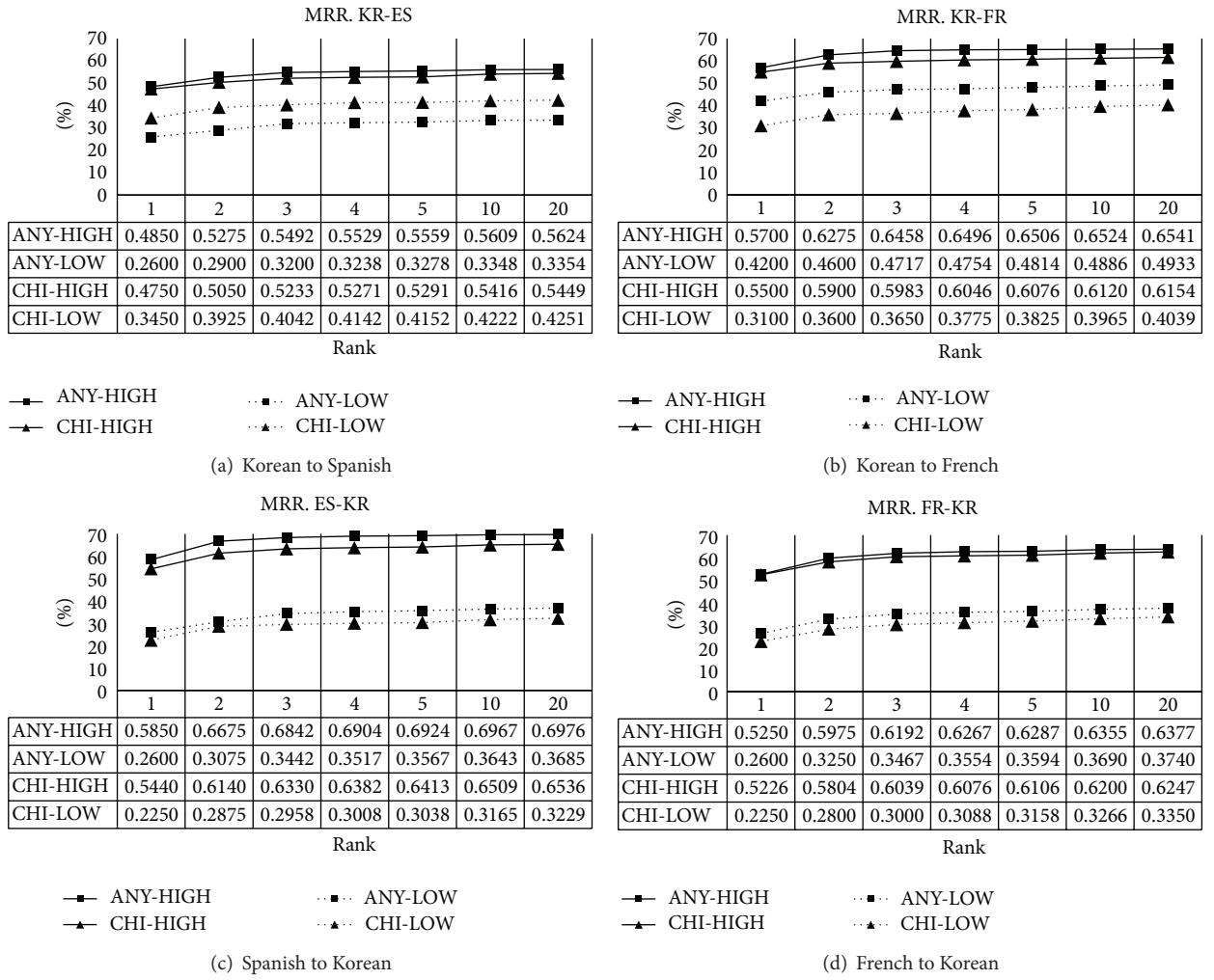
its induced translation candidates. More formally, top-k ACC, ACC_k is defined as

$$\mathrm{ACC}_k = \frac{1}{N}\sum_{i=1}^{N}\max_{1\le j\le k} a_{ij}, \quad \text{where } a_{ij} = \begin{cases} 1 & \text{if } t_{ij} \in A_i \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where N is the number of source words which are evaluated, A_i is a set of translation equivalents for a source word s_i, and t_{ij} for s_i is the jth translation candidate, which is induced by a BLE system.

Basically, we can show the performance for at least one correct answer through ACC as mentioned before. The results are shown in Figure 2. Overall, ANY a little outperforms CHI in accuracy, but the two methods show similar performance at the top 20. Although it is different in language pair, the minimum and maximum accuracy for HIGH is 48.5% and 58.5% at the top 1, respectively. This means that our system finds the correct translation equivalents for half of high frequent words in a source language at the top 1. As for the ACC in the top 5 for HIGH, the minimum and maximum accuracy is 66.5% and 83.5%, respectively. According to this result, most of the dominating translation equivalents for words are found in the top 1 through 5. Consider the performance in the case of LOW. ACC for LOW is much lower than that of HIGH as you can predict. The maximum at the top 1, however, is 42% in the case of KR-FR. This states that it is not bad even if a word is rare. We believe that the most important thing in BLE is that a system has good performance for rare words as the so-called *hapax legomena*.

4.2.2. Performance of Mean Reciprocal Rank. Mean reciprocal rank (MRR) is derived from question answering [34] and the average of the reciprocal ranks of translation candidates that are correct translations for source words. This definition does not specify what to do if none of the proposed results are correct (use mean reciprocal rank 0), or if there are multiple correct answers in the list (use the best one). Consequently MRR accentuates translation candidates at higher ranks more than others at lower ranks so that the correct translation candidates at higher ranks could be treated



	1	2	3	4	5	10	20
ANY-HIGH	0.4850	0.5275	0.5492	0.5529	0.5559	0.5609	0.5624
ANY-LOW	0.2600	0.2900	0.3200	0.3238	0.3278	0.3348	0.3354
CHI-HIGH	0.4750	0.5050	0.5233	0.5271	0.5291	0.5416	0.5449
CHI-LOW	0.3450	0.3925	0.4042	0.4142	0.4152	0.4222	0.4251

(a) Korean to Spanish

	1	2	3	4	5	10	20
ANY-HIGH	0.5700	0.6275	0.6458	0.6496	0.6506	0.6524	0.6541
ANY-LOW	0.4200	0.4600	0.4717	0.4754	0.4814	0.4886	0.4933
CHI-HIGH	0.5500	0.5900	0.5983	0.6046	0.6076	0.6120	0.6154
CHI-LOW	0.3100	0.3600	0.3650	0.3775	0.3825	0.3965	0.4039

(b) Korean to French

	1	2	3	4	5	10	20
ANY-HIGH	0.5850	0.6675	0.6842	0.6904	0.6924	0.6967	0.6976
ANY-LOW	0.2600	0.3075	0.3442	0.3517	0.3567	0.3643	0.3685
CHI-HIGH	0.5440	0.6140	0.6330	0.6382	0.6413	0.6509	0.6536
CHI-LOW	0.2250	0.2875	0.2958	0.3008	0.3038	0.3165	0.3229

(c) Spanish to Korean

	1	2	3	4	5	10	20
ANY-HIGH	0.5250	0.5975	0.6192	0.6267	0.6287	0.6355	0.6377
ANY-LOW	0.2600	0.3250	0.3467	0.3554	0.3594	0.3690	0.3740
CHI-HIGH	0.5226	0.5804	0.6039	0.6076	0.6106	0.6200	0.6247
CHI-LOW	0.2250	0.2800	0.3000	0.3088	0.3158	0.3266	0.3350

(d) French to Korean

Figure 3: Mean reciprocal rank of k best candidates with different ranks.

as more important. MRR at the top k, MRR_k, is formally defined as

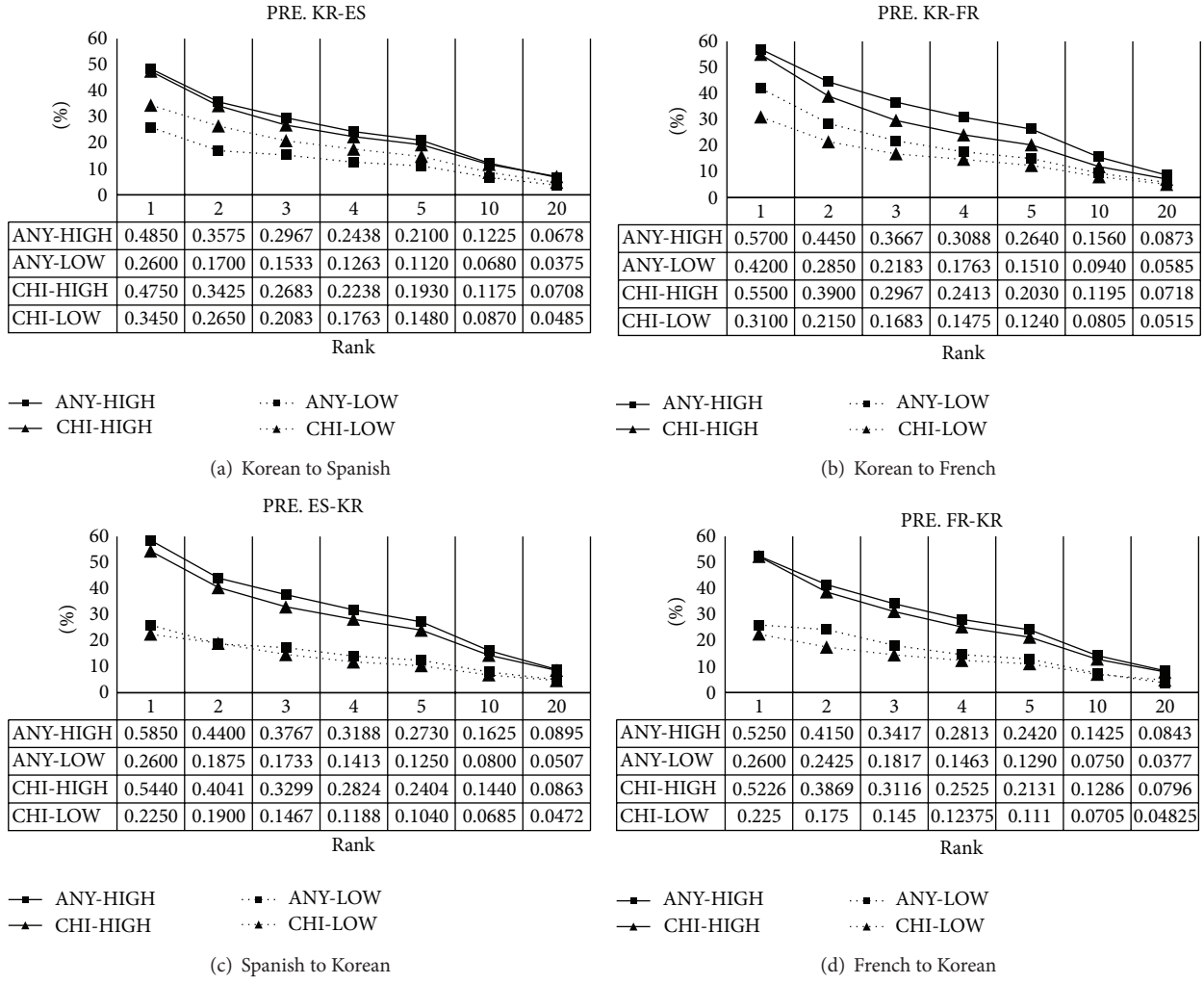
$$\mathrm{MRR}_k = \frac{1}{N}\sum_{i=1}^{N}\max_{1\le j\le k} r_{ij}, \quad \text{where } r_{ij} = \begin{cases} \frac{1}{j} & \text{if } t_{ij} \in A_i \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Basically, like ACC, MRR also reveals the performance for at least one correct answer, but unlike ACC, MRR prefers correct answers at the lower rank (i.e., the top 1) to those at the higher rank. The results are shown in Figure 3. ANY also outperforms CHI in MRR because MRR shows the same characteristics with ACC in general. Consider the slopes of graphs minutely. For all over the graphs, the slopes are steep in the top 5 below and the slopes are gentle in the top 5 above. 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. In this respect, we can tell that the PBSA is very promising.

4.2.3. Performance of Precision. Precision (PRE) (also called positive predictive value) is widely used in information retrieval and is the fraction of extracted translation candidates contained in the evaluation dictionary. Unlike ACC and MRR, PRE can show the performance for multiple correct answers. Formally, PRE at the top k, PRE_k, is defined as

$$\mathrm{PRE}_k = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{k}\sum_{j=1}^{k} a_{ij}, \quad \text{where } a_{ij} = \begin{cases} 1 & \text{if } t_{ij} \in A_i \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

The results are shown in Figure 4. For HIGH, ANY outperforms CHI generally like ACC, but the gap of the performance is not big. Same as in ACC, the minimum and the maximum PREs for HIGH are 48.5% and 58.5% at the top 1, respectively. This means that our system finds the correct translation equivalents for half of high frequent words in a source language in the top 1. The minimum and maximum PREs, however, are 6.78% and 8.95% at the top 20, respectively. This indicates that correct translation equivalents comprise about 7% of extracted translation candidates at the top 20. At a glance, the performance is very low. See the number of ambiguity as the number of multiple answers in Table 4. The average number is in between 5 and 11 and then



PRE. KR-ES	1	2	3	4	5	10	20
ANY-HIGH	0.4850	0.3575	0.2967	0.2438	0.2100	0.1225	0.0678
ANY-LOW	0.2600	0.1700	0.1533	0.1263	0.1120	0.0680	0.0375
CHI-HIGH	0.4750	0.3425	0.2683	0.2238	0.1930	0.1175	0.0708
CHI-LOW	0.3450	0.2650	0.2083	0.1763	0.1480	0.0870	0.0485

(a) Korean to Spanish

PRE. KR-FR	1	2	3	4	5	10	20
ANY-HIGH	0.5700	0.4450	0.3667	0.3088	0.2640	0.1560	0.0873
ANY-LOW	0.4200	0.2850	0.2183	0.1763	0.1510	0.0940	0.0585
CHI-HIGH	0.5500	0.3900	0.2967	0.2413	0.2030	0.1195	0.0718
CHI-LOW	0.3100	0.2150	0.1683	0.1475	0.1240	0.0805	0.0515

(b) Korean to French

PRE. ES-KR	1	2	3	4	5	10	20
ANY-HIGH	0.5850	0.4400	0.3767	0.3188	0.2730	0.1625	0.0895
ANY-LOW	0.2600	0.1875	0.1733	0.1413	0.1250	0.0800	0.0507
CHI-HIGH	0.5440	0.4041	0.3299	0.2824	0.2404	0.1440	0.0863
CHI-LOW	0.2250	0.1900	0.1467	0.1188	0.1040	0.0685	0.0472

(c) Spanish to Korean

PRE. FR-KR	1	2	3	4	5	10	20
ANY-HIGH	0.5250	0.4150	0.3417	0.2813	0.2420	0.1425	0.0843
ANY-LOW	0.2600	0.2425	0.1817	0.1463	0.1290	0.0750	0.0377
CHI-HIGH	0.5226	0.3869	0.3116	0.2525	0.2131	0.1286	0.0796
CHI-LOW	0.225	0.175	0.145	0.12375	0.111	0.0705	0.04825

(d) French to Korean

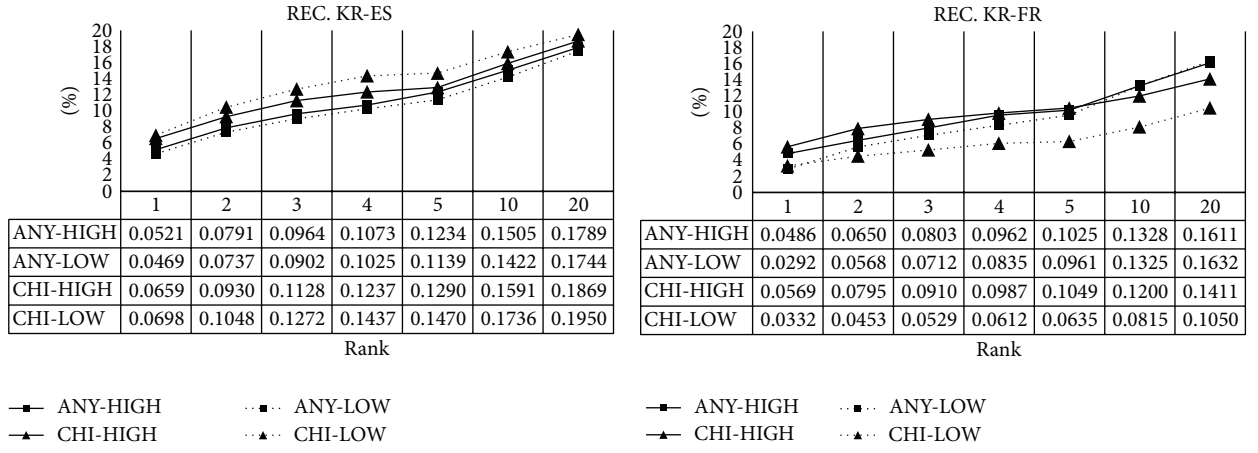
Figure 4: Precision of k best candidates with different ranks.

false alarms in many cases cannot help turning on among the top 20. For LOW, CHI outperforms ANY for KR-ES, but ANY outperforms CHI for the rest. The reason has not been figured out so far, so we need to examine the strange phenomenon in fact.

4.2.4. Performance of Recall. Recall (REC) (also known sensitivity) is also widely used in information retrieval and is defined as the fraction of translation equivalents that the system extracted. Unlike ACC and MRR, REC can show the performance for multiple correct answers like PRE. Formally, REC at the top k, REC_k, is defined as

$$\mathrm{REC}_k = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|A_i|}\sum_{j=1}^{k}a_{ij}, \quad \text{where } a_{ij} = \begin{cases} 1 & \text{if } t_{ij} \in A_i \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

The results are shown in Figure 5. CHI outperforms ANY in REC generally. This is caused by the size of context vectors and the size of context vectors for CHI is generally larger than that for ANY. This phenomenon tells us that CHI is favorable to multiple answers. Also as for REC, slightly different inclination is shown as in Figure 5(a). The performance of LOW is not lower than that of HIGH. Generally words in HIGH are very ambiguous, which means that the words have many translation equivalents in the evaluation dictionary (see Table 5). This has been revealed by Figure 5. On the whole, the performance is quite low from about 4% to 7% at the top 1 and from about 10% to 20% at the top 20. To tell the truth, most of translation equivalents in the evaluation dictionary are not used as the meaning in the corpora. For example, translation equivalents for the Korean word *jeonryak* are *strategy*, *tactic*, and *stratagem* in English according to a machine-readable dictionary of Korean-English, but the first one is mainly used in the real text (the frequency of *strategy*, *tactic*, and *stratagem* is 6026, 415, and 48, resp., in the BYU-BNC (http://corpus.byu.edu/bnc/)). Consider the slopes of graphs minutely. For all over the graphs, the gradients are continuously increased although the magnitude

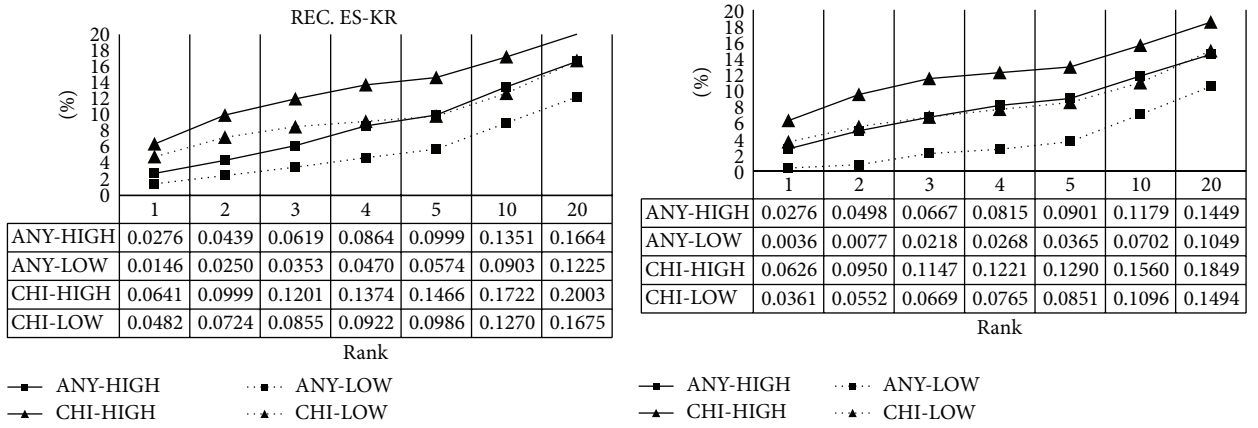


	1	2	3	4	5	10	20
ANY-HIGH	0.0521	0.0791	0.0964	0.1073	0.1234	0.1505	0.1789
ANY-LOW	0.0469	0.0737	0.0902	0.1025	0.1139	0.1422	0.1744
CHI-HIGH	0.0659	0.0930	0.1128	0.1237	0.1290	0.1591	0.1869
CHI-LOW	0.0698	0.1048	0.1272	0.1437	0.1470	0.1736	0.1950

(a) Korean to Spanish

	1	2	3	4	5	10	20
ANY-HIGH	0.0486	0.0650	0.0803	0.0962	0.1025	0.1328	0.1611
ANY-LOW	0.0292	0.0568	0.0712	0.0835	0.0961	0.1325	0.1632
CHI-HIGH	0.0569	0.0795	0.0910	0.0987	0.1049	0.1200	0.1411
CHI-LOW	0.0332	0.0453	0.0529	0.0612	0.0635	0.0815	0.1050

(b) Korean to French



	1	2	3	4	5	10	20
ANY-HIGH	0.0276	0.0439	0.0619	0.0864	0.0999	0.1351	0.1664
ANY-LOW	0.0146	0.0250	0.0353	0.0470	0.0574	0.0903	0.1225
CHI-HIGH	0.0641	0.0999	0.1201	0.1374	0.1466	0.1722	0.2003
CHI-LOW	0.0482	0.0724	0.0855	0.0922	0.0986	0.1270	0.1675

(c) Spanish to Korean

	1	2	3	4	5	10	20
ANY-HIGH	0.0276	0.0498	0.0667	0.0815	0.0901	0.1179	0.1449
ANY-LOW	0.0036	0.0077	0.0218	0.0268	0.0365	0.0702	0.1049
CHI-HIGH	0.0626	0.0950	0.1147	0.1221	0.1290	0.1560	0.1849
CHI-LOW	0.0361	0.0552	0.0669	0.0765	0.0851	0.1096	0.1494

(d) French to Korean

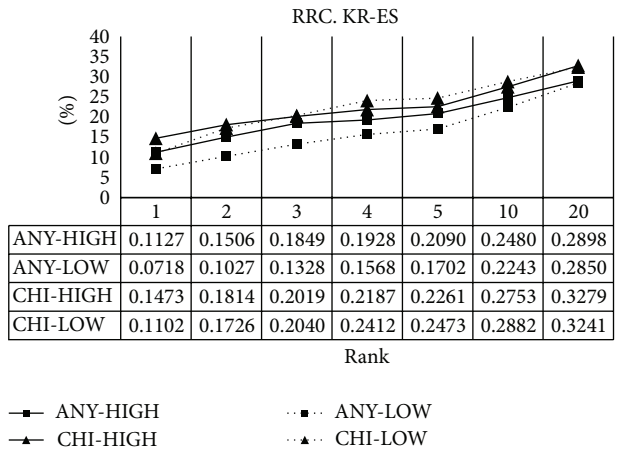
Figure 5: Recall of k best candidates with different ranks.

is low. It indicates that the higher the rank is the more the correct translation equivalents are found.

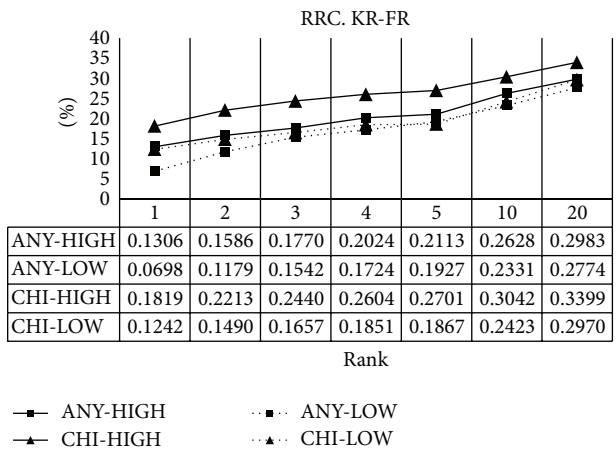
4.2.5. Performance of Rated Recall. First of all, consider an evaluation dictionary as a gold standard for evaluating a BLE system. Generally an entry of the dictionary is composed of a pair of source words and one or more translation equivalents (target words). Any translation equivalents do not appear in the training corpus at all and some translation equivalents account for a great part of the training corpus. In the case of the former, there is no way of extracting them under BLE. In the case of the latter, it is important that the high frequent translation equivalents should be found as soon as possible. That is, the frequency of translation equivalents in the training corpus is not reflected at all in performance evaluation so far. In order to supplement such problems, rated recall (RRC) is proposed by Seo et al. [36]. RRC is the proportion of extracted translation equivalents in the training corpus. Formally, RRC at the top k, RRC_k, is defined as

$$\mathrm{RRC}_k = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{k} a_{ij} r\left(t_{ij}\right), \quad \text{where } a_{ij} = \begin{cases} 1 & \text{if } t_{ij} \in A_i \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

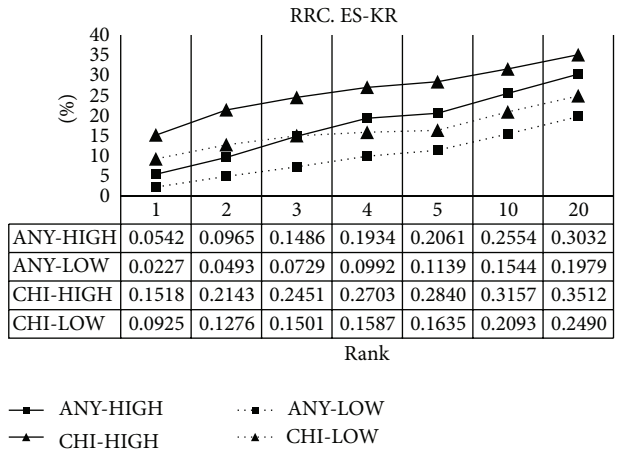
where $r(t_{ij})$ is the relative frequency of translation equivalents t_{ij} for a source word s_i in the training corpus and $\sum_{t \in A_i} r(t) = 1$. An example of how to calculate RRC is described in more detail as below. Table 6 shows Korean translation equivalents for the French word *décision* and their relative frequency. As you can see in Table 6, the Korean translation equivalent *geol-dan-ryuk* cannot be extracted in any case because its relative frequency is zero where it does not appear in the training corpus. We can calculate the RRC_k for the French



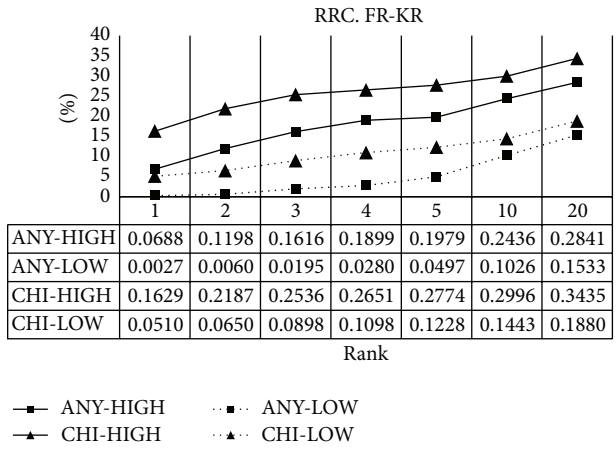
(a) Korean to Spanish



(b) Korean to French



(c) Spanish to Korean



(d) French to Korean

Figure 6: Rated recall of k best candidates with different ranks.

Table 6: A list of Korean translation equivalents for the French word *décision* gotten from the FR-KR evaluation dictionary.

Korean translation equivalent	Frequency[a]	$r(t)$
gyuol-jung (*decision*)	6,007	0.754
jae-jung (*finance*)	880	0.110
pan-jung (*judgment*)	414	0.052
gyeol-ui (*resolution*)	369	0.046
gyeol-sim (*determination*)	173	0.022
gyeol-dan (*determination*)	130	0.016
geol-dan-ryuk (*the strength of one's mind*)	0	0.000
Total	7,973	1.000

[a]The frequency is counted from the Korean part of the KR-EN parallel corpus.

word *décision* under Table 6. If *gyuol-jung* and *gyeol-ui* are extracted among the top 5, RRC_5 is 80% (0.754 + 0.046).

The results are shown in Figure 6. As you can see in Figure 6, the performance is much higher compared to the recall shown in Figure 5. As for HIGH, the performance is from approximately 5% to 13% at the top 1 and from approximately 28% to 35% at the top 20. As for LOW, the performance is from approximately 5% to 13% at the top 1 and from approximately 1% to 32% at the top 20. The range is quite big. Overall, CHI outperforms ANY for HIGH and LOW.

4.3. Error Analysis. Tables 7 and 8 show Korean translation candidates for the Spanish word *estrategia* (*strategy*) and for the French word *monde* (*world*), respectively. Each table contains correct translation equivalents and some errors on the ranked list. In this section, we consider errors. We present

Table 7: Examples of Korean translation candidates for the Spanish word *estrategia* (strategy).

Korean (gloss)	Similarity	Correct answer	Error class
sa-eop-jeon-ryak (business strategy)	0.732	False	Compound segmentation
jeon-ryak (strategy)	0.725	True	
su-rip (establishment)	0.573	False	True negative
gi-bon-jeon-lyak (basic strategy)	0.366	False	Compound segmentation
tu-ja-jeon-ryak (investment strategy)	0.362	False	Compound segmentation

Table 8: Examples of Korean translation candidates for the French word *monde* (world).

Korean (gloss)	Similarity	Correct answer	Error class
se-gye (world)	0.929	True	
se-sang (world)	0.534	True	
weol-deu (world)	0.395	False	Transliterated
jeon-se-gye (worldwide)	0.390	False	False positive (de facto true, not in gold standard)
gak-guk (each country)	0.307	False	Ditto

characteristics of Korean word formation by analyzing errors in BLE.

First, many Korean words are derived from Chinese, for example, the Korean word *bu-eob* (its translation equivalents in English are *side job*, *by-job*, *sideline*, *side business*, *subsidiary work*, *auxiliary occupation*, *minor occupation*, and *sideline occupation* (from http://endic.naver.com/)) derived from a sequence of the Chinese characters 副業. A Chinese character has a meaning because a Chinese character is an ideogram. In the previous example, the meaning of the Chinese character of 副 is *secondary*, *auxiliary*, *subsidiary, and so forth* and that of 業 is *job*, *work*, *occupation, and so forth*. These characteristics make words abundant and consequently the ambiguity of Korean words could be reduced to a certain degree (see Table 5). These characteristics are one of the main reasons of errors in BLE (word-to-word mapping). As you can see in the previous example, two translation equivalents, *sideline* and *by-job*, are comprised of only one word and the rest are comprised of two words. The former can be extracted in BLE, but the latter cannot exactly and only a part of them can be found. In the end, they make errors in bilingual word-to-word lexicon extraction like this work. This is the reason why multi-word expressions should be treated in BLE. In this paper, multi-word expressions as translation equivalents are not included in the evaluation dictionary and then the performance is not influenced by this characteristic.

Second, word spacing is unclear in Korean. Basically a Korean word is separated by a space, so compound words, especially compound nouns, should contain one or more spaces between words. In many case, compound words in Korean real text may be used without any spaces. These characteristics cause errors in BLE. Such errors can be found in Table 7. See the first row for composing *sa-eop* (*business*) and *jeon-ryak* (*strategy*), the fourth row for making up *gi-bon* (*basic*) and *jeon-lyak* (*strategy*), and the fifth for forming *tu-ja* (*investment*) and *jeon-ryak* (*strategy*). All of these contain the basic meaning of *jeon-ryak* (*strategy*), but they are counted as errors. If a tokenizer would separate a compound into basic words, these could be counted as errors. Consequently a good tokenizer (or a good POS tagger) has to be used in order to reduce such errors. Any tokenizer, however, cannot help generating a few errors no matter how good the tokenizer is and they can also affect the performance.

Third, many transliterated words in Korean are used in real text, but not include in the MRD which we used as the evaluation dictionary (gold standard). For example, *weol-deu* (*world*) at the third row of Table 8 is transliterated by substituting the English letters world for Korean letters. This characteristic also causes errors in BLE for Korean.

Fourth, every MRD as a gold standard is incomplete. See the fourth and the fifth row of Table 8. The basic meaning of *se-gye* is the same as that of *jeon-se-gye*. The difference is that the former is usually used as noun and the latter as adjective. The word *gak-guk* involves *every country*, namely, *whole world*. These kinds of words can be treated as correct. To solve this problem, human judgments can be used, but it is too expensive and tedious and it is not effective. At the end, different evaluating method to take into account similar words should be considered.

4.4. Discussion on Experiments. The purpose of this work is not to try to find the best results by looking for the best tuning of each variation like model parameters. Here, the main interest is to show the validity of pivot-based context vectors and the usability of the PBSA for resource-poor language pairs. The validity of context vectors in BLE was described by Gaussier et al. [37]. In the same context, pivot-based context vectors are represented by words in a pivot language (as for standard approach, a target language) and can be interpreted as the same geometric view, and our experimental results are its evidence. Also through experimental results, we have observed that the PBSA is useful for resource-poor language pairs. In fact, any linguistic resources between Korean and Spanish can barely be obtained in the public domain, but we were able to build a bilingual lexicon between two languages. It is important to say that the PBSA is conceptually naive. The process of looking for translation equivalents does not take into account any linguistic resources like seed dictionaries;

Table 9: Comparison of the performance in two estimation methods, ANY and CHI.

Frequency	Dictionary	ACC	MRR	PRE	REC	RRC
HIGH	KR-ES	ANY	ANY	ANY	CHI	CHI
	KR-FR	ANY	ANY	ANY	CHI	CHI
	ES-KR	ANY	ANY	ANY	CHI	CHI
	FR-KR	ANY	ANY	ANY	CHI	CHI
LOW	KR-ES	CHI	CHI	CHI	CHI	CHI
	KR-FR	ANY	ANY	ANY	ANY	CHI
	ES-KR	ANY	ANY	ANY	CHI	CHI
	FR-KR	ANY	ANY	ANY	CHI	CHI

it is just based on publicly available parallel corpora that are sharing a resource-rich language as a pivot language. But, this approach in performance is worth comparing with other approaches in BLE.

Consider two estimation methods, ANY and CHI. Table 9 demonstrates comparison of the performance in two estimation methods, ANY and CHI. In the table, the cell marked ANY shows that ANY outperforms CHI, on the contrary, the cell marked CHI shows that CHI outperforms. As you can see in the table, on the whole, although there are some exceptions, ANY outperforms CHI in accuracy like ACC, MRR, and PRE, while CHI outperforms ANY in recall like REC and RRC. Unusually for KR-ES of LOW, CHI is always superior in performance. We are seriously thinking about the reasons. We believe that the distinction between ANY and CHI results from the size of context vectors as mentioned in Section 4.2.4. Context vectors for ANY are short in size and noise elements in the vectors are small in number compared with CHI.

In BLE, effective evaluation methods should be developed because a MRD as a gold standard is not sufficient as mentioned in Section 4.3. In general the MRD which is publicly available does not reflect real text of the day, but also we cannot say that the dictionary is complete.

Almost always, systems for Korean as a source language outperform those for Spanish and French. We believe that this is due to the ambiguity of translation equivalents. As you can see in Table 4, the ambiguity for Korean as a source language is lower than ambiguity of others.

5. Conclusions and Perspectives

A PBSA for bilingual lexicon extraction is adapted from the standard approach and is based on similarity of context vectors represented by words in a pivot language like English. The PBSA uses two parallel corpora sharing an intermediary language as a pivot language and does not employ any linguistic resources like seed dictionaries. In this paper, we evaluated two different methods for estimating the context vectors under the PBSA in BLE. One estimates them from two parallel corpora on the basis of word association between source words (resp., target words) and pivot words and the other one on the basis of word alignment tools. Though the results were applied to only two language pairs (e.g., Korean-Spanish and Korean-French), the PBSA was quite attractive where public bilingual corpora between two languages are directly unavailable, but public parallel corpora such as English are available. As a result, the method is very promising for resource-poor languages. Furthermore, our methods perform quite well for words with low frequency.

For future works, we plan to extend the pivot-based standard approach to multi-word expressions as well because they play an important role in BLE as mentioned before. Furthermore, evaluation methods against similar words should be considered and more translation equivalents in bilingual dictionary should be added for a larger coverage. .

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work was supported by the IT R&D program of MSIP/KEIT (10041807, Development of Original Software Technology for Automatic Speech Translation with Performance 90% for Tour/International Event Focused on Multilingual Expansibility and Based on Knowledge Learning).

References

[1] D. Andrade, T. Matsuzaki, and J. Tsujii, "Statistical extraction and comparison of pivot words for bilingual lexicon extension," *ACM Transactions on Asian Language Information Processing*, vol. 11, no. 2, article 6, 31 pages, 2012.

[2] P. Fung, "Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus," in *Proceedings of the 3rd Workshop on Very Large Corpora (VLC '95)*, pp. 173–183, June 1995.

[3] R. Rapp, "Identifying word translations in nonparallel texts," in *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics (ACL '95)*, pp. 320–322, 1995.

[4] M. Sahlgren and J. Karlgren, "Automatic bilingual lexicon acquisition using random indexing of parallel corpora," *Natural Language Engineering*, vol. 11, no. 3, pp. 327–341, 2005.

[5] J. Klavans and E. Tzoukermann, "Combining corpus and machine-readable dictionary data for building bilingual lexicons," *Machine Translation*, vol. 10, pp. 1–34, 1996.

[6] P. Fung, "A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora," in *Proceedings of the Parallel Text Processing*, pp. 1–17, 1998.

[7] P. F. Brown, J. Cocke, S. D. Pietra et al., "A statistical approach to machine translation," *Computational Linguistics*, vol. 16, no. 2, pp. 79–85, 1990.

[8] F. J. Och and H. Ney, "Improved statistical alignment models," in *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL '00)*, pp. 440–447, October 2000.

[9] A. B. Ismail, *Minimally supervised techniques for bilingual lexicon extraction [Ph.D. dissertation]*, Department of Computer Science, The University of York, 2012.

[10] R. Rapp, "Automatic identification of word translations from unrelated English and German corpora," in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL '99)*, pp. 519–526, College Park, Md, USA, June 1999.

[11] P. Koehn and K. Knight, "Learning a translation lexicon from monolingual corpora," in *Proceedings of the Association for Computational Linguistics on Unsupervised Lexical Acquisition (ULA '02)*, pp. 9–16, July 2002.

[12] H. Kaji, S. Tanamura, and D. Erdenebat, "Automatic construction of a Japanese-Chinese dictionary via English," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC '08)*, pp. 699–706, Marrakech, Morocco, May 2008.

[13] J. H. Kim, H. W. Seo, and H. S. Kwon, "Bilingual lexicon induction through a pivot language," *Journal of the Korean Society of Marine Engineering*, vol. 37, no. 3, pp. 300–306, 2013.

[14] H. S. Kwon, H. W. Seo, and J. H. Kim, "Bilingual lexicon extraction via pivot language and word alignment tool," in *Proceedings of the 6th Workshop on Building and Using Comparable Corpora*, pp. 11–15, 2013.

[15] K. Tanaka and K. Umemura, "Construction of a bilingual dictionary intermediated by a third language," in *Proceedings of the International Conference on Computational Linguistics (COLING '94)*, pp. 297–303, Kyoto, Japan, August 1994.

[16] T. Tsunakawa, N. Okazaki, and J. Jsujii, "Building a bilingual lexicon using phrase-based statistical machine translation via a pivot language," in *Proceedings of the International Conference on Computational Linguistics (COLING '08)*, pp. 127–130, 2008.

[17] G. S. Mann and D. Yarowsky, "Multipath translation lexicon induction via bridge languages," in *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies (NAACL '01)*, pp. 1–8, Pittsburgh, Pa, USA, June 2001.

[18] D. Bouamor, N. Semmar, and P. Zweigenbaum, "Towards a generic approach for bilingual lexicon extraction from comparable corpora," in *Proceedings of the 14th Machine Translation Summit*, pp. 143–150, 2013.

[19] P. Fung and K. McKeown, "Finding terminology translation from non-parallel corpora," in *Proceedings of the 5th Annual Workshop on Very Large Corpora (VLC '97)*, pp. 192–202, Hong Kong, 1997.

[20] I. Vulić and M. F. Moens, "A study on bootstrapping bilingual vector spaces from non-parallel data (and nothing else)," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '13)*, pp. 1613–1624, Seattle, Wash, USA, October 2013.

[21] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 23, pp. 146–162, 1954.

[22] M. Apidianaki, N. Ljubešić, and D. Fišer, "Disambiguating vectors for bilingual lexicon extraction from comparable corpora," in *Proceedings of the 8th Language Technologies Conference*, pp. 10–15, Ljubljana, Slovenia, October 2012.

[23] 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.

[24] P. Fung and K. W. Church, "K-vec: a new approach for aligning parallel texts," in *Proceedings of the 15th Conference on Computational Linguistics*, pp. 1096–1102, Kyoto, Japan, August 1994.

[25] S. Evert, *The statistics of word co-occurrences: word pairs and collocations [Ph.D. thesis]*, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Stuttgart, Germany, 2013.

[26] P. Pecina, "Lexical association measures and collocation extraction," *Language Resources and Evaluation*, vol. 44, no. 1-2, pp. 137–158, 2010.

[27] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, Mass, USA, 1999.

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

[29] W. Gale and K. W. Church, "Identifying word correspondences in parallel texts," in *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, pp. 152–157, 1991.

[30] H. W. Seo, H. Kim, H. Y. Cho, J. H. Kim, and S. W. Yang, "Automatically constructing English-Korean parallel corpus from web documents," in *Proceedings of the 26th on Korea Information Processing Society Fall Conference*, vol. 13, pp. 161–164, 2006.

[31] P. Koehn, "Europarl: a parallel corpus for statistical machine translation," in *Proceedings of the 10th Conference on the Machine Translation Summit*, pp. 79–86, 2005.

[32] W. Lee, S. Kim, G. Kim, and K. Choi, "Implementation of modularized morphological analyzer," in *Proceedings of the 11th Annual Conference on Human and Cognitive Language Technology (HCLT '99)*, pp. 123–136, 1999.

[33] D. Chatterjee, S. Sarkar, and A. Mishra, "Co-occurrence graph based iterative bilingual lexicon extraction from comparable corpora," in *Proceedings of the 4th International Workshop on Cross Lingual Information Access (COLING '10)*, pp. 35–42, 2010.

[34] M. Voorhees, "TREC-8 question answering track report," in *Proceedings of the 8th Text Retrieval Conference*, pp. 77–82, Gaithersburg, Md, USA, November 1999.

[35] A. Haghighi, P. Liang, T. Berg-Kirkpatrick, and D. Klein, "Learning bilingual lexicons from monolingual corpora," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL '08)*, pp. 771–779, June 2008.

[36] H. W. Seo, H. S. Kwon, and J. H. Kim, "Rated recall: evaluation method for constructing bilingual lexicons," in *Proceedings of the 25th Annual Conference on Human and Cognitive Language Technology (HCLT '13)*, pp. 146–151, October 2013.

[37] E. Gaussier, J. M. Renders, I. Matveeva, C. Goutter, and H. Déjern, "A geometric view on bilingual lexicon extraction from comparable corpora," in *Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics (ACL '04)*, pp. 526–533, Barcelona, Spain, 2004.