Automatic Construction of Bilingual Lexicon for Language where Word Boundaries are Unclear

WANG XINZHU (1510064)

School of Information Science,
Japan Advanced Institute of Science and Technology

February 10, 2016

Keywords: Bilingual Lexicon, Parallel Corpus, Word Segmentation, Knowledge Acquisition.

A bilingual lexicon is knowledge for natural language processing, which consists of a set of words and their translation in a different language. It is essential for multilingual information processing such as machine translation, cross lingual information retrieval and so on. However, it is difficult to manually make a bilingual lexicon including all words and all possible translation in the world. Therefore, a method to automatically construct the bilingual lexicon is indispensable. There are many previous studies that aim at acquiring the bilingual lexicon from a parallel corpus. However, there is a common problem on them. When we acquire the bilingual lexicon of the language where word boundaries are not clearly denoted in the text, such as Chinese, Japanese, Korean, Thai etc., sentences in the parallel corpus should be divided into the words by a morphological analysis tool. But the errors of the word segmentation influence the acquisition of the bilingual lexicon. Although the performance of the current morphological analysis tool is generally high, no tool is available or the performance is not good for low resource languages. Furthermore, the performance of the morphological analysis tool is not good when it is applied to the text of the specific domain (e.g. medical domain).

This thesis proposes a new method to automatically construct a bilingual lexicon from a large parallel corpus. The proposed method is less

influenced by the errors of word segmentation given by the existing morphological analysis tool. The target languages in this thesis are Chinese and English, that is, we try to construct Chinese-English bilingual lexicon. First, as preprocessing of the sentences, Chinese sentences are divided into the words, 1-gram of character, or 2-gram of character. Thus three parallel corpora are built by these three different preprocessing. Then, the translation pairs are obtained from each parallel corpus. In the previous methods that used the morphological analysis tool, the errors of word segmentation cause a problem. In the previous methods that did not use the tool but divided Chinese sentences into the characters, there is another problem that the words are not always correctly restored when several characters are merged into the words. In the proposed method, by combining these two approaches, the accuracy of extraction of translation pairs can be improved. Because the errors caused by one method could be ignored in another method. Furthermore, it is important to acquire new translation pairs that have not been compiled in the existing Chinese-English bilingual lexicons. The proposed method is evaluated from this point of view.

The details of the proposed method are as follows. First, we prepare a parallel corpus of Chinese and English. The parallel corpus should be a sentence aligned corpus, that is, the corpus is a set of a Chinese sentence and its translated sentence in English. Second, preprocessing is performed for both Chinese and English sentences. For English sentences, lemmatization is performed: each word in the sentences is converted to its base form. For Chinese sentences, three types of preprocessing are performed: word segmentation using the morphological analysis tool, division into character 1-gram and division into character 2-gram. Third, word alignment in each parallel corpus obtained by one of preprocessing is automatically determined by GIZA++, which is a public tool for statistical machine translation. Fourth, aligned Chinese and English words are extracted as temporal translation pairs. From three parallel corpora, three sets of the temporal translation pairs are obtained. Finally, since many incorrect pairs are included in the sets of temporal translation pairs, only the correct ones are chosen by heuristic rules and scoring. Four heuristic rules are introduced as follows. (1)When an English word is in a list of stop words, the translation pair is discarded. (2)When a Chinese word is only one char-

acter, the translation pair is discarded. Because words consisting of one character are rare in Chinese. (3)When an English word is a number, the translation pair is discarded, since the pair of the numbers in Chinese and English is not so meaningful. (4)When an English word appeared in six or more temporal translation pairs, all pairs are discarded. It means that one English word corresponds to many Chinese words. Such a situation is not common and they may be incorrect. After filtering out incorrect pairs by the rules, the score of each temporal translation pair is calculated. It is defined based on the relative frequency of the translation pairs in the parallel corpora. When the translation pair frequently appears in the corpora, the higher score is given for it. The translation pair is chosen if it is extracted from two or more parallel corpora and its score is greater than a certain threshold. A Chinese-English bilingual lexicon is obtained in this way.

An experiment is conducted to evaluate the proposed method. Two parallel corpora are used in this experiment: BBC news corpus and Parallel Corpus of China's Law Documents (PCCLD corpus). They are the parallel corpora of newspaper articles and law texts, respectively. The proposed method is compared with three baselines: methods using one type of preprocessing of Chinese sentences (word segmentation, division into character 1-gram or division into character 2-gram). The top 100 ranked translation pairs of each method are judged whether they are correct by human judgment. The accuracy of the proposed method is 0.94 for the news corpus and 0.95 for the law corpus. They are higher than that of all three baselines. Next, "new word ratio" is defined as proportion of the number of the translation pairs that is not compiled in an existing Chinese-English bilingual lexicon to the number of the translation pairs that appears five times or more in the training corpus. LDC English Chinese bilingual word lists is used as the existing bilingual lexicon. It consists of about 110,000 pairs of Chinese and English words. The new word ratio of the proposed method is 0.935 for the news corpus and 0.934 for the law corpus, respectively. They are 0.02 to 0.08 points higher than the baselines. Comparing the news and law corpus, the difference between the proposed method and the baseline was larger in the law corpus than the news. This is because there exists more unknown words in the law corpus than the news, and

the proposed method can acquire such unknown translation pairs than the baselines. The above results indicate that our proposed method is effective to construct the Chinese-English bilingual lexicon.

Future work of this study is as follows. Currently, only the top 100 translation pairs are evaluated. The quality of translation pairs with the lower scores should also be evaluated. In addition, it is also necessary to evaluate the recall of the proposed method. Word boundaries are not identified in other languages such as Japanese and Korean. The effectiveness of the proposed method should be verified for the language except for Chinese.