A Fast Method to Filter Noisy Parallel Data WMT2023 Shared Task on Parallel Data Curation

Minh-Cong Nguyen-Hoang[1] **Vinh Nguyen Van**[2] **Le-Minh Nguyen**[1]

[1]*Japan Advanced Institute of Science and Technology, JAIST*
[2]*University of Engineering and Technology, VNU, Hanoi, Vietnam*
{congnhm, nguyenml}@jaist.ac.jp
vinhnv@vnu.edu.vn

Abstract

The effectiveness of a machine translation (MT) system is intricately linked to the quality of its training dataset. In an era where websites offer an extensive repository of translations such as movie subtitles, stories, and TED Talks, the fundamental challenge resides in pinpointing the sentence pairs or documents that represent accurate translations of each other. This paper presents the results of our submission to the shared task WMT2023 (Sloto et al., 2023), which aimed to evaluate parallel data curation methods for improving the MT system. The task involved alignment and filtering data to create high-quality parallel corpora for training and evaluating the MT models. Our approach leveraged a combination of dictionary and rule-based methods to ensure data quality and consistency. We achieved an improvement with the highest 1.6 BLEU score compared to the baseline system. Significantly, our approach showed consistent improvements across all test sets, suggesting its efficiency.

1 Introduction

Neural Machine Translation (NMT) has revolutionized the field of machine translation by utilizing deep learning algorithms to learn from large amounts of data and generate high-accurate translations (Sennrich et al., 2016; Vaswani et al., 2017). However, the success of NMT models heavily depends on the quantity and quality of data used for training. On low-resource language pairs, the NMT architectures perform poorly (Koehn and Knowles, 2017; Khayrallah and Koehn, 2018) and are more sensitive to noisy data than statistical machine translation (SMT) methods (Belinkov and Bisk, 2017; Koehn et al., 2018). Therefore, access to vast cleaned corpus can significantly improve the performance of NMT models, allowing them to learn and produce more accurate translations (Bojar et al., 2017).

Fortunately, very large text sources offer a massive collection of data for various types of content, including movie subtitles, stories, and TED Talks. These resources have not been fully exploited for NMT training due to the lack of alignment between the source and target languages. Furthermore, the parallel data which movie subtitles also could be noisy with poor accuracy (Khayrallah and Koehn, 2018). To address this challenge, WMT2023 introduced a shared task on Parallel Data Curation for the Estonian-Lithuanian (et-lt) language pair, focusing on finding the best possible training data set within the web-crawled data to train a downstream NMT model (Sloto et al., 2023).

Among the popular solutions, Thompson and Koehn (2019) introduced using Vecalign to embed sentences and compute the cosine similarity of sentence pairs. Following this method, the shared task provides participants with extensive cosine similarity files and LASER embeddings generated by the LASER model (Heffernan et al., 2022). Participants are tasked with identifying the most optimal parallel data to train the MT models. Although this approach performs efficiently in many cases, Zhou et al. (2022) has shown that the cosine similarity has several limitations. Because the sentence representation in vector space could be impacted by word frequency. To tackle this problem, we build a pipeline to improve the quality of the parallel corpus. Our contributions focus on:

- using the phrased base dictionary to distill the high-quality sentences.
- making the pipeline to re-ranking the top-K cosine similarity.
- analyzing the influence of cosine similarity thresholds on corpus size and MT Models.

The related work is presented in section 2. The detail of our method is described in section 3, experi-

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 359–365
December 6–7, 2023.

ments and results are shown in section 4. Finally, the analysis is presented in section 4.5.

2 Related work

The WMT2023 shared task builds upon previous shared tasks focused on document alignment (WMT 16) and sentence filtering (WMT 18, 19, 20) (Buck and Koehn, 2016; Koehn et al., 2018, 2019, 2020). Previously, several researchers proposed a method to align documents, such as Gomes and Pereira Lopes (2016) used the phrase table to align in the search space and then fill in and refine alignments. Moreover, Thompson and Koehn (2019) employed the Vecalign to gain the sentence embeddings. Nevertheless, Sentence alignments based on cosine similarity have some limitations because the cosine scores could be dense in the range of 0.5 to 1 (Zhou et al., 2022). And with the same query sentence, the higher score could not determine the quality of the parallel sentence.

In addition, for the filtering shared task, participants applied filtering rules to eliminate noisy data, including removing too long/short sentences, using language identification for source and target (Kejriwal and Koehn, 2020) or fine-tuning pre-trained models such as BERT, XLM to re-score sentence pairs (Yang et al., 2019; Bernier-Colborne and Lo, 2019; Açarçiçek et al., 2020). Besides, Xu and Koehn (2017) created artificially noisy data by generating inadequate and nonfluent translations. They used this noisy data to train a classifier to distinguish between high-quality and low-quality sentence pairs within a corpus containing noise.

We found the related ideas from (Lu et al., 2020; Xu et al., 2020). Both of these approaches only focus on the alignment rules and adopt the other pre-trained models. Junczys-Dowmunt (2018) trained an NMT model to filter data and became standard for the high-resource case. Nevertheless, when training an original NMT model with low or noisy resources, the NMT model could face certain limitations. In our work, we utilize the phrase table to compute *edit distance* and extract the superior sentences. Furthermore, we introduce a pipeline to re-rank sentences based on their top-K cosine similarity scores and extract the best compact corpus for training purposes. The detail of our method is presented in section 3.

3 Methodology

3.1 LASER Similarity Scores

The LASER2 similarity scores are produced for the WMT23 shared task. These files are an intermediate output from our baseline submission. The laser embeddings applied L2 normalization and added them to a flat inner product index, such that the resulting scores are equivalent to cosine similarity. And query each index with all L2 normalized embeddings in the target sentences and store the top-8 results (locally, per chunk). Finally, the data is aggregated and meticulously sorted across unique IDs.

3.2 Building Dictionary

Our proposed method incorporates several innovative techniques to enhance the accuracy and efficacy of the filtering process. Initially, we train a phrased table based on MGiza++[1] (Gao and Vogel, 2008), a widely utilized algorithm for learning phrase tables from parallel corpora. Given a source string $X^I = \{x_1...x_i...x_I\}$ and a target string $Y^J = \{y_1...y_j...y_J\}$. In the context of statistical alignment, the probability of a source sentence given a target sentence is formulated as follows:

$$P(X^I|Y^J) = \sum_{i=1}^{J} P_\theta(X_i^I, a_i^J|Y_i^J), \qquad (1)$$

Where a_i^J represents the alignment of the sentence pair. The parameters θ can be estimated using maximum likelihood estimation (MLE) on a training corpus to represent the statistical probability with the best alignment of the sentence pair:

$$a_1^J = \arg\max_{a_1^J} p_\theta(x_i^I, a_i^J|y_i^J), \qquad (2)$$

These steps enable us to establish connections between words in the source and target languages. After that, we extract the dictionary from the phrase base table. This stage helps to remove unnecessary or redundant words and sentences, streamlining the dictionary and improving its quality.

3.3 Edit Distance

In this work, we utilized the dictionary to translate source sentences to target sentences called candidate strings. To identify sentences where the source and target are similar, we compute the edit

[1] https://github.com/moses-smt/mgiza

distance between a pair of candidate and reference sentences.

$$score = \sigma(C_i^I, Y_i^J), \quad (3)$$

Here, C_i^I is the candidate sentence and Y_i^J indicates the target sentence. The σ function employs the Damerau-Levenshtein distance (Miller et al., 2010). We set each insertion, deletion, and substitution as one step, but the transposition (swapping) of two words is computed as $\frac{1}{2}$ step. We opted for option $\frac{1}{2}$ in the swapping step due to the limitation of using a dictionary to translate strings, which neglects word positions. Finally, we choose the sentences that have scores greater than or equal to $\frac{N}{2}$, with N being the max length of the candidate sentence and target sentence.

3.4 Accumulative Filtering

Because the word frequency impacts the cosine score, we produce a filtering pipeline to improve the corpus quality and apply it to the larger corpus. The implementation method is described in Algorithm 1.

Algorithm 1: Accumulative Filtering

Data: The raw parallel corpora: $(X^I, Y^J) \in D$, dictionary, and threshold_values in range {0.7-0.9}
Result: The cleaned data: $(X_f, Y_f) \in D$
`/* Initialize the NMT model` θ `*/`
$t \leftarrow 0.9$;
$X, Y \leftarrow filter(D, t)$;
`/* Filter data via top1 cosine score with the threshold of t */`
$C' \leftarrow translate(X, dictionary)$;
`/* Translate source using dictionary */`
$X_f, Y_f \leftarrow select(C', Y)$;
`/* Select sentences based on edit-distance score */`
$\theta \leftarrow train(X_f, Y_f)$;
`/* Loop t with the step as 0,5 */`
for $t \in threshold\_values$ **do**
  $X, Y \leftarrow filter\_topK(D, t, 8)$;
  $X' \leftarrow translate(X, \theta)$;
  $X_f, Y_f \leftarrow select(C', Y)$;
  $\theta \leftarrow train(X_f, Y_f)$;
end

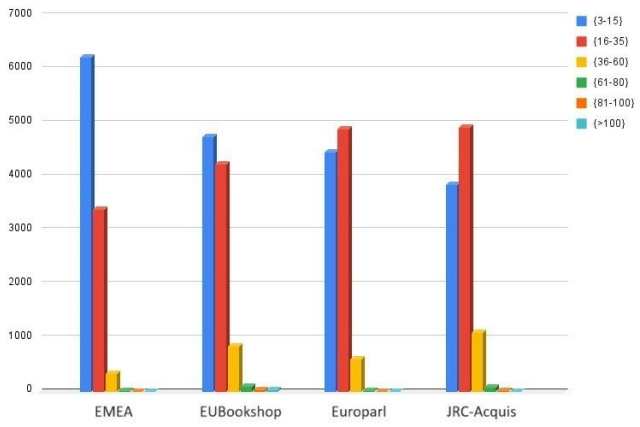

Figure 1: The statistics of the sentence length of each test set are used to evaluate the cleaned corpus. We separate the length of sentences into five levels, with 3-15 as the total sentences exhibit lengths that fall within the 3 to 5 range. The 16-35, 36-60, 61-80, 81-100, and >100 are the 16 to 35, 36 to 60, 61 to 80, 81 to 100, and greater 100 correspondingly.

4 Experiments

In this section, we describe the experimental setup for our system, including the data, training tools, and baseline system.

4.1 Data

In this shared task, the corpus was gathered from a recent snapshot of CommonCrawl[2].

Training Data

From the crawled data, the data is smoothed to some steps such as extracting plain texts from HTML documents, using the identifier language to hold the Estonian and Lithuanian documents, and removing the unsafe and offensive content. Besides, each sentence is assigned a distinct, randomly generated unique ID. These identifiers are uniform within their language datasets but diverge between two languages. This allows quick access and operation with data. Table 2 depicts the total number of collected data.

Testing Data

To assess the quality of the cleaned corpus, we train NMT models and evaluate them in four test sets, including EMEA, EU-Bookshop, Europarl, and JRC-Acquis. The statistics of the sentence length of each test set are exhibited in Figure 1, with the

[2]https://commoncrawl.org/blog/jan-feb-2023-crawl-archive-now-available

	EMEA BLEU	EU-Bookshop BLEU	Europarl BLEU	JRC-Acquis BLEU	EMEA chrF	EU-Bookshop chrF	Europarl chrF	JRC-Acquis chrF
LASER (Baseline)	18.3	19.1	18.1	24.3	49.7	52.3	51.8	55.2
Dictionary +Edit-Dist	18.3	19.1	18.5	24.3	49.7	52.3	51.8	54.9
Accummulative Filtering:								
Threshold-0.9	18.1	20.0	18.3	25.1	49.6	52.2	51.9	55.1
Threshold-0.85	18.5	**20.3**	19.1	25.4	49.7	52.7	**52.2**	55.2
Threshold-0.8*	**19.2**	20.1	**19.2**	25.7	**49.9**	**52.8**	52.1	55.4
Threshold-0.75	19.0	20.2	18.9	**25.9**	49.8	52.7	52.0	**55.6**

Table 1: The evaluation of BLEU scores and chrF scores for the filtering and alignment corpus.

No.	Estonian	Lithuanian
Num of Sents	53,279,844	63,556,320

Table 2: The statistics of sentences are available in the corpus for the Estonian-Lithuanian.

total number of sentences in each test set being 10,000.

4.2 Training Tools

We utilize the training scripts[3] provided by organizers to run the evaluation for the Shared Task. To observe the effect of filtered datasets, we use the same hyper-parameters for the whole experiment to compare results equally. In more detail, the Transformer architecture (Vaswani et al., 2017) is used in the training tool with the default 8 heads, 6 layers, and the model size is 512. Besides, the training pipeline employs the subword segmentation tool provided by (Sennrich et al., 2016) for tokenization. We use the sacreBLEU (Post, 2018) and ChrF++ (Popović, 2015) score to evaluate whole experiments.

4.3 Baseline

Following scripts provided by organizers, we present briefly how to create a simple baseline. Firstly, we collect the whole provided cosine similarity files. Secondly, we extract sentence alignments with the threshold of 0.9 and only select the top highest similarity scores. And finally, we run the end-to-end evaluation to produce BLEU scores from the extracted data.

[3]https://github.com/awslabs/sockeye/tree/wmt23_data_task

4.4 Our system

In the first place, we obtain the cosine files that are computed from Laser embeddings. From these files, We extract the sentence pairs by considering the highest cosine similarity score, specifically the top-1 score, and we set a threshold of 0.9. In the following phase, we remove longer sentences having 200 tokens and more and utilize the dictionary to perform word-by-word translation of these source sentences into the target language. After that, we compute the edit distance and eliminate poor-quality sentences. Finally, we employ the cumulative filtering algorithm discussed in section 3.4 to acquire the expanded corpus, opting for thresholds of 0.9, 0.85, and 0.8, respectively.

4.5 Results

In this section, we present the results obtained through a comparative analysis of different methods within the context of our works. Table 1 illustrates the results attained in the development system while preparing the submission. The system responsible for generating the scores for our final submission is shown in underline. We consider the reported results as the LASER baseline, and the outperforming results are indicated in bold.

Our investigation reveals that the LASER baseline provides a starting point for evaluation and moderate levels of performance across a range of metrics. However, the Accumulative Filtering approach, particularly when applying lower threshold values (0.85, 0.8, and 0.75), showcases significant improvements in various metrics. Notably, the choice of threshold within the Accumulative Filtering method influences performance, with lower thresholds yielding higher results. These findings

	top1_0.95	top1_0.9	top1_0.85	top1_0.8	top1_0.75	top1_0.7
Corpus size	173,239	1,230,810	4,194,132	12,918,719	27,811,424	32,568,712
BLEU	16.8	23.9	25.1	25.4	25.0	24.4

Table 3: The influence of sentences on the corpus size and BLEU score. The evaluation of the BLEU score is conducted specifically on the JRC-Acquis test set. Sentence selection is based solely on the cosine score threshold, with additional criteria involving the removal of excessively short or long sentences.

underscore the importance of threshold selection and methodological considerations in achieving optimal outcomes. Further analysis and task-specific considerations are required to determine the most suitable approach for our specific research objectives. We analyze the impact of the cosine similarity score thresholds on the corpus size and quality of NMT models. The details are described in section 4.5.

5 Analysis

In this section, we delve deeply into our approaches and the scale of our data corpora. Firstly, we conduct some experiments to find the best threshold when selecting the top-K highest cosine similarity score. For every source sentence, our approach involves selecting a single target sentence from a set of eight candidates based on the highest cosine similarity score provided. Table 3 illustrates the impact of different sentence selection criteria, denoted by the cosine similarity thresholds (top1_0.95, top1_0.9, top1_0.85, top1_0.8, top1_0.75, top1_0.7), on both the corpus size and BLEU score. The corpus size varies significantly depending on the threshold, ranging from 173,239 sentences to 32,568,712 sentences. Simultaneously, the BLEU score, evaluated on the JRC-Acquis test set, fluctuates, with the highest score of 25.4 achieved at the top1_0.8 threshold. These findings underscore the delicate balance between corpus size and translation quality, highlighting the importance of threshold selection in the context of machine translation evaluation.

Secondly, we conduct a statistical analysis to determine the number of sentence pairs that achieve the highest cosine score but are not considered parallel sentences. Table 4 shows the statistics for the number of sentences that do not have the highest cosine similarity score but are regarded as parallel sentences. The table indicates a total of 5,981,148 sentences in cleaned data and 353,642 sentences are considered re-ranking parallel sentences.

No.	Cleaned Data	Re-ranking
Num of Sents	5,981,148	353,642

Table 4: Number of sentences that are not in top-1 cosine similarity score, but are considered parallel sentences.

6 Conclusion

In conclusion, our study has provided valuable insights into the performance of different methods employed in our research on WMT2023 parallel data curation shared tasks. Our findings reveal that while the LASER baseline and using the dictionary method exhibited moderate and consistent performance across several metrics, the accumulative filtering approach, particularly when adopting lower threshold values (0.85, 0.8, and 0.75), demonstrated notable improvements in various aspects. Notably, the selection of the threshold played a pivotal role in influencing performance outcomes. Furthermore, our analysis also encompassed the identification of sentence pairs that exhibit parallel characteristics, even if they may not always possess the highest cosine similarity scores. In the future, further investigation and task-specific considerations will be essential in finding the smallest possible set of training data and achieving the highest result.

References

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

Yonatan Belinkov and Yonatan Bisk. 2017. Synthetic and natural noise both break neural machine translation.

Gabriel Bernier-Colborne and Chi-kiu Lo. 2019. NRC parallel corpus filtering system for WMT 2019. In *Proceedings of the Fourth Conference on Machine*

Translation (Volume 3: Shared Task Papers, Day 2), pages 252–260, Florence, Italy. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Christian Buck and Philipp Koehn. 2016. Findings of the WMT 2016 bilingual document alignment shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 554–563, Berlin, Germany. Association for Computational Linguistics.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, Ohio. Association for Computational Linguistics.

Luís Gomes and Gabriel Pereira Lopes. 2016. First steps towards coverage-based document alignment. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 697–702, Berlin, Germany. Association for Computational Linguistics.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages.

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.

Ankur Kejriwal and Philipp Koehn. 2020. An exploratory approach to the parallel corpus filtering shared task WMT20. In *Proceedings of the Fifth Conference on Machine Translation*, pages 959–965, Online. Association for Computational Linguistics.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.

Philipp Koehn, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. 2020. Findings of the WMT 2020 shared task on parallel corpus filtering and alignment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 726–742, Online. Association for Computational Linguistics.

Philipp Koehn, Francisco Guzmán, Vishrav Chaudhary, and Juan Pino. 2019. Findings of the WMT 2019 shared task on parallel corpus filtering for low-resource conditions. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 54–72, Florence, Italy. Association for Computational Linguistics.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018. Findings of the WMT 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 726–739, Belgium, Brussels. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Jun Lu, Xin Ge, Yangbin Shi, and Yuqi Zhang. 2020. Alibaba submission to the WMT20 parallel corpus filtering task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 979–984, Online. Association for Computational Linguistics.

F.P. Miller, A.F. Vandome, and J. McBrewster. 2010. *Damerau-Levenshtein Distance*. Alphascript Publishing.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting bleu scores.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units.

Steve Sloto, Brian Thompson, Huda Khayrallah, Tobias Domhan, Thamme Gowda, and Philipp Koehn. 2023. Findings of the WMT 2023 shared task on parallel data curation. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore. Association for Computational Linguistics.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348, Hong Kong, China. Association for Computational Linguistics.

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

Hainan Xu and Philipp Koehn. 2017. Zipporah: a fast and scalable data cleaning system for noisy web-crawled parallel corpora. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2945–2950, Copenhagen, Denmark. Association for Computational Linguistics.

Runxin Xu, Zhuo Zhi, Jun Cao, Mingxuan Wang, and Lei Li. 2020. Volctrans parallel corpus filtering system for WMT 2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 985–990, Online. Association for Computational Linguistics.

Yinfei Yang, Gustavo Hernandez Abrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax.

Kaitlyn Zhou, Kawin Ethayarajh, Dallas Card, and Dan Jurafsky. 2022. Problems with cosine as a measure of embedding similarity for high frequency words. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 401–423, Dublin, Ireland. Association for Computational Linguistics.