Effect of data filtering techniques on neural machine translation performance for English-Polish biomedical domain

Kamil Kojs
IT University of Copenhagen
kako@itu.dk

Janos Mate
IT University of Copenhagen
janma@itu.dk

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

Mikołaj Barański
IT University of Copenhagen
mikba@itu.dk

Abstract

This paper investigates the impact of data filtering techniques on the efficiency and performance of Neural Machine Translation (NMT) models, focusing on English-Polish translations in the biomedical domain. With the rapid expansion of Large Language Models (LLMs) and the consequent increase in training times, our study emphasizes the need for high-quality training data. We propose selective data filtering as a method to enhance training efficiency without compromising the model's performance. We explore if applying filtering techniques to NMT LLMs can significantly reduce training dataset size while maintaining or even improving translation quality. We use LASER, MUSE, LaBSE, and BERT for filtering the UFAL Medical Corpus to create varying dataset sizes to fine-tune the mBART50 model. We focus on filtering out low-quality data and its impact on training time and model performance, assessed using the SacreBLEU metric. We show that certain filtering methods, notably using LASER embeddings, demonstrate a capacity to reduce dataset size significantly while either maintaining or improving the translation quality.

1 Introduction

Recent advancements in LLMs have resulted in a notable increase in the size of model architectures. State-of-the-art models that are publicly accessible, such as mBart-50, reach parameter sizes upwards of 600 million (Ushio et al., 2023). This escalation in parameter size has consequentially led to increased training times 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 established, that the inclusion of such low-quality data in the training process contributes minimally to the overall quality and performance of LLMs (Koehn et al., 2018). Considering the recent developments in model parameter sizes and the associated increased training times, we propose that it is imperative to carefully inspect the training data. By limiting the training dataset to only high-quality data, we anticipate a substantial reduction in training time without detrimentally impacting model performance.

In this study, we hypothesize that when fine-tuning NMT LLMs for a specific domain and language pair, the application of filtering techniques can significantly reduce the size of the training dataset, without adversely affecting the resulting model's performance. We introduce the following research questions related to our task: *RQ1: "Can we reduce dataset size - and training time - using data filtering without compromising performance when fine-tuning for in-domain biomedical translation for ENG-PL language pair?"*, *RQ2: "Will data filtering reduce the performance drop of using a significantly reduced dataset when compared to random sampling?"*. To investigate this, we look at domain adaptation for biomedical translation from English to Polish using a pre-trained mBART50. We fine-tune it on filtered smaller subsets — using different methods — from a large biomedical corpus and compare performance to that of using a whole corpus or randomly sampled equally sized subsets.[1]

[1]Our results, implemented methods, and other code can be found in the project's GitHub repository: jorgedelpozolerida/KDS_AdvancedNLP_FinalProject

2 Related Work

Neural Machine Translation models require data of large quantity and high quality to be adapted to new domains (Koehn et al., 2018). To increase the quality of large web-scraped corpuses, automated filtering methods have been studied by the NMT field. Approaches including outlier detection (Taghipour et al., 2011), discriminator models trained on synthetically noisy translations (Xu and Koehn, 2017) graph-based unsupervised models (Cui et al., 2013), and more recently LLM-based classifiers or scorers (Açarçiçek et al., 2020). With the advent of LLMs, encoders specifically designed to create language-agnostic embeddings such as, LASER / LaBSE / MUSE and more, allowed researchers to directly score bi-lingual sentence similarity for dataset filtering. Recent research indicates this approach can be competitive with more complex classifier-based models (Chaudhary et al., 2019).

Bane and Zaretskaya (2021), explored the use of filtering based on the quality of sentence pairs in English-Japanese and English-German. The models used for filtering included a pre-trained Marian-scorer and LASER, MUSE, and XLM-R embedding models. The authors found that all models, which limited the dataset size to 54%-73% of the original size, achieved comparable translation BLEU scores. However, the majority of filtering methods couldn't outperform a random dataset downsizing. Only the Marian-based filtering consistently achieved higher BLEU scores than random. Another finding was the wide variance of performance between the two language pairs for the MUSE approach, which performed best for German while significantly reducing model performance for the Japanese translation. The findings of Bane and Zaretskaya (2021), motivated our study design while indicating that results are likely to be nontransferable to different language pairs or topic domains.

Bane et al. (2022) conducted further experimentation on the strengths and weaknesses of specific filtering methods. The authors collected and generated a dataset of translation pairs with 10 different types of noise or errors (artificially induced where needed). The authors found a custom-trained Marian-Scorer achieved the highest cleaning performance, while embedding-based methods such as XLM-R, MUSE, and LASER performed worse but still at a good level. The embedding-based methods were specifically good at finding sentence pairs with number mismatches and missing segments, while also catching a large share of spelling and word order permutations in texts. Research of Bane et al. (2022), indicates that embedding-based methods successfully tackle noise in translated segments without the need for additional costly calibration such as the Marian-Scorer.

3 Data

The selected model was fine-tuned on the UFAL Medical Corpus (ufa). Given the focus of our research question, only the Polish-English sentence pairs were used for the project. The dataset comprised 1,116,773 sentence pairs in English and Polish. The sentences come from the documents of the European Centre for Disease Prevention and Control (2.3k sentences), documents of the European Medicines Agency (1,111.6k sentences), and subtitles (3.0k). While the medical corpus had already undergone preprocessing to remove duplicates, further refinements were made to enhance overall quality.

We implemented basic data cleaning: excluding untranslated sentences (identical Polish and English counterparts), removing sentence pairs shorter than 15 or longer than 200 characters in either language (due to GPU-memory constraints), and eliminating sentences in other languages, identified by containing unique characters from Russian, French, Greek, Bulgarian, and Romanian alphabets. The inclusion of these cleaning measures allows us to test the effect of filtering on noise that cannot be easily picked up by such basic rules. These adjustments yielded a final training dataset of 711,720, from which 700k were randomly sampled.

As the test set, a dataset from Khresmoi (Dušek et al., 2017) was employed. This publicly available dataset consists of 1,500 high-quality Polish-English sentence pairs in the medical domain, and the quality was found to be good, thus it did not require any specific data-cleaning procedures. This dataset was chosen as it is unrelated to the training data while being a clean in-domain dataset that allows us to evaluate our model in an unbiased and more representative manner.

4 Methodology

We employed four widely used multilingual embedding models — LASER, MUSE, LaBSE, and BERT — to filter and subset the medical-domain

corpus. Each subset was used to separately fine-tune a pre-trained mBART50 model, and the evaluation was conducted on an independent dataset.

4.1 Experimental setup

The sentence pairs of the in-domain training data, were scored by the four methods using cosine similarity. That is, for each sentence an embedding was generated for its English and Polish language versions as an average of the token embeddings in either sentence. Next, cosine similarity was calculated between these fixed-size vector pairs. The cosine similarity metric is bounded between 0 and 1, where 1 represents full similarity. Based on these scores we kept 20% (∼150k pairs) and 60% (∼ 420k pairs) of the highest scoring sentences for each method. These eight (two sizes per method) filtered datasets were used to create fine-tuned mBART50 models.

To allow us to compare the effect of filtering, a baseline model was trained on the full unfiltered dataset (*Base-all*)[2] and further models were trained on three randomly selected 20% and 60% subsets of the data (*Base-20%* and *Base-60%*). That is, to reduce the impact of random variance, three models were trained for every subset size sampled with different seeds.

As presented in Table 1, in total we tested 16 models, out of which 15 were fine-tuned by various cuts of our in-domain data. As well as, the untouched pre-trained model, which was our non fine-tuned baseline (*Base-none*). Regardless of the size of the subset, a stratified split of 80% and 20% was made, preserving original proportions for different sources of data in the corpus.

To test the performance of the fine-tuned models we evaluated their performance on an independent validation dataset - the Khresmoi Summary Translation dataset (Dušek et al., 2017). In line with relevant literature (see Sec. 2), we used the BLEU metric (Papineni et al., 2002) implemented via the SacreBLEU method (Post, 2018). The evaluation is conducted in the following steps - (1) one of the tested mBART50 models is used to translate the English sentences into Polish, (2) the SacreBLEU score is calculated between the predicted Polish sentences and the ground truth. Finally, we also report the training time, to indicate the reduced computational intensity of the fine-tuning on reduced dataset sizes.

[2]Note that it is fine-tuned on the pre-processed and pre-cleaned medical dataset (see Sec. 3).

All computations were done in LUMI supercomputer, which uses Slurm as a job scheduler. More specifically, we used the *standard-g* partition with AMD MI250x GPUs. Each model took a different time to train due to varying train set sizes, and combined with evaluation computing time, we utilized a total of 646 GPU hours. For training, we used a 16-bit precision, a batch size of 15 for training and 20 for evaluation, a linear learning rate scheduler using the first 100 steps, and an AdamW optimizer. All models were trained for 3 epochs with consistently declining training and evaluation loss, hence the last checkpoint of each model was used.

4.2 NMT model - mBART50

For machine translation, mBART50, developed by Facebook AI, was used (Tang et al., 2020). The model was chosen because it is one of the state-of-the-art models that can be employed for various tasks and a wide variety of languages, which makes it ideal for the comparability and reproducibility of our results. The model is a multilingual Sequence-to-Sequence model pre-trained using the Multilingual Denoising Pretraining (Figure 1). During this pre-training, the source documents are noised in two ways. Firstly, it masks 35% of the words in each instance by randomly sampling a span length according to a Poisson distribution ($\lambda = 3.5$). Secondly, it uses sentence permutation to change the original sentences' order. After that, the model has to reconstruct the original text. The decoder input is the original text shifted by one position. It uses an initial token called the language id (<LID>) to predict the sentence.

For our fine-tuning, the previously introduced UFAL Medical Corpus (ufa) was used, including its Polish and English translations. To tokenize the text, the MBart50Tokenizer (MBa) class from the Hugging Face transformers library was employed, using pre-trained from "facebook/mbart-large-50-one-to-many-mmt" with source and target language codes "en_XX" and "pl_PL" respectively.

4.3 Evaluation metric

The evaluation was based on the SacreBLEU implementation of the BLEU metric of Polish sentences translated from English. In SacreBLEU, the transition quality is measured on a scale from 0 to 100 where higher scores mean better translations. It calculates the number of overlapping n-grams between the generated translation and the source translations. SacreBLEU, compared to BLEU, en-

Dataset	Size	Training time	Baseline	LaBSE	MUSE	LASER	BERT
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 1: Model specifications and average training times. *Base-none* is the raw pre-trained model without any fine-tuning. *Base-all*, *Base-20%*, *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.

sures consistent reporting and enables the comparison of scores across research works (Post, 2018).

4.4 Filtering methods

4.4.1 BERT

The BERT (Bidirectional Encoder Representations from Transformers) model is a popular model used for a wide variety of tasks. It was developed by Google and was published in the paper "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" (Devlin et al., 2019). As it employs bidirectional training, it achieves a deeper understanding of language context compared to single-direction language models. For the filtering tasks, we use the bert-base-multilingual-cased version provided by Hugging Face (ber), which was pre-trained on 104 languages (Devlin et al., 2019).

We included this model in our study to provide a comparison between the performance of filtering based on the raw multilingual-embedding approach of BERT versus the language-agnostic approach of the other models.

4.4.2 LaBSE

LaBSE - Language-agnostic BERT Sentence Embeddings, is a method developed by a Google research team for creating BERT-based cross-lingual sentence embeddings for over 109 languages (Feng et al., 2022). We implemented the version of the model from Hugging Face (LaB). This model was created to address the shortcomings of the original BERT model's multilingual embeddings, which are not inherently language-agnostic. The model was trained as a dual-encoder, which used a pre-trained BERT model to generate embeddings for two translations of the same sentence. The model's training loss was set as the difference between the two embeddings, leading towards the convergence of the embedding models to create a cross-lingual embedding space.

The model's authors use a variety of methods state-of-the-art training methods including masked language modeling, translation language modeling, dual encoder translation ranking, and the use of the additive margin softmax (Feng et al., 2022).

This model was included in our research due to the promising results indicated in its paper, as well as the fact that it has not been studied in previous data filtering literature.

4.4.3 LASER

The LASER (Language-Agnostic SEntence Representations) model (Schwenk and Douze, 2017) (LAS), developed by Facebook AI Research, is an advanced method for generating language-independent sentence embeddings. The core of the LASER model is a BiLSTM (Bidirectional Long Short-Term Memory) neural network architecture, which is trained on a large multilingual corpus. This BiLSTM encoder reads input sentences and encodes them into a fixed-size vector, capturing the deep semantic content of the sentence.

A key feature of LASER is its language-agnostic design. LASER is trained to handle input from over 90 languages. This is achieved by training the model on parallel corpora, where the same sentence in multiple languages is used to teach the model to generate similar embeddings for semantically equivalent sentences, regardless of the language. Furthermore, the model includes a Byte Pair Encoding (BPE) based tokenizer, allowing the model to handle diverse languages and aids in generalizing across languages by breaking down words into common subword units shared across different languages.

Model	Filter	Avg. BLEU	Min	Max	Δ Base-all	Δ Base 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
Filtered-20%	BERT	16.273	-	-	-1.129	-0.528

Table 2: Experimental Results. The BLEU score is based on the translations of sentences from the independent Khresmoi validation dataset (Dušek et al., 2017). I.e., overlap between fine-tuned mBART-50 English to Polish translations and the Polish ground truth. Average scores are reported for the Base-60% and Base-20% models, which were trained based on three random subsets of the data. For all other methods, only a single BLEU score was calculated and reported.

This model was chosen for our analysis given its proven performance in data filtering in literature from other domains and languages (Chaudhary et al., 2019).

4.4.4 MUSE

MUSE - Multilingual Unsupervised and Supervised Embeddings, is a model by Meta's team, which facilitates the development and evaluation of cross-lingual word embeddings (Conneau et al., 2017). The model incorporates unsupervised alignment of monolingual word embeddings involves three steps: adversarial training to create a linear mapping from the source to the target embedding space, extracting a synthetic dictionary from the shared space, and refining the alignment using the Procrustes solution. This enables alignment without the need for cross-lingual annotated data or parallel corpora.

The original model was no longer available; however, we applied its pre-made embeddings dictionary (Meta). The model was included in our study as it has performed well in other language pairs and domains in previous data filtering literature (Bane and Zaretskaya, 2021).

5 Results

As expected based on related research (see Sec. 2), our results in Table 2 indicate that fine-tuning on in-domain data improves performance when testing on an independent in-domain dataset. The pre-trained and not fine-tuned *Base-none* shows a BLEU score of 14.936, worse than any of the fine-tuned models. The opposite is represented by *Base-all*, which has been trained on the whole corpus and shows a BLEU of 17.402, an increase of 2.466 compared to *Base-none*. When training or random subsets of the data, *Base-20%* and *Base-60%* show average performance increases of 1.931 and 2.298 respectively, showing less performance increase than when using the whole corpus, as would be 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[3]. 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 Appendix A.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*. We still see LASER and MUSE methods showing beneficial results when compared to *Base-20%*, specifically showing 0.246 and 0.203 increases respectively. Our manual verification

[3]Due to a technical error we were unable to run the 60% version of the dataset filtered by BERT.

of *MUSE-20%*'s translations indicates it struggles with medical terminology, while *LASER-20%* produces consistently high-quality text. Again, LaBSE does not seem to help but rather does the opposite and the same holds for the BERT method.

6 Discussion

Altogether, the relation seems to hold that **LASER_n > MUSE_n > Baseline_n > LaBSE_n > BERT_n** for **n** being the size of the subset. **LASER is the best-performing filtering method** since its subset of 60% shows even better performance than using the whole corpus. The latter is very significant, since by reducing your corpus by 40%, you can obtain even better performance and reduce your computing time by almost half. Additionally, while further halving training time, models trained on data filtered down by 80% using LASER and MUSE achieved acceptably lower validation BLEU scores (-0.288 and -0.331 respectively) compared to *Base-all*, and significantly higher scores compared to *Base-none* (2.178 and 2.135 respectively). However, the difference between MUSE & LASER is larger here, as *MUSE-20%*'s translations appear less accurate in specialized medical terminology (see Appendix A.2). This was partly expected as short sentences full of medical terminology were given a low score by the MUSE filtering. This indicates the power of *LASER-20%* in resource-limited scenarios.

Results for LaBSE and BERT methods are underwhelming as they result in lower validation scores than worst worst-performing randomly selected dataset. This indicates that the filtering methods give high scores to sentence pairs with low training value. BERT's low performance was expected due to it being a model with language-specific embeddings, i.e., its embeddings were not explicitly trained to be similar across languages. The result for LaBSE is surprising, as we expected all language-agnostic methods to outperform or be equally good as the random sampling. Moreover, we have found sentence scores based on LaBSE to be highly correlated with LASER scores (r = 0.81), due to which we initially hypothesized that LaBSE and LASER will perform similarly. It appears; however, that the sentences where the two models provided differing scores gave LASER an edge. The weakness of LaBSE may be particularly rooted in its use in our language pair or domain.

A limitation of our research is that we have evaluated all models after being trained for 3 epochs, and reported results for the final model. Further research could explore the performance of the models with more or less epochs of training. This could further indicate how training resources could be optimized via filtering.

It must be noted that filtering was done using only cosine similarity, which is meant to represent semantic similarity between sentences. A possible extension of our approach could be to use some filtering methods that do not only rely on quality but also on how much in-domain the sentences are. We observed in our corpus that some sentences were not so specific to the medical domain, containing just a few medical terms and the rest being pretty general language. Filtering based on how specific to the domain the language is, could further increase the performances of fine-tuned models.

Another possible line for future work to improve our analysis would be to manually evaluate the performance with input from a translation expert. For instance, we could select top k predictions of the models fine-tuned on the filtered dataset and from the model fine-tuned on an unfiltered dataset. Then a translation expert evaluates the quality being blinded on the prediction model origin. Also, a larger sample of tested models and significance testing could improve further the robustness of our results.

7 Concluding remarks

In our project, we analyzed four multilingual NLP models for data filtering, on English to Polish translations in the biomedical domain. After fine-tuning mBART50 on differently sized filtered subsets of a bigger in-domain corpus, we observed that the LASER filtering method can yield similar or better performance with significantly reduced training times as compared to using the whole corpus. Furthermore, this method allowed us to reduce the performance drop caused when using even smaller subsets. Thus for LASER, we can answer our initial research questions *R1* and *R2* with confirmation. MUSE was also a strong performer and confirms *R2*; however, it provided less fluent translations. Lastly, LaBSE and BERT showed significantly weaker performance compared to the other methods and even randomly decreased datasets. Overall, we indicate the benefits of data filtering for this specific language pair and domain and recommend using the LASER method.

References

Bert multilingual base model (cased). [Accessed 10-12-2023].

Github - facebookresearch/laser: Language-agnostic sentence representations — github.com. https://github.com/facebookresearch/LASER. [Accessed 10-12-2023].

Labse. Accessed [2023.10.21].

Mbart50tokenizer. Accessed [2023.10.21].

Ufal medical corpus. Accessed [2023.10.21].

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

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.

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.

Meta. Muse. [Accessed 16-12-2023].

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 of the Association for Computational Linguistics*, pages 311–318.

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

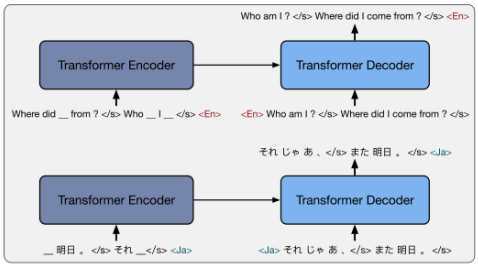

Figure 1: Multilingual Denoising Pre-training (mBART)

A.1 Group contributions

The work was fairly distributed among team members with no significant imbalance in workload. All team members contributed to writing the report.

- Kamil Kojs: LASER, data cleaning
- Janos Mate: MUSE, BERT, data cleaning
- Mikołaj Baranski: LaBSE, data cleaning
- Jorge del Pozo Lerida: mBart50 training, LUMI setup

A.2 Examples of manual translation evaluation

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.
Prediction LaBSE-60	Meningooka jest ciężkim zakażeniem bakteryjnym, które może powodować obrzęk mózgu i rdzenia kręgowego oraz zakażenie krwi i innych narządów.
Prediction LASER-60	Choroba meningokokowa jest ciężkim zakażeniem bakteryjnym, które może powodować obrzęk mózgu i rdzenia kręgowego oraz zakażenie krwi i innych narządów.
Prediction MUSE-60	Choroba meningokokowa jest ciężkim zakażeniem bakteryjnym, które może powodować obrzęk mózgu i rdzenia kręgowego oraz zakażenie krwi i innych narządów.
Prediction Base-60_1	Choroba meningokokowa jest ciężkim zakażeniem bakteryjnym, które może powodować obrzęk mózgu i rdzenia kręgowego oraz zakażenie krwi i innych narządów.
Prediction Base-60_2	Choroba meningokokowa jest ciężkim zakażeniem bakteryjnym, które może powodować obrzęk mózgu i rdzenia kręgowego oraz zakażenie krwi i innych narządów.
Prediction Base-60_3	Choroba meningokokowa jest ciężkim zakażeniem bakteryjnym, które może powodować obrzęk mózgu i rdzenia kręgowego oraz zakażenie krwi i innych narządów.
Prediction LaBSE-20	Meningokoczka jest poważnym zakażeniem bakteryjnym, które może powodować obrzęk mózgu i rdzenia kręgowego oraz zakażenie krwi i innych narządów.
Prediction LASER-20	Choroba meningokokowa jest poważnym zakażeniem bakteryjnym, które może powodować obrzęk mózgu i rdzenia kręgowego oraz zakażenie krwi i innych narządów.
Prediction MUSE-20	Choroba gruczołu krokowego jest poważnym zakażeniem bakteryjnym, które może powodować obrzęk mózgu i rdzenia kręgowego oraz zakażenie krwi i innych narządów.
Prediction BERT-20	Meningookoka jest poważnym zakażeniem bakteryjnym, które może powodować obrzęk mózgu i rdzenia kręgowego oraz zakażenie krwi i innych narządów.
Prediction Base-20_1	Choroba meningokokowa jest ciężkim zakażeniem bakteryjnym, które może powodować obrzęk mózgu i rdzenia kręgowego oraz zakażenie krwi i innych narządów.
Prediction Base-20_2	Choroba meningokokowa jest ciężkim zakażeniem bakteryjnym, które może powodować obrzęk mózgu i rdzenia kręgowego oraz zakażenie krwi i innych narządów.
Prediction Base-20_3	Choroba meningokokowa jest ciężkim zakażeniem bakteryjnym, które może powodować obrzęk mózgu i rdzenia kręgowego oraz zakażenie krwi i innych narządów.
Prediction Base-all	Choroba meningokokowa jest ciężkim zakażeniem bakteryjnym, które może powodować obrzęk mózgu i rdzenia kręgowego oraz zakażenie krwi i innych narządów.
Prediction Base-none	Chorób gruczołu krokowego jest poważnym zakażeniem bakteryjnym, które może powodować obrzęk mózgu i rdzenia kręgowego, i zakażenie krwi i innych organów.

Table 3: Example of evaluation translations. In this specific set of translations we can see models filtered by LaBSE, BERT, and not fine-tuned Base-none, all fail to capture the correct medical term for the disease. Additionally, *MUSE-20* also acts in the same way as the un-fine-tuned model. The fine-tuned baseline models all show good performance, with the best translation provided by LASER.

Type	Sentence
English	The chest tube has drained 300 ml.
Ground Truth	Przez dren klatki piersiowej odprowadzono 300 ml.
Prediction LaBSE-60	Tuba klatki piersiowej wysuszyła 300 ml.
Prediction LASER-60	Tuba klatki piersiowej wyciekła 300 ml.
Prediction MUSE-60	Tuba klatki piersiowej odwodniła 300 ml.
Prediction Base-60_1	Tuba do klatki piersiowej zawiera 300 ml odtworzonego płynu.
Prediction Base-60_2	Tuba do klatki piersiowej zawiera 300 ml odkażonego płynu.
Prediction Base-60_3	Wypłukany wkład w klatce piersiowej wynosi 300 ml.
Prediction LaBSE-20	Tuby piersiowe wysuszyły 300 ml.
Prediction LASER-20	Tuba piersiowa wysuszyła 300 ml.
Prediction MUSE-20	Tłok w klatce piersiowej odwodnił 300 ml.
Prediction BERT-20	Strumień w klatce piersiowej nasączył 300 ml płynu.
Prediction Base-20_1	Tuba w klatce piersiowej wyodrębniła 300 ml.
Prediction Base-20_2	Tuba w klatce piersiowej wysiękła 300 ml.
Prediction Base-20_3	Tuba w klatce piersiowej pokryta jest odwodnieniem w ilości 300 ml.
Prediction Base-all	Tuba klatki piersiowej posiada 300 ml odprowadzanego płynu.
Prediction Base-none	Trubka piersiowa wyparła 300 ml.

Table 4: Example of evaluation translations. This is an example of a specifically difficult translation for the systems. Interestingly, all translations have at least some level of grammatical or meaning inconsistency. The *Base-all* and *LASER-60* model achieves the closest translation in terms of used vocabulary; however, its sentence changes the meaning of the English phrase. As in the table above, *MUSE-20* has severely reduce performance compared to *MUSE-60*.