JAIST Repository
https://dspace.jaist.ac.jp/

Title	琉日機械翻訳のための対訳コーパスの自動拡張について
Author(s)	久高, 優也
Citation	
Issue Date	2020-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/16394
Rights	
Description	Supervisor: 白井 清昭, 先端科学技術研究科, 修士（情報科学）

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

On Automatic Expansion of Parallel Corpus
for Ryukyu Japanese Machine Translation

1810056 Yuya Kudaka

Several studies on machine translation aim at translation between languages used in the same country, such as a dialog and a standard language of a country. One of them is machine translation between Ryukyu dialect (Okinawa dialect) and Japanese standard language (Japanese). The linguistic features of Ryukyu dialect and Japanese are quite different. Therefore, Ryukyu-Japanese machine translation is required for people who are non-native speakers of Ryukyu dialect to easily understand it.

In recent years, most methods of machine translation are based on statistical machine translation or neural machine translation, which learn translation models from a large amount of a bilingual corpus. The performance of these machine translation methods heavily depends on the amount of a training bilingual corpus. Therefore, the performance of them is not good for low-resource languages such as the Ryukyu dialect.

A method has been proposed to increase the amount of a bilingual corpus by automatically generating translation pairs to improve the performance of machine translation for low-resource languages. An attempt has also been made to apply this method to Ryukyu-Japanese Statistical Machine Translation (Ryukyu-Japanese SMT). However, the previous study didn't consider the quality of automatically generated translation pairs, so the performance of machine translation was not so good. Therefore, this thesis aims at improving performance of Ryukyu-Japanese SMT by expanding a bilingual corpus considering the quality, variety and amount of the translation pairs. Consideration of the quality means to generate natural translation pairs. Consideration of the variety means to construct an expanded bilingual corpus so that it contains not only similar sentences but a wide variety of sentences. Consideration of the amount means to avoid excessive expansion of a bilingual corpus. When too many sentences are added to an expanded bilingual corpus, many unnatural translation pairs are likely to be added too. In this thesis, we implement the proposed method and conduct experiments to translate Ryukyu dialect to Japanese to confirm that the above three ideas can contribute to improve the performance of Ryukyu-Japanese SMT.

Our proposed method to expand a bilingual corpus consists of two steps: generation of translation pair candidates and selection of translation pair candidates.

In the generation of translation pair candidates, new translation pairs are generated to enlarge an initial (small) bilingual corpus. For a given

translation pair in an initial bilingual corpus, if a word pair compiled in a Ryukyu-Japanese bilingual lexicon is included in the sentences of both source and target languages, new translation pairs are generated by replacing those words with other words, whose parts of speech are the same, in the bilingual lexicon.

In the selection of translation pair candidates, appropriate translation pairs are chosen among candidates from a point of views of their quality, variety and amount, then they are added to the initial bilingual corpus to make an expanded dialog corpus. To consider the quality of translation pairs, a score to evaluate fluency of a Japanese sentence in a translation pair is calculated, then translation pairs with high scores are selected. We propose two kinds of the scores: one is the generation probability of the sentence given by the probabilistic language model, the other is the difference of the generation probability between the derived (newly generated) sentence and its original sentence. To consider the variety, the translation pairs are selected so that the same number of translation pairs are generated from all the sentences in the initial bilingual corpus. In this way, words and contexts included in the initial bilingual corpus can be uniformly transferred in the expanded bilingual corpus. To consider the amount of translation pairs, the number of newly generated translation pairs is controlled. It enables us to prevent unnatural translation pairs from being excessively added to the bilingual corpus.

In the experiments, the following methods were evaluated and compared: no expansion (using only the initial bilingual corpus), the method of previous study, random selection, our method considering the quality (with two types of scores), our method considering the quality and variety, and our method combining quality-based selection and random selection. We trained an SMT model using the expanded bilingual corpus constructed by each proposed method or the baseline, translated test sentences of Ryukyu dialect into Japanese, and evaluated their performance using BLEU and RIBES. By considering both the quality and variety, BLEU improved up to 1.24 points and RIBES improved up to 2.54 points comparing with the random selection. By considering the variety, BLEU improved from 7 to 10 points and RIBES improved from 5 to 8 points comparing with the method considering the quality only. In addition, we examined the changes in BLEU and RIBES when the amount of the expanded bilingual corpus was changed. It was found that BLEU and RIBES decreased when more sentences were added to the expanded bilingual corpus. In addition, our method considering both the quality and variety outperformed no expansion method only when the number of expanded translation pairs was 2,000.

From the above results, it was found that the translation performance was

improved by expanding the bilingual corpus considering both the quality and variety. It was not necessary to add many expanded translation pairs, but it was important to optimize the amount of the expanded bilingual corpus appropriately. Especially, the method that considers only the quality but not the variety achieved poorer performance than other methods. This may be because sentences with high probability of the probabilistic language model tended to be short, and the number of words in the expanded bilingual corpus became small. Therefore, the obtained expanded bilingual corpus might not contain sufficient words and contexts. We can conclude that the basic idea of the proposed method to keep the variety of the expanded bilingual corpus is effective.

BLEU and RIBES of the proposed methods were improved comparing to the baseline, but the difference was small. Furthermore, when the amount of translation pairs in the expanded bilingual corpus was increased, BLEU and RIBES decreased. It may be caused by a naive method to generate translation pair candidates, where they are generated by randomly replacing the words. Most of generated translation pairs are unnatural, and only the small number of natural translation pairs are expanded. In the future, instead of using the naive method by word replacement, we will explore a method to generate translation pair candidates by paraphrase sentences with sophisticated natural language techniques as a better method of expansion of the bilingual corpus.