# Expanding the Applicability of Machine Translation

もりした　まこと
氏　　名　森下　　睦
研究科，専攻の名称　東北大学大学院情報科学研究科（博士課程）システム情報科学専攻
学 位 論 文 題 目　Expanding the Applicability of Machine Translation
（機械翻訳の適応先拡張）
論 文 審 査 委 員　主査　東北大学教授　鈴木 潤
東北大学教授　北村 喜文　　東北大学教授　乾 健太郎
東北大学教授　大町 真一郎


# 論文内容要約

**Chapter 1 Introduction**

In this thesis, we address the following research issues:

How can we make the Japanese-English machine translation model more accurate for general domains?

Recent machine translation algorithms mainly rely on parallel corpora.

However, since the availability of parallel corpora remains limited, only some resource-rich language pairs can benefit from them.

We focus on the Japanese-English language pair and found that the publicly available parallel corpora for that language pair are still limited compared to other European language pairs.

This thesis investigates how we can create a large parallel corpus for Japanese-English and push up translation accuracy for that language pair.

How can we build an accurate machine translation model for a specific domain?

We sometimes need to train a machine translation model for a specific domain.

Although a machine translation model trained with a large in-domain parallel corpus achieves comparable performance to professional translators, it still works poorly when large in-domain data are unavailable.

To adapt the model for the specific domain, we need a small pair of parallel sentences in that domain.

This limits the applicability of the machine translation when the target domain's data are limited, such as the case of COVID-19, which has appeared relatively recently, and thus no parallel corpus is available.

In this thesis, we discuss two situations: (1) when we have a small amount of in-domain parallel corpus or (2) when we have no in-domain parallel sentences and have to collect them.

Methodologies for translating document with larger context.

The current machine translation model outputs a sentence given a sentence and does not consider the context outside of the sentence.

However, it is crucial to provide an inter-sentence context in neural machine translation models for higher-quality translation.

We propose a simple but effective methodology for context-aware machine translation.

**Chapter 2 JParaCrawl: A Large Scale Web-Based English-Japanese Parallel Corpus**

Since current machine translation (MT) approaches are mainly data-driven, one key bottleneck has been the lack of parallel corpora.

This problem continues with the recent neural machine translation (NMT) architecture.

Our goal is to create large parallel corpora to/from Japanese.

In our first attempt, we focused on the English-Japanese language pair.

Currently, ASPEC is the largest publicly available English-Japanese parallel corpus, which contains 3.0 million sentences for training.

Unfortunately, this is relatively small compared to such resource-rich language pairs as French-English.

Also, available domains remain limited.
We address this problem, which hinders the progress of English-Japanese translation research, by crawling the web to mine for English-Japanese parallel sentences.

Current NMT training requires a great deal of computational time, which complicates running experiments with few computational resources.
We alleviate this problem by providing NMT models trained with our corpus.
Since our web-based parallel corpus contains a broad range of domains, it might be used as a pre-trained model and fine-tuned with a domain-specific parallel corpus.

Through this chapter, we created a new English-Japanese parallel corpus, named JParaCrawl, containing more than 8.7 million sentence pairs.
Our experiments showed how JParaCrawl contains a broader range of domains and can be used for general purposes.
We also drastically reduced the training time by fine-tuning the JParaCrawl pre-trained NMT models and in doing so maintained or even boosted the performance.
Finally, we showed that JParaCrawl also improved the performance of a specific domain when training a model with an existing corpus from an initial state.

In future work, we will crawl more websites and make the dataset larger.
We also plan to improve the bitext aligner and cleaner, especially for Japanese.
We only focused on English-Japanese in the initial release, but we hope to eventually add more language pairs to/from Japanese.

JParaCrawl and the NMT models pre-trained with it are freely available online for research purposes.

**Chapter 3 JParaCrawl v3.0: An Updated Large-scale English-Japanese Parallel Corpus**
The current neural machine translation models are generally trained by supervised approaches, denoting reliance on parallel corpora.
However, since publicly available parallel corpora remain limited, training a model for many language pairs is difficult.
Thus, constructing a parallel corpus is crucial for expanding the applicability of machine translation.

This paper introduces a new large-scale web-based parallel corpus for English-Japanese for which only limited parallel corpora are available.
One of the current largest parallel corpora for this language pair is JParaCrawl, which is constructed by crawling the web and automatically aligning parallel sentences.
However, this corpus contains around 10 million sentence pairs, which is still limited compared to the other resource-rich language pairs, and it is somewhat outdated because it was created two years ago.

We re-crawled parallel websites by analyzing the latest CommonCrawl archive and extended the crawl target to PDF and Word documents.
After filtering out noisy sentences, the new JParaCrawl v3.0 included more than 21 million unique sentence pairs.
We empirically confirmed that the new corpus boosts the translation accuracy on various domains, especially on the trendiest news articles.
Our future work will update the JParaCrawl corpus and propose better alignment/filtering techniques.

Our new corpus, named JParaCrawl v3.0, will be publicly available through our website.
We expect that JParaCrawl v3.0 will support future research and products.

**Chapter 4 Domain Adaptation to the News Task**

This chapter describes our participation in the WMT 2018 news translation task and organizing of Japanese-English and English-Japanese translation tasks in 2020.
In 2018, we participated in English-to-German (En-De) and German-to-English (De-En) translation tasks.
The starting point of our system is the Transformer model, which recently established better performance than conventional RNN-based models.
We incorporated a parallel corpus cleaning technique and a right-to-left n-best re-ranking technique and also used a synthetic corpus to exploit monolingual data.
To maintain the quality of the synthetic corpus, we checked its back-translation BLEU scores and filtered out the noisy data with low scores.

Through experiments, we found that careful parallel corpus cleaning for the provided and synthetic corpora largely improved accuracy, and we confirmed that R2L re-ranking works well even with the Transformer model.
Our comparison between the Transformer and RNN-based models suggests that the latter models might surpass the former when the training data are not enough large.
This result sheds light on the importance of large, clean data for training the Transformer model.

In 2020 and 2021, we organized the Japanese-English and English-Japanese news translation tasks.
For these tasks, we provided JParaCrawl parallel corpus, described in the previous section, as a training data.
We describe more details of these tasks in this chapter.

**Chapter 5 Domain Adaptation of Machine Translation with Crowdworkers**

Recent Neural Machine Translation (NMT) has achieved remarkable performance; however, its translation quality drastically drops when the input domain is not covered by the training data.
One typical approach to translating these inputs is adapting the machine translation model to the domain with a small portion of in-domain parallel sentences.
These in-domain parallel sentences are normally extracted from a large existing parallel corpus or created synthetically from a monolingual corpus.
However, these methods are sometimes difficult to apply because the existing parallel/monolingual data may not include sentences relevant to the target domain.
For example, it is difficult to adapt a model to the COVID-19 domain because this issue has arisen very recently and current available data do not sufficiently cover this domain.

This has created a serious technical obstacle in the recent development of machine translation.
In the case of COVID-19, announcements from government bodies and health authorities are usually in their local language, and it may not be possible to quickly translate such announcements as they are updated.
This limits the ability of foreign residents, not speaking the local language, to access the latest information, and it may even put these people in danger.

To alleviate such a digital divide imposed on foreign residents, we propose a method to rapidly adapt a machine translation model to any domain at a reasonable cost and time period.
We hypothesize a small number of in-domain parallel sentences of the target domain available on the web, and we ask crowdworkers to report these web URLs as a task of web mining.
This is inspired by the success of the ParaCrawl project, which creates a large parallel corpus from the web automatically.
After this data collection, we adapt the machine translation model with the collected target-domain parallel sentences.
Our method has the advantage of being applicable to any domain, in contrast to previous works that use existing parallel/monolingual data.

Through experiments, we empirically confirmed that our framework could significantly improve translation performance for a

target domain within a few days of crowdsourcing and at a reasonable cost.
Our method can be used for any domain, and it is especially effective for an urgent situation when rapid domain adaptation is required, such as handling COVID-19.

**Chapter 6 Context-aware Neural Machine Translation with Mini-batch Embedding**
Current standard neural machine translation (NMT) models translate sentences in a sentence-by-sentence manner.
However, some have argued that it is critical to consider the inter-sentence context in handling discourse phenomena, which include coherence, cohesion, coreference, and writing style.
To correctly translate these linguistic features, some works provide additional context information to an NMT model by concatenating the previous sentence, applying a context encoder, or using a cache-based network.

Most of the previous studies have considered only a few previous context sentences.
Several methods, such as the cache-based network, consider long-range context but heavily modify the standard NMT models and require additional training/decoding steps.
Our goal is to make a simple but effective context-aware NMT model, which does not require heavy modification to standard NMT models and can handle a wider inter-sentence context.
To this end, we propose a method to create an embedding that represents the contextual information of a document.
To create this embedding, we focused on the mini-batch, which is commonly used in NMT training and decoding for efficient GPU computation.
We modified the mini-batch creation algorithm to choose sentences from a single document and created an embedding that represents the features of the mini-batch.
We call this embedding mini-batch embedding (MBE) and incorporate it in the NMT model to exploit contextual information across the sentences in the mini-batch.

We incorporated MBE in the NMT model, which enabled it to outperform competitive baselines.
We found that our NMT model could choose the appropriate word and writing style to match the document context.
An analysis showed that our model's performance improves with a large context, but it still achieves comparable or even better performance than that of the baseline when translating a single sentence.
Our future work includes applying MBE to other applications and improving the method to generate embeddings from a mini-batch.

**Chapter 7 Conclusion**
Key contributions of this thesis can be summarized as follows:

Creating a large-scale Japanese-English parallel corpus from the web:
We constructed a parallel corpus for English-Japanese, for which the amount of publicly available parallel corpora is still limited.
We constructed the parallel corpus by broadly crawling the web and automatically aligning parallel sentences.
Our collected corpus, called JParaCrawl, amassed over 8.7 million sentence pairs in v1.0, and over 21 million in v3.0, which is the largest publicly available parallel corpus for this language pair.
We released JParaCrawl and the pre-trained models publicly for research purposes.

Adapting a machine translation model to the specific domain:
We propose a method to adapt a machine translation model to the specific domain with a small amount of in-domain parallel data.
We create a synthetic corpus by largely translating the in-domain monolingual data and training the NMT model.
We show how our synthetic data improves translation accuracy through experiments in the news domain.

Establishing a new method for collecting parallel sentences of specific domain:
We propose a framework to efficiently and effectively collect parallel sentences in a target domain from the web with the help of crowdworkers.
With our collected parallel data, we can quickly adapt a machine translation model to any target domain.
Our experiments show that the proposed method can collect target-domain parallel data over a few days at a reasonable cost.

Proposing a new model for context-aware machine translation:
We propose a simple method for context-aware machine translation.
With the aim of using a simple approach to incorporate inter-sentence information, we propose mini-batch embedding as a way to represent the features of sentences in a mini-batch.
With the new mini-batch embedding, our model can consider a whole document context and consistently outperforms the strong context-aware baselines.