氏　　名	Aizhan Imankulova（アイジャン イマンクロヴァ）
所　　属	システムデザイン研究科 システムデザイン専攻
学位の種類	博士（工学）
学位記番号	シス博　第 134 号
学位授与の日付	令和 3 年 3 月 25 日
課程・論文の別	学位規則第 4 条第 1 項該当
学位論文題名	A Study on Exploiting Additional Resources for Low-resource Neural Machine Translation (低リソースのニューラル機械翻訳のための追加リソースの活用に関する研究)
論文審査委員	主査　准教授　小町　守 委員　教　授　山口　亨 委員　教　授　高間　康史 委員　准教授　須藤　克仁（奈良先端科学技術大学院大学）

【論文の内容の要旨】

Machine translation (MT) is the task of translating input text from a source language into a target language. The practical use of MT will enable smooth communication between different languages. In the real world, the MT research results are applied as various services such as Google translation and DeepL. One of the breakthroughs in MT in recent years is the arrival of neural machine translation (NMT). NMT models have been reporting significant performance improvements. On the other hand, neural MT models require a large number of parallel sentences for training.

The biggest issue with low-resource languages is the extreme difficulty of obtaining enough resources. MT has proven successful for several language pairs. However, each language comes with its challenges. Low-resource languages have largely been left out of the MT revolution. For instance, there are often very few written texts, and even the languages that have monolingual text do not always have a parallel text in another language.

I research to what extent it is possible to improve MT systems' performance in a low-resource scenario using other pseudo-parallel data, other helping language pairs, and other modality data to increase the training data size for different language pairs and domains.

Previously, additional training data has been augmented by pseudo-parallel corpora obtained by using MT models to translate monolingual corpora into the source language. However, in low-resource language pairs, in which only low-accurate MT systems can be used, translation quality degrades when a pseudo-parallel corpus is naively used. Therefore, I consider data selection and filtering of the generated pseudo-parallel corpora using different similarity metrics.

Another way to improve low-resource MT would be to use out-of-domain data. However, merely using MT systems trained on out-of-domain data for in-domain translation is known to perform poorly. To effectively use large-scale out-of-domain data for low-resource tasks, we need to utilize domain adaptation and multilingual transfer approaches. In order to do that, I propose a multistage fine-tuning method, which combines two types of transfer learning, i.e., domain adaptation and multilingual transfer from other language pairs with conventional fine-tuning, where an NMT system trained on out-of-domain data is fine-tuned only on in-domain data, or mixed fine-tuning, where pre-trained out-of-domain NMT system is fine-tuned using a mixture of in-domain and out-of-domain data.

Different from conventional full-sentence MT, simultaneous MT is also considered to be one of the low-resource scenarios due to involving translating a sentence before the speaker's utterance is completed in order to realize real-time understanding. This task is significantly more laborious than the general full sentence translation because of the shortage of input information during decoding. To alleviate this shortage, I propose to leverage visual clues as an additional modality to help MT systems predict translations from richer information.

The main contribution of this thesis is improving MT performance for low-resource language pairs by effectively using additional information from different resources.

To improve MT performance with low-resource language pairs, I propose methods to effectively expand the training data via filtering the pseudo-parallel corpus based on back-translation and round-trip translation. Furthermore, I propose a novel multilingual multistage fine-tuning approach for low-resource NMT, taking a challenging Japanese-Russian pair for benchmarking.

By using additional modality to simultaneous MT, I verified the importance of visual information during decoding by performing throughout the evaluation and analyzing its effect on different low-resource language pairs.

This thesis is organized as follows:

Chapter 1 introduces the motivation, aim, and objectives of creating and filtering pseudo-parallel corpora by back-translation and round-trip translation.

Chapter 2 introduces methods of creating and filtering pseudo-parallel corpora by back-translation and round-trip translation. Here, I show that using filtered pseudo-parallel corpora as additional training data improves NMT performance compared to using unfiltered pseudo-parallel corpora for both back-translation and round-trip translation methods. The proposed method achieved up to 3.46 BLEU points in the Russian→Japanese translation and up to 5.25 BLEU points in the Japanese→Russian translation.

Chapter 3 addresses the research questions of the advantages and disadvantages of out-of-domain data for low-resource language pairs. To effectively exploit out-of-domain parallel data, I propose a multistage fine-tuning method, which combines domain adaptation multilingual transfer approaches. The proposed method achieved up to 2.72 BLEU points in the Russian→Japanese and up to 3.06 BLEU points in the Japanese→Russian translation.

Chapter 4 introduces a novel technique of utilizing different modality for low-resource simultaneous MT. In this chapter, I propose to combine multimodal and simultaneous NMT to enrich incomplete text input information using a visual clue. As a result, the proposed method significantly outperformed text-only baselines in all experimented language-pairs, especially for language pair with different word orders such as English→Japanese.

Chapter 5 concludes this thesis, discusses insights and limitations, and describes potential future work for low-resource MT.