早稲田大学大学院情報生産システム研究科

博士論文概要

論文題目

A Study of Translation Equivalences at Various Levels of Granularity for Chinese-Japanese Technical Translation

申請者
Wei YANG

情報生産システム工学専攻
用例翻訳・言語処理研究

2017 年 5 月

China and Japan are producing a large amount of scientific articles and patents in their respective languages. The World Intellectual Property Organization (WIPO) Indicators show that China was the first country for patent applications in 2013. Japan was the first country for patent grants in 2013. Much of current scientific development in China or Japan is not readily available to non-Chinese or non-Japanese speaking scientists. Making Chinese patents and scientific texts available in Japanese, and Japanese patents and scientific texts in Chinese is a key issue for increasing economical development in Asia.

The crucial process of translation of technical texts should be helped by the use of machine translation (MT) as recognized by the Japan Science and Technology Agency (JST) project and State Intellectual Property Office (SIPO) project in China. Sentence-level aligned parallel corpora are an essential resource for data-driven statistical machine translation (SMT). For Chinese-Japanese SMT, the first problem is that there are almost no Chinese-Japanese parallel corpora publicly freely available in any domain. Another problem is that Chinese and Japanese do not have typographic boundaries in their writing system. Thus, for machine translation, word segmentation (tokenization), i.e., breaking sentences down into individual characters/words (tokenizes) is normally treated as the first step of preprocessing. But different segmentation conventions lead to different segmentation results at different levels of granularity (like a sentence composed of characters or words) between Chinese and Japanese that lead to inconsistencies in alignment that negatively affect the accuracy of translation. In Chinese-Japanese technical machine translation, a corpus may contain large amounts of domain-specific terms in words or multi-word expressions. Reasonable word/multi-word alignment in terms is an important processing task for machine translation in order to keep higher translation accuracy.

The main focus of this dissertation is exploiting several freely available linguist resources to address the scarcity problem of linguistic data and technical term translation in specific domains between Chinese and Japanese. The organization of the dissertation is as follows.

Chapter 1 [Introduction] describes the background and the basic knowledge of the research. Moreover, it gives an overview of the approach adopted in the thesis and presents the contributions of the thesis.

Chapter 2 [Chinese and Japanese characters] addresses the problem of scarcity of bilingual lexica between Chinese and Japanese. Extracting bilingual lexica from Chinese-Japanese parallel or comparable corpora are proposed in previous works (R.Raap, 1999; I.Vulić et al., 2011). However, the scarcity of parallel corpora and the parallelism of comparable corpora are still problems. We propose a method to construct a Chinese-Japanese lexicon by combining several automatic techniques on several freely available resources. Our method elaborates on the classical pivot language technique. With this method the quality lies below 45% of correct entries in our experiments. To improve the quality, we propose to combine three additional techniques: one time inverse consultation (76% of correct entries); Japanese kanji to Chinese hanzi character conversion (98.5% of correct entries); expansion through a Chinese synonyms table (98.5% of correct entries). The three additional methods allowed us to increase the quality of the Chinese-Japanese lexicon from less than 45% to 85% and get 45,386 entries in total. By comparison with a reference dictionary, 83% of the word pairs in our lexicon do not appear in a large reference dictionary, the EDR dictionary constructed by NICT (about 300,000 entries). We make use of our kanji-hanzi conversion method through out our work, because there exist a large amount of characters shared with the same meaning in the Chinese and Japanese writing systems. Our results show that they can be safely used as clues to align words or multi-word expressions.

Chapter 3 [Monolingual and bilingual term extraction] addresses the problem of the scarcity of digitalized terminological banks between Chinese and Japanese. The identification and translation of terms in patents and scientific texts is of course crucial in technical translation. We propose a method to improve Chinese-Japanese technical translation of patents and scientific texts by re-tokenizing the training corpus with aligned bilingual multi-word terms. We extract bilingual multi-word terms from the training corpus. These extracted terms are used in adjusting and balancing the tokenization between Chinese and Japanese technical data. We propose two experimental protocols to make use of the extracted bilingual terms in Chinese-Japanese statistical machine translation (SMT) experiments so as to select the better one. We obtain quality of correspondence of 80% in bilingual term extraction and a significant improvement of 1 BLEU score (p-value < 0.01) in translation accuracy. We combine using the kanji-hanzi conversion method (Chapter 2), and obtain better result in correspondence of bilingual terms (93%) and BLEU with 1.5 BLEU point improvement (p-value < 0.01). We also consider the cases where one side is a single-word term and the other side is a multi-word term. We obtain even better results with 95% in correspondence of terms and 2 BLEU point improvement (p-value < 0.01) in translation accuracy. Our pre-processing on terms has the effect of reducing the problem of different segmentation conventions across languages.

Chapter 4 [Quasi-parallel data construction] addresses the problem of scarcity of bilingual corpora between Chinese and Japanese. In SMT, the translation knowledge is acquired from the parallel sentences. Consequently, the quantity and the quality of the translation relations extracted between words or phrases between the source language and the target language depends on the quantity and the quality of the parallel sentences. We propose a method to construct a quasi-parallel corpus by using analogical associations based on large amounts of monolingual data and a small amount of parallel data, so as to improve Chinese-Japanese SMT quality. We generate amounts of new candidate sentences using analogical associations. We filter over-generated sentences using two filtering methods: one based on BLEU (used in Chapter 3 for evaluation) and the second one based on N-sequences. We also combine these two filtering methods. The N-sequence method allows us to keep sentences may be considered grammatically correct in 99% of the cases. The constructed quasi-parallel corpora are added to an existing training corpus to address the shortage of parallel corpora between Chinese and Japanese. The best result that we obtain is a very significant improvement of 6 BLEU points (p-value < 0.01) over a Chinese-Japanese baseline system. This kind of quasi-parallel sentences used as additional training data in SMT helps in acquiring more potential useful translation knowledge from the inflated training corpus. We also combine several proposed techniques and results described in previous chapters and this chapter for improving the translation accuracy for statistical machine translation in technical domains. The combination of these works lead to statistically significant improvement of 1.8 BLEU point with p-value < 0.01.

Chapter 5 [Conclusion and future work] summarizes and concludes the dissertation. Future directions on bilingual lexicon construction, bilingual multi-word term extraction for SMT and quasi-parallel corpus construction for solving the problem of scarcity of bilingual corpus are also presented.