Rakuten's Participation in WAT 2022: Parallel Dataset Filtering by Leveraging Vocabulary Heterogeneity

Alberto Poncelas, Johanes Effendi, Ohnmar Htun, Sunil Kumar Yadav, Dongzhe Wang, Saurabh Jain
Rakuten Institute of Technology
Rakuten Group, Inc.
{alberto.poncelas,johanes.effendi,ohnmar.htun,
sunilkumar.yadav,dongzhe.wang,saurabh.b.jain}@rakuten.com

Abstract

This paper introduces our neural machine translation system's participation in the WAT 2022 shared translation task (team ID: *sakura*). We participated in the Parallel Data Filtering Task. Our approach based on Feature Decay Algorithms achieved +1.4 and +2.4 BLEU points for English→Japanese and Japanese→English respectively compared to the model trained on the full dataset, showing the effectiveness of FDA on in-domain data selection.

1 Introduction

This paper introduces our neural machine translation (NMT) systems' participation in the 9th Workshop on Asian Translation (WAT-2022) shared translation tasks (Nakazawa et al., 2022). We participated in the Parallel Corpus Filtering Task[1] and our team id is *sakura*.

The task consists of domain specific data selection out of noisy parallel corpus mined from the web. The goal is to build English→Japanese and Japanese→English models with better performance on scientific domain. The constraint is to select the data from JParaCrawl v3.0 (Morishita et al., 2020). Only a subset of the data should be extracted and no other actions, such as transformation or augmentation, are allowed. The models built with this data are evaluated using the test set from ASPEC (Nakazawa et al., 2016), a scientific domain parallel corpus.

In our submissions, we used two independent techniques viz. feature decay algorithms (Biçici and Yuret, 2011; Biçici, 2013; Biçici and Yuret, 2015) (FDA) and log-likelihood scores. FDA based submission achieved +1.4 and +2.4 BLEU for English→Japanese and Japanese→English respectively. Log-likelihood based submission achieved +0.5 BLEU for Japanese→English direction only. Our submission related scripts can be accessed through following public repository. [2]

2 Data Selection

In this section we detail our approach to select domain specific sentences. Our approach aimed to extract the sentences from JParaCrawl (Morishita et al., 2020) that were in-domain, based on the train set of ASPEC (Nakazawa et al., 2016).

2.1 Feature Decay Algorithms

FDA is an *n*-gram based data selection technique. It has shown a better performance when compared to other word-based data selection methods (Silva et al., 2018).The selection is based on *n*-grams, and has demonstrated good performance when used to train NMT models (Poncelas et al., 2018, 2019).

The strength of this technique is that it aims to find a balance between the number of *n*-grams that are present in the in-domain data and the heterogeneity of the *n*-grams. This is achieved by considering not only the relevance of each *n*-gram in the domain but also how frequently it has been selected already.

The technique iteratively selects the sentence s (from a set of candidates, initially being the full JParaCrawl set) with the highest score according to the Equation (1):

$$score(s, S_{ASPEC}, S_{sel}) = \frac{\sum_{ngr \in \{s \bigcap S_{ASPEC}\}} 0.5^{count(ngr, S_{sel})}}{\text{len(s)}} \quad (1)$$

and adding it to a set of selected sentences S_{sel}.

The in-domain *n*-grams of s are obtained by finding $\{s \bigcap S_{ASPEC}\}$ (i.e. the intersection with the in-domain set S_{ASPEC}). Each *n*-gram has a contribution towards the score inversely proportional to the number of instances in the selected set

[1] https://sites.google.com/view/wat-filtering/

[2] https://github.com/sukuya/wat2022-parallel-data-filtering

Proceedings of the 9th Workshop on Asian Translation, pages 68–72
October 17, 2022.

S_{sel}. By default, this is conducted by computing $0.5^{count(ngr,S_{sel})}$. In our system, we decided to follow this configuration although it is not necessarily the optimal (Poncelas et al., 2017; Poncelas, 2019). We leave for future work exploring different configurations and finding a better selection criterion.

The selection was executed considering the *n*-grams of order up to 3 on the English side only. Configurations where the selection is based on both source and target sides have been reported to achieve good results (Poncelas et al., 2022). However, on the Japanese side, it is unclear what should be considered as *n*-gram (e.g. character-wise or token-wise) to achieve the best performance.

Another important question is the number of sentences that should be selected. For our system, we selected 5M sentences. This decision is based on the scores of FDA presented in Figure 1. In the plot, there is a relatively sharp decrease in FDA scores after 10M. From top 10M sentences we selected 5M based purely on empirical observations by carrying out experiments using 1M, 3M, 5M and 7M sentences and ASPEC Dev set performance (see Figure 2). We were mainly focused on minimising the number of selected sentences without compromising on model performance in terms of BLEU.

2.2 Normalised Log Probability Scores

Our second submission for the task involves using the normalised log-probability scores, inspired by dual conditional cross entropy filtering (Junczys-Dowmunt, 2018). We train two separate models on ASPEC Train, one for each direction. We calculate normalised (by number of output tokens) log-probability scores using marian-scorer (Junczys-Dowmunt et al., 2018) for entire JParaCrawl using these models. We calculate the final score for a parallel sentence by summing these two log-likelihood scores. Finally, we sort the sentences based on final scores and take the top 5M (4.634M unique) sentences for our submission.

3 Model

We trained both English→Japanese and Japanese→English models. We follow the details from the organizers[3] and build transformer (Vaswani et al., 2017) models using Fairseq (Ott et al., 2019) framework. The sentences were tokenized with a SentencePiece model (Kudo and Richardson, 2018) based on BPE method with 32000 operations. We use train and dev test from ASPEC only. The size of the different datasets are reported in Table 1.

[3] https://github.com/MorinoseiMorizo/wat2022-filtering

Dataset	Size
JParaCrawl	25.7M
ASPEC Train	3M
ASPEC Dev	1.8K
ASPEC Devtest	1.8K
ASPEC Test	1.8K

Table 1: Size (number of lines) of the datasets.

Submission	BLEU		Adequacy
	Dev	Test	Test
FDA(5M)	28.8	28.4	4.31
Baseline(26M)	27.4	27.0	4.18
Marian-Score(5M)	26.7	26.1	xx

Table 2: Results: English to Japanese.

4 Results

In Table 2 and 3, we show the evaluation scores for our submissions based on BLEU metric (Papineni et al., 2002) and human evaluation (on 200 sentences selected by organizers) from ASPEC Corpus (official results). Our FDA based system is submitted for official human evaluation and is labeled as *sakura-fda* in the figures.

Figure 3 and 4 shows the detailed breakdown of adequacy scores from the organisers. No other team participated in the task, so top system summary is detailed in the Table 2 and 3.

We see that our best submission achieved +1.4 BLEU improvement over the system trained with the full JParaCrawl set (the baseline) in the English→Japanese direction and +2.4 points for Japanese→English. This is achieved by selecting 5M sentences, approximately 20% of the data. The second submission based on sum of normalised log-likelihood scores shows minor improvement of +0.5 BLEU on Japanese→English direction but

Submission	BLEU		Adequacy
	Dev	Test	Test
FDA(5M)	21.3	21.8	4.49
Marian-Score(5M)	19.5	19.9	xx
Baseline(26M)	20.6	19.4	4.35

Table 3: Results: Japanese to English.

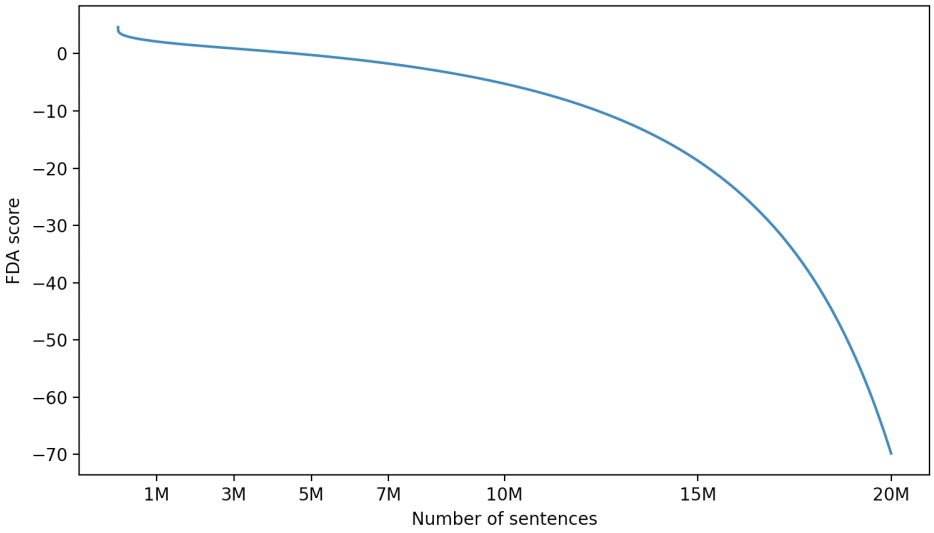

Figure 1: FDA scores of the top-20M sentences (in log scale).

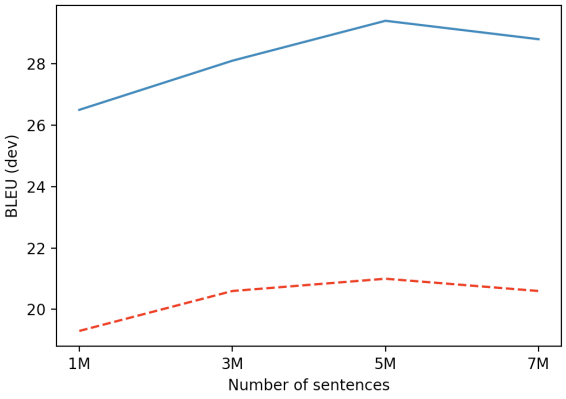

Figure 2: Evaluation of the NMT model (on dev set) built using different amount of sentences selected using FDA. The plot shows the BLEU scores for English→Japanese (blue line) and Japanese→English (dotted red line) models.

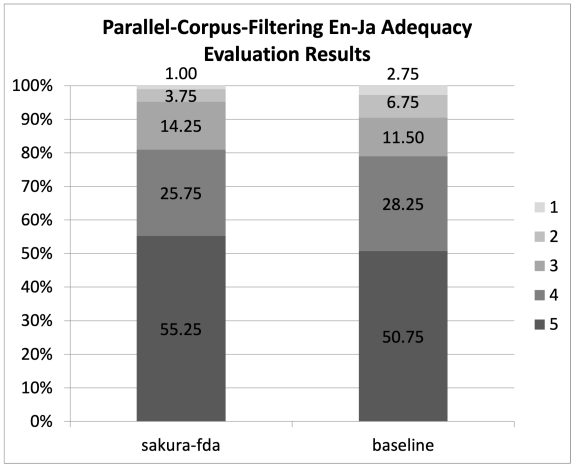

Figure 3: Adequacy evaluation results for English → Japanese.

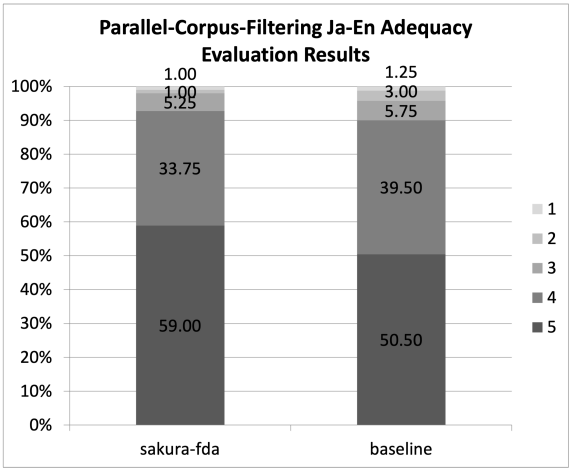

Figure 4: Adequacy evaluation results for Japanese → English.

underperforms in other direction as well as dev test.

5 Conclusion

We presented our submissions (team ID: *sakura*) to the WAT 2022 Parallel Data Filtering Task in this paper. We described our data selection system based on FDA and log-probability scores. FDA based filtering showed effectiveness in finding a subset of parallel sentences that were more useful to train a scientific-domain NMT model than using all the sentences. Our system was trained just on a 20% of the data and achieved +1.4 BLEU improvement over the baseline in the English→Japanese direction and +2.4 for Japanese→English.

As a future work, we want to use FDA in combination with the normalised log-probability scores. The work of Soto et al. (2020) demonstrated that

the inclusion metrics such as lexical richness can boost the performance of FDA. More generally, we plan to explore how can we improve a scientific domain NMT model, by using limited amount of ASPEC data along with JParacrawl. The motivation is to gauge the efficacy of FDA based approach in data selection where very less in-domain data is available along with lot of noisy mixed domain data.

References

Ergun Biçici. 2013. Feature decay algorithms for fast deployment of accurate statistical machine translation systems. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 78–84, Sofia, Bulgaria.

Ergun Biçici and Deniz Yuret. 2011. Instance selection for machine translation using feature decay algorithms. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 272–283, Edinburgh, Scotland.

Ergun Biçici and Deniz Yuret. 2015. Optimizing instance selection for statistical machine translation with feature decay algorithms. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(2):339–350.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Belgium, Brussels. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium.

Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2020. JParaCrawl: A large scale web-based English-Japanese parallel corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3603–3609, Marseille, France.

Toshiaki Nakazawa, Hideya Mino, Isao Goto, Raj Dabre, Shohei Higashiyama, Shantipriya Parida, Anoop Kunchukuttan, Makoto Morishita, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, Yusuke Oda, and Sadao Kurohashi. 2022. Overview of the 9th workshop on Asian translation. In *Proceedings of the 9th Workshop on Asian Translation (WAT2022)*, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. Aspec: Asian scientific paper excerpt corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2204–2208, Portorož, Slovenia.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA.

Alberto Poncelas. 2019. *Improving transductive data selection algorithms for machine translation*. Ph.D. thesis, Dublin City University.

Alberto Poncelas, Gideon Maillette de Buy Wenniger, and Andy Way. 2017. Applying n-gram alignment entropy to improve feature decay algorithms. *The Prague Bulletin of Mathematical Linguistics*, 108(1):245–256.

Alberto Poncelas, Gideon Maillette de Buy Wenniger, and Andy Way. 2018. Feature decay algorithms for neural machine translation. In *21st Annual Conference of the European Association for Machine Translation*, pages 239–248, Alicante, Spain.

Alberto Poncelas, Gideon Maillette de Buy Wenniger, and Andy Way. 2019. Transductive data-selection algorithms for fine-tuning neural machine translation. In *Proceedings of The 8th Workshop on Patent and Scientific Literature Translation*, pages 13–23, Dublin, Ireland.

Alberto Poncelas, Gideon Maillette de Buy Wenniger, and Andy Way. 2022. Improved feature decay algorithms for statistical machine translation. *Natural Language Engineering*, 28:71–91.

Catarina Cruz Silva, Chao-Hong Liu, Alberto Poncelas, and Andy Way. 2018. Extracting in-domain training corpora for neural machine translation using data selection methods. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 224–231, Brussels, Belgium.

Xabier Soto, Dimitar Shterionov, Alberto Poncelas, and Andy Way. 2020. Selecting Backtranslated Data from Multiple Sources for Improved Neural Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 3898–3908, Seattle, USA.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, Long Beach, USA.