# LING 506 Final Report

**Shinka Mori**
Department of Linguistics
University of Illinois
shinkam2@illinois.edu

## Abstract

In this proposal, I will describe the system I will implement for the WMT20 News Shared Task for the English-Japanese pair. Since this is the first year that a Japanese-English pair is included, the findings of WAT20 and WAT19 are used to guide the project. The model is a transformer model created with OpenNMT(Klein et al. 2017). It is trained to perform both Japanese - English(Ja-En) and English - Japanese(En-Ja) translation. The data is provided by the WMT20 shared task.

I will be producing a model that performs Ja-En and En-Ja translation using OpenNMT. Byte-pair encoding will be done using `sentencepiece`. The model will be trained on a general corpora, then fine-tuned on a domain specific dataset. After completing a model that can translate from end to end, I will be adding other methods such as back-translation and ensemble methods as time permits. Once I implemented those methods, the method described by Wang et al.(2016) will be programmed to create a system that recalls missing pronouns in translation, to handle the drop-pronoun problem commonly seen in Japanese.

## 1 Proposed MT project

The project delineated in this paper is based on the WMT20 shared task, for the Ja-En/En-ja language pair. The corpora provided by WMT 20's shared task will be used. Specifically, this dataset consists of data collected from ParaCrawl, Japanese-English Subtitle Corpus, The Kyoto Free Translation Task Corpus, and TED Talks. The task is to translate News texts in the provided language pair.

Some important aspects of Japanese to highlight is that it is a non-segmented language, meaning there is no whitespace to indicate word boundaries. It is also a pro-drop language, so it is possible to have sentences with missing subjects. These features create some obstacles for Ja-En translation. I will describe some methods that can be used to improve Ja-En/En-Ja translation.

In this research, right to left re-ranking, byte pair encoding, domain adaptation and ensemble method will be utilized to handle the peculiarities of Japanese text.

Byte pair encoding, described in Seinrich et al. (2015), is helpful for translating technical words or rare words; BPE is especially effective for Japanese, as demonstrated by previous findings(Morishita, Suzuki, and Nagata 2020)(Yang and Ogata 2019), since many Japanese words are produced by compounding multiple words. `sentencepiece` provides a module which does byte-pair encoding.

Right to left re-ranking has been shown to improve translation quality(Liu et al. 2016). Translation is traditionally produced in the direction of the language; however, since translation is done sequentially, errors can accumulate easily and lead to an incorrect translation. This problem is alleviated by Right to left re-ranking, which trains a left-to-right and a right-to-left model, and creates a translation that the two model agrees the most on. The models are used to produce two k-best lists, which are then unioned and scored again, from which the best candidate is chosen.

Backtranslation is a method proposed in Sennrich, Haddow, and Birch 2015. It consists of creating a target-to-source model which is used to translate authentic data into the source language. The output is then fed into the source-to-target model as synthetic data. In Poncelas et al. 2018 it was noted that using a combination of authentic and synthetic data is the most effective, as compared to just using synthetic or authentic data. To create a robust model that can handle specific cases seen in News texts, domain adaptation can be used, which was proposed in Chu, Dabre, and Kurohashi 2017. The model will be trained across the various corpora provided for the shared task–ParaCrawl, The Kyoto Free Translation Task Corpus, WikiMatrix corpus, Japanese-English Subtitle Corpus, and TED Talks– and be fine tuned on News-related corpora: News Commentary v25.

Lastly, ensemble method will be used to explore different hyper-parameters, training sets, and the result of combining these different models. Multiple models will be trained on different hyper-parameters and training sets, and will compete against each other to produce the best model.

The tokenization of Japanese sentences will be done with the python-wrapped version of Juman++(Morita, Kawahara, and Kurohashi 2015)(Tolmachev, Kawahara, and Kurohashi 2018). English sentences will be tokenized with the python wrapped version of Moses(Koehn et al. 2007).

After implementing a fully working model, other methods can be applied to improve the performance. Ameliorating the issues of the pro-Drop feature of Japanese may affect the quality of the translation. I plan to implement the method outlined by Wang et al.(2016) to supplicate the dropped pronouns in the original Japanese dataset with predicted pronouns using RNNs and a multi-layer perceptron classifier.

## 2 Prior Work

I will be exploring methods used by the state-of-the-art systems for Ja-En and En-Ja translation, and evaluating their efficacy in this section. The shared tasks that I used as reference for my research are summarized below. **(**Workshop for Asian Translation 2019 This workshop is an ”open evaluation campaign focusing on Asian languages”)(Nakazawa et al. 2019). For the 2019 task, ASPEC, Japanese Patent Corpus, Timely Disclosure Documents Corpus, and JIJI were provided as possible corpora for teams to produce models for the Japanese and English pair. The details of the corpus is described below:

- ASPEC - an abstract corpus of scientific papers
- JPC - a collection of patent corpus for various language pairs
- TDDC - corpus of timely disclosure documents
- JIJI - news text colected from JIJI press

**MT Robustness** The goal of this shared task is to improve MT system's robustness for orthographic variations, grammatical errors, and other linguistic phenomena(Li et al. 2019). It provides the MTNT dataset, which contains noisy social media texts. The objective of the shared task is to effectively handle noisy data and produce a coherent translation.

### 2.1 Methods in state-of-the-art Models

**Byte Pair Encoding/ Sub-word encoding** Byte Pair Encoding(BPE) and Sub-word encoding were employed for many of the submissions for the WAT 2019 Shared Task, namely Yang and Ogata 2019, Park et al. 2019, and Morishita, Suzuki, and Nagata 2019 used this technique. BPE and sub-word encoding is effective for Japanese because it can break long Japanese Kanji words into individual clusters, so that the translation of the clusters can be produced easily. This facilitates the translation of long phrases by allowing it to be translated in parts.

**Base Model** The systems presented in WAT 2019 and MT Robustness 2019 for the Japanese and English pair were all neural network models, with Transformers being the most common one. Many of the participants used OpenNMT(Yang and Ogata 2019)(Park et al. 2019) or Tensor2Tensor(Mino et al. 2019). OpenNMT and Tensor2Tensor are regarded as state-of-the-art networks for Neural Machine Translation.

**Right to Left re-ranking** Right to Left re-ranking(R2L) is a method where two models are trained, one that translates from source to target, and another that translates from target to source(Liu et al. 2016). The L2R mode generates $n$ best list, and is scored with the $n$ best list produced by R2L, which are then re-ranked to choose the best translation. R2L reranking prevents mistranslation resulting from generating translations in only one direction, which can potentially allow translation errors to accumulate. This technique was used by Park et al. 2019, and Morishita, Suzuki, and Nagata 2019 and was observed to increase translation performance.

**Backtranslation/Data Augmentation** For some of the corpora provided for WAT 2019, such as the JIJI corpus, the data size was small. To compensate for this, backtranslation and data augmentation was used. Backtranslation is implemented by creating a translator that translates from the target to the source language, feeding input into the translator, and using the result as synthetic data for the source-to-target translator. The method was proposed by Sennrich, Haddow, and Birch 2015, and further analysis by Poncelas et al. 2018 showed that using a mix of synthetic and authentic data improves the performance. Backtranslation was used for Park et al. 2019, Imamura and Sumita 2019, Mino et al. 2019, and Morishita, Suzuki, and Nagata 2019. Backtranslation was ineffective for Ja-En model for Morishita, Suzuki, and Nagata 2019. To amend for this, they made synthetic data using forward translation, which essentially sets the synthetic corpus as the target corpus, which slightly improved the performance.

Data augmentation is when additional data is added to the training set, which is considered a regularization method and prevents overfitting for small datasets. Park et al. 2019 used multi-source translation, where they combined other language pairs with the same target language to increase the data size. They found that this improved the translation quality of the En-Ja model. Mino et al. 2019 produced four different versions of the JIJI corpus(Equivalent, Aligned, back-translated, Aligned Yomiuri shinbun corpus) and trained their model on a combination of the corpora to find the optimal combination. Morishita et. al produced their own corpora, ”JParaCrawl”, by crawling from the web. Using a combination of ”JParaCrawl” and the ASPEC dataset was found to increase the translation quality.

It should be noted that some teams chose to add a tag to the generated dataset(Park et al. 2019)(Mino et al. 2019). This helps make the system more generalized and robust towards different input text, and was shown to increase translation performance for the Naver lab submission for the MT Robustness shared task(Berard, Calapodescu, and Roux 2019).

**Domain Adaptation** Domain adaptation is essentially when a model is trained on a general dataset, then fine-tuned for a specific domain type. This approach works well when there is insufficient data for a particular language task. Morishita, Suzuki, and Nagata 2019 fine-tuned their model with specific tagged category data, which increased the BLEU scores.

**Ensemble methods** Ensemble methods allows models to be combined to potentially improve performance. This technique was used by Yang and Ogata 2019, Park et al. 2019, Mino et al. 2019, and Morishita, Suzuki, and Nagata 2019. For Yang and Ogata 2019, ensemble methods yielded the highest BLEU socres for En-Ja and Ja-En. Morishita, Suzuki, and Nagata 2019. achieved the highest score for their model with a combination of fine-tuning, ensemble method, and R2L reranking.

**Pre-processing** In the WAT 2019 submissions, the submissions did not go in depth about pre-processing the data prior to tokenization; however, pre-processing may potentially affect translation quality. The best submission of the MT Robustness shared task, Naver labs, found that filtering out identical sentences on source and target, sentences with length mismatch, and sentences belonging to a language other than the source or target language significantly improved their BLEU scores(Berard, Calapodescu, and Roux 2019).

## 2.2 Other Methods

The methods described above were based on the state-of-the art systems presented at WAT 2019; however, other approaches to Ja-En and En-Ja translation has been attempted.

**Handling dropped pronouns** In addition to the works summarized above, I researched possible approaches to predict gender for pro-drop languages. One of the obstacles in Ja-En trasnlation is the dropping of pronouns in Japanese, requiring the need to infer pronouns and its gender. Wang et al. 2016 introduces an approach where they aligned the target and source sentence, predicting the missing pronoun by projecting the misaligned pronoun in the non-pro-drop language to the pro-drop language. They then train an RNN on the resulting aligned data to create a classifier that predicts the missing pronoun. Out of the resulting sentences containing the predicted pronouns, n-best results are used to create a corpus with predicted pronouns. A Statistical Machine Translator is then trained on this corpus. This method was shown to increase accuracy over the baseline performance. The Ja-En set had a lower agreement compared to the Chinese-English pair, possibly because Japanese is a SVO language(Wang et al. 2016). Webster and Pitler also explored this issue, but focused on handling the gender imbalance in the mistranslation of gender pronouns from MT results. A corpus was produced from scraping Wikipedia articles and aligning the results. Pronoun tagging was performed to copy the gender of the source to the target sentence. Finally, the corpus was resampled so that the ratio of masculine and feminine example was 1 : 1. A BERT model was then trained on this corpus to create a gender classifier for pronouns. Using this classifier, the training set was tagged for gender. Some noise was introduced by randomly flipping whether a sentence contains a target pronoun 2% of the time. This was shown to increase the accuracy remarkably and using the baseline and the gender-tagged data gave the best performance. While this research was done for Spanish and Engish, the approach is language-agnostic and is highly applicable to Japanese and English translation(Webster and Pitler 2020).

### 2.3 Observations

Note that many of the techniques employed by the state-of-the-art systems described above are not explainable. To put it more clearly, the translations are all done in a blackbox that utilizes technique such as R2L, backtranslation, and such and does not provide information regarding *why* it works for certain tasks but not others. This is problematic, as explainability is a key component of Translation Systems. In the case where a system makes multiple mistranslations for a certain word or topic, there should ideally be a simple process to address them. For the NMT systems described above, this is not the case.

Another point to be made is that these methods listed above are not specifically curated for the language pair and are general methods applied to achieve state-of-the-art models. Many of the submissions to the shared tasks did not implement specific methods aimed at improving Ja-En or En-Ja pair. For example, Japanese is a pro-drop pronoun, and the pronoun as well as the gender must be inferred from the context. These models do not address that issue, which may lead to translation errors that may be amendable with an approach targeted at addressing pro-drop languages. The motivation behind not using a language specific approach could be that the abundance of the dataset and the effectiveness of the state-of-the-art Transformers allows for an easy implementation of a high-performing model. It would be interesting to see the effects of applying the methods proposed by Wang et. al or Webster and Pitler for this task.

## 3 Approach

The progress of the project and the methods and data involved in it are described in this section. Specifically, I will introduce the techniques used to process the data, as well as the design aspects of the model and the methods involved to create the model. The model's parameters and the results of fin-tuning the hyperparameters will also be discussed in this section.

### 3.1 Data

The amount of sentences used for training, validation, and testing varied greatly among the submissions for the 6th WAT shared task, but many of them used at least 1M sentences for training. Imamura and Sumita(2019) used 3,007,754 sentences from the ASPEC dataset, and 1,398,184 sentences from the TDDC dataset for training, and 1,812 sentences from the ASPEC dataset and 1,148 sentences from the TDDC dataset for testing(Imamura and Sumita 2019). Based on this, I plan on using around 3M sentences for training, and 2,000 sentences for testing.

At checkpoint 1, I have tokenized the The Kyoto Free Translation Task Corpus as well as 5000 sentences from the WikiMatrix corpus(in total, 445,289 sentences) to be used as the training set. The WikiMatrix corpus was filtered using the code provided by the developers of WikiMatrix, to extract sentence pairs that has a margin score higher than 1.04, as recommended by the developers(Schwenk et al. 2019).

I plan on creating the corpus from ParaCrawl, The Kyoto Free Translation Task Corpus(KFFT), WikiMatrix corpus, Japanese-English Subtitle Corpus, and TED Talks and be fine tuned on the News commentary corpus. It should be noted that the level of formality differs on the Japanese corpus. KFFT and WikiMatrix corpus uses a more formal level of Japanese, whereas the Japanese-English Subtitle corpus and ParaCrawl often contains casual Japanese. The topics relating to the KFFT are strictly Kyoto related. Since Japanese news texts tend to have high level of formality, I will be using a combination of the different corpus to create a large corpora of about 3M sentences. Three variations of the corpus will be produced to have a diverse range of language formality and vocabulary. By checkpoint 2, I plan on tokenizing the other datasets and combining them into a larger corpus to total approximately 3M sentences and create train, development, and test sets out of them.

## 3.2 Method

### 3.2.1 Preprocessing

I used `Juman++` to tokenize Japanese texts, and used Moses to tokenize English texts. In addition, I used the `sentencepiece` library to apply `sentencepiece` encoding to the dataset(Kudo and Richardson 2018). Applying `sentencepiece` greatly decreases the amount of unknown tokens on the model and improves the overall quality of the translation. Right-to-Left encoding will be implemented to decrease the bias resulting from one-directional translation.

### 3.2.2 Model

By checkpoint 1, I have successfully implemented a simple neural machine translator model capable of encoding with `sentencepiece` and producing a translation from English to Japanese on a small corpus of 445,289 sentences, consisting of sentences from the Kyoto Free Translation Task and WikiMatrix sentences. I used `OpenNMT-py` to create a transformer model, using the default configurations recommended in the tutorial to run the model. I used `OpenNMT-py`'s built in transformer model. The model used 3,000 steps for training, and 1,000 steps for validation. The vocabulary size was 3,200 for both languages. It evaluates the model outputs using the BLEU scores calculated using `SacreBLEU`.

Since the main task is to translate news-related texts, I plan on producing several corpora with varying amounts of formal and casual Japanese sentences. Since the size of the News Commentary task is limited(1.k sentences), domain adaptation will be used by training the model on a larger dataset and fine-tuning using the News Commentary Dataset. , and ensemble method will be used to find the best combination of the corpora.

In addition, I will be referring to previous literature, particularly ones from the WAT Shared task, to find the optimal hyperparameters and fine-tune them for the model.

## 3.3 Data

To create a corpus for the model, I sampled sentences from the Kyoto Translation Task corpus, Subtitles corpus, News Commentary corpus and Wikimatrix corpus.

| Corpus | Train | Test |
|---|---|---|
| KFFT | 330k | 1,160 |
| Subtitles | 300k | 0 |
| News Commentary | 1811 | 0 |
| Wikimatrix | 300k | 1,000 |
| Mixed corpus | 3811 | 0 |

Table 1: Corpus Description

The Mixed corpus is a combination of $1,000$ sentences from the Subtitle corpus and Wikimatrix corpus, combined with the News Commentary Corpus.

Japanese-English Subtitle corpus and ParaCrawl often contains casual Japanese. The topics relating to the KFFT are strictly Kyoto related. Since Japanese news texts tend to have high level of formality, I will be using a combination of the different corpus to create a large corpora of 3M sentences that has a diverse representation of formality in Japanese. Three variations of the corpus will be produced to have a diverse range of language formality and vocabulary.

I omitted the Subtitles corpus from being used in testing, since by nature it contains dialogues, which does not resemble the type of sentences seen in a news corpus. The news corpus was also omitted due to its scarcity. Using it for testing purposes would limit the amount of news-related sentences available for training

## 3.4 Method

### 3.4.1 Model

By checkpoint 1, I have successfully implemented a simple neural machine translator model capable of encoding with `sentencepiece` and producing a translation from English to Japanese on a small corpus of 445,289 sentences, consisting of sentences from the Kyoto Free Translation Task and WikiMatrix sentences. I used `OpenNMT-py` to create a transformer model, using the default configurations recommended in the tutorial to run the model. I used `OpenNMT-py`'s built in transformer model. The model used 3,000 steps for training, and 1,000 steps for validation. The vocabulary size was 3,200 for both languages. Sentences with more than 150 sequences of tokens are filtered from the corpus. The output is then evaluated by `SacreBLEU`. By checkpoint 2, I developed a more sophisticated model. The details of the model are summarized in Table 1.

| | |
|---|---|
| epochs | 10,000 |
| model type | Adam |
| $\beta_1$ | 0.9 |
| $\beta_2$ | 0.98 |
| learning rate | 0.0002 |
| attention heads | 8 |
| encoder layers | 6 |
| decoder layers | 6 |
| label smoothing | 0.1 |

Table 2: Table 2

I use a vocabulary size of $30,000$, omitting words that occur less than 10 times.

#### 3.4.2 Byte Pair encoding

Each of the corpus was tokenized by Byte-Pair encoding with the use of `sentencepiece`(Kudo and Richardson 2018). `sentencepiece` uses a vocabulary size of $30,000$ and uses a character coverage of $0.9995$ as recommended by the developers for Japanese texts. Character coverage determines the amount of characters covered by the model.

#### 3.4.3 Domain Adaptation

I trained the model with the mixed fine tuning method(Chu, Dabre, and Kurohashi 2017). This is done by creating an in-domain and an out-of-domain corpus as well as a blend of in-domain and out-of-domain corpus, which I refer to as the mixed corpus. Each sentence in the corpus is tagged with the tag `<out-of-domain>` if it is out of domain, and `<in-domain>` if it is in-domain. The model is then trained on the out-of-domain corpus until convergence, and re-trained on the mixed corpus.

#### 3.4.4 Backtranslation

I created synthetic data for backtranslation by concatenating $100k$ sentences each from the KFFT corpus, WikiMatrix corpus, and the Subtitle corpus. I then fed the resulting corpus as well as the authentic training corpus into the Ja-En translator, and used the outputs as training samples for the En-Ja translator. Based on the findings of Poncelas et al. 2018, using a mixture of synthetic back-translated corpus and authentic corpus achieved the best results.

## 4 Preliminary results

To evaluate the results of the model, I used `SacreBLEU`(Post 2018). For the simple model, I used a smaller corpus of $445,289$ sentences, tested on $1,160$ sentences. The resulting sentences contained many unknown words, which can be explained by the small corpus size and the tendency of the corpus to have contents detailing on specific topics, since they were from Wikipedia as well as articles relating to Kyoto. Using `sentencepiece` to tokenize the inputs significantly decreased the amount of unknown words, which is explained by the nature of Japanese to compound words and BPE's ability to break up commonly appearing word clusters.
For the larger model, I used the corpus specified in Table 1, and used BPE and Domain Adaptation. The motivation of using Domain Adaptation was to accommodate for the severe lack of domain-specific corpus. The News Commentary corpus is very limited in size($1,811$ sentences) so mixed fine tuning was used to tune the model on News Commentary corpus. After training for $100,000$ steps on the large model, at the final epoch the `SacreBLEU` score for En-Ja was $24.4$. The model was trained for an additional $10,000$ epochs on the mixed corpus to perform mixed fine-tuning, which gave a score of 32.8, a clear improvement over the model trained on out-of-domain corpus. I also trained the backtranslation model by combining the synthetic corpus and the authentic corpus, again training for $100,000$ steps. The model output has a `SacreBLEU` score of 21.6 which is a drop from the baseline result. This may be attributed to the imbalance in synthetic and authentic corpus, and changing the ratio to include more authentic corpus may help. Applying domain adaptation and training for additional $50,000$ epochs increased the score to $27.6$.

To further improve the model, I plan on exploring alternative Domain Adaptation techniques, namely Ensemble Decoding of two different models, and implementing Right to Left re-ranking, which was shown to improve translation quality for WAT 2019 submissions(Morishita, Suzuki, and Nagata 2019)(Park et al. 2019).

For the Ja-En model, I trained on the same dataset for $100,000$ iterations, applying the same techniques(BPE). The `SacreBLEU` score was 27.4. Retraining with Domain Adaptation improves the score up to $39.2$. Training with the back-translated corpus gives. score of 27.6, which is higher than baseline but not a significant improvement. Retraining with Domain Adaptation brought the score up to 35.0.

The results show that for both En-Ja and Ja-En translation, Domain Adaptation significantly improves the score; however, backtranslation is not very effective. This may be due to the balance of synthetic and authentic data, and rebalancing the back-translated data to contain more authentic text may alleviate this issue.

## 5 Conclusion

In this paper, I described my implementation of an end-to-end neural machine translator for Ja-En and En-Ja. I applied common techniques employed in state-of-the-art models: Domain Adaptation, Backtranslation, and Byte-pair Encoding. Implementing R2L reranking may further improve the model's performance.

## References

Koehn, Philipp et al. (June 2007). "Moses: Open Source Toolkit for Statistical Machine Translation". In: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Prague, Czech Republic: Association for Computational Linguistics, pp. 177–180. URL: https://www.aclweb.org/anthology/P07-2045.

Morita, Hajime, Daisuke Kawahara, and Sadao Kurohashi (Sept. 2015). "Morphological Analysis for Unsegmented Languages using Recurrent Neural Network Language Model". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pp. 2292–2297. DOI: 10.18653/v1/D15-1276. URL: https://www.aclweb.org/anthology/D15-1276.

Sennrich, Rico, Barry Haddow, and Alexandra Birch (2015). *Improving Neural Machine Translation Models with Monolingual Data*. arXiv: 1511.06709 [cs.CL].

Liu, Lemao et al. (June 2016). "Agreement on Target-bidirectional Neural Machine Translation". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, pp. 411–416. DOI: 10.18653/v1/N16-1046. URL: https://www.aclweb.org/anthology/N16-1046.

Wang, Longyue et al. (2016). "A novel and robust approach for pro-drop language translation". In: *Machine Translation* 31, pp. 65–87.

Chu, Chenhui, Raj Dabre, and Sadao Kurohashi (July 2017). "An Empirical Comparison of Domain Adaptation Methods for Neural Machine Translation". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 385–391. DOI: 10.18653/v1/P17-2061. URL: https://www.aclweb.org/anthology/P17-2061.

Klein, Guillaume et al. (2017). "OpenNMT: Open-Source Toolkit for Neural Machine Translation". In: *Proc. ACL*. DOI: 10.18653/v1/P17-4012. URL: https://doi.org/10.18653/v1/P17-4012.

Kudo, Taku and John Richardson (2018). *SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing*. arXiv: 1808.06226 [cs.CL].

Poncelas, Alberto et al. (2018). *Investigating Backtranslation in Neural Machine Translation*. arXiv: 1804.06189 [cs.CL].

Post, Matt (Oct. 2018). "A Call for Clarity in Reporting BLEU Scores". In: *Proceedings of the Third Conference on Machine Translation: Research Papers*. Belgium, Brussels: Association for Computational Linguistics, pp. 186–191. URL: https://www.aclweb.org/anthology/W18-6319.

Tolmachev, Arseny, Daisuke Kawahara, and Sadao Kurohashi (Nov. 2018). "Juman++: A Morphological Analysis Toolkit for Scriptio Continua". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, pp. 54–59. DOI: 10.18653/v1/D18-2010. URL: https://www.aclweb.org/anthology/D18-2010.

Berard, Alexandre, Ioan Calapodescu, and Claude Roux (Aug. 2019). "Naver Labs Europe's Systems for the WMT19 Machine Translation Robustness Task". In: *Proceedings of the Fourth Conference on Machine Translation (Volume

*2: Shared Task Papers, Day 1)*. Florence, Italy: Association for Computational Linguistics, pp. 526–532. DOI: 10.18653/v1/W19-5361. URL: https://www.aclweb.org/anthology/W19-5361.

Imamura, Kenji and Eiichiro Sumita (Nov. 2019). "Long Warm-up and Self-Training: Training Strategies of NICT-2 NMT System at WAT-2019". In: *Proceedings of the 6th Workshop on Asian Translation*. Hong Kong, China: Association for Computational Linguistics, pp. 141–146. DOI: 10.18653/v1/D19-5217. URL: https://www.aclweb.org/anthology/D19-5217.

Li, Xian et al. (Aug. 2019). "Findings of the First Shared Task on Machine Translation Robustness". In: *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Florence, Italy: Association for Computational Linguistics, pp. 91–102. DOI: 10.18653/v1/W19-5303. URL: https://www.aclweb.org/anthology/W19-5303.

Mino, Hideya et al. (Nov. 2019). "Neural Machine Translation System using a Content-equivalently Translated Parallel Corpus for the Newswire Translation Tasks at WAT 2019". In: *Proceedings of the 6th Workshop on Asian Translation*. Hong Kong, China: Association for Computational Linguistics, pp. 106–111. DOI: 10.18653/v1/D19-5212. URL: https://www.aclweb.org/anthology/D19-5212.

Morishita, Makoto, Jun Suzuki, and Masaaki Nagata (Nov. 2019). "NTT Neural Machine Translation Systems at WAT 2019". In: *Proceedings of the 6th Workshop on Asian Translation*. Hong Kong, China: Association for Computational Linguistics, pp. 99–105. DOI: 10.18653/v1/D19-5211. URL: https://www.aclweb.org/anthology/D19-5211.

Nakazawa, Toshiaki et al. (Nov. 2019). "Overview of the 6th Workshop on Asian Translation". In: *Proceedings of the 6th Workshop on Asian Translation*. Hong Kong, China: Association for Computational Linguistics, pp. 1–35. DOI: 10.18653/v1/D19-5201. URL: https://www.aclweb.org/anthology/D19-5201.

Park, Cheoneum et al. (Nov. 2019). "KNU-HYUNDAI's NMT system for Scientific Paper and Patent Tasks onWAT 2019". In: *Proceedings of the 6th Workshop on Asian Translation*. Hong Kong, China: Association for Computational Linguistics, pp. 81–89. DOI: 10.18653/v1/D19-5208. URL: https://www.aclweb.org/anthology/D19-5208.

Schwenk, Holger et al. (2019). *WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia*. arXiv: 1907.05791 [cs.CL].

Yang, Wei and Jun Ogata (Nov. 2019). "Our Neural Machine Translation Systems for WAT 2019". In: *Proceedings of the 6th Workshop on Asian Translation*. Hong Kong, China: Association for Computational Linguistics, pp. 159–164. DOI: 10.18653/v1/D19-5220. URL: https://www.aclweb.org/anthology/D19-5220.

Morishita, Makoto, Jun Suzuki, and Masaaki Nagata (May 2020). "JParaCrawl: A Large Scale Web-Based English-Japanese Parallel Corpus". In: *Proceedings of The 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 3603–3609. ISBN: 979-10-95546-34-4. URL: https://www.aclweb.org/anthology/2020.lrec-1.443.

Webster, K. and Emily Pitler (2020). "Scalable Cross Lingual Pivots to Model Pronoun Gender for Translation". In: *ArXiv* abs/2006.08881.