LETTER

Automatically Extracting Parallel Sentences from Wikipedia Using Sequential Matching of Language Resources

Juryong Cheon and Youngjoong Ko

SUMMARY In this paper, we propose a method to find similar sentences based on language resources for building a parallel corpus between English and Korean from Wikipedia. We use a Wiki-dictionary consisted of document titles from the Wikipedia and bilingual example sentence pairs from Web dictionary instead of traditional machine readable dictionary. In this way, we perform similarity calculation between sentences using sequential matching of the language resources, and evaluate the extracted parallel sentences. In the experiments, the proposed parallel sentences extraction method finally shows 65.4% of F1-score.

key words: Automatic Parallel Corpus Construction, Language Resources, Sentence Similarity Calculation, Wikipedia.

1. Introduction

Parallel corpora with many parallel sentences are used as an essential resource in language translation and analysis. In particular, cross-language information retrieval (CLIR) with bilingual or multilingual language needs a large amount of parallel corpora with good quality [1][2][3][4]. However, the task of building a parallel corpus is required lots of time and cost. These days, Wikipedia has been used to automatically build up a parallel corpus because it has comparable documents with the same title [5][6]. In this way, the constructed parallel corpus can contribute to natural language processing with machine translation, cross-language named entity recognition and CLIR.

In order to build parallel corpora of good quality, we selected document pairs with the same title in Wikipedia as comparable documents and regarded sentences in those documents as parallel sentence candidates. The main task of this research is to extract a pair of parallel sentences describing same contents from the parallel sentence candidates. Therefore, we explored new sentence similarity calculation methods using a variety of language resources. Conventionally, the sentence similarity measure methods have used term’s matching techniques using a machine readable dictionary (MRD). However, this method has two main problems. First, most of the comparable sentences of Wikipedia includes out-of-vocabulary (OOV). Second, MRD has a translation ambiguity problem because many words have a lot of different meanings in MRD. To solve the former problem, we constructed a bilingual dictionary using inter-wiki links, redirect pages and disambiguation pages from Wikipedia, called Wiki-Dic. Since the Wiki-Dic contains many named entities with multi-terms, such as the name of a person, location and a title of the movie, it can be very useful to detect and resolve OOVs. And then we also used bilingual example sentence pairs from Web dictionary1 to solve translation ambiguity. We estimated translation probabilities, which improve to select translated words, using bilingual example sentence pairs from the Web dictionary.

In the experiments, we evaluated our proposed method using the F1-score. The performance of our proposed method showed 65.4% of F1-score. This result is to extract better quality of the parallel sentences than 46.0% with the MRD.

The remainder of this paper is organized as follows. The next section describes some related work. In section 3, we then explain our proposed method in detail. In section 4, experimental results and comparison of result are presented. The paper is finally concluded in section 5.

2. Related Work

These days, tasks for building parallel corpora have been done because of increasing the importance of CLIR and machine translation. The traditional methods to extract parallel sentences first extract comparable document pairs and then they judge parallel sentence by comparing sentences in each document. But these tasks for building parallel corpora are required a lot of time and cost.

Methods to find similar documents have been widely studied before finding similar sentences. Resnik and Smith proposed a method to extract similar documents from Web based on the HTML [7]. Talvensaari et al. proposed a method to find similar documents based on translated words in the source language to the target language by using the main keywords [4]. Munteanu and Marcu [8] and Thuy et al. [9] found important words by using meta-information of documents and bilingual dictionary, and then they judged similar documents by frequency of words.

For studies to find similar sentences to construct a parallel corpus, Gale and Church proposed a method using the number of words and document length gap

† The author is with Dong-A University, 840, Hadan 2-dong, Saha-gu, Busan, 604-714, Republic of Korea.
a) E-mail: balendia@gmail.com
b) E-mail: youngjoong.ko@gmail.com (Corresponding author)

1 http://dic.daum.net/

3. Automatically Extracting Parallel Sentences

Wikipedia has a huge number of articles written in more than 200 languages and it is one of the most varied language resources in the world. The current version of Wikipedia expresses multilingual relationships by only inter-links. Therefore, Wikipedia can be used to extract document pairs of same topic as a comparable corpus and to extract sentence pairs from the document pairs as parallel sentences. The whole processes to extract parallel sentences from Wikipedia are shown in Fig. 1.

If there are inter-links of connection between document titles for each language, content of these documents could be almost similar. Thus sentence pairs in similar document pairs are regarded as parallel sentence candidates and the final parallel sentence pairs are selected from these candidates by using the proposed sentence similarity measure based on language resources.

To calculate sentence similarity for the parallel sentence extraction, term matching methods are traditionally used with a machine readable dictionary (MRD). However, this method has two problems. First, most of the comparable sentences of Wikipedia contains out-of-vocabulary (OOV). Second, MRD has a translation ambiguity problem because many words have a lot of different meanings in MRD. To solve these problems, we first constructed the Wiki-Dic on inter-wiki links, redirect pages and disambiguation pages from Wikipedia. This dictionary has many named entities with multi-terms, such as a title of the movie. And we also used bilingual example sentence pairs of a Web dictionary to estimate translation probabilities, which improve to select translated words. The bilingual example sentence pairs are extracted from sample sentences of each dictionary word to show its real usage. In addition, the multi-term pairs of Wiki-Dic are added to the bilingual example sentence pairs, because only a part of multi-term named entities occurs in many Wikipedia sentences as a kind of abbreviations, such as “Obama” for “Barack Obama.” Eventually, translation probabilities were estimated by the GIZA++ toolkit on the bilingual example sentence and multi-term pairs.

A sequential word matching method is developed for measuring similarity between sentences using the language resources described above. According to importance of the language resources, the matching order of a word, number, and translation probability is determined, as shown in Fig. 2. Since many multi-terms are in Wikipedia sentences and they are named-entities in most cases, multi-term matching with Wiki-Dic has the highest priority. As a second priority, numbers are another important evidence that two sentences tell the similar story. Translation probability that is given the last priority is for solving translation ambiguity in common words.

A modified Jaccard similarity based on the sequential word matching is used for similarity calculation to select parallel sentences from parallel sentence candidates, such as Eq. (1).

$$J(A,B) = \frac{M_2 + M_3}{M_1 + M_2 + M_4}$$

Step 1: Multi-terms matching is conducted with the Wiki-Dic between two sentences. Multi-terms occurred in source and target candidate sentences are extracted using the Wiki-Dic. The number of matching multi-terms

---

2 GIZA++ is an extension of the program GIZA (part of the SMT toolkit EGYPT).
is denoted by $M_i$ in Eq. (1) and then the matched multi-terms are removed from the sentences.

**Step 2**: Number matching is between two sentences. As in the **Step 1**, all number pairs occurred in both candidate sentences are extracted from each sentence. And the number of matched numbers is added to $M_i$. After matched numbers are removed from the sentences, the numbers of remaining words in source and target candidate sentences are assigned to $M_i$ and $M_k$, respectively.

**Step 3**: Finally, single-terms matching is conducted using translation probabilities between two sentences. If there are plural single-terms matched with translation probabilities, single-term with the highest translation probability is selected. The matching score $M_2$ is calculated by the sum of translation probabilities of matched single-terms.

An example of sentence similarity calculation using sequential word matching is shown in Fig. 3. The matching scores $M_1$, $M_2$, $M_3$, and $M_4$ are 2, 1.52, 6, and 6, respectively, and the final sentence similarity score is calculated as 0.251 by Eq. (1). We judged sentence pairs with 0.1 or over similarity scores to be the final parallel sentence pairs.

### 4. Experiments

In this section, we present results from several experiments that demonstrate the effectiveness of our proposed method.

#### 4.1 Data sets and evaluation measures

106,582 document pairs with the same title in English-Korean Wikipedia were used as comparable documents. 100 document pairs with comparatively long content and a little gap in length of each other were randomly selected for our experiments. Correct parallel sentence pairs that exist in these document pairs were labeled by five annotators manually. The proposed method was evaluated by using a precision, recall and F1-score.

#### 4.2 Experimental results of each language resources

To evaluate each language resources for parallel sentence extraction, we conducted first several experiments, including only MRD (baseline), only Wiki-Dic (WD), only translation probability from Web's bilingual example sentence pairs (TP-PS), only translation probability from Widi-Dic (TP-WD), only translation probability from Web example sentence pairs and Wiki-Dic (TP-PSWD). The experimental results are shown in Table 1.

<table>
<thead>
<tr>
<th>Language Resources</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRD (baseline)</td>
<td>43.9</td>
<td>48.3</td>
<td>46.0</td>
</tr>
<tr>
<td>WD</td>
<td>30.6</td>
<td>20.1</td>
<td>24.3</td>
</tr>
<tr>
<td>TP-PS</td>
<td>50.2</td>
<td>70.8</td>
<td>58.7</td>
</tr>
<tr>
<td>TP-WD</td>
<td>53.5</td>
<td>61.6</td>
<td>57.3</td>
</tr>
<tr>
<td>TP-PSWD</td>
<td>55.1</td>
<td>70.2</td>
<td>61.7</td>
</tr>
</tbody>
</table>

In this result, TP-PSWD achieved the best performance of F1-score 61.7%. Moreover, MRD seems like quite higher performance than WD. However, WD showed better results in the sequential combination with other resources in the next subsection.

#### 4.3 Experimental results of sequential matching with language resources

Experimental results that combine sequentially each language resource are presented in Table 2. In Table 2, the method using combination of MRD and TP-PS is denoted by MRD+TP-PS and the method using combination of WD and TP-PS is by WD+TP-PS. The matching of numbers is conducted in all the methods.
Table 2. Experimental results using combined language resources (%)

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+TP-PS</td>
<td>54.5</td>
<td>63.1</td>
<td>58.5</td>
</tr>
<tr>
<td>+TP-WD</td>
<td>54.4</td>
<td>62.7</td>
<td>58.3</td>
</tr>
<tr>
<td>+TP-PSWD</td>
<td>55.5</td>
<td>69.5</td>
<td>61.7</td>
</tr>
<tr>
<td>WD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+TP-PS</td>
<td>58.2</td>
<td>62.8</td>
<td>60.4</td>
</tr>
<tr>
<td>+TP-WD</td>
<td>54.1</td>
<td>61.6</td>
<td>57.6</td>
</tr>
<tr>
<td>+TP-PSWD</td>
<td>60.2</td>
<td>71.6</td>
<td>65.4</td>
</tr>
</tbody>
</table>

While the method using only WD showed lower performance than only MRD, WD combined with several TPs outperformed MRD+TPs in Table 2. It is because parallel sentence candidates from Wikipedia contain a lot of named entities that can be matched with WD and their matching is almost correct. Moreover, our proposed method using TP-PSWD can effectively match common words as well as single-term of named entities. It can be calculated as a score of translation probabilities of translated common words for ambiguity resolution. As a result, WD+TP-PSWD achieved the best performance of 65.4% of F1-score. Comparison of overall experimental results is shown in Fig. 4.

![Fig. 4. Comparison of overall experimental results by combined language resources (%)](image)

5. Conclusions and Future Work

Parallel corpora have been used as an essential resource in language translation and analysis. We proposed the method for automatically extracting parallel sentence between English and Korean Wikipedia. As a result, our proposed method achieved the significant high performance for extracting parallel sentences. Contribution of our method is to use only the language resources that are easily utilized around, and it can be further improved with passage of time by using Wikipedia.

For the future work, we will continue to study for getting the higher performance of automatically extracting parallel sentences. Moreover, we will conduct experiments not only English-Korean but also other languages for extracting parallel sentences.

Acknowledgments

References


IEICE TRANS. INF. & SIS., VOL.XX-X, NO.X XXXX XXXX