Analyzing Errors in Bilingual Multi-word Lexicons Automatically Constructed through a Pivot Language

Hyeong-Won Seo¹  ·  Jae-Hoon Kim†
(Received November 4, 2014  ·  Revised December 17, 2014  ·  Accepted December 17, 2014)

Abstract: Constructing a bilingual multi-word lexicon is confronted with many difficulties such as an absence of a commonly accepted gold-standard dataset. Besides, in fact, there is no everybody’s definition of what a multi-word unit is. In considering these problems, this paper evaluates and analyzes the context vector approach which is one of a novel alignment method of constructing bilingual lexicons from parallel corpora, by comparing with one of general methods. The approach builds context vectors for both source and target single-word units from two parallel corpora. To adapt the approach to multi-word units, we identify all multi-word candidates (namely noun phrases in this work) first, and then concatenate them into single-word units. As a result, therefore, we can use the context vector approach to satisfy our need for multi-word units. In our experimental results, the context vector approach has shown stronger performance over the other approach. The contribution of the paper is analyzing the various types of errors for the experimental results. For the future works, we will study the similarity measure that not only covers a multi-word unit itself but also covers its constituents.

Keywords: Bilingual lexicon, Multi-word units, Parallel corpora, Pivot language, Error analysis

1. Introduction

A bilingual lexicon is broadly used for many natural language processing (NLP) domains. Especially, such a lexicon is helpful to improve a performance of statistical machine translation (SMT) system [1]. There are still many challenges in this area, while lots of studies have been proposed. Furthermore, constructing a bilingual multi-word lexicon is more complicated than a single-word lexicon.

Some studies [2]-[5] have proposed bilingual multi-word extraction methods from parallel corpora. These studies extract multi-word units (MWUs) in resource-rich language pairs such as English–French (EN–FR) and English-Chinese (EN–CH). In general, collecting datasets such as parallel corpora in EN–FR is much easier than collecting in resource-poor language pairs such as Korean–French (KR–FR). Under these circumstances, Seo et al. [6] have proposed the method of constructing bilingual multi-word lexicons by using a pivot language in a resource-poor language pair, e.g., KR–FR. However, they did not compare the method with other general methods because of an absence of gold-standard datasets or other evaluation benchmarks.

In this paper, we focus on evaluating performance and on analyzing errors by comparing with one of general researches using phrase-tables by GIZA++. The remaining parts of the paper are organized as follows: Section 2 represents several works related with methods for constructing bilingual multi-word lexicons. Section 3 presents our experiments and Section 4 analyzes the results. Finally, Section 5 draws conclusions and gives future works that we have planned.

2. Related Works

The method for constructing a bilingual multi-word lexicon can be split into two stages, the identification and the alignment. The identification stage identifies multi-word candidates from
monolingual (i.e., source and target) corpora. We assume that all identified candidates are truly multi-word units, although the identification method described in this paper is not able to catch 100% of multi-word units. The alignment stage aligns the identified candidates with their translation equivalents. The overall structure of the method of constructing a bilingual multi-word lexicon is described in Figure 1.

2.1 Multiword Identification

The multi-word identification stage is to extract source multi-word candidates (resp. target multi-word candidates) from the source language–pivot language ($L_s$–$L_p$) parallel corpora (resp. $L_t$–$L_p$ parallel corpora). In this stage, multi-word units in a pivot language are unnecessary because these words may give rise to another of errors. Therefore, we assume that pivot single-words are enough to represent context vectors and to play the role of bridges that connect two languages (i.e., source and target).

Identifying multi-word candidates is summarized as follows: Firstly, all kinds of word bi-/tri-grams are extracted from the source monolingual corpus (resp. target monolingual corpus). Before the word bi-/tri-grams are extracted, all stop-words such as punctuations are removed. Secondly, a co-occurrence metric like pointwise mutual information (PMI) is computed in order to leach out bad multi-word candidates (i.e., rare phrases). The metric evaluates whether the co-occurrence is purely by chance or statistically significant. As a result, highly related multi-word candidates (i.e., frequent phrases) are selected by using the metric. Finally, specific POS patterns are used to remove irrelevant multi-word candidates. Just several simple regular expressions are enough to achieve this.

In this paper, we only concern about noun phrases because there is no commonly accepted definition about a MWU. All multi-word candidates identified by those steps are passed to the next step, i.e., the alignment stage.

2.2 Bilingual Multi-word Alignment

In this section, two approaches for aligning words are represented. To simplify the way to deal with multi-word units, an extracted multi-word candidate is made by putting together its component words as a single-word via a special character.

2.2.1 Context Vector Approach

Seo et al. [6] proposed a method of constructing a bilingual multi-word lexicon for a resource-poor language pair. The proposed method (denoted as the context vector approach) builds context vectors that representing the meaning of words as points in vector spaces. The approach is summarized as follows: Firstly, all punctuations and stop-words satisfying specific POS patterns except nouns, verbs, adjectives and adverbs are removed from each sentence in $L_s$–$L_p$ parallel corpora (resp. $L_t$–$L_p$ parallel corpora). Secondly, co-occurrence metric such as Chi-square test is computed to see how two words are related to each other. Thirdly, context vectors are built with the computed scores. All source words (resp. target words) are represented by the scores demonstrating the relationship between source words (resp. target words) and pivot words. Finally, vector distance measure such as cosine similarity is computed to see how close these context vectors are. And then, target vectors are sorted and ranked for each source word. Top k words are represented as translations of a source word. Each source word (resp. target word) could be a single-word or a multi-word.

As mentioned before, in this paper, only multi-word units in a source language (resp. target language) are focused in our experiments.

2.2.2 Phrase-table Approach

Tsunakawa et al. [7] proposed a method of increasing the size of a bilingual lexicon by using two other lexicons built from phrase-tables by GIZA++. The method (denoted as the phrase-table approach) combines two lexicons into one by calculating phrase translation probabilities.
Let $\tilde{a}_x$ be a phrase of language $x$ (e.g., source, target or pivot) $L_x$ and $(\tilde{a}_p, \tilde{a}_s)$ be a phrase pair of source–target language. The phrase translation probability $P(\tilde{a}_s|\tilde{a}_p)$ can be described as Equation (1).

$$P(\tilde{a}_s|\tilde{a}_p) = \frac{\sum_{a_p} P(\tilde{a}_s|a_p) P(a_p|\tilde{a}_p)}{\sum_{\tilde{a}_s} \sum_{a_p} P(a_p|\tilde{a}_s) P(\tilde{a}_s|a_p)}$$ (1)

All source and target words in two parallel corpora are aligned with probability scores. Aligning algorithm by phrase-tables is described as follows: Firstly, same pivot words (e.g., English) are matched by using a string matching algorithm. For matched pivot words, source and target words are paired by the scores. Finally, by using the scores, the top $k$ target words are sorted and ranked for each source word.

### 3. Experiments

Unfortunately, there is no gold-standard dataset for the method of extracting KR–EN multi-word units, (i.e., especially noun-phrases). Moreover, there are no similar cases or evaluation benchmarks for the data used in this study. Therefore, we compare the context vector approach with the phrase-table approach by using one common dataset. Through the comparison, we evaluate how useful the context vector approach is.

In this paper, we just focus on measuring the accuracy of the top 20 for KR–FR (resp. KR–ES), and take only a noun-phrase into account to concern multi-word units.

#### 3.1 Data

The KR–EN parallel corpus $^1$ (433,151 sentence pairs) $^8$ and the FR–EN (resp. ES–EN) parallel corpus (each $^8$–EN corpus contains 500,000 sentence pairs) that randomly extracted from the Europarl parallel corpora $^2$ $^9$ are used for experiments. All words are tokenized and POS-tagged by the following tools: U-\texttt{tagger}$^3$ for Korean, TreeTagger$^4$ for both English and French. After all words are POS-tagged, light POS filters $^2$$^10$ for noun phrases are used and the patterns are described in Table 1.

Additionally, word bi-/tri-grams appear less than 3 times in corpora are eliminated (i.e., the number of unique word bi-/tri-

---

1. https://sites.google.com/site/nlpatkmu/Resources/Corpora
4. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

---

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 http://dic.naver.com
Analyzing Errors in Bilingual Multi-word Lexicons Automatically Constructed through a Pivot Language

specific POS patterns described in Table 1, and complete evaluation sets are represented in Table 3.

The number of all POS-tagged/filtered source words are represented as collected. The number of collected words appear more than 2 times in each corpus are represented as evaluation. This table presents only the number of source words and the number of their translations are 1 on average.

Table 1: The list of noun phrases (N: noun, J: adjective, P: preposition, V: verb, E: ending-modification, g: genitive case marker)

<table>
<thead>
<tr>
<th>Korean</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-N</td>
<td>N-N</td>
</tr>
<tr>
<td>N-g-N</td>
<td>J–N / N–J</td>
</tr>
<tr>
<td>J-E-N</td>
<td>N–N-J / J–N-N</td>
</tr>
<tr>
<td>N-N-N</td>
<td>N–P–N</td>
</tr>
</tbody>
</table>

Table 2: The statistics of parallel corpora

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentences</td>
<td>433,151</td>
<td>500,000</td>
<td>500,000</td>
</tr>
<tr>
<td>SWUs</td>
<td>43,550/41,626</td>
<td>22,364/18,299</td>
<td>28,722/18,126</td>
</tr>
<tr>
<td>MWUs</td>
<td>3,640/0</td>
<td>1,606/0</td>
<td>1,346/0</td>
</tr>
</tbody>
</table>

Table 3: The statistics of extracted evaluation set.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Korean</td>
<td>15,287</td>
<td>8,489</td>
</tr>
<tr>
<td>French</td>
<td>28,961</td>
<td>15,540</td>
</tr>
<tr>
<td>Collected</td>
<td>754</td>
<td>426</td>
</tr>
<tr>
<td>Evaluation</td>
<td>630</td>
<td>529</td>
</tr>
</tbody>
</table>

3.2 Experimental Result

The accuracies with the evaluation sets described in Table 3 are represented in Figure 2. In the case of KR→FR, the context vector approach shows 48.7 percent accuracy and the phrase-table approach method shows 17.7 percent accuracy when top 20 candidates are considered. In the opposite case (i.e., FR→KR), the context vector approach shows 43.8 percent accuracy and the phrase-table approach shows 20.8 percent accuracy. The case of KR→ES is similar with the case of KR→FR. In the case of KR→ES, the context vector approach shows 56.8 percent accuracy and the phrase-table approach shows 31.9 percent accuracy when top 20 candidates are considered. In the case of ES→KR, the context vector approach shows 45.6 percent accuracy and the phrase-table approach shows 20.4 percent accuracy.

The experimental environments of two methods are not able to be compared directly (language pairs in the context vector approach: Korean–French–Spanish, the phrase-table approach: Chinese–Japanese). Nevertheless, the reason why such comparison is meaningful is that both methods use English as a pivot language. As a result, the context vector approach exceeds the performance of the phrase-table approach. This result looks quite meaningful, but there are also several defects.

Figure 2: The accuracies of two methods (x: rank, y: accuracy).
4. Error Analysis

The statistics of error frequencies are represented in Table 4 and Table 5. The interesting fact about error frequencies is that nearly half of evaluated source words are very low-frequent words, i.e., they appear less than or equal to 10 times in each corpus and 10 of the 500,000 sentences is extremely low number. Consequently, almost 60% (233 of 387) of Korean error words in the KR–FR corpus have only a few context factors to represent vectors. As for the case of KR–ES, it is pretty the same. This scarcity problem derives some sort of errors (more details below).

Table 5: The statistics of errors frequencies in Korean-French

<table>
<thead>
<tr>
<th>Case</th>
<th>Source word</th>
<th>Gloss</th>
<th>Translation</th>
<th>Top 1</th>
<th>Top 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>point de vue</td>
<td>point of view</td>
<td>[kwan-jeon]</td>
<td>(shi-gak; point of view)</td>
<td>(gyun-gi; point of view)</td>
</tr>
<tr>
<td>II</td>
<td>régime fiscal</td>
<td>tax system</td>
<td>[se-geum-je-do]</td>
<td>(se-geum-gong-je; tax deduction)</td>
<td>(jip-jeop-se; direct tax)</td>
</tr>
<tr>
<td>III</td>
<td>jugement de valeur</td>
<td>value of judgement</td>
<td>[ga-chi-pan-dan]</td>
<td>(pan-dan-ryuk; judgment, discernment, sense)</td>
<td>(pan-dan; judgement, adjudication, decision)</td>
</tr>
</tbody>
</table>


The case I means that synonyms are extracted but they are not included in an evaluation dictionary. As we mentioned before, the evaluation dictionary has one translations in average. As you can see, the example in Table 7, the French multi-word point de vue “point of view” has the Korean translation [kwan-jeon] in the FR→KR evaluation dictionary. The Korean word [kwan-jeon] also has meanings of (shi-gak) and (gyun-gi), but these words are not included in the dictionary. For this reason, the two translations, (shi-gak) and (gyun-gi), are marked as incorrect. Total 2.3% (KR→FR: 9), 15.3% (KR←FR: 54), 6.0% (KR→ES: 11), and 6.6% (KR←ES: 19) of errors are belong to this case.

Secondly, the case II means that extracted translation is incorrect but is a part of a same topic. For example, the French multi-word régime fiscal “tax system” means (se-geum-je-do) in Korean. However, another words (se-geum-gong-je) “tax deduction” (jip-jeop-se)

Several types of the errors can be described as follows, and the statistics of these error types are represented in Table 6.

Table 6: The statistics of errors frequencies in Korean-Spanish

<table>
<thead>
<tr>
<th>Case</th>
<th>Source word</th>
<th>Gloss</th>
<th>Translation</th>
<th>Top 1</th>
<th>Top 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>point de vue</td>
<td>point of view</td>
<td>[kwan-jeon]</td>
<td>(shi-gak; point of view)</td>
<td>(gyun-gi; point of view)</td>
</tr>
<tr>
<td>II</td>
<td>régime fiscal</td>
<td>tax system</td>
<td>[se-geum-je-do]</td>
<td>(se-geum-gong-je; tax deduction)</td>
<td>(jip-jeop-se; direct tax)</td>
</tr>
<tr>
<td>III</td>
<td>jugement de valeur</td>
<td>value of judgement</td>
<td>[ga-chi-pan-dan]</td>
<td>(pan-dan-ryuk; judgment, discernment, sense)</td>
<td>(pan-dan; judgement, adjudication, decision)</td>
</tr>
</tbody>
</table>


The case I means that synonyms are extracted but they are not included in an evaluation dictionary. As we mentioned before, the evaluation dictionary has one translations in average. As you can see, the example in Table 7, the French multi-word point de vue “point of view” has the Korean translation [kwan-jeon] in the FR→KR evaluation dictionary. The Korean word [kwan-jeon] also has meanings of (shi-gak) and (gyun-gi), but these words are not included in the dictionary. For this reason, the two translations, (shi-gak) and (gyun-gi), are marked as incorrect. Total 2.3% (KR→FR: 9), 15.3% (KR←FR: 54), 6.0% (KR→ES: 11), and 6.6% (KR←ES: 19) of errors are belong to this case.

Secondly, the case II means that extracted translation is incorrect but is a part of a same topic. For example, the French multi-word régime fiscal “tax system” means (se-geum-je-do) in Korean. However, another words (se-geum-gong-je) “tax deduction” (jip-jeop-se)
"direct tax" are extracted as results. These three Korean words are about a tax. This is caused by a lack of context vectors representing words. If each word has an enough context vector, e.g., the size of corpus is bigger than now, this kind of problems would be solved. Total 20.7% (KR→FR: 80), 44.1% (KR→FR: 156), 32.1% (KR→ES: 59), and 30.9% (KR→ES: 89) of errors are belong to the case II.

Thirdly, the case III means that several translations corresponding to components are extracted. For example, as represented in Table 7, two Korean translations are related with the component word jugement of French multi-word jugement de valeur “value of judgement”. This is because the whole multi-word jugement de valeur has a poor context vector, while its component word jugement has a rich context vector. This means that we need to get components involved in a whole word when vector similarity scores are calculated. In other words, the total similarity score should be a sum of the similarity scores of all involved words, i.e., a whole word plus its components words. For example, for a certain source word “x”, let “x” be a similarity score between “x” and the target multi-word jugement de valeur, and “y” be the score for jugement, and “z” be the score for valeur. And the total similarity score “x” is the “x+y+z”. If this calculation method is conducted, better performances are expected. Total 37.2% (KR→FR: 144), 27.1% (KR→FR: 96), 34.8% (KR→ES: 64), and 16.7% (KR→ES: 48) of errors are belong to the case III.

5. Conclusions

This paper evaluated performance of the context vector approach that constructs bilingual multi-word lexicons using a pivot language and context vectors and also analyzed errors. The method built context vectors from two parallel corpora, and compared them to find out the top k target vectors that were highly related with a source vector. To evaluate the method, we compared the method with the phrase-table approach. This task was a meaningful comparison because both methods used a common pivot language, English. Both methods took common multi-word candidates and aligned them in a respective way. In our experimental results, the context vector approach has shown stronger performance over the phrase-table approach.

As we mentioned in Section 4, most errors came from a lack of context vectors. It could be a problem of sentence alignments for parallel corpora or a size of corpora. Besides, if a domain of two parallel corpora is the same, overall accuracy would be higher. Moreover, evaluation dictionaries have a problem. Most source words in the dictionaries have one translation. If source words get more translations (i.e., synonyms), overall accuracy would be higher. This problem can be fixed by manually extending translations in the dictionaries. However, considering component words are much heavier problem. As mentioned before, calculating similarity scores of related component words together is needed. To solve this problem, the novel measure that deals all components consisting a whole word should be considered. Otherwise, a size of corpora should be larger than now to have abundant context vectors.

For the future work, we will extend the evaluation dictionaries by experts. Furthermore, we will study about the similarity measure that not only covers a whole word but also covers all component words.

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]

This paper is extended and updated from the short version that appeared in the Proceedings of the International Symposium on Marine Engineering and Technology (ISMT 2014), held at BEXCO, Busan, Korea on October 17-19, 2014.

References


