Filtering of a Web-Crawled Corpus to Achieve a Strong MT Model: a Case Study on the Japanese-Bulgarian Language Pair

Iglika Nikolova-Stoupak Shuichiro Shimizu Chenhui Chu Sadao Kurohashi
Graduate School of Informatics, Kyoto University
{iglika,sshimizu,chu,kuro}@nlp.ist.i.kyoto-u.ac.jp

Abstract

This paper considers a low-resource language pair (Japanese-Bulgarian), whose largest corpus currently available is CCMatrix (4.1M parallel sentences). Unfortunately, due to the imperfect web-crawling process in which the corpus is assembled, MT models directly trained on it do not provide optimal results without further work. This paper seeks to train MT models on a portion of CCMatrix that provide a compromise between size and performance. Two main filtering criteria are utilised in the process: the original margin score that has been used to assemble CCMatrix and a score based on the application of a successful classifier model as presented by [1] for a related WMT shared task. The BLEU scores achieved by several derived models are promising.

1 Introduction

1.1 Low-Resource Language Pairs

Low-resource language pairs currently present a major challenge to natural language processing and specifically machine translation (MT). As Dabre et al. [2] illustrate, several techniques have been developing in parallel in attempts to improve the performance of MT models involving such languages. For instance, transfer learning consists in training a "parent" model on a higher-resourced language pair and applying the ensuing embeddings to the original model instead of resorting to random initialisation. Another method, zero-shot translation, involves training the encoder on a number of languages that permits eventual translation among language pairs that the model has not explicitly encountered beforehand [3]. Another line of experiments dealing with the problem of under-resourced languages relates to dataset creation, which in turn can be automated to different degrees. With web-crawling, predefined portions of the web are scanned in search of parallel text. In this manner, OPUS [4] has assembled corpora in a number of language pairs, including the one discussed herein. For instance, WikiMatrix is obtained through exclusive crawling of Wikipedia pages, whilst CCMatrix involves broad portions of the web [5]. In both cases, the cosine-distance-based "margin" metric is applied in the comparison of LASER[1)] sentence embeddings.

1.2 Corpus Filtering

CCMatrix, the largest corpus in the Japanese-Bulgarian language pair currently available (4.1M parallel sentences), is the focus of this research. If trained directly on the corpus, a state-of-the-art MT model is far from achieving optimal performance due to the abundance of noise that is typical to automatically crawled corpora [6].

Between 2018 and 2020, WMT introduced shared tasks that aimed at the filtering of large noisy corpora. Participants were asked to provide scores for each sentence pair in the provided corpus, allowing for solely the highest-scoring pairs to be used in the training of MT models. In this paper, one of the most successful models issuing from the task, Acarcicek et al. [1], is utilised to provide an additional score to the afore-mentioned margin one. Following preprocessing of the corpus based on heuristic rules, the two scores are tested separately as well as in different combinations in order to train optimal Transformer MT models in terms of both achieved BLEU score and training size. The highest BLEU score in the Japanese-Bulgarian direction (43.40) comes from a model trained on 500k parallel sentences based on the margin metric. A compact model of 200k parallel sentences, based on a combination of the margin and classifier scores, has a comparable result of

1) https://github.com/facebookresearch/LASER

42.82. The best model in the Bulgarian-Japanese direction, also based on a combined metric and trained on 200k sentences, has a score of 40.77.

This paper extends the work of Nikolova-Stoupak et al. [7], which also applies the mentioned classifier model in the selection of high-quality parallel sentences which are then used in the training of MT models of three increasing sizes (200k, 500k, and 1M sentence pairs). The classifier-based models' performance comes short of that of margin-based counterparts. Therefore, further work has been required in using the classifier in the filtering process. Improvements introduced in this paper include the construction of a new classifier model, trained on the Flores-200 evaluation dataset [8], additional training sizes, a combination of the two scoring mechanisms, and the assembly of an alternative dataset for evaluation of MT models.

2 Related Work

WMT organised three corpus-filtering shared tasks between the years 2018 and 2020, the last two centred specifically on low-resource language pairs. A number of successful projects were produced and, in particular, classifiers were utilised in several discrete ways. For instance, Sánchez-Cartagena [9] applied the classification tool Bicleaner to the provided noisy corpus, achieving an estimation of the parallelism of each original sentence pair. Acarcicek et al.'s [1] model, which is used in this paper, involves a classifier ("proxy-filter"), which is placed on top of a multilingual RoBERTa-Large model [10] and relies on "fuzzy (approximate) string matching" to select challenging negative examples of sentence parallelism.

3 Noise in CCMatrix

As recounted in [7], noise in the investigated corpus comes in a variety of forms, including mismatched numbers and dates, redundancy, lack of parallelism in meaning and machine-translated text. Importantly, the assembly of an evaluation dataset to score the derived MT models (described in section 5.1) has motivated additional observations concerning the types of noise that are still persistent in highly scoring sentences according to margin- and classifier-based filtering. Complete lack of correspondence within margin-derived sentence pairs is not uncommon. Automatically translated text is often observed and is typically reflected in mistakes in grammar, style and the use of gender and number in Bulgarian sentences. The highest scoring classifier-based sentences are in their vast majority parallel in terms of general meaning. Machine-translated text is, although less common than in the case of margin-based sentences, still present.

4 Methodology

Please refer to figure 1 for an overview of the full filtering process.

4.1 Preprocessing

The pipeline of initial preprocessing is identical to the one described in [7]. This stage provides an opportunity to consider the specificities of the two investigated languages; for instance, at determining the acceptable proportions between the sentences in a pair. Such an emphasis is not found in WMT's shared tasks as languages are selected solely based on the relative availability of associated corpora. Consequently, many models, including the regarded one by [1], do not involve a preprocessing stage based on heuristic rules.

4.2 Filtering Criteria

4.2.1 Margin Scores

The margin score that readily comes with the CCMatrix corpus as result of the web-crawling process is used in subsequent experiments with subcorpora selection in isolation or in addition to the score derived following application of [1]'s classifier model. The margin between a pair of sentences equals the ratio between their cosine distance and the cosine distances with their nearest neighbours in both directions in terms of LASER representations [5].

4.2.2 Proxy-Filter Classifier

As described in [7], the main reasons behind the selection of [1]'s model include its high performance, reproducibility and use of state-of-the-art neural translation models. In its quest for negative examples, the classifier selects sentences that are close to the correct translation of the source sentence in terms of Levenshtein distance. Ultimately, each original sentence pair receives a score denoting the estimated probability of it being parallel.

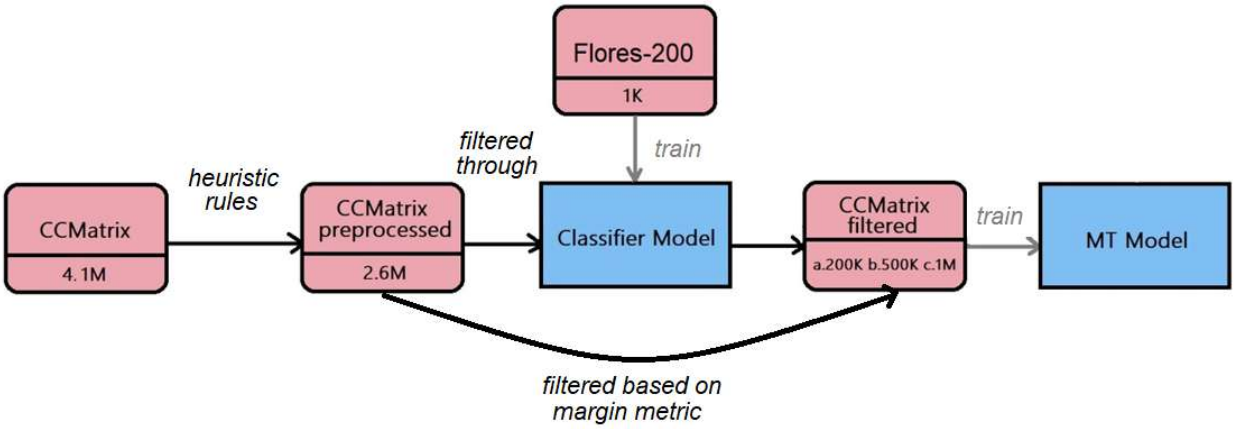

Figure 1 Broad pipeline of the presented work.

4.3 MT Models

Once the CCMatrix corpus is filtered and training sets of different sizes assembled, Transformer MT models are trained on them. The Transformer, a state-of-the-art model whose main strength is a reliance on attention mechanisms without the need for recurrence, is opted for due to its simplicity and high translation results [11].

5 Experiments

5.1 Data

This paper pertains to the cleaning of the CCMatrix corpus in the Japanese-Bulgarian language pair, whose original size is 4.1M parallel sentences. After the corpus is preprocessed as per section 4.1, its size is reduced to 2.5M sentences.

The evaluation dataset used for the MT models (divided into equal-sized test and validation sets) consists of 1k sentence pairs taken from the original CCMatrix corpus. Half of them are taken at random from the top-scoring 10k sentences based on the margin metric, and the other half are a classifier-based counterpart. In this way, a bias for a particular filtering method is avoided. In addition, the sentences are manually edited in order to ensure their quality.

In [7], the "proxy-filter" classifier is trained on OPUS's Open Subtitles corpus. However, due to its significantly more limited domain and register than CCMatrix, it has been replaced with Flores-200 [3]. This multilingual and rigorously assembled dataset is based on a variety of Wikipedia pages and has undergone professional human translation followed by several manual checks, thus providing both much higher relevance to the task at hand and higher overall quality. On the negative side, the dataset is meant for use in the evaluation of trained MT models and is therefore limited in size (1k parallel sentences). In contrast, Acarcicek et al. [1] explicitly state that their model has optimal quality when trained on a minimum of 2k sentence pairs.

5.2 Application of Scores

The presented translation models differ in terms of the size of training data as well as the ways in which they have been filtered from the original CCMatrix corpus. Like in [7], three main dataset sizes are considered: 200k, 500k and 1M sentence pairs. In addition, an alternative sizing is introduced that is based on the number of tokens following tokenization with sentencepiece[2)] (see Appendix A).

Combinations of the two described metrics have also been sought in order to optimise performance. The newly-introduced "margin-classifier" score is derived through averaging of the two scores, preceded by min-max normalisation. An additional experiment is carried out with doubling the weight of the margin-based score.

5.3 Translation Models

Once supcorpora of CCMatrix are selected based on classifier scores (please refer to Appendix B for a detailed description of the classifier model and the derived scores), Transformer models are trained on them. In addition, in order to provide a bigger picture of the impact filtering mechanisms and training sizes have on MT, models are

2) https://github.com/google/sentencepiece

Table 1 Performance of the full vs preprocessed CCMatrix corpus, JA to BG

Training Corpus	Size	BLEU
CCMatrix Full	4.1M	31.11
CCMatrix Preprocessed	2.5M	**41.39**

Table 2 BLEU scores for each filtering method for the three subcorpus sizes based on number of sentences, JA to BG

	200k	500k	1M
Random	26.02	31.90	37.36
Classifier-Based	25.53	31.66	37.11
Margin-Based	42.25	**43.40**	43.11
Margin-Classifier 1:1	42.82	41.75	41.35
Margin-Classifier 2:1	42.10	39.15	39.72

also trained on the full CCMatrix corpus and on the corpus following the preprocessing stage. All models are evaluated with the dataset described in 5.1 and the derived BLEU scores are juxtaposed.

The original idea behind the project is to investigate translation from Japanese to Bulgarian, as the latter language is the lower-resourced one and such a translation system would allow for direct translation into it of unique content originally composed in the Japanese language. However, for purposes of comparison and further insight into the work of the trained MT models, the opposite direction is also experimented with.

6 Results

6.1 Without Filtering

Comparing the BLEU scores achieved by MT models trained on the full vs the preprocessed CCMatrix corpus (Table 1) reveals that the preprocessing stage has caused a significant improvement while largely reducing the training size.

6.2 Following Filtering

As can be observed in Table 2, MT performance can be further improved following filtering of the preprocessed CCMatrix corpus. The most successful model, with a BLEU score of 43.40, is trained on 500k parallel sentences as filtered based on the margin metric. Overall, margin-based models achieve the highest results. The ones that follow are based on the 1:1 margin-classifier metric; notably, achieving a promising high-scoring compact model of 200k sentence pairs (BLEU 42.82). Like in the case of [7], last come the classifier-based models, whose weaker performance is not overcome despite the use of the more appropriate Flores-200 dataset to train the underlying classifier model. It is possible that overfitting is produced.

Table 3 BLEU scores for each filtering method for the three subcorpus sizes based on number of sentences, BG to JA

	200k	500k	1M
Random	23.2	30.11	32.87
Classifier-Based	22.19	29.05	34.52
Margin-Based	38.16	36.90	36.75
Margin-Classifier 1:1	**40.77**	37.75	36.71
Margin-Classifier 2:1	38.53	35.76	35.07

Going back to the results achieved by [7], the strongest one of which is 28.49 and comes from the 1M margin-based model, one can witness a large general increase of scores. In the case of margin-based models, a training size limit has now been reached, leading the 500k-sentence model to score higher than the 1M-sentence one.

6.3 Bulgarian-Japanese Direction

In the Bulgarian to Japanese translation direction, the best models are once again the margin- and margin-classifier-based (1:1) ones (Table 3), notably the latter outperforming the former. It can also be noticed that models using the strongest selection methods perform better in their more compact training sizes.

7 Conclusion and Future Work

This paper presents the training of MT models based on portions of the large but noisy CCMatrix corpus in the under-resourced Japanese-Bulgarian language pair. Filtering is based on the margin metric using which the corpus is web-crawled as well as on scores provided by a classifier model following the work of [1]. MT performance is significantly improved as compared with that of the full CCMatrix model as well as of a smaller version based on preprocessing with heuristic rules. The strongest model in the Japanese to Bulgarian direction (BLEU 43.40) is based on margin distance and is trained on 500k sentence pairs. In the opposite direction, best scoring is the model trained on 200k sentence pairs based on averaging of the two metrics. Future improvements of the filtering process to address the prevalence of machine-translated text may include sentence pre-ordering and the explicit inclusion of morphological information along with sentences.

Acknowledgements

This work was supported by Samsung, SDS.

References

[1] Haluk Açarçiçek, Talha Çolakoğlu, Pınar Ece Aktan Hatipoğlu, Chong Hsuan Huang, and Wei Peng. Filtering noisy parallel corpus using transformers with proxy task learning. In **Proceedings of the Fifth Conference on Machine Translation**, pp. 940–946, Online, November 2020. Association for Computational Linguistics.

[2] Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. A survey of multilingual neural machine translation. **ACM Comput. Surv.**, Vol. 53, No. 5, sep 2020.

[3] Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. The flores-101 evaluation benchmark for low-resource and multilingual machine translation, 2021.

[4] Jörg Tiedemann. Parallel data, tools and interfaces in OPUS. In **Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)**, pp. 2214–2218, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA).

[5] Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. CCMatrix: Mining billions of high-quality parallel sentences on the web. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 6490–6500, Online, August 2021. Association for Computational Linguistics.

[6] Roland Schäfer, Adrien Barbaresi, and Felix Bildhauer. Focused web corpus crawling. 04 2014.

[7] Iglika Nikolova-Stoupak, Shuichiro Shimizu, Chenhui Chu, and Sadao Kurohashi. Filtering of noisy web-crawled parallel corpus: the Japanese-Bulgarian language pair. In **Proceedings of the 5th International Conference on Computational Linguistics in Bulgaria (CLIB 2022)**, pp. 39–48, Sofia, Bulgaria, September 2022. Department of Computational Linguistics, IBL – BAS.

[8] Alexandre Magueresse, Vincent Carles, and Evan Heetderks. Low-resource languages: A review of past work and future challenges, 2020.

[9] Víctor M. Sánchez-Cartagena, Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez. Prompsit's submission to WMT 2018 parallel corpus filtering shared task. In **Proceedings of the Third Conference on Machine Translation: Shared Task Papers**, pp. 955–962, Belgium, Brussels, October 2018. Association for Computational Linguistics.

[10] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

Table 4 The hyperparameters used in the grid search for the optimal "proxy-filter" classifier model

Hyperparameter	Definition	Values
Number of training epochs		2, 5, 10
Learning rate		2e-6, 2e-4, 2e-2, 0.2
Negative random sampling	the ratio of negative examples in the classifier	2, 5, 8, 10
Fuzzy ratio	the number of similar sentences taken	1, 2, 5
Fuzzy max score	a threshold for the permitted similarity of sentences	30, 60, 100
Positive oversampling	oversampling of the classifier's positive examples	1, 2, 5

Table 5 BLEU scores for each filtering method for the three token-based subcorpus sizes, JA to BG

	Small	Medium	Large
Random	28.53	34.53	37.07
Classifier-Based	25.23	34.77	38.77
Margin-Based	42.25	**43.40**	43.11

A Token-Based Training Sizes

In one set of experiments, the number of tokens in random and classifier-based models is deliberately made to match the one in the three main sizes of margin-based models. The reason behind this experiment lies in the perceived and consequently proven via statistical analysis significant difference between sentence sizes associated with the different filtering methods. Whilst the margin score favours sentences that are longer than the CCMatrix corpus's average sentence size, the classifier model strongly prefers short sentences.

The results deriving from token-size-based experiments are recorded in Table 5. Note that the margin-based results are identical to the ones presented in Table 4, as the 200k-, 500k- and 1M-sentence sizes of margin-based models are used to generate random and classifier-based models of the same token sizes. As expected, performance of the random and classifier-based models is increased with the increase of training size. The only exception is the largest random model, whose performance (BLEU 37.07) is lower than that achieved with the slightly smaller sentence-based model (37.36). This observation implies that a limit has been reached for the size of random-based models, too.

B Classifier Model

The particular classifier model to be applied in the filtering of CCMatrix is selected following grid searching, as in [7]. Based on previous results, the range of hyperparameters is slightly modified in the following way: the number of training epochs is increased to include 10, positive oversampling is slightly reduced and a decision is made to retain the possibility for a fuzzy max score equal to 100 (i.e. high similarity between sentences is not discouraged as the Flores-200 dataset is free or redundancy). Table 4 shows the content of the grid search as well as the definitions of project-specific hyperparameters. The models are trained on a single TITAN RTX GPU.

Despite the limited data classifier models were trained on, the highest scoring model based on the grid search achieved a surprisingly high F1 score of 0.95. In contrast, its counterpart in [7], achieved an F1 score of 0.58. The model was trained for 5 epochs with negative random sampling of 2 and positive oversampling of 5, a proportion common to the highest-scoring models. The model' s fuzzy ratio is 2, and the fuzzy max score is 100.

After the winning classifier model is applied to the preprocessed CCMatrix corpus, each sentence pair receives a score representing the certainty of the two sentences being mutual translations. The values of the derived scores range between 0.0563 and 0.9996, the median coming at 0.9992.