Journal of the Korean Society of Marine Engineering, Vol. 38, No. 5 pp. 557~565, 2014 ISSN 2234-7925 Print
http://dx.doi.org/10.5916/jkosme.2014.38.5.557 ISSN 2234-8352 Online

Extended pivot-based approach for bilingual lexicon extraction

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

(Received January 15, 2014 ; Revised March 12, 2014 ; Accepted March 26, 2014)

Abstract: This paper describes the extended pivot-based approach for bilingual lexicon extraction. The basic features of the approach can be described as follows: First, the approach builds context vectors between a source (or target) language and a pivot language like English, respectively. This is the same as the standard pivot-based approach which is useful for extracting bilingual lexicons between low-resource languages such as Korean-French. Second, unlike the standard pivot-based approach, the approach looks for similar context vectors in a source language. This is helpful to extract translation candidates for polysemous words as well as lets the translations be more confident. Third, the approach extracts translation candidates from target context vectors through the similarity between source and target context vectors. Based on these features, this paper describes the extended pivot-based approach and does various experiments in a language pair, Korean-French (KR-FR). We have observed that the approach is useful for extracting the most proper translation candidate as well as for a low-resource language pair.

Keywords: Bilingual lexicon extraction, Pivot language, Rated Recall

1. Introduction

Bilingual lexicon is an important resource used for various research domains such as natural language processing (NLP), machine translation (MT), information retrieval (IR) **[1]**, sentiment classification **[2]** and so on. Bilingual corpora, comparable or parallel, are a key to extract it, but so far, a bilingual lexicon for a low-resource language pair like Korean-French has suffered from a difficulty to get it because constructing the bilingual corpora is time-consuming process.

In spite of such a problem, many researchers studied to collect the lexicon in various ways. Some researchers represent words by a context vector based on its lexical context **[3]-[5]**. In fact, they achieve 80 to 91% accuracy for single terms, even though the performance drops to 60% when specialized small corpora are used by some researchers **[6][7]**. Other researchers **[8]-[10]**, on the other hand, focused on a size of an initial seed dictionary (basically, bilingual lexicon extraction needs an initial bilingual seed dictionary to translate context vectors). Some researchers **[11][12]** focused on similar words sharing similar lexical circumstances and one of the approaches **[11]** will be denoted as the extended context-based approach in the rest of the paper. Alternatively, Seo et al. **[13]** proposed the standard pivot-based approach that is useful for only a low-resource language pair, and extracts bilingual lexicons using a pivot language such as English.

To evaluate these approaches, usually the accuracy or the mean reciprocal rank (MRR) is used **[9][13]**.

† Corresponding Author: Dept. of Computer Engineering, Korea Maritime and Ocean University, 727 Taejong-ro, Yeongdo-gu, Busan 606-791, Republic of Korea, e-mail: jhoon@kmou.ac.kr, Tel: 051-410-4574
1 Dept. of Computer Engineering, Korea Maritime and Ocean University, e-mail: wonn24@gmail.com, Tel: 051-410-4896
2 Dept. of Computer Engineering, Korea Maritime and Ocean University, e-mail: hong8c@naver.com, Tel: 051-410-4896

These measures focus on finding correct translation candidates (equivalences). Even though rare words as unknown words are much more founded by far than frequent words from a system, we cannot judge whether the system is good or not. To understand easily this problem, this paper also discuss about Rated Recall **[14]** that considers how many times translation candidates appears in the system. In this paper, we will evaluate the related recall as well as the accuracy and the MRR of the proposed system.

The rest of the paper is organized as follows: Section 2 discusses related works, the standard pivot-based approach and the extended context-based approach. Section 3 represents a extended pivot-based approace and Section 4 reports and discusses some experiments. Finally, Section 5 concludes and addresses future works.

2. Related Work

All approaches described in this section are based on the context-based approach that is represented by Rapp **[4]**. The first motivated work, the standard pivot-based approach, uses a high-resource language like English as a pivot language to extract bilingual lexicons from two sets of parallel corpora (e.g., English-*). Another work, the extended context-based approach, uses comparable corpora in deference to the coverage of the initial bilingual dictionary which is used to translate a source language into a target language **[11]**.

2.1 Standard Pivot-Based Approach

The standard pivot-based approach aims to extract a bilingual lexicon from a low-resource language pair such as Korean-French without any external resource like an initial bilingual dictionary and a parallel corpus between two languages. This approach was proposed by some researchers **[13][15]** to improve some disadvantages of the context-based approach that builds context vectors and then compares them to identify correct translation candidates from comparable corpora **[4]**. The context-based approach, in particular, uses comparable corpora and an initial bilingual dictionary to translate all words, s, in a source text into target words, s''. In this case, the performance of the system may depend on a coverage of the initial bilingual dictionary. To overcome this problem, some researchers have proposed various studies **[8]** for the dictionary coverage.

The standard pivot-based approach, however, uses a pivot language to represent words in a common language. As a result, the dimension of two vectors becomes as same, and are comparable each other. For this reason, the initial bilingual dictionary is no longer needed for the translation task. On the other hand, just parallel corpora of a high-resource language pair like English-* should be needed. The standard pivot-based approach follows the next three steps:

I. Building the context vectors from two sets of parallel corpora (Korean-English and French-English; hereinafter KR-EN and FR-EN, respectively). All context vector values are represented with association scores, and these values show how much they are related to each other. To calculate the association scores, co-occurrence counts are collected in a source text (resp. a target text) by source-pivot word matrix (resp. target-pivot word matrix). The co-occurrence counts come from a frequencies of parallel sentences that contain both source (resp. target) and pivot words together. In this case, all sentences are treated as documents. The association scores can be measured as a point-wise mutual information, a chi-square score.

II. Calculating similarity scores between each source context vector, $\boldsymbol{s}$, and all target context vectors, $\boldsymbol{t}$, on the basis of vector distance measures such as a cosine similarity and a weighted Jaccard coefficient.

III. Sorting the top k translation candidates based on

their similarity scores.

The approach is useful where a public parallel corpus between two languages are directly unavailable but a corpus for some language pairs like English-* is publically available. Furthermore, the approach is very simple because any linguistic resource such as an initial bilingual dictionary or a word alignment tool is not required.

2.2 Extended Context-Based Approach

Déjean and Gaussier (2002) **[11]** have proposed an extended method (hereinafter, the extended context-based approach) to the fundamental context-vector approach reported earlier by few researchers **[3][16]**. The goal of the approach is a less dependency for the coverage on the initial bilingual dictionary. The approach also identifies the second-order affinities **[17, p. 280]** in a source language as follows:

> "*Second-order affinities show which words share the same environments. Words sharing second-order affinities need never appear together themselves, but their environments are similar.*"

The main ideas of the extended approach **[18]** can be described as follows:

I. The k nearest words in a source text according to the similarity score, $\mathrm{sim}(\boldsymbol{s},\boldsymbol{s}')$ between two words, $\boldsymbol{s}$ and $\boldsymbol{s}'$, are identified.
II. The similarity score, $\mathrm{sim}(\boldsymbol{s}'',\boldsymbol{t})$ for each word, t, in a target language with the respect to the translated k nearest words, $\boldsymbol{s}''$, is calculated.

This extended context-based approach is motivated to the extended pivot-based approach, and it will be described in more details in the next section.

3. Extended Pivot-Based Approach

As described in the previous section, the extended pivot-based approach is pretty motivated from the extended context-based approach represented by Déjean and Gaussier (2002) **[11]**. That is, a base system is same with the standard pivot-based approach, but the conceptual idea from the extended context-based approach is added. The idea is finding the k nearest words in a source text to identify more confident translation candidates as described in the previous section.

The big question is how the extended context-based approach can be apply to the standard pivot-based approach. As we mentioned before, the step that is finding k nearest words from the extended context-based approach is conducted where all words are from comparable corpora. To compare all context vectors, therefore, k nearest words, $\boldsymbol{s}'$, in a source text should be translated into target words, $\boldsymbol{s}''$, with an initial bilingual dictionary. As a result of the translation task, all words (e.g., $\boldsymbol{s}''$ and $\boldsymbol{t}$) are represented in a common target language so the dimensions of the context vectors are changed to be equal.

However, on the other hand, the situation of the standard pivot-based approach is little bit different. There is a pivot language that connect two different languages (e.g., source and target) so the step for calculating similarity scores can be conducted in much simpler way. The overall structure of the extended pivot-based approach is described in **Figure** 1.

As you can see **Figure** 1, two sets of parallel corpora consists of three different language, source language (SL), target language (TL), and pivot language (PL), are used to build context vectors. An important thing of the overall structure for the extended pivot-based approach is that the SL-SL context vectors are indirectly used when the similarity score sim($\boldsymbol{s}$, $\boldsymbol{t}$) is calculated. The overall structure can be describe in more details as follows:

Three types of context vectors, SL-SL (1.a), SL-PL (1.b) and TL-PL (1.c), are built by parallel corpora, respectively. All context vector values are represented with association scores like a point-wise mutual information (PMI) or a chi-square score.

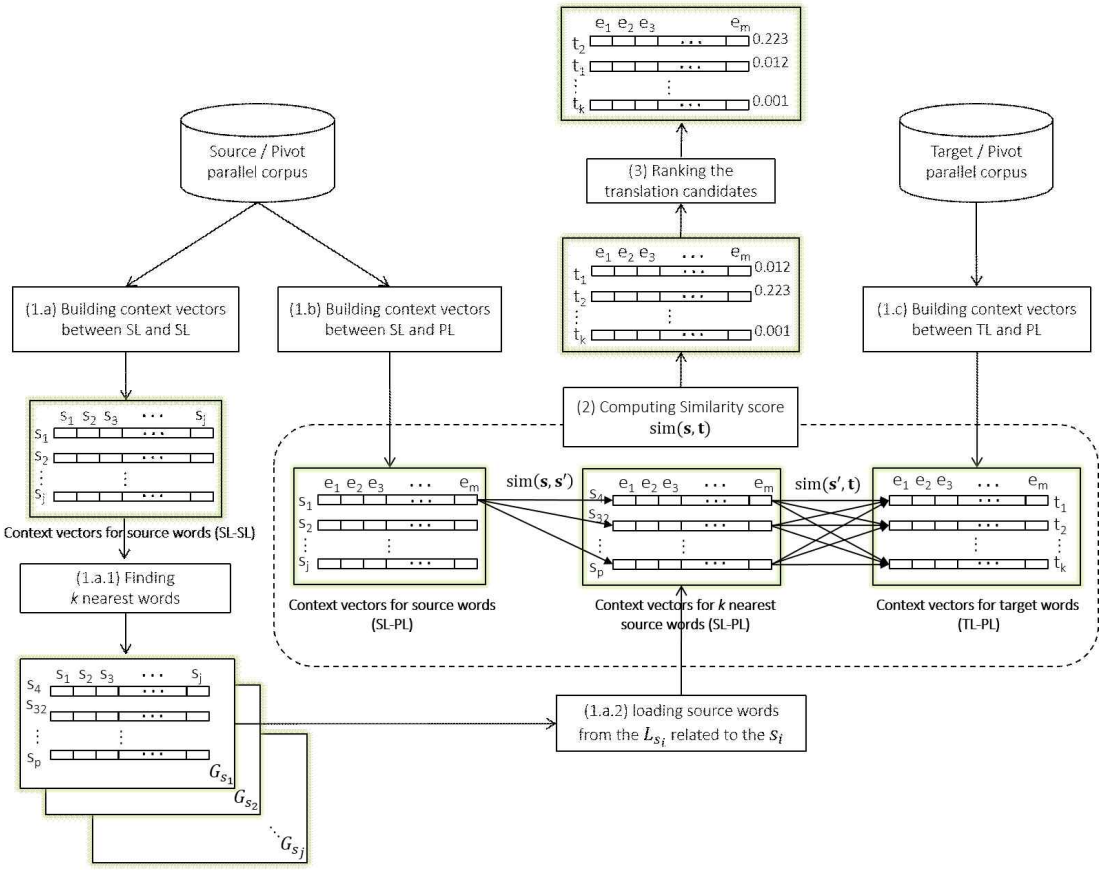

Figure 1: Overall structure of the extended pivot-based approach

1.a.1. Let s_i be a word in a source language, and then the group G_{s_i}, a set of k nearest words related to a source word s_i, is collected based on vector distance measures such as a cosine similarity. With these similarity scores, all words include a source word s_i itself are then grouped and filtered out by a threshold.

1.a.2. A bunch of nearest source words $\boldsymbol{s'}$ from G_{s_i} with respect to the source word s_i is extracted to compute the similarity score $\mathrm{sim}(\boldsymbol{s},\boldsymbol{t})$. In this case, only words s' in the group G_{s_i} are loaded but context vectors are from (1.b).

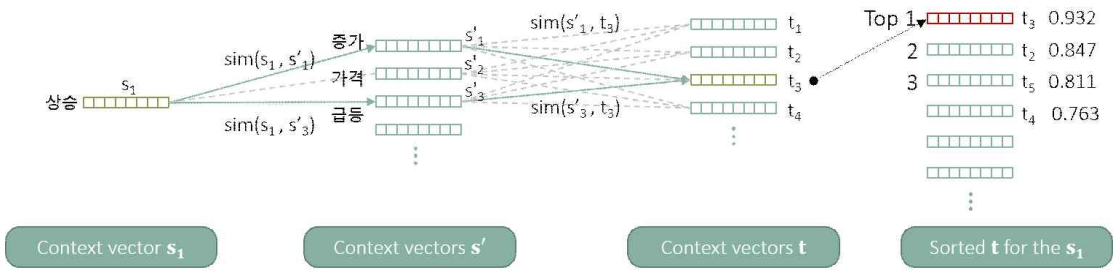

Figure 2: An example of the calculation for the similarity score sim($\mathbf{s_1}$, $\mathbf{t_3}$)

2. Two similarity scores, $\mathrm{sim}(\boldsymbol{s},\boldsymbol{s'})$ and $\mathrm{sim}(\boldsymbol{s'},\boldsymbol{t})$, are calculated for the $\mathrm{sim}(\boldsymbol{s},\boldsymbol{t})$. The similarity score, $\mathrm{sim}(\boldsymbol{s},\boldsymbol{t})$, can be described as the formula (1).

$$\mathrm{sim}(s,t) = \sum_{s \in k\ \mathrm{nlw}} \mathrm{sim}(s,s') \times \mathrm{sim}(s',t) \quad (1)$$

The similarity score, $\mathrm{sim}(\boldsymbol{s},\boldsymbol{t})$, is the sum of all possible (collectable) similarity scores sharing k nearest words (called klw), s', related with two words, s and t. The example of the formula **(1)** is represented in **Figure** 2 in more details. This example assumes that the word, t_3, in a target language is the closest translation candidate with the Korean word, s_1, '상승' (means rise; ascension). In general, all context vectors, s, only related with t_3 should be considered to measure the similarity score. If two nearest words, s'_1 '증가' (means augmentation; gain; rise) and s'_3 '급등' (means jump; a sudden rise), for the s_1 are closely related to the word t_3, the similarity score, sim($\mathbf{s_1}$, $\mathbf{t_3}$), should be measured as follows:

$$\begin{aligned}\mathrm{sim}\ (s_1,t_3) = \mathrm{sim}\ (s_1,s'_1) \times \mathrm{sim}\ (s'_1,t_3) \\ + \mathrm{sim}\ (s_1,s'_3) \times \mathrm{sim}\ (s'_3,t_3)\end{aligned}$$

After measuring all the similarity scores, all target context vectors, **t**, are then sorted based on the final similarity scores. This job is recursively conducted as much as the size of the source context vectors, **s**.

4. Experiments and Discussions

In order to evaluate the performance of the extended pivot-based approach, we use the KR-EN parallel corpus (433,151 sentences) **[13]** and the EUROPARL parallel corpus (500,000 sub-sentences) [19] for the language pair FR-EN. On the average, each sentence contains 42.36(KR of KR-EN), 36.02(EN of KR-EN), 31.17(FR from FR-EN), 28.68(EN from FR-EN) words per sentence. All content words (nouns, main verbs, adjectives and adverbs) are POS-tagged and all stop-words fitted to each language (EN, KR and FR) are eliminated. Moreover the same KR-FR bilingual lexicon is used for the evaluation. Each language set (e.g., KR or FR) contains 100 frequent words (denoted as High) and 100 rare words (denoted as Low).

In order to measure the performance, three different measures, the accuracy, MRR (Mean Reciprocal Recall) and RR (Rated Recall), are considered in this paper. Especially, MRR **[20]** accentuates translation candidates at higher ranks more than the others at lower ranks, so that the correct translation candidates at higher ranks could be treated as more important. The RR, one of the other measuring method, means that how many correct translation candidates treated as IMPORTANT (as much important as it occurs in a document) are extracted by a system, while the basic recall means that how many correct translation candidates treated as NORMAL (each candidate has a same weight) are extracted by a system. The RR can be represented as the formula **(2)**.

$$RR_n = \frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{n} C_{jk}\, r(t_{jk}), \quad (2)$$

$$C_{jk} = \begin{cases} 1, & \mathrm{if}\ t_{jk} \in C_j \\ 0, & \mathrm{otherwise} \end{cases}$$

N means the total number of the lexical unit, i, in a source text and n (and also denoted as $|C_j|$) means the number of correct translations in a bilingual lexicon. It goes without saying that each lexical unit, i, has the different number of correct translations or senses. C_{jk} is also known as the Kronecker delta function. If j-th lexical unit, i_j, in a source text has the k-th translation candidate, t_{jk}, the C_{jk} is 1 or 0. $r(t_{jk})$ is the rate means how many times t_{jk} occurs in a target text. The sum of all properties related with C_j should be 1 (e.g., $\sum_{t \in C_j} r(t) = 1$). An example about how to calculate the RR is described in more details as below.

Table 1: Correct Korean translations of the French word “décision” (e.g., source: FR, target: KR)

Translations (Romanization)	Gloss	Freq.	$r(t)$
결정(Gyeol-jung)	decision; conclusion	6,007	0.752
결심(Gyeol-sim)	determination; resolution	173	0.022
결의(Gyeol-ui)	a resolution; a decision; a vote	369	0.046
결단(Gyeol-dan)	determination; resolution; determine	130	0.016
결단력 (Gyeol-dan-ryuk)	the strength of one's mind; resolution	10	0.001
재정(Jae-jung)	finance(s); financial affairs	880	0.110
판정(Pan-jung)	Judgment; decision; judge	414	0.052
Total		7,983	1.000

As you can see **Table** 1, all translations occur at least 10 times, and especially ‘결정’ is the word that appears the most times (6,007) in a target text. Each translation in a bilingual dictionary has the rate, $r(t)$. It means that other senses are eliminated unless it occurs in a target text.

Table 2 describes the example of Korean translation candidates of French word “décision” (e.g., system’s output). All translation candidates occurs in a target text but some of them would not in a bilingual dictionary (e.g., it is an incorrect translation candidate). The correct translation candidates have the right $r(t)$ as well. In such a case, RR and recall can be calculated as **Table** 3.

As you can see **Table** 3, RR represents how much important translation candidate occurs in a target text. If the most frequent word is yielded by a system, it would be better than a case of a lower frequent word (rare word). In order to compare the extended pivot-based approach and the standard pivot-based approach, this paper represents three different evaluation results in this section.

Table 2: The system output: Translation candidates

Rank	Translation candidate	Gloss	$r(t)$
1	결정 (Gyeol-jung)	decision; conclusion	0.752
2	통합 (Gyeol-sim)	Unification; unity; combination	
3	결단 (Gyeol-dan)	determination; resolution; determine	0.016
4	여부(Yeo-bu)	yes or no; whether or not; if	
...			
7	판정 (Pan-jung)	Judgment; decision; judge	0.052
...			
15	결심 (Gyeol-sim)	determination; resolution	0.022
...			

Table 3: Rated Recall and Recall at each rank

Rank	RR (Rated Recall)	Recall
1	0.752	1/7 = 0.143
3	0.752 + 0.016 = 0.768	2/7 = 0.286
7	0.768 + 0.052 = 0.820	3/7 = 0.429
15	0.820 + 0.022 = 0.842	4/7 = 0.571

Firstly, **Figure** 4 shows the accuracy of the performance for two approaches, the standard and extended pivot-based approach. All graphs describe the average score of two experimental cases, KR to FR and FR to KR. As you can see **Figure** 4, the extended approach slightly yields higher performance at Low, and especially at top 1 in case of both Low and High. Besides, the entire slope for the accuracy decreases when top 20 is considered. Nevertheless, it can be interpreted as finding k nearest words helps the performance to have a higher score practically. Actually, these figures are hard to persuade us to see the huge advantage of the extended approach.

Secondly, **Figure** 5 shows the MRR scores within

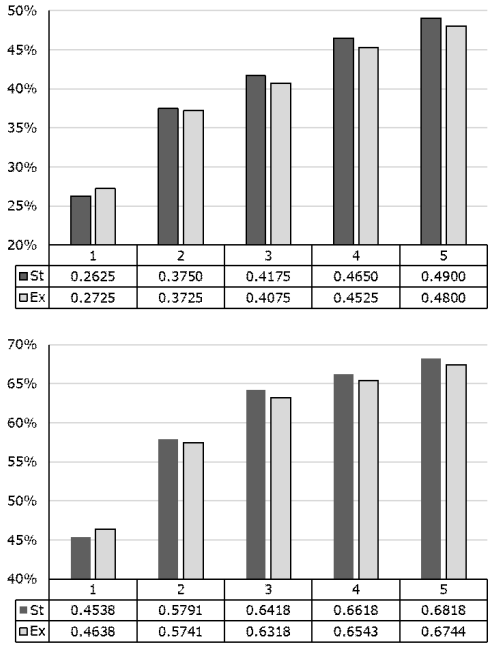

Figure 4: Accuracy at Low and High (High is at bottom)

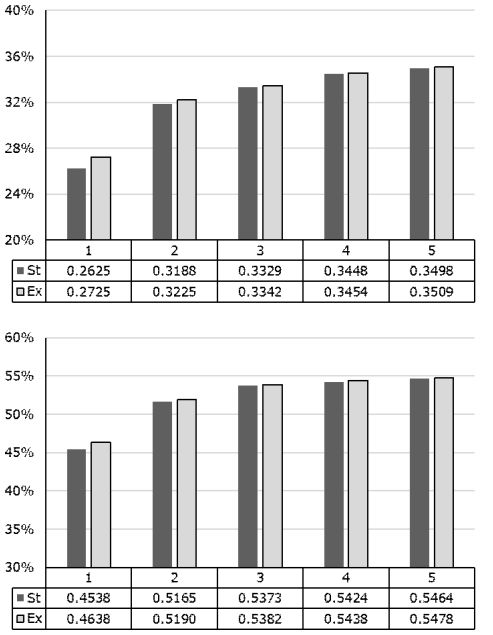

Figure 5: MRR at Low and High (High is at bottom)

top 5 for two approaches. Showing only within top 5 can explain how much frequent translation candidates appear in higher ranks. As you can see the figure, the extended pivot-based approach slightly surpasses the standard pivot-based approach, especially at the top 1. Therefore we can say that the most frequent words appear at higher rank, and it explains how much the extended approach can affect the performance. The third evaluation measure, RR, also proves this.

Thirdly, **Figure** 6 shows the RR scores within top 5 for two approaches. This figure also shows results only within top 5. As you can see, only the top 1 at the high rank slightly surpasses for the extended pivot-based approach. On the other hand, the RR scores of the extended pivot-based approach within top 3 at the low rank surpasses the standard pivot-based approach. This result explains that the extended pivot-based approach can extract better outcomes for rare frequent words than the standard pivot-based approach. Finding k nearest words, s', can be considered as reinforcing some rare words, s, in a source text to be noticed. Even though the extended pivot-based approach could not cover the whole rank sections, the approach is useful when low frequent words are considered.

5. Conclusion and Future work

In this paper, we evaluate the extended pivot-based approach for improving the performance of the standard pivot-based approach by finding k nearest words sharing similar contexts. Sometimes these k nearest words can help rare words but frequently appears in a document to be noticed. The three different evaluation results have shown that the extended pivot-based approach had a good effect on the final results as well.

For the future works, a little more various experiments should be conducted for highly frequent words. Furthermore, the way to improve the performance for rarely frequent words also should be executed. This is

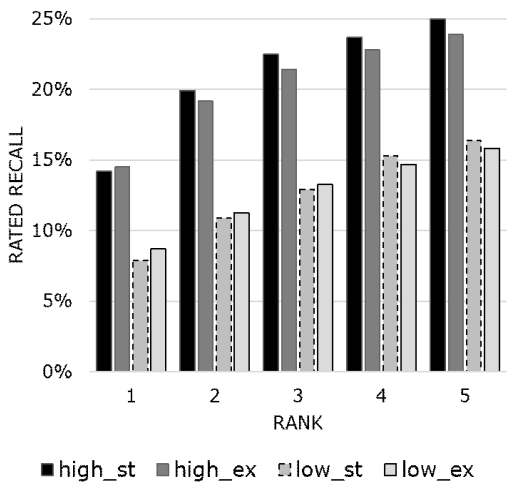

Figure 6: Rated Recall scores within the top 5 for four cases: Low and High, Extended and Standard

caused by our experimental results for rare words. In the other hand, automatic marine terminology extraction using bilingual lexicons can be considered as a future work. Finally, comparable corpora for domain problem and more translation equivalences in bilingual lexicons should be considered for a large coverage.

Acknowledgement

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]

Reference

[1] P. Fung, “A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora”, In Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas, pp. 1-16, 1998.

[2] K.-J. Lee, J.-H. Kim, H.-W. Seo, and K.-S. Ryu, "Feature weighting for opinion classification of comments on news articles", Journal of the Korean Society of Marine Engineering, vol. 34, no. 6, pp. 871-879, 2010.

[3] P. Fung and K. McKeown, “Finding terminology translations from non-parallel corpora”, Proceedings of the 5th Annual Workshop on Very Large Corpora, pp. 192-202, 1997.

[4] R. Rapp, “Automatic identification of word translations from unrelated English and German corpora” Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pp. 519-526, 1999.

[5] Y. Cao and H. Li, “Base noun phrase translation using web data and the EM algorithm” Proceedings of the 19th International Conference on Computational Linguistics, pp. 127-133, 2002.

[6] Y. Chiao and P. Zweigenbaum, “Looking for candidate translational equivalents in specialized, comparable corpora”, Proceedings of the 19th International Conference on Computational Linguistics, pp. 1208-1212, 2002.

[7] H. Déjean and E. Gaussier, “Une nouvelle approcheà léxtraction de lexiques bilingues à partir de corpus comparables”, Lexicometrica, Alignement Lexical Dans les Corpus Multilingues, pp. 1-22, 2002.

[8] P. Koehn and K. Knight, “Learning a translation lexicon from monolingual corpora”, Proceedings of the Association for Computational Linguistic on Unsupervised Lexical Acquisition, pp. 9-16, 2002.

[9] T. Tsunakawa, N. Okazaki, and J. Tsujii, “Building a bilingual lexicon using phrase-based statistical machine translation via a pivot language”, Proceeding of the 22nd International Conference on Computational Linguistics, Posters Proceedings, pp. 18-22, 2008.

[10] P. Koehn, F. Och, and D. Marcu, “Statistical

phrase-based translation", Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, pp. 48-54, 2003.

[11] H. Déjean, F. Sadat, and E. Gaussier, "An approach based on multilingual thesauri and model combination for bilingual lexicon extraction", Proceedings of the 19th International Conference on Computational Linguistics, pp. 218-224, 2002.

[12] B. Daille and E. Morin, "French-English terminology extraction from comparable corpora", Proceedings of the 2nd International Joint Conference on Natural Language Processing, pp. 707-718, 2005.

[13] H.-W. Seo, H.-S. Kwon, and J.-H. Kim, "Context-based bilingual lexicon extraction via a pivot language", Proceedings of the Conference of the Pacific Association for Computational Linguistics, 2013.

[14] H.-W. Seo, H.-S. Kwon, and J.-H. Kim, "Rated recall: Evaluation method for constructing bilingual lexicons", Proceedings of the 25th Annual Conference on Human and Cognitive Language Technology, pp. 146-151, 2013.

[15] 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.

[16] R. Rapp, "Identify word translations in non-parallel texts", Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, pp. 320-322, 1995.

[17] G. Grefenstette, "Corpus-derived first, second and third-order word affinities", Proceedings of the 6th Congress of the European Association for Lexicography, pp. 279-290, 1995.

[18] A. Hazem, E. Morin, and S. Saldarriaga, "Bilingual lexicon extraction from comparable corpora as metasearch", Proceeding of the 4th workshop on Building and Using Comparable Corpora, pp. 35-43, 2011.

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

[20] E. Voorhees, "The TREC-8 question answering track report", Proceedings of the 8th Text Retrieval Conference, pp. 77-82, 1999.