Advanced Science and Technology Letters
Vol.54 (HCI 2014), pp.95-99
http://dx.doi.org/10.14257/astl.2014.54.24

# Constructing Bilingual Multiword Lexicons for a Resource-Poor Language Pair

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

Computer Engineering Institute, Korea Maritime and Ocean University,
Dongsam-dong, Yeongdo-gu, Busan, South Korea
wonn24@gmail.com, hong8c@naver.com, minah014@outlook.com, jhoon@kmou.ac.kr

**Abstract.** This paper presents a method for constructing bilingual multiword lexicons for a resource-poor language pair such as Korean–French. For this, at first, we identify multiword candidates from parallel corpora, and then use the *pivot context approach* [1] to align those candidates. Our empirical study shows encouraging results (e.g., accuracy), even though this study is ongoing.

**Keywords:** Bilingual Lexicon, Multiword Units, MWU Identification, MWU Alignment, Pivot Language, Parallel Corpora, Comparable Corpora.

## 1 Introduction

A bilingual lexicon is broadly used for many natural language processing (NLP) domains. For instance, the lexicon is useful to improve a performance of statistical machine translation (SMT) system [2]. While a lot of studies about the lexicon have been proposed, there are a lot of challenges in this field. Especially, extracting bilingual multiword units (MWU) is even more complicated than single-word units.

Some studies [3, 4] have proposed bilingual MWU extraction methods from bilingual parallel corpora. Both studies extract such units in resource-rich language pairs such as English–*. In general, collecting data sets in EN–* such as parallel corpora is much easier than resource-poor language pairs like Korean–French (KR–FR). In contrast with the former resource-rich case, we construct bilingual multiword lexicons for a resource-poor language pair. Besides, we explore multiwords, especially noun phrases, in parallel corpora for our experiments.

The remaining parts of the paper are organized as follows: section 2 describes the main method to construct bilingual multiword lexicons. Section 3 presents experimental results with some discussions. Finally, section 4 concludes and mentions future works that we have planned.

## 2 Bilingual multiword lexicon construction

In this section, we describe a method for constructing bilingual multiword lexicons for a resource-poor language pair (e.g., KR–FR) using one pivot language (e.g., EN). The proposed method can be split into two stages, the *multiword identification* and

ISSN: 2287-1233 ASTL

*alignment*. In this paper, we use two parallel corpora in the source language–pivot language ($L_s$–$L_p$) and the target language–pivot language ($L_t$–$L_p$) for both stages.

### 2.1 Multiword identification

Multiword candidates are respectively identified from two monolingual corpora in $L_s$ and $L_t$. Since we assume that the pivot language $L_p$ is just used for connecting both languages $L_s$ and $L_t$, multiword candidates in $L_p$ are unnecessary. The algorithm for identifying multiword candidates in each language is described as follows: Firstly $n$-grams ($2 \leq n \leq 3$) (except stop-words and punctuations) are collected from two monolingual corpora in $L_s$ and $L_t$, respectively. And then, co-occurrence metrics such PMI (threshold $\geq 0$) are measured between $n$-grams. Finally, matched $n$-grams with specific POS patterns (e.g., noun-phrase) are taken into account for final multiword candidates.

### 2.2 Bilingual multiword alignment

After multiword candidates are collected, we use the *pivot context approach* [1] to align those candidates. We just add multiword candidates to parallel corpora in order to deal with the candidates as a single word, because the pivot context approach only handle single tokens. As a result, except two pivot corpora (e.g., EN from KR–EN and FR–EN), both monolingual corpora in $L_s$ and $L_t$ include not only single words but also multiwords.

All words in parallel corpora are represented to context vectors by pivot single words, and are weighted by the word association score (e.g., chi-square test) between source single/multi words (resp. target single/multi) and pivot single words. Then each source vector is compared to all target vectors to get a vector similarity score based on the cosine similarity. Finally, the top $k$ translation candidates for one source word are sorted.

## 3 Experimental result

In this section, we show preliminary experimental results with discussion. Unfortunately there are no evaluation standards for extracting KR–FR noun phrases as multiwords. Furthermore, there are no similar cases with this method. Therefore, we were not able to compare with other researches by now. In this paper, we focus on measuring the accuracy on rank 20 in the KR–FR language pair.

### 3.1 Data

In order to evaluate the method, both the KR–EN parallel corpus[1] (433,151 sentence pairs) [5] and the FR–EN parallel sub-corpus (500,000 sentence pairs) that randomly selected from the Europarl parallel corpus[2] [6] are used for experiments. All words were tokenized and POS-tagged (part-of-speech-tagged) by the following tools: U-tagger[3] for Korean, and TreeTagger[4] for both English and French. After all POS-tags are attached, light POS filters[5] [2, 7] extract noun phrases for experiments. Lastly, $n$-grams appear more than 2 times in corpora are filtered out (e.g., KR: 54,742 of 2,621,316 and FR: 27,712 of 517,461).

To evaluate the method, we manually built a set of evaluation dictionaries (KR to FR, FR to KR) by using the Web dictionary[6]. Each dictionary contains 94 (KR) and 138 (FR) source multiwords and its translations. For the first time, we gather top 200 high-frequent words from the parallel corpora. Those source words that its translations are also in the parallel corpora are used to evaluate in this paper. These translations are manually collected from the dictionary. In case of the target translations, some of the translations could be rare in the target corpus, even though their corresponding source multiwords is frequent in the source corpus.

### 3.2 Result and discussion

The first experiment explores the FR→KR. When PMI are considered as an association metric (more details in section 2.1), 20.2% of FR candidates are correctly aligned at rank 1. The accuracy increases from 20.2% to 68.8% when the rank 20 is considered. Alternatively, we compute the pointwise KL-divergence [8] for comparison with PMI. In this case, only 12.3% of FR candidates are correctly aligned with KR translations. In our experimental environment, most KR translations consist of a combination of single words or several morphemes have independent meanings such as a compound word. For example, "지문" (ji-mun; finger print) is not able to be split into two separated words, but it has a combination of two senses "지" (ji; finger) and "문" (mun; print). This is caused by Korean characteristics. Another example "여론조사" (yeo-ron-jo-sa; opinion poll) is able to be split into two separate words each having its own sense, "여론 (yeo-ron; opinion)" and "조사 (jo-sa; poll)". The latter example is caused by the Korean POS tagger. If the Korean POS tagger divides such words into separated words, the form of many-to-many will be shown on the result.

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

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

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

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

[5] Korean POS filter: N+N, N+g+N, V+E+N, J+E+N, N+N+N, French POS filter: N+N, J+N, N+J, J+J+N, J+N+J, N+J+J, N+N+J, J+N+N, N+P+N (N: noun, J : adjective, P : preposition, V : verb, E : ending-modificaion, g : genetive case marker)

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

In fact, the translations in the experimental result are not quite frequent than its source words. It means that its contexts are not enough to connect both source and target words. Therefore, the way considering a measurement of similarities between its components when vector similarity scores are computed must improve accuracies of experimental results.

The second experiment explores the KR→FR, and it is almost similar with the former case, but the experimental result shows much better performance than the former. Accuracies (47.9% to 71.2%) have been shown for the PMI are considered (24.5% to 39.3% by the pointwise KL-divergence). In this case, the accuracy is generally higher than the opposite case (e.g., the FR→KR).

## 5 Conclusion

In this paper, we describe the method for collecting bilingual multiword lexicons for a resource-poor language pair. The algorithm uses *pivot context approach* for the alignment. Unfortunately, the approach are not able to compare directly with other multiword extraction methods. However, constructing pivot context vectors helps to align multiword candidates with its translations. The experimental results are successful in both case (KR to FR and FR to KR). However, some problems are shown as a result. The context vectors of translations are rare than its source words. In case of the KR target language, splitting single compositional words into two separate words can improve the performance. Also a novel similarity measurement that take its components into account can helpful.

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

**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]

## References

1. Seo, H.-W., Kwon, H.-S., Kim, J.-H.: Context-based Lexicon Extraction via a Pivot Language. In: Conference of the Pacific Association for Computational Linguistics, Japan (2013)
2. Bouamor, D., Semmar, N., Zweigenbaum, P.: Automatic Construction of a Multiword Expressions Bilingual Lexicon: A Statistical Machine Translation Evaluation Perspective. In: 3rd Workshop on Cognitive Aspects of the Lexicon, 95--108. India (2012)
3. Koehn, P.: Europarl: a parallel corpus for statistical machine translation. In: Conference on the 10th Machine Translation Summit, 79--86. Thailand (2005)

4. Kang, B.-M, Kim, H.-G.: Sejong Korean Corpora in the Making. In: 4th International Conference on Language Resources and Evaluation (LREC 2004), 5:1747--1750. Portugal (2004)
5. Seo, H.-W., Kim., H.-C., Cho, H.-Y., Kim, J.-H., Yang, S.-I.:, Automatically Constructing English–Korean Parallel Corpus from Web Documents. In: 26th on Korea Information Processing Society Fall Conference, 13(2):161--164. South Korea (2006)
6. Daille, B., Samuel, D.K., Morin, E.: French–English multi-word terms alignment based on lexical content analysis. In: 4th International Conference on Language Resources and Evaluation (LREC 2004), 3: 919--922. Portugal (2004)
7. Wu, D., Xia, X.: Learning an English–Chinese Lexicon from a Parallel Corpus. In: 1st Conference of the Association for Machine Translation in the Americas, 206--213. USA (1994)
8. Tomokiyo, T., Hurst, M.: A Language Model Approach to Keyphrase Extraction. In: the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment, 18: 33--40. USA (2003)