Neural Machine Translation for Low-Resource Languages Based on Iterative Data Augmentation of Source and Target Languages

2230028 Yoshifumi Hanada

In recent neural machine translation, which is the mainstream of current machine translation research, high translation accuracy is achieved by training translation models using large parallel corpora. However, in low-resource languages where available language resources are limited, the application of state-of-the-art neural machine translation technology has been behind. Nevertheless, several attempts have been made to improve neural machine translation performance for low-resource languages. For example, Iterative Back-Translation is a method that iteratively enhances translation performance by repeating training of a back-translation model (a model of translation from a target language to a source language) and expansion of a parallel corpus using this back-translation model. However, it has not been sufficiently explored whether such existing methods are effective for translation in all low-resource languages.

This research focuses on machine translation from Hokkaido Ainu, one of the extreme low-resource languages in Japan, to Japanese, and aims to enhance its performance. The existing Iterative Back-Translation is first applied to Ainu-Japanese translation and its effectiveness is empirically verified. Furthermore, we propose "Iterative Cyclic-Back-Forward-Translation" (hereafter we call it "Iterative Cyclic-Translation" for simplicity) as an improvement to Iterative Back-Translation.

First, we built an Ainu-Japanese parallel corpus by extracting texts from several documents. The Ainu language materials used for the construction of this parallel corpus were limited to Hokkaido Ainu. As of 2024, there is no standardized orthography for Hokkaido Ainu, and spelling and morpheme segmentation are different across documents, so the notation of Ainu sentences in the documents is manually converted into the uniform notation when they are compiled in the parallel corpus. Finally, we constructed a parallel corpus consisting of 23,337 translation pairs, which serves as our initial parallel corpus.

Our proposed Iterative Cyclic-Translation is carried out as follows. First, data augmentation is performed by cyclic translation on the initial parallel corpus. Sentences in the target language (Japanese) in the parallel corpus are translated to sentences in an auxiliary language and then translated back to Japanese, resulting in generating paraphrases of the sentences in the target language. These paraphrases are combined with the original sentences in the source language (Ainu) to acquire new translation pairs, which are then

added to the parallel corpus. In this research, English is chosen as the auxiliary language, as Japanese-English machine translation is well-developed and existing high-performance translation systems are available. Next, data augmentation is performed by back-translation. A back-translation model is trained using the augmented parallel corpus and is used to translate target language sentences into source language sentences. New translation pairs obtained by this procedure are added to the parallel corpus. Then, data augmentation is performed by forward translation. The parallel corpus is further augmented by training a translation model using the augmented parallel corpus and translating source language sentences into target language sentences using this translation model. The above procedures are repeated several times to obtain the final Ainu-to-Japanese translation model. In the existing Iterative Back-Translation, data augmentation by the back-translation, which generates new source language sentences, and data augmentation by the forward translation, which generates new target language sentences, are repeated. In contrast, in the proposed Iterative Cyclic-Translation, additional data augmentation by the cyclic translation, which generates new target language sentences, is incorporated. Throughout the iterative training process, the addition of cyclic-translation-based data augmentation results in a larger final parallel corpus compared to Iterative Back-Translation, which is expected to improve the translation performance of the machine translation model trained on it.

In the evaluation experiments, three models were compared: an Ainu-to-Japanese translation model trained with the initial parallel corpus only (baseline model), a model trained with Iterative Back-Translation, and a model trained with Iterative Cyclic-Translation. For both Iterative Back-Translation and Iterative Cyclic-Translation, the maximum number of iterations was set to 2. Ainu test sentences were translated into Japanese and the accuracy of the translation was evaluated using BLEU and CHRF^{++} metrics.

As for the model trained with Iterative Back-Translation, after one iteration, the BLEU and CHRF^{++} were improved by 8.95 and 6.48 points, respectively, compared to the baseline model. However, after two iterations, the improvements were only 0.66 points for BLEU and 1.47 points for CHRF^{++}, which were lower than the scores achieved by the model after one iteration. On the other hand, when applying our proposed Iterative Cyclic-Translation method, there was no improvement in either BLEU or CHRF^{++} scores compared to the baseline after one and two iterations.

Furthermore, we conducted manual evaluation of several sentences. The sentences translated by the Iterative Back-Translation model after one iteration were generally comprehensible and acceptable. However, the model using Iterative Cyclic-Translation demonstrated a tendency to produce mis-

translations of Ainu cultural terms and omissions of low-frequency words. Regarding the translation of the Ainu folklore specific expressions such as parallelism and periphrasis, while all models could generate translations that captured the essential meaning, the model obtained by Iterative Cyclic-Translation often generated simplified expressions by inadequately summarizing redundant expressions in the folklore style.

In automatic evaluation, our proposed Iterative Cyclic-Translation showed lower BLEU and CHRF^{++} scores compared to the existing Iterative Back-Translation. This may be because the existing Japanese-English machine translation system used in cyclic translation is inappropriate for translating sentences containing Ainu cultural terms, resulting in generating low-quality augmented translation pairs. Additionally, the current initial Ainu-Japanese parallel corpus is both qualitatively and quantitatively insufficient. The number of Ainu language documents used in this research was limited, and many Japanese sentences in the initial parallel corpus were not easy to understand. Furthermore, to prepare translation pairs as much as possible, we created a unified parallel corpus that consists of Ainu sentences in different domains, such as folktales and colloquial styles, and different dialects. Training a single translation model that handles different domains and dialects may have resulted in the failure to train a sophisticated translation model.

In future work, in addition to expanding and improving the documents included in the Ainu-Japanese parallel corpus, we aim to add English translations for Ainu-Japanese translation pairs, train models between target and auxiliary languages for improvement of cyclic translation, and verify whether translation accuracy is improved. Furthermore, given that folktales and colloquial texts exhibit differences in expressions and vocabulary across domains and dialects, we verify whether our proposed method achieves similar improvements in translation accuracy when a sufficient parallel corpus is prepared for each dialect using more Ainu documents. Additionally, to ensure the quality of augmented Ainu sentences, we plan to explore a method that incorporates a grammatical check of Ainu sentences by verifying whether the number of arguments of a verb in a sentence is correct.