Contemporary Engineering Sciences, Vol. 7, 2014, no. 23, 1225 - 1233
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ces.2014.49152

Bilingual Multi-Word Lexicon Construction via a Pivot Language

Hyeong-Won Seo, Hong-Seok Kwon, Min-ah Cheon, Jae-Hoon Kim

Dept. of Computer Eng., Korea Maritime and Ocean University
Busan, South Korea

Abstract

Bilingual multi-word lexicons are helpful for statistical machine translation systems to improve their performance. In this paper we present a method for constructing such lexicons in a resource-poor language pair such as Korean-French. By using two parallel corpora sharing one pivot language we can easily construct such lexicons without any external language resource like a seed dictionary. The experimental results for the KR to FR have shown that the accuracy is quite promising, even though this research is ongoing.

Keywords: Multi-word Units, MWU Identification, MWU Alignment, Pivot language, Parallel Corpora, Comparable Corpora

1. Introduction

The multi-word unit (MWU) extraction from bilingual corpora is an important task for many natural language processing (NLP) domains. Especially, such MWU bilingual lexicons are useful to improve the performance of statistical machine translation (SMT) [1]. While various researches for MWUs have been proposed, there are many challenges for the bilingual MWU extraction.

Several researches have been proposed by using collocation and association metrics [2], n-grams, and so on. Other researches [3, 4] have been done from a bilingual parallel corpus. However, a parallel corpus is plentiful for some resource-rich language pairs such as English–French or English–Chinese.

Moreover, collecting a large-scale of bilingual parallel corpora, especially for resource-poor language pairs such as Korean–French (KR–FR), is very expensive and time-consuming. An alternative research using comparable corpora [5] has been proposed to solve this problem. However, even if comparable corpora are available, some external knowledge such as an initial seed dictionary should be used to deal with these comparable corpora.

A pivot language can be one of solutions preventing us from constructing any bilingual corpora for a resource-poor language pair. Tsunakawa *et al*. [6] combine two different bilingual lexicons as one of resources of SMT systems into one bilingual lexicon via a pivot language. Seo *et al*. [7] build context vectors from two different parallel corpora sharing English, and then calculate their similarities to align them. However, both pivot-based approaches are applied to only single-word units (SWUs).

In this paper, we present a method for constructing bilingual multi-word lexicons using a pivot language for a resource-poor language pair, e.g., KR–FR. It is much easier way to build English (EN)–* parallel corpora than to do parallel corpora in resource-poor language pairs. Furthermore, using parallel corpora is much simpler than using comparable corpora, because usual methods using comparable corpora should translate context vectors into another language with an initial seed dictionary. In this point of view, we use two parallel corpora (e.g., KR–EN and FR–EN) to prevent us from constructing bilingual corpora directly between KR–FR. Moreover we bring the pivot-based approach from Seo *et al*. [7] to deal with MWUs for KR and FR. The MWUs we consider in this paper are noun phrases such as *taxi driver*, although there are a lot of MWU types [8] such as collocations (e.g., *alcoholic drink*), frozen expressions (e.g., *kick the bucket*), named entities (e.g., *statue of liberty*), etc. Finally we use a number of techniques to identify MWU candidates from monolingual texts, and use comparable corpora to support *phraseness* [9] of the candidates.

The rest of the paper is as follows: Section 2 describes about related works and Section 3 describes the main method to construct bilingual multi-word lexicons. Section 4 presents experimental results with some discussions. Finally, Section 5 draws conclusions and mentions future works that we have planned.

2. Related work

2.1. MWU identification

Several characteristic and linguistic features of MWUs were represented by Kunchukuttan [10]. Based on these features, following three tasks should be considered before MWUs are aligned. Firstly, potential MWU candidates are extracted by using a co-occurrence metric such as pointwise mutual information

(PMI). After then, secondly, the linguistic validity of the candidates is established via specific part-of-speech (POS) filters or dependency relation filters. Thirdly, a semantic non-compositionality of the candidates is measured in various ways.
Piao and McEnery [11] have proposed a hybrid algorithm for aligning nominal MWUs. This algorithm extracts n-grams to make MWU candidates, and then extracts a shortlist of the candidates by using specific POS filters. After that, a co-occurrence metric is used to find out true nominal MWU candidates. These techniques are used for our approach in concert, when potential monolingual MWU candidates are extracted from monolingual corpora. We use two different measures, PMI and pointwise KL-divergence [9], for co-occurrence metrics to see whether the candidates are general.

2.2. Pivot-based approach

The *pivot-based approach* [7] extracts bilingual lexicons without any external linguistic resources such as an initial bilingual dictionary. The approach is based on the *standard approach* proposed by Rapp [12]. Therefore, it builds context vectors from parallel corpora, and then compares them to extract correct translation candidates. The context vectors are constructed in a common pivot language such as English. As a result, source and target context vectors are comparable to each other. Those vectors, finally, are compared and sorted based on similarity measures such as cosine similarity.

The approach is useful when parallel corpora between a resource-poor language pair are publically unavailable but parallel corpora that share one pivot language are available. Through several experiments, the approach has been shown promising results for a resource-poor language pair (e.g., KR–FR), although the performance may vary depending on language pairs.

3. Bilingual multi-word lexicon construction

In this section, we describe a method for constructing bilingual multi-word lexicons using two parallel corpora as same as the *pivot-based approach*. The method can be divided into two stages, e.g., *identification* and *alignment*. Firstly, multi-word candidates are independently identified from two monolingual corpora, e.g., source and target corpora, of the parallel corpora in the source–pivot language (L_s–L_p) and the target–pivot language (L_t–L_p). Secondly, bilingual multi-word alignment is performed by using the *pivot-based approach* which takes two kinds of word types (e.g., single-words and multi-words) as an input at the same time.

In this paper, context vector is only represented in single-words because multi-words in the pivot language may give rise to another errors. Therefore, we

assume that the single words in the pivot language are enough to represent context vectors and to play the role of a bridge which connects two different languages.

3.1. Multi-word identification

The multi-word identification stage is to extract source (and target) multi-word candidates from two monolingual corpora in L_s and L_t. The algorithm for identifying multi-word candidates comes from the approach described in section 2.1, and it has following three steps:

I. Extracting n-grams ($2 \leq n \leq 3$) from monolingual corpora in L_s and L_t
II. Measuring co-occurrence metrics between the n-grams
III. Filtering the n-grams with specific POS patterns to extract candidates

First of all, all kinds of *bi*/*tri*-grams are independently extracted from the two monolingual corpora. Before all *bi*/*tri*-grams are extracted, all stop-words such as punctuation are eliminated first. Secondly, co-occurrence metrics such as PMI are used for highly co-related candidates. Finally, specific POS patterns are adopted to remove irrelevant multi-word candidates. All kinds of *bi*/*tri*-grams that mismatch with the POS patterns are eliminated.

In this paper, we consider only noun phrases, so multi-word candidates which are supposed to be a noun phrase are passed to the next step. We assume that these "identified" multi-word candidates are truly reasonable to be aligned well by the *pivot context approach*.

3.2. Bilingual multi-word alignment

After multi-word candidates are identified, context vectors for single-/multi-word representation can be constructed from two parallel corpora in L_s–L_p and L_t–L_p. As same as the *pivot-based approach*, the context vectors are represented in pivot single-words and are weighted with word association scores (e.g., PMI) between source single-/multi-words (resp. target single-/multi-words) and pivot single-words. Both of the context vectors are represented in a common pivot language, so the dimensions of the two context vectors are comparable to each other. This method is very useful, because we can use two parallel corpora for resource-rich language pairs, L_s–L_p or L_t–L_p, and also it reduces the effort to translate the context vectors into another language. Then we compute the similarity scores between each source vector and all target vectors on the basis of vector similarity measures such as the cosine similarity. For the last step, we sort the top *k* translation candidates corresponding to the source words.

4. Experimental results

In this section, we show preliminary experimental results. Unfortunately, the proposed method is not exactly able to be compared with other general researches

by now. This is caused by our experimental environment, e.g., evaluation data. Moreover, there are even no similar cases or evaluation benchmarks to compare with the data. Therefore, in this paper, we just focus on measuring the accuracy on rank 20 in the KR–FR language pair. As mentioned before, we use both PMI and pointwise-KL divergence to compare which measure shows better performance to our experiments.

4.1. Data

In order to evaluate the proposed method, two sets of parallel corpora (KR–EN and FR–EN) and two comparable corpora (KR and FR) are used together. Basically the parallel corpora are just fine to embody the proposed method. Nonetheless, comparable corpora are additionally used for measuring the pointwise KL-divergence mentioned in section 2.1. That is, at the identification stage, two monolingual corpora in L_s and L_t from two parallel corpora are used for the measuring PMI. Moreover, the two monolingual corpora and other two comparable corpora in L_s and L_t are used together for measuring the pointwise KL-divergence. On the other hand, in case of the alignment stage, only the two parallel corpora are only used.

The parallel corpora are consist of the KR–EN parallel corpus[1] compiled by Seo *et al*. [13] and the FR–EN parallel sub-corpus from the Europarl parallel corpus[2] [14]. The KR–EN parallel corpus and KR/FR comparable corpora come from several news articles[3]. All words were tokenized and POS-tagged by the following tools: U-tagger[4] for Korean, and TreeTagger[5] for English and French. Table 1 shows the statistics about the parallel corpora and the comparable corpora that used at our experiments.

Table 1. The statistics for KR–EN/FR–EN parallel corpora and KR/FR comparable corpora. The symbol "-" means N.A.

#	Parallel		Comparable	
	Korean/English	**French/English**	**Korean**	**French**
Sentences	433,151	500,000	418,474	426,341
Single-word types	54,742/218,114	21,712/81,608	214,484	153,083
Multi-word types	4,407/ -	2,055/ -	-	-

As you can see Table 1, word types of single-/multi-words for Korean have

[1] https://sites.google.com/site/nlpatkmu/Resources/Corpora

[2] http://www.statmt.org/europarl/

[3] http://www.naver.com (KR), http://www.abc.es (ES), http://www.lemonde.fr (FR)

[4] http://nlplab.ulsan.ac.kr/

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

more various types than French. This is caused by the Korean characteristic that Korean single-words basically have at least one morpheme. That is, the multi-word types of Korean consist of several morphemes or words (e.g., namely a compound word). Occasionally, a single Korean compound word has two or three morphemes.

All French n-grams are filtered by 9 POS patterns[6] (for *bi*/*tri*-grams) that come from the approach by Bouamor *et al*. [1], and all Korean n-grams are filtered by several patterns[7] [15] with considering the Korean characteristics. The n-grams that appear more than 2 times in the corpus are then extracted (e.g., KR: 54,742 of 2,621,316 and FR: 21,712 of 517,461).

To evaluate the proposed method, we manually built a set of evaluation dictionaries (KR to FR, FR to KR). Each dictionary contains 94 (KR) and 138 (FR) source multi-words and its translations. The evaluation dictionaries are constructed by following steps: Firstly, we collect top 200 high-frequent words from the parallel corpora. After that, several source words that its translations are not in the parallel corpora are removed. These translations are manually collected from the Web dictionary[8]. In case of the target translations, some translations could be rare in the target corpus, even though their corresponding source multi-words are frequent in the source corpus.

4.2. Result

The first experiment concerns the case of FR to KR. As mentioned before, we evaluate 138 FR multi-word candidates. The accuracies of 20.2 ~ 68.8% have been shown when PMI is calculated (e.g., described at step II in section 3.1). This means that 20.2% of the 138 FR multi-word candidates have at least one correct KR translation when the PMI is calculated. On the contrary, only 12.3% of source candidates are found when the pointwise KL-divergence is calculated. These metrics are for leaching out bad multiword candidates (e.g., rare phrases), also for seeing whether those n-grams are compositional or non-compositional units. While PMI usually computes *phraseness* [9] of co-related units (e.g., collocations) in the monolingual corpora, pointwise KL-divergence takes both *phraseness* and *informativeness* [9] at the same time. As we can see the Figure 1, the method using PMI is much higher than the method using the pointwise KL-divergence. It comes from the probabilities of *uni*-grams in the comparable corpora, and those *uni*-grams affect to the total probabilities to leach out most general terms. Consequentially, pointwise KL-divergence makes general terms are eliminated from the corpora.

---

[6] N-N, J-N, N-J, J-J-N, J-N-J, N-J-J, N-N-J, J-N-N, N-P-N (J : adjective, N: noun, P : preposition)

[7] N-N, N-g-N, V-E-N, J-E-N, N-N-N (E : ending-modification, g : genitive case marker, V : verb)

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

This phenomenon is also shown at the opposite case, e.g., the case of the KR to FR. In this case, finding correct translations have shown much better performance than the FR to KR. We evaluate 94 KR multi-word candidates with at least one correct FR translation. Accuracies of 47.9% to 71.2% have been shown for the PMI are considered (24.5% to 39.3% with the pointwise KL-divergence). In fact, the translations are not quite frequent than its source words. It means that its contexts cannot be enough to connect both source and target words. Therefore, considering a measurement of the similarity between its components can expect higher performance.

In our experimental environment, most target translations (e.g., Korean words) consist of a single-word or a compound word. Such compound words can be caused by the Korean POS tagger. For example "여론조사" (yeo-ron-jo-sa; opinion poll) can be split into two separate words having its own sense, "여론"

(yeo-ron; opinion) and "조사" (jo-sa; poll). If the POS tagger divides those words into several separate words, the form of many-to-many will be more shown on the evaluation result.

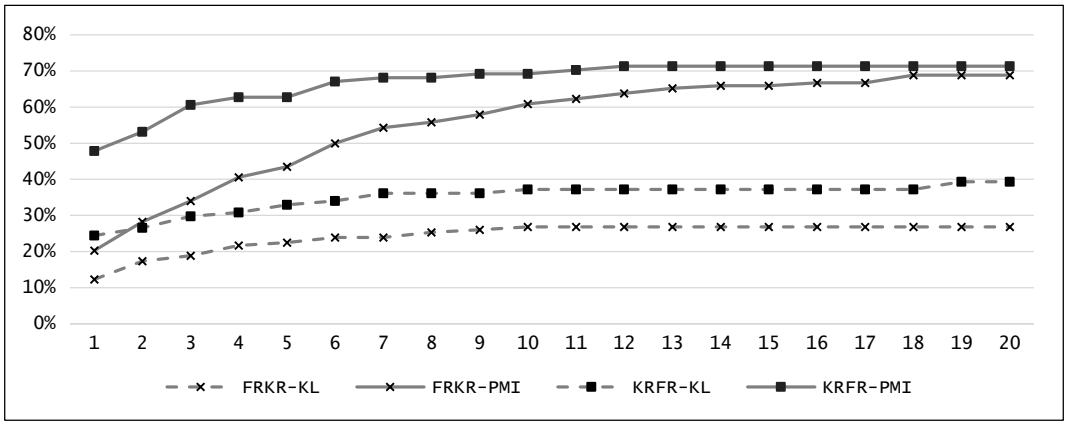

Fig. 1. Accuracy in rank 20 (*x*: rank, *y*: accuracy).

5. Conclusion

In this paper, we explore the method about constructing bilingual multi-word lexicons via a pivot language. The proposed method uses mainly parallel corpora in a resource-poor language pair. It builds context vectors in English to compare two different languages (e.g., source and target). We test the proposed method in both cases, KR to FR and FR to KR, and the experimental results are successful.

In future works, we plan to focus on collocations to align multi-word candidates and a measurement to take components into account. Furthermore, we will study about the term-hood of collocations to adopt this system.

Acknowledgements. 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. Bouamor, N. Semmar, and P. Zweigenbaum, Automatic Construction of a Multiword Expressions Bilingual Lexicon: A Statistical Machine Translation Evaluation Perspective, *Proc. of 3rd Workshop on Cognitive Aspects of the Lexicon (CogALex-III)*, Mumbai, India (2012), pp. 95-108.

[2] F. Smadja, K. R. McKeown, and V. Hatzivassiloglou, Translating collocations for bilingual lexicons: a statistical approach, *Computational Linguistics*, California, USA (1996), Vol. 22, No. 1, pp. 1-38.

[3] B. Daille, D.-k. Samuel, and M. Emmanuel, French–English multi-word terms alignment based on lexical content analysis, *Proc. of 4th Int. Conf. on Language Resources and Evaluation* (*LREC 2004*), Lisbon, Portugal (2004), Vol. 3, pp. 919-922.

[4] D. Wu, and X. Xuanyin, Learning an English–Chinese Lexicon from a Parallel Corpus, *Proc. of 1st Conf. on Association for Machine Translation in the Americas*, Columbia, USA (1994), pp. 206-213.

[5] B. Lu, and B. K. Tsou, Towards Bilingual Term Extraction in Comparable Patents, *Proc. of 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC 2009)*, Hong Kong, China (2009), pp. 755-762.

[6] T. Tsunakawa, N. Okazaki, and J. Tsujii, Building Bilingual Lexicons using Lexical Translation probabilities via pivot languages, *Proc. of 6th Int. Conf. on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco (2008), pp. 1664-1667.

[7] H.-W. Seo, H.-S. Kwon, and J.-H. Kim, Context-based Lexicon Extraction via a Pivot Language, *Proc. of 13th Conf. on Pacific Association for Computational Linguistics (PACLING 2013)*, Tokyo, Japan, (2013).

[8] I. A. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger, Multiword Expressions: a pain in the neck for nlp. *Proc. of 3rd Int. Conf. on Intelligent Text Processing and Computational Linguistics*, Mexico city, Mexico (2002), pp. 1-15.

[9] T. Tomokiyo, and M. Hurst, A Language Model Approach to Keyphrase Extraction, *Proc. of Workshop on Advances in ACL for Multiword expressions: analysis, acquisition and treatment*, Sapporo, Japan (2003), Vol. 18, pp. 33-40.

[10] A. Kunchukuttan, Multiword Expression Recognition, *Indian Institute of Technology*, Mumbai, India (2007).

[11] S. S. Piao, and T. EcEnery, Multiword Unit Alignment in English–Chinese Parallel Corpora, *Proc. of Conf. on Corpus Linguistics 2001*, Lancaster, United Kingdom (2001), pp. 466-475.

[12] R. Rapp, Automatic Identification of Word Translations from Unrelated English and German Corpora, *Proc. of 37th Annual Meeting on Association for Computational Linguistics (ACL99)*, Maryland, USA (1999), pp. 519-526.

[13] H.-W. Seo, H.-C. Kim, H.-Y. Cho, J.-H. Kim and S.-I. Yang, Automatically Constructing English–Korean Parallel Corpus from Web Documents, *Proc. of 26th Conf. on Korea Information Processing Society Fall Conference*, Seoul, South Korea (2006), Vol. 13, No. 2, pp. 161-164.

[14] P. Koehn, Europarl: a parallel corpus for statistical machine translation, *Proc. of 10th Conf. on Machine Translation Summit*, Phuket, Thailand (2005), pp. 79-86.

[15] B.-M. Kang, and H. Kim, Sejong Korean Corpora in the Making, *Proc. of 4th Int. Conf. on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal (2004), Vol. 5, pp. 1747-1750.

[16] L. Qiu, Verb Classification Using Bilingual Lexicon and Translation Information Tibetan Language, Int. Journal of Multimedia and Ubiquitous Engineering (2014), Vol. 9, No. 2, pp. 45-62

Received: August 15, 2014