A comparison of data filtering techniques for English-Polish neural machine translation in the biomedical domain

Jorge del Pozo Lérida
IT University of Copenhagen
jord@itu.dk

Kamil Kojs
IT University of Copenhagen
kako@itu.dk

Janos Mate
IT University of Copenhagen
janma@itu.dk

Mikołaj Antoni Barański
IT University of Copenhagen
mikba@itu.dk

Christian Hardmeier
IT University of Copenhagen
chrha@itu.dk

Abstract

Large Language Models (LLMs) have become the state-of-the-art in Neural Machine Translation (NMT), often trained on massive bilingual parallel corpora scraped from the web, which contain low-quality entries and redundant information, leading to overall significant computational challenges. Various data filtering methods exist to reduce dataset sizes, but their effectiveness largely varies based on specific language pairs and domains. This paper evaluates the impact of commonly used data filtering techniques—LASER, MUSE, and LaBSE—on English-Polish translation within the biomedical domain. By filtering the UFAL Medical Corpus, we created varying dataset sizes to fine-tune the mBART50 model, which was then evaluated using the SacreBLEU metric on the Khresmoi dataset, having quality of translations assessed by bilingual speakers. Our results show that both LASER and MUSE can significantly reduce dataset sizes while maintaining or even enhancing performance. We recommend the use of LASER, as it consistently outperforms the other methods and provides the most fluent and natural-sounding translations.

1 Introduction

Recent advancements in LLMs have resulted in a notable increase in the size of model architectures used for NMT, with publicly accessible models such as mBart-50 reaching parameter sizes up to 600 million (Ushio et al., 2023). This escalation in parameter size has consequentially led to increased computational requirements for these models, necessitating the use of the most advanced GPUs available for their training. A common approach in training LLMs involves utilizing the entirety of the data present in the prepared training dataset. However, it is often the case that these datasets are not entirely cleansed of low-quality entries. Given that training datasets frequently comprise over 1 million samples, often scraped from the web, it is impractical to manually inspect and purify the entire dataset of such substandard samples. Consequently, it is typical that only a minimal fraction of the dataset is of high quality, with the majority being of poor quality.

Past research has demonstrated that including low-quality data in the training process minimally contributes to the overall quality and performance of LLMs (Koehn et al., 2018), underscoring the imperative to refine training datasets to contain only high-quality entries. In this study, we hypothesize that the application of filtering techniques can significantly reduce the size of the training dataset in NMT with LLMs, when fine-tuning for a specific domain and language pair, without compromising or potentially improving performance.

Specifically, we investigate domain adaptation for biomedical translation from English to Polish, performing a systematic comparison of common filtering methods to identify the best candidate for this particular setup. We achieve this by filtering a large in-domain biomedical corpus into smaller datasets of different sizes using the different methods. We then fine-tune mBART50 model on these filtered subsets and evaluate on an independent dataset, comparing performance to that of using a whole corpus or randomly sampled equally sized subsets. Our main contribution is to provide the first systematic comparison of data filtering methods for English-Polish biomedical NMT and provide specific recommendations. For the sake of reproducibility all code is made available[1]

2 Related Work

NMT models depend on large quantities of high-quality data for domain adaptation (Koehn et al., 2018). To enhance the quality of web-scraped corpora, various automated filtering methods have

[1] https://github.com/jorgedelpozolerida/Biomed-NMT-EngPol

Model	**Filter**	**Avg. BLEU**	**Min**	**Max**	**∆ All**	**∆ 60/20**
Base-none	-	14.936	-	-	-2.466	-
Base-all	-	17.402	-	-	-	-
Base-60%	-	17.234	17.174	17.296	-0.168	-
Filtered-60%	LASER	**17.411**	-	-	**0.009**	**0.177**
Filtered-60%	MUSE	17.239	-	-	-0.163	0.005
Filtered-60%	LaBSE	17.151	-	-	-0.251	-0.083
Base-20%	-	16.801	16.516	17.041	-0.601	-
Filtered-20%	LASER	**17.114**	-	-	**-0.288**	**0.313**
Filtered-20%	MUSE	17.071	-	-	-0.331	0.270
Filtered-20%	LaBSE	16.376	-	-	-1.026	-0.425

Table 1: Evaluation results on the Khresmoi test dataset using SacreBLEU. For the Base-60% and Base-20% models average scores between random seeds are reported.

been explored within the NMT field, including outlier detection (Taghipour et al., 2011), discriminator models (Xu and Koehn, 2017), graph-based unsupervised models (Cui et al., 2013), and LLM-based classifiers or scorers (Açarçiçek et al., 2020). With the advent of LLMs, Language-agnostic encoders like LASER, LaBSE, and MUSE have enabled direct scoring of bilingual sentence similarity for dataset filtering, proving competitive with more complex classifier-based models (Chaudhary et al., 2019).

Research by Bane and Zaretskaya (2021) evaluated filtering effectiveness on English-Japanese and English-German sentence pairs using models like Marian-scorer, LASER, MUSE, and XLM-R. Findings showed limited dataset reduction (54%-73%) with comparable BLEU scores to random selection, though Marian-based filtering consistently outperformed random downsizing, and MUSE showed variable performance by language. The latter indicates that filtering results might not be universally applicable across different language pairs or topics, which motivated our study design to investigate Polish-English translation specifically.

Further exploration by Bane et al. (2022) assessed the strengths and weaknesses of specific filtering methods. They developed a dataset with ten types of noise or errors to test these methods. Results indicated that a custom-trained Marian-Scorer had the best cleaning performance, while embedding-based methods like XLM-R, MUSE, and LASER, although less effective, still performed adequately and were particularly effective at identifying issues like number mismatches and spelling errors without requiring the computationally costly calibration needed for the Marian-Scorer.

3 Data

The selected model was fine-tuned on the Polish-English sentence pairs from the UFAL Medical Corpus [2], consisting of 1,116,773 pairs sourced from documents of the European Centre for Disease Prevention and Control, the European Medicines Agency, and Open subtitles. Extensive preprocessing included removing duplicates, untranslated sentences, those under 15 or over 200 characters, and sentences containing characters from non-target languages, resulting in a refined dataset of 711,720 sentences, with 700k randomly sampled for training. For testing, we used the Khresmoi dataset (Dušek et al., 2017), which comprises 1,500 high-quality Polish-English medical sentence pairs.

4 Methodology

We employed three widely used multilingual embedding models — LASER, MUSE, LaBSE — to help us filter the medical-domain corpus. Each embedding method was used to generate a sentence representation of each sentence in a pair, either by averaging all token embeddings in the sentence or by taking the sentence representation from the method if already provided. We then utilized cosine similarity to score sentence pairs from the in-domain training data, retaining 20% (approximately 150k pairs) and 60% (approximately 420k pairs) of the highest-scoring sentences from each method to create eight different filtered training datasets for all combinations of embedding methods and sizes. Each subset was used to separately fine-tune a pre-trained mBART50 model, and the evaluation was conducted on an independent dataset. Additionally, KK and MB — native

[2] https://ufal.mff.cuni.cz/ufal$_{m}edical_{c}orpus$

in polish and fluent in English — examined qualitatively the translations to determine which method sounded more natural.

4.1 Filtering methods

4.1.1 LaBSE

LaBSE (Language-agnostic BERT Sentence Embeddings) is a method developed by Google for generating BERT-based cross-lingual sentence embeddings in over 109 languages (Feng et al., 2022). This model, available on Hugging Face [3], addresses the limitations of the original BERT's multilingual embeddings by employing a dual-encoder framework. It uses a pre-trained BERT to produce embeddings for two translations of the same sentence, with the training loss calculated as the difference between these embeddings, facilitating the development of a unified cross-lingual embedding space. LaBSE was chosen for our analysis due to its demonstrated potential in new applications and its effectiveness in cross-lingual settings, and because it has not been studied in previous data filtering literature.

4.1.2 LASER

LASER (Language-Agnostic SEntence Representations) (Schwenk and Douze, 2017), developed by Facebook AI Research, utilizes a BiLSTM (Bidirectional Long Short-Term Memory) architecture to create language-agnostic sentence embeddings [4]. This model is trained on a vast multilingual corpus with parallel corpora, enabling it to generate consistent embeddings for semantically equivalent sentences across more than 90 languages. LASER employs a Byte Pair Encoding (BPE) tokenizer to process various languages by segmenting words into shared subword units, enhancing its language-generalization capability. It encodes input sentences into a fixed-size vector. Due to its robust performance in data filtering across different domains and languages, LASER was selected for our analysis (Chaudhary et al., 2019).

4.1.3 MUSE

MUSE (Multilingual Unsupervised and Supervised Embeddings) is a model developed by Meta to foster the creation and evaluation of cross-lingual word embeddings (Conneau et al., 2017). This model uses an unsupervised approach for aligning monolingual word embeddings, which includes adversarial training to establish a linear mapping between source and target embedding spaces, synthesizing a dictionary from the mapped space, and refining the alignment with the Procrustes solution, allowing for cross-lingual alignment without annotated data or parallel corpora. Though the original model is no longer available, we utilized its precompiled embeddings dictionary [5]. MUSE was selected for our research due to its demonstrated effectiveness across various language pairs and domains in data filtering tasks (Bane and Zaretskaya, 2021).

4.2 Experimental setup

We compared the performance of 10 models, with 9 of them fine-tuned on variously sized subsets of in-domain corpus and one remaining being the untouched pre-trained baseline (*Base-none*). The later was fine-tuned on the full unfiltered dataset to obtained fine-tuned baseline (*Base-all*). We employed a stratified split of 80% training and 20% validation for all experiments to maintain consistent proportions of the three different data sources. To be able to observe the effect of filtering against randomness, baseline models were trained on randomly selected subsets of 20% and 60% (*Base-20%* and *Base-60%*), each trained three times with different seeds to average out random variance in performance evaluation.

We utilized the publicly available mBART50 model (Tang et al., 2020)[6], developed by Facebook AI and available for use in Polish and English, with text tokenization performed by the MBart50Tokenizer. Evaluation was conducted on an independent test dataset using the BLEU metric implemented via SacreBLEU (Post, 2018), ensuring unbiased assessment. Training time was also reported to highlight the efficiency gains from fine-tuning on smaller data subsets.

Computations were conducted on the LUMI supercomputer [7], employing its *standard-g* partition with AMD MI250x GPUs, totaling 646 GPU hours. Training involved using Trainer function from transformers library, with 16-bit precision for weights, batch sizes of 15 for training and 20 for evaluation, a linear learning rate scheduler for the initial 100 steps, and an AdamW optimizer. All models underwent exactly three training epochs.

[3] https://huggingface.co/sentence-transformers/LaBSE

[4] https://github.com/facebookresearch/LASER

[5] https://ai.meta.com/tools/muse/

[6] https://huggingface.co/facebook/mbart-large-50

[7] https://www.lumi-supercomputer.eu/about-lumi

Type	Sentence
English	Meningococcal Disease is a serious bacterial infection that can cause swelling of the brain and spinal cord, and infection of the blood and other organs.
Ground Truth	Infekcja meningokokowa jest poważną chorobą bakteryjną, która może spowodować obrzęk mózgu i rdzenia, infekcję krwi i innych narządów.
LaBSE-60	Meningooka jest ciężkim zakażeniem bakteryjnym, które może powodować obrzęk mózgu i rdzenia kręgowego oraz zakażenie krwi i innych narządów.
LASER-60	Choroba meningokokowa jest ciężkim zakażeniem bakteryjnym, które może powodować obrzęk mózgu i rdzenia kręgowego oraz zakażenie krwi i innych narządów.
MUSE-60	Choroba meningokokowa jest ciężkim zakażeniem bakteryjnym, które może powodować obrzęk mózgu i rdzenia kręgowego oraz zakażenie krwi i innych narządów.
Base-60	Choroba meningokokowa jest ciężkim zakażeniem bakteryjnym, które może powodować obrzęk mózgu i rdzenia kręgowego oraz zakażenie krwi i innych narządów.
LaBSE-20	Meningokoczka jest poważnym zakażeniem bakteryjnym, które może powodować obrzęk mózgu i rdzenia kręgowego oraz zakażenie krwi i innych narządów.
LASER-20	Choroba meningokokowa jest poważnym zakażeniem bakteryjnym, które może powodować obrzęk mózgu i rdzenia kręgowego oraz zakażenie krwi i innych narządów.
MUSE-20	Choroba gruczołu krokowego jest poważnym zakażeniem bakteryjnym, które może powodować obrzęk mózgu i rdzenia kręgowego oraz zakażenie krwi i innych narządów.
Base-20	Choroba meningokokowa jest ciężkim zakażeniem bakteryjnym, które może powodować obrzęk mózgu i rdzenia kręgowego oraz zakażenie krwi i innych narządów.
Base-all	Choroba meningokokowa jest ciężkim zakażeniem bakteryjnym, które może powodować obrzęk mózgu i rdzenia kręgowego oraz zakażenie krwi i innych narządów.
Base-none	Choirób gruczołu krokowego jest poważnym zakażeniem bakteryjnym, które może powodować obrzęk mózgu i rdzenia kręgowego, i zakażenie krwi i innych organów.

Table 2: Example of evaluation translations for different models. Only one seed is shown for Base-60 and Base-20.

5 Results

Unsurprisingly, when looking at evaluation results in Table 1 we observe that fine-tuning on in-domain data improves performance, as *Base-none* shows a worse performance than any of the fine-tuned models, whereas *Base-all* shows a BLEU of 17.402, a 2.466 increase from *Base-none*. When training on smaller random subsets of the data, *Base-20%* and *Base-60%* show less performance increase than when using the whole corpus, as was expected.

The benefit of non-random filtering is especially visible in *LASER-60%*, where performance is higher than *Base-60%* (an increase of 0.177) and even than *Base-all* (an increase of 0.009), meaning that removing 40% of the "worst quality data" yielded marginally increased performance. The case of *MUSE-60%* is also positive since it only meant a decrease of 0.163 compared to *Base-all*, and performance was higher than *Base-60%* by 0.005. The case of *LaBSE-60%* is different, since it decreased the performance by 0.083 compared to *Base-60%*, indicating its use was not beneficial. These results are reflected in the manual verification of test translations, where *LASER-60%* together with *Base-all* produces the most accurate and well-sounding translations (see example in Table 2).

When comparing the smaller sizes of subsets of 20% of the data, none of the filtering methods helped obtain a model that was better than *Base-all*, but we observe that LASER and MUSE outperform *Base-20%*. Our human qualitative assessment of translations showed that *MUSE-20%* struggles with medical terminology, while *LASER-20%* produces consistently high-quality, fluent and natural-sounding text. Here again, the use of LaBSE is not beneficial.

6 Discussion

Altogether, our evaluation of data filtering methods on English to Polish translations in the biomedical domain reveals a performance hierarchy: $\mathbf{LASER_n > MUSE_n > Baseline_n > LaBSE_n}$, with **n** indicating the subset size. LASER proved to be the most effective, enhancing performance even more than the full corpus when using only 60% of the data, reducing computing time by nearly half. Furthermore, using 80% filtered data, LASER and MUSE achieved relatively lower validation BLEU scores (-0.288 and -0.331 respectively) compared to the baseline model (*Base-all*), but significantly better than the unfiltered baseline model (*Base-none*), with scores of 2.178 and 2.135 respectively. However, *MUSE-20%*'s translations appear qualitatively less accurate in specialized medical terminology based on our qualitative inspection. This was partly expected as short sentences full of medical terminology were given a low score by the MUSE filtering.

Surprisingly, LaBSE, expected to perform comparably to LASER, did not meet expectations despite high score correlations (r = 0.81) between the two methods. Differences in scoring specific sentences might explain LASER's superior performance. In summary, our findings validate the efficiency of LASER in reducing dataset size without compromising, and sometimes enhancing, model performance, thus affirmatively answering our research questions *R1* and *R2*. MUSE, while effective, was less consistent in translation quality. LaBSE, despite its expected potential, fell short in this specific setting.

Overall, we recommend LASER as the most effective method for data filtering for the task NMT from English to Polish in the biomedical domain.

Ethical Considerations

In compliance with the ACL Ethics Policy[8], this study upholds the principles of ethical research. Our work involves investigating data filtering techniques within NLP for machine translation, utilizing solely the publicly available UFAL Medical Corpus, and no extra private data is used that could contain private or sensitive information. Additionally, our methodologies are designed to be reproducible, with significant details shared publicly to facilitate verification by the broader research community.

Limitations

A primary limitation of our study is that all models were trained for only 3 epochs, with results reported for the final model state. This approach may not fully capture the potential of the models if they were subjected to more or fewer training epochs. Future research could allow to vary the number of epochs and perhaps use a different stopping criteria to explore how it impacts model performance and efficiency, particularly assessing whether fewer epochs could suffice in achieving optimal results with filtered data, thus optimizing training resources.

Another limitation of this study is that our filtering process relied solely on cosine similarity to evaluate semantic similarity between sentence pairs. Other similarity scores could be investigated for this task. Moreover, no filtering method that assessed the domain specificity of the sentences was used in the study, which led to the inclusion of sentences that are might not be entirely pertinent to the medical domain, e.g. because only containing some medical proper nouns.

Additionally, our analysis did not include any quantitative human evaluation of model predictions by a translation expert. This would involve selecting top predictions from models fine-tuned on both filtered and unfiltered datasets and having an expert assess them without knowledge of their origin to provide unbiased quality evaluations. Furthermore, expanding the sample size of tested models and implementing significance testing would be needed to fully bolster the robustness and generalizability of our results, offering a more detailed understanding of the models' performance across various settings.

[8] https://www.aclweb.org/portal/content/acl-code-ethics

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.

Fred Bane, Celia Soler Uguet, Wiktor Stribiżew, and Anna Zaretskaya. 2022. A comparison of data filtering methods for neural machine translation. In *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 2: Users and Providers Track and Government Track)*, pages 313–325.

Fred Bane and Anna Zaretskaya. 2021. Selecting the best data filtering method for nmt training. In *Proceedings of Machine Translation Summit XVIII: Users and Providers Track*, pages 89–97.

Vishrav Chaudhary, Yuqing Tang, Francisco Guzmán, Holger Schwenk, and Philipp Koehn. 2019. Low-resource corpus filtering using multilingual sentence embeddings. *arXiv preprint arXiv:1906.08885*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Lei Cui, Dongdong Zhang, Shujie Liu, Mu Li, and Ming Zhou. 2013. Bilingual data cleaning for smt using graph-based random walk. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 340–345.

Ondřej Dušek, Jan Hajič, Jaroslava Hlaváčová, Jindřich Libovický, Pavel Pecina, Aleš Tamchyna, and Zdeňka Urešová. 2017. Khresmoi summary translation test data 2.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. *Preprint*, arXiv:2007.01852.

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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. *CoRR*, abs/1804.08771.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167, Vancouver, Canada. Association for Computational Linguistics.

Kaveh Taghipour, Shahram Khadivi, and Jia Xu. 2011. Parallel corpus refinement as an outlier detection algorithm. In *Proceedings of Machine Translation Summit XIII: Papers*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Asahi Ushio, Yi Zhou, and Jose Camacho-Collados. 2023. An efficient multilingual language model compression through vocabulary trimming.

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.

A Appendix

A.1 More training details

Dataset	**Size**	**Training Time**	**Baseline**	**LaBSE**	**MUSE**	**LASER**
Base-none	n/a	n/a	✓			
Base-all	700k	17H:20M	✓			
Base-60%	420k	10H:30M	3 seeds			
Filtered-60%	426k	11H:00M		✓	✓	✓
Base-20%	150k	03H:05M	3 seeds			
Filtered-20%	158k	03H:20M		✓	✓	✓

Table 3: Model specifications and average training times. *Base-none* is the raw pre-trained model without any fine-tuning. *Base-all*, *Base-20%*, and *Base-60%* are models fine-tuned on all, randomly selected 20%, and randomly selected 60% of the training data respectively. Equivalently, the filtered models were fine-tuned on subsets of the data selected as the highest-scored sentence pairs for each filtering method. The reported training times are based on 3 epochs of training on the LUMI supercomputer. Their values are indicative based on representative training runs.