Multilingual Data Filtering using Synthetic Data from Large Language Models

Anonymous ACL submission

Abstract

Filtering data, especially when the data has been scraped from the Internet, has long been known to improve model performance. Recently, it has been shown that an effective filter can be created by using large language models (LLMs) to create synthetic labels, which are then used to train a smaller neural model. However, this approach has mainly been tested in English. Our paper extends this approach to languages beyond English, including languages not officially supported by LLMs. We validate our results on the downstream task of NMT and demonstrate that our approach is effective at both filtering parallel text for translation quality and filtering for domain specificity. Additionally, we find that using a classification objective is more performant and robust than a regression objective at low data thresholds when training our filtering models.

1 Introduction

Increasing model scale and larger pre-training datasets have fueled recent advances in the world of LLMs. Beyond scale, other pre-training data characteristics also significantly impact downstream tasks, such as de-duplication and removing low-quality examples (Touvron et al., 2023; Young et al., 2024). An interesting approach that has recently been proposed is training filtering models on synthetic labels, which are generated by prompting LLMs (Grattafiori et al., 2024; Abdin et al., 2024; Penedo et al., 2024a; Lozhkov et al., 2024). Such filtering models can be efficiently run on very large corpora, such as pre-training data, to select the most appropriate examples for training. Due to the flexibility of designing prompts, this pipeline is especially appealing, enabling data to be filtered on criteria beyond quality without requiring labelled data and thereby tailoring the selected pre-training data to the eventual downstream task.

The FineWeb project by Penedo et al. (2024a) observed that by filtering pre-training data towards educational content, they were able to not only obtain a 4% improvement on the MMLU benchmark (Hendrycks et al., 2021) but also to converge quicker when compared to non-filtered baseline. The educational content filter was a classifier based on synthetic LLM-labeled data, and the approach was validated via training a 1.71B model on 350 billion tokens; however, the study was centred on enhancing performance exclusively in English. Although the experiment validates the methodology's effectiveness for English downstream tasks, the technique could also be beneficial for other languages, where data quality is even more crucial given the overall scarcity of resources. This work attempts to unravel one unexplored axis of synthetic filtering: the method's efficacy beyond English. From here on, we refer to this approach as MDFS (Multilingual Data Filtering using Synthetic Data).

We investigate and evaluate MDFS via the Neural Machine Translation (NMT) task. NMT is an excellent downstream task for a series of reasons. First, it has an established history of a range of data filtering WMT shared tasks (Conference on Machine Translation, Koehn et al., 2018, 2019, 2020). Secondly, NMT models are reasonably cheap to train compared to LLMs, allowing us to run a suite of experiments investigating different setups for filtering multilingual data using MDFS, which would be prohibitively expensive if done with LLMs. Additionally, NMT has a history of neural QE (Quality Estimation) metrics such as COMET-KIWI or BLEURT (Rei et al., 2022; Sellam et al., 2020), which are effective at filtering training data (Peter et al., 2023). Hence, we can employ such QE models trained on hand-made, high-quality annotations as a robust filtering baseline. We use MDFS as an instance of a synthetic, LLM-labeled quality estimator and validate the approach under different NMT setups that range from general translation tasks to domain adaptation in various languages.

As we initially stated, the most significant appeal of MDFS is the flexibility to filter based on any criteria simply by adjusting the prompt. We, therefore, run two sets of experiments to establish the efficacy of MDFS for non-English languages. Firstly, we train En→De and En→Ar NMT systems filtered only for translation quality to analyse the MDFS pipeline for non-English languages when compared to QE filtering using models trained on human annotations. Secondly, we train En→Ar and En→Ro NMT systems which are trained with data filtered for medical content.

We summarise our contributions as follows:

- We explore LLM-based data filtering techniques for multiple languages and validate them on machine translation - showing that they work for both filtering on the source and target sides.
- We show that LLM-based filtering is effective beyond pure quality filtering by allowing us to filter for domain. We show that LLM filtering has benefits over baseline keyword filtering.
- We explore filtering the LLM scores as classes or regression models. We find that classification is superior as it is more robust for non-English languages at very small cutoff thresholds.

2 Related Work

Penedo et al. (2024a) introduce FineWeb-Edu, and demonstrate a 4% increase on MMLU and a 11% increase on the ARC benchmark (Clark et al., 2018). Similar approaches were also used when training the Llama and Phi family models (Grattafiori et al., 2024; Abdin et al., 2024). Our work also experiments with filtering models trained from synthetic labels. However, unlike these works, we investigate filtering in non-English contexts and experiment with different approaches for the filtering models.

Since the advent of NMT, it has been known that low amounts of noise in the training data can lead to erroneous translations (Koehn et al., 2018). As such, NMT has a history of data filtering, especially for scraped corpora such as ParaCrawl (Bañón et al., 2020). A series of cleaning tasks for parallel data (Koehn et al., 2018, 2019, 2020) resulted in the development of several cleaning models for NMT, including LASER (Schwenk and Douze, 2017) embedding based models and BICLEANER (Sánchez-Cartagena et al., 2018; Ramírez-Sánchez et al., 2020). Later Zaragoza-Bernabeu et al. (2022) released an updated BICLEANER that incorporates a neural model. BICLEANER is used to filter public corpora such as ParaCrawl. Compared to our work, these models all focus on removing training examples that are not mutual translations of each other rather than picking the best translations and can only filter for quality.

Peter et al. (2023) compare filtering training data using BICLEANER (Zaragoza-Bernabeu et al., 2022) to filtering using COMET-KIWI, a QE model for NMT. The authors filter 50% the WMT 23 (Kocmi et al., 2023) training data for three language pairs and show that filtering with COMET-KIWI leads to improved COMET scores. They highlight that filtering with QE metrics discriminates in a more fine-grained manner. Our approach can also be used to filter for criteria beyond quality and can also be used to filter only monolingual data. Additionally, we experimented with filtering at different thresholds.

3 Filtering Pipeline

We begin by describing the outline of the MDFS pipeline in the context of both the translation quality and medical domain NMT experiments before discussing each pipeline stage in more detail.

3.1 MDFS

We adopt the pipeline introduced by Penedo et al. (2024a), which consists of three stages. First, we use an LLM to score approximately 500,000 sentences based on the task criteria. Similarly to Penedo et al. (2024a), we follow Yuan et al. (2024) and use an additive prompt. The filtering criteria are divided into a 5-point scale, and the LLM is instructed to determine a score on a point-by-point basis; the total score is the sum of the points awarded. The translation quality and medical domain task prompts are given in Appendix A. We use `Llama-3.1-70B-Instruct`[1] to generate the synthetic labels. As the primary benefit of this approach is using out-of-the-box LLMs to create synthetic training data, we avoid using specifically multilingual LLMs such as Tower (Alves et al., 2024), which are trained on human-labelled DA (Direct Assessment) and MQM (Multidimensional Quality

[1] `https://huggingface.co/meta-llama/Meta-Llama-3.1-70B-Instruct`

Metrics) data.

The next step is training our MDFS filtering models using the synthetic labels generated from the LLM. The lightweight models are based on pre-trained encoder models, specifically XLMR (Conneau et al., 2020). Our experiments explore training the models with either linear regression or classification as an objective. During training, we finetune all model parameters with a classification or regression head architecture similar to the COMET models (Rei et al., 2020).

Finally, we filter our NMT training data with the filtering models trained in stage two. In order to evaluate the performance, we threshold our training set according to the number of sentences used to train the NMT models. For each threshold, we then select the best sentences according to the scores assigned by the filtering model. We use the continuous scores for models trained with a linear objective function to select the best sentences. As the classification objective function only gives us a categorical ranking, we select categories of sentences until we exceed the threshold; when we exceed the threshold, we use a random sample of the current category to make up the training data.

3.2 Translation Quality

These experiments aim to understand the best pipeline for filtering multilingual data. Using parallel data, we train the MDFS filtering models by concatenating the source and target sentences. Therefore, the model can access both English and non-English sentences when scoring an example. We select one high-resource language pair, En-De, which `Llama-3.1-70B-Instruct` fully supports and is also part of the human-labeled DA data used to train COMET-KIWI. En-Ar is not officially supported by `Llama-3.1-70B-Instruct` or in the COMET-KIWI training and in a non-Latin script.

3.3 Medical Domain

Unlike the translation quality experiments, we filter only the source or the target side, the reasons for which are twofold. Firstly, this makes the setup more comparable to filtering LLM training data for task-specific monolingual data. Secondly, it allows us to evaluate the differences observed when filtering on the English and the non-English side. We select En→Ar and En→Ro as both target languages are not supported by `Llama-31-70B-Instruct` and have available medical data to evaluate the NMT models. As we can directly compare it to filtering the English sentences, we did not include another officially supported language.

4 MDFS Models

All of our experiments use NMT as a downstream task; however, the specific setup varies for the translation quality and medical domain experiments.

4.1 Training Data

All experiments start with a set of training data of which a small portion is removed to train an MDFS filtering model for the given downstream task. For the En→De translation quality experiments, we use ParaCrawl data used in the WMT23 campaign as training data (Kocmi et al., 2023; Esplà et al., 2019). For the En→Ar translation quality experiments, we use the CCMatrix dataset (Schwenk et al., 2021).

For En→Ar we use CCMatrix and ELRC Wikipedia-Health[2] corpus comprising of 15,130 sentences. For En→Ro, we combine CCMatrix, ParaCrawl and 783,742 sentences from ELRC-EMEA.[3] All the training data was downloaded from OPUS (Tiedemann, 2012).

4.2 LLM Labeling

In order to train our filtering models, we label a small subset of our training datasets with `Llama-3.1-70B-Instruct`. We randomly remove a small amount of the parallel training data for the translation quality experiments, which the LLM then labels. Randomly selecting from the entire training data for the domain filtering task is problematic as the medical sentences constitute only a small proportion of the training data. Hence, the sampled data would be significantly unbalanced. For En→Ar, we realistically address this by filtering the datasets using a curated list of 30 English medical keywords (Appendix B). We then sample 50%.

4.3 Filtering Models

Having obtained synthetic labels for 400,000-500,000 sentences, we use 1000 sentences as a validation set and 10,000 sentences as a test set for each experiment, with the rest being used to train the MDFS models. We also removed all sentences for which the LLM either did not generate a score, or the score was in the wrong format. Based on higher validation F1-scores for the translation quality task, we run all further experiments with full

[2] https://elrc-share.eu/elrc-wikipedia-health
[3] https://elrc-share.eu/elrc-emea

pre-training and first expand the hidden dimension in the classification head similar to the COMET models. We report results for both a linear regression and a classification objective function. As the test set is created with labels from the LLM, we are only evaluating how well our filtering models can replicate the scores generated by the LLM; in the case of linear regression, we follow the Fine-Web Edu authors (Penedo et al., 2024a) and truncate and round the continuous scores to obtain ordinal scores. We train for 20 epochs and select the best model using the macro-averaged F1-score on the validation set. We base our hyper-parameter selection on the COMET-KIWI paper (Rei et al., 2022). All models are trained with data mixed from both scoring directions, resulting in bidirectional scoring models.

4.4 Filtering Approaches

We compare the following approaches to filtering the NMT training set.

RANDOM: Our first baseline randomly selects sentences from the training data for filtering.

COMET-KIWI: Our second baseline uses COMET-KIWI scores to filter the data. COMET-KIWI is a QE model trained on human direct assessment data, which has been shown to improve NMT metrics when used for filtering training data (Peter et al., 2023). Additionally, COMET-KIWI is a compelling baseline because it uses the same pre-trained model as our MDFS models, XLMR. We only use this baseline for the translation quality experiments where we filter on bilingual text.

KEYWORDS: For the medical domain experiments, our second baseline filters the English side of the training corpus with a curated list of 30 medical keywords. Keywords are a quick and simple method for filtering domain-specific data but could be less effective in morphologically richer languages than English.

MDFS-LINEAR: Linear refers to our filtering model trained on the synthetic LLM labels by finetuning all parameters and training with a linear regression objective function.

MDFS-CLASSIFICATION: Classification refers to our filtering model trained on the synthetic LLM labels by finetuning all parameters and training with a classification objective function.

4.5 MDFS Results

Table 1 and 2 give the F1-scores evaluated on the LLM labelled test set for the MDFS models. Results are given when thresholding at scores of 3, 4 and 5, where the linear scores are clipped and rounded to obtain ordinal values. Hence, F1-scores show how well the MDFS models can replicate the labels generated by LLM.

Model	MDFS-LINEAR			MDFS-CLASS		
Thresh	3	4	5	3	4	5
En→De	**0.908**	0.777	0.640	**0.908**	**0.782**	**0.644**
De→En	**0.920**	**0.673**	0.381	**0.890**	0.670	**0.430**
En→Ar	**0.920**	**0.757**	**0.398**	0.918	0.745	0.385
Ar→En	**0.934**	**0.804**	0.570	0.929	0.791	**0.571**

Table 1: F1-scores for MDFS-LINEAR and MDFS-CLASS for the translation quality experiments. Bold numbers indicate the higher F1-score when comparing MDFS-LINEAR and MDFS-CLASS for the same threshold and scoring direction.

When thresholding at 3, the lowest F1-score observed for either experiment is 0.890, for the De→En translation quality classification model. Demonstrating that in our approach, the MDFS models can reproduce the distribution of scores generated by `Llama-3.1-70B-Instruct` to a sufficient level to differentiate between "good" and "bad" examples. We take this as evidence that MDFS models are able to filter for the same criteria as the `Llama-3.1-70B-Instruct` in non-English via transfer learning using synthetic labels. Additionally, we observe that, even though filtering for the quality of translation using parallel data results in lower F1-scores when compared to the monolingual domain filtering results, our method is robust across different filtering requirements and inputs. The lowest F1-scores in Table 1, (0.381 for De→En and 0.385 for En→Ar) occur at a threshold of 5, indicating that whilst MDFS models effectively distinguish between high and low scores, they struggle to rank the best examples accurately.

Table 2 demonstrates that filtering the non-English side of the translation results in comparable F1-scores to filtering the English sentences. When thresholding at 3, the F1-scores for both Arabic and Romanian are higher, with the former being 0.038 higher than the English MDFS-LINEAR model. However, both Arabic and Romanian fall short of filtering the English when selecting the

Model	MDFS-LINEAR			MDFS-CLASS		
Thresh	3	4	5	3	4	5
Ar	**0.950**	**0.854**	0.658	0.947	0.853	**0.670**
En	0.912	0.853	**0.744**	**0.917**	**0.870**	0.734
Ro	0.974	0.948	0.754	**0.976**	**0.952**	**0.779**
En	**0.964**	**0.938**	0.812	**0.964**	**0.938**	**0.826**

Table 2: F1-scores for MDFS-LINEAR and MDFS-CLASS for the medical domain experiments. Bold numbers indicate the higher F1-score when comparing MDFS-LINEAR and MDFS-CLASS for the same threshold and scoring direction.

highest quality sentences, suggesting that there is an element of degradation when trying to identify the best sentences in a non-English language.

4.6 Domain Filtering Analysis

We focus on the medical domain experiments to analyse the properties of the filtered datasets as they enable a more direct comparison between English and non-English languages. Table 3 shows the percentage of medical sentences in the NMT training data, where we take all sentences with a score greater or equal to 3 as having a degree of medical content.

	Medical Percentage	
	Arabic	**Romanian**
KEYWORD	4.35	4.52
MDFS-CLASS (En)	4.54	8.32
MDFS-CLASS	7.12	10.54
MDFS-LINEAR (En)*	4.68	8.75
MDFS-LINEAR*	7.56	11.04

Table 3: percentage of medical sentences in the training data. Medical sentences for MDFS models are taken as those with a score greater than 3.*LINEAR scores are clipped and rounded.

For En→Ar, we obtain a similar number of medical sentences when filtering on the English side and when compared to the KEYWORD baseline. In contrast, for En→Ro, filtering in either language identifies a larger proportion of medical sentences than KEYWORD. Across both experiments, MDFS models predict a greater number of medical sentences when using non-English than English. The overall low number of medical sentences is due to the corpora we are filtering, which consists largely of data scraped from the internet and hence has a low proportion of medical content.

In order to analyse the diversity of the filtered NMT datasets, we adopt an n-gram-based approach introduced by (Li et al., 2016). First, we tokenise the 1M threshold datasets using the XLMR tokeniser before counting the unique token 1-grams in the first 1M tokens to measure the lexical diversity in each filtered dataset. Table 4 demonstrates that filtering for medical data leads to reduced lexical data and increased sentence length. Datasets created with MDFS exhibit a lower lexical diversity than the KEYWORD baseline; we propose this is due to keyword filtering selecting a larger proportion of sentences outside the medical domain. Furthermore, when filtering En→Ro, we note that MDFS-LINEAR results in a lower lexical diversity and longer sentences compared to MDFS-CLASS. Finally, filtering the non-English side of the datasets results in lower lexical diversity, especially for the En→Ar data.

	Arabic		Romanian	
	Unique 1-gram	**Length**	**Unique 1-gram**	**Length**
RANDOM	32319	27	36455	21
KEYWORD	24125	37	31409	32
MDFS-CLASS (En)	21779	39	29672	39
MDFS-CLASS	20953	44	29638	36
MDFS-LINEAR (En)	21643	40	27898	43
MDFS-LINEAR	20688	45	26779	44

Table 4: Unique token 1-grams and median sentence lengths for the first 1M tokens at a threshold of 1M sentences for Arabic and Romanian.

5 Machine Translation as a Downstream Task

In all NMT experiments, we translate from English. We train encoder-decoder standard transformer models with ∼63M parameters. All models are trained for 100,000 updates using FAIRSEQ (Ott et al., 2019). For the translation quality experiments, we evaluate on the FLORES-200 (NLLB Team, 2022; Goyal et al., 2022) test set comprising 1,007 sentences. The En→Ar medical domain experiments use the TICO-19 (Anastasopoulos et al., 2020) dataset; we use 1,000 sentences as the validation set and the remaining 2,701 as the test set. Finally, for the En→Ro experiments, we use the HIML[4] (Health in My Language) and WMT18 (Bojar et al., 2018) Biomedical test sets. We take 500 sentences of the HIML NHS 24 data as the validation set and combine the 467 Cochrane sentences with the 278 WMT18 biomedical sentences as the test data.

[4] https://www.himl.eu/test-sets

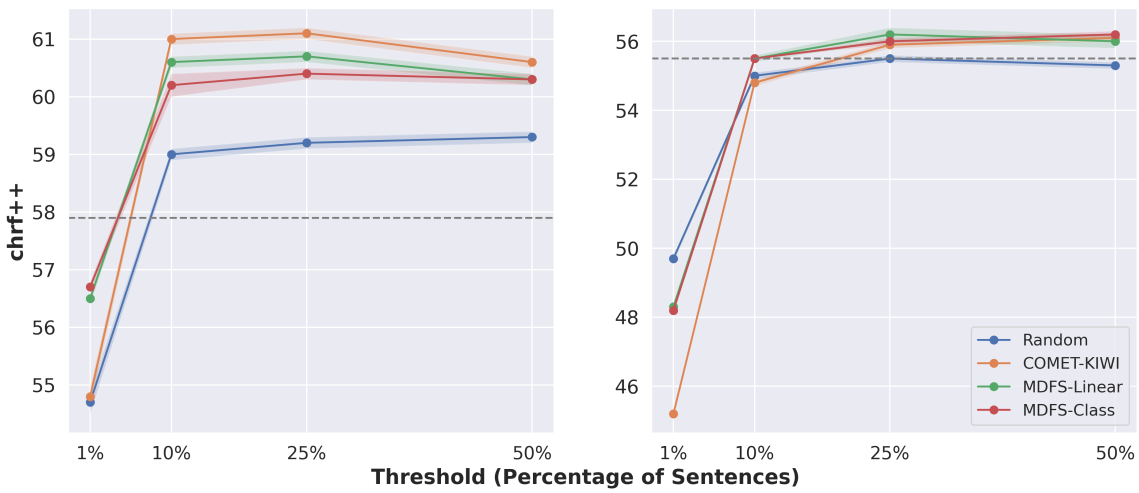

Figure 1: **Left**: Mean chrf++ scores En→De. **Right**: Mean chrf++ scores En→Ar. Results are reported on the Flores-200 test set using three different random seeds. The dashed horizontal line represents the result when running on the entire training data. The errors are calculated using the Standard Error of the Mean.

We train the NMT models using the training data outlined in Section 4.1 with the MDFS training data removed. For the translation quality experiments, we filter to thresholds of 1%, 10%, 25% and 50% of the original training dataset size. Meanwhile, we have a threshold of 1, 2.5, 5, and 10 million sentences for the medical domain experiments. Unless otherwise stated, all results are generated using beam search with a beam size of 5. We report chrf++ (Popović, 2015), a lexical metric as neural metrics have been shown to be less sensitive to wrongly named entities, insertions and deletions (Amrhein and Sennrich, 2022; Alves et al., 2022). As medical content often focuses on a small number of technical terms surrounded by more general language, we believe a lexical metric is more appropriate. We ran each experiment three times with random seeds of 42, 962 and 2025 and reported mean metrics, estimating error using the Standard Error of the Mean. For data filtering techniques that involve random sampling, we also generate three data sets with different seeds.

5.1 Translation Quality Results

Figure 1 presents the mean chrf++ scores from three different random seeds thresholding at 1%, 10%, 25% and 50% of the total training data for the translation quality experiments. Apart from the 1% threshold for En→Ar MDFS results in higher mean chrf++ scores compared to the RANDOM baseline. The largest improvement for En→De over the best RANDOM result is 1.4 chrf++ for MDFS-LINEAR using 25% of the training data, with a 2.8 chrf++ improvement compared to training with the entire dataset. The maximal improvement over RANDOM for En→Ar is lower at 0.7 chrf++ by MDFS-LINEAR at 25% and MDFS-CLASS at 50% of the training data. We hypothesise that this lower improvement is due to the pre-filtered dataset having a large proportion of high-quality sentences, as evidenced by the comparable chrf++ score achieved when training on the entire dataset. These results support that MDFS models effectively filter the training data and, by extension, that the filtering pipeline is effective for non-English languages.

The mean chrf++ scores for MDFS-LINEAR and MDFS-CLASS do not show much variation with a largest observed difference of 0.4 chrf++ for En→De whilst retaining 10% of the total training data, which is also supported by the comparable F1-scores for En→De and En→Ar in Table 1.

Both En→De and En→Ar demonstrate that COMET-KIWI results in worse translations at 1%, and for En→Ar, this also holds true at 10%. For En→De MDFS performs worse than COMET-KIWI for the other thresholds, whereas for En→Ar it achieves comparable chrf++ scores at 25% and 50% of the data. This result is likely due to the fact that COMET-KIWI has been trained with human DA data for En→De but not for En→Ar. Overall, the results suggest that MDFS is better at selecting small amounts of data, whereas COMET-KIWI improves with the size of the filtered dataset.

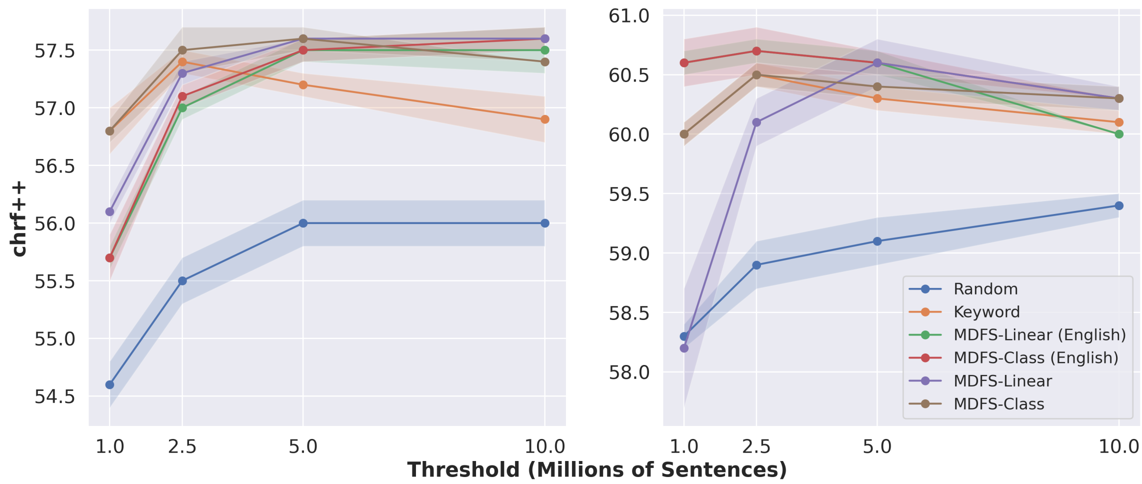

Figure 2: **Left**: Mean chrf++ scores En→Ar. **Right**: Mean chrf++ scores En→Ro. Results are reported on the TICO-19 test set for En→Ar and a combination of HIML and WMT18 data for En→Ro using three random seeds. Errors are calculated using the Standard Error of the Mean.

5.2 Medical Domain Results

Figure 2 shows the mean chrf++ plotted against the threshold. In comparison to RANDOM, all MDFS models achieve a higher chrf++ apart from the Romanian MDFS-LINEAR dataset containing 1M sentences. For En→Ar, Arabic Classification is the strongest MDFS model according to the chrf++ scores. This is true especially when training with only 1M sentences where Arabic MDFS-CLASS scores 0.7 chrf++ higher than any other MDFS model and 2.2 chrf++ higher than RANDOM. The strongest En→Ro model according to the chrf++ scores in Figure 2 is the MDFS-CLASS English model, resulting in the joint highest chrf++ at all thresholds. However, English MDFS-LINEAR equals the chrf++ scores for the three lowest thresholds, and Romanian MDFS-LINEAR does so for the two largest thresholds. Compared to the best score for RANDOM (at 10M sentences), both English MDFS methods improve by 1.3 chrf++ when trained with 2.5M million sentences.

Overall, we find further evidence that the MDFS pipeline achieves comparable results when applied to non-English and English languages. The major exception to this observation is for MDFS-LINEAR Romanian, which has lower chrf++ scores than the other MDFS models at 1M and 2.5M sentences. Romanian MDFS's chrf++ score at 1M is comparable to the RANDOM baseline. We suggest that the low score is related to the lower lexical diversity exhibited by the MDFS-LINEAR model in Romanian at low thresholds rather than the LLM labels as MDFS-CLASS obtains a chrf++ of 60.0 at 1M sentences.

The KEYWORD baseline is competitive with all non-English MDFS baselines at the lower thresholds, whereas it achieves slightly lower chrf++ scores at higher thresholds. All keyword-containing data has been selected at higher thresholds and must be supplemented with a random selection. The strong chrf++ score for KEYWORD filtering demonstrates the effectiveness of hand-written rules, especially for terminology-heavy fields such as medicine. Additionally, as we are training models from scratch, the higher lexical diversity will likely lead to stronger translation systems when it comes to non-medical content.

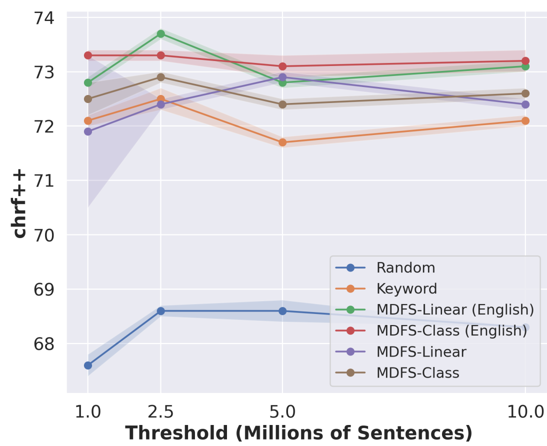

Figure 3: Mean chrf++ scores En→Ro reported on 1,000 best sentences according to COMET from the held-out test set labelled with `Llama-3.1-70B-Instruct` using three different random seeds. The errors are calculated using the Standard Error of the Mean.

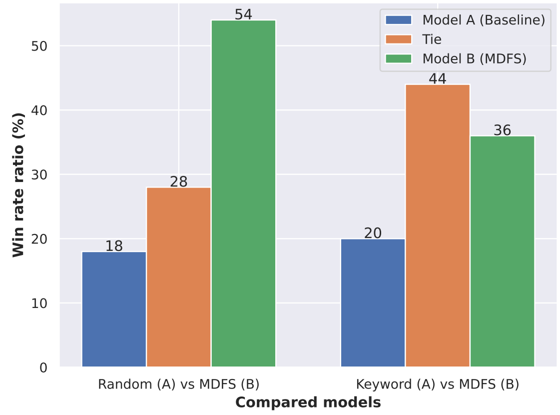

Figure 4: Win rates % of models in terms of terminology translation. Comparison of models trained with different filtering in terms of capability to correctly translate domain-specific terms.

We also construct an additional test set for En→Ro from the 10,000 sentence test set (originally extracted from the training data) labelled with the LLM and used to evaluate the MDFS filtering models. Similar to results in Section 4.5, we use this test set to see if our filtering pipeline improves the translation of those sentences that the LLM labels as being of high quality. We create the test set by taking all sentences that receive a score of 4 or higher from the LLM in the En→Ro direction and selecting the top 1,000 according to COMET. The chrf++ scores for these are given in Figure 3. The results evidence a larger improvement of the MDFS methods, with English MDFS-LINEAR improving the chrf++ by 1.2 and Romanian MDFS-CLASS improving by 0.4 compared to the KEYWORD baseline at 2.5M sentences.

5.3 Domain-specific terminology evaluation

The filtering techniques in our experiments select different subsets of parallel corpora that may cause a downstream model to exhibit patterns that we are unable to capture via a system-level metric. Therefore, given a medical domain adaptation task, we decided to focus on an important aspect of domain adaptation - terminology translation. Given the flexibility of our approach, we decided to check if the filtering is robust compared to baselines and whether our approach translates into the capability to focus on domain nuances.

We set up our experiment as follows. Given our medical evaluation dataset for En→Ro, we sample 100 examples using the *subset2evaluate*[5] library (Zouhar et al., 2025) to establish the most efficient evaluation subset. Afterwards, we employ LLM-based evaluation (Qian et al., 2024) to assess NMT systems pair-wise, i.e. a baseline against MDFS. Rather than focus on overall translation quality, we rank the systems based on the accurate translation of medical terminology, as judged by the LLM. We provide this experiment's prompt and more details in Appendix C.

The evaluation results are presented in Figure 4. Although the chosen evaluation data point (i.e. the threshold of 2.5, see Figure 3) did not indicate a substantial difference between KEYWORD and MDFS in system-level metric, in terms of terminology translation, MDFS denotes 16 percentage points more wins, which showcases the robustness of the approach over hand-written rules. Compared to the random baseline, MDFS provides even more benefits, reaching 54% wins overall.

6 Conclusion

We trained classification and linear regression data filtering models from labels generated by `Llama-3.1-70B-Instruct` to filter NMT data based on translation quality and medical relevance. Our findings show that such a filtering pipeline extends beyond English languages, effectively filtering data. For our medical domain experiments, we report comparable NMT results when filtering English or non-English data. Furthermore, these findings support that LLMs can effectively generate labels for languages they do not officially support, even when compared to a model like COMET-KIWI, which was trained using manually annotated data.

We find that training with a classification objective is preferable when filtering data for low cutoffs and in non-English languages, whereas linear regression might perform slightly better at larger data sizes. We make this observation, not only base on the low chrf++ scores of En→Ro but also the fact that the Arabic MDFS-CLASS model is the best at lower thresholds. We suggest this indicates that the continuous ranking of sentences provided by the linear regression models is not effective at selecting the very best sentences, possibly due to the inability of the MDFS models to correctly distinguish between "good" and "excellent" sentences. Hence, such filtering models may not be suitable for selecting the best examples available in a dataset for annealing LLMs, but they may be better suited for pre-training.

[5]Used parameters: method="diversity", metric="lm"

Limitations

A natural extension to our work would be to evaluate multilingual filtering on a large-scale LLM pre-training dataset such as FineWeb 2 (Penedo et al., 2024b). Whilst such an experiment is more directly related to pre-training multilingual LLMs, it also comes at a much more significant computational cost. Additionally, focusing on parallel data in NMT allows a more direct comparison of filtering the same data in different languages.

As we actively chose to select languages for the medical domain experiments that `Llama-3.1-70B-Instruct` does not officially support, we did not have much choice regarding available test sets. Those that are available tend to use more general language than scientific medical writing. Hence, the results may be slightly different scientific translations. We also acknowledge that both translation pairs for the medical domain experiments are unsupported by `Llama-3.1-70B-Instruct`, but we argue that we are comparing our method to filtering the English side, which is supported.

All our experiments focus on training small NMT models from scratch rather than finetuning larger models. Our reasoning is that our work is most applicable to filtering large amounts of pre-training data rather than selecting the best examples from a smaller subset of data for pre-training. However, to address this shortcoming, we present the results of finetuning `nllb-600` and `nllb-1.3B` (Team et al., 2022) in Appendix E.

The major risk for filtering data using neural models may lead to the reinforcement of biases in the filtering training data. This is especially true of linear regression models that exhibit the lowest lexical diversity after filtering. Such bias may also be exacerbated by a distribution shift between the data used to train the filtering model and the data to which the filtering model is applied.

Lastly, we would like to point out that the prompt used to generate the LLM scores for the translation quality experiments has some minor spelling mistakes.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Qin Cai, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, and 96 others. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv preprint*. ArXiv:2404.14219 [cs].

Duarte Alves, Ricardo Rei, Ana C Farinha, José G. C. de Souza, and André F. T. Martins. 2022. Robust MT evaluation with sentence-level multilingual augmentation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 469–478, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *Preprint*, arXiv:2402.17733.

Chantal Amrhein and Rico Sennrich. 2022. Identifying weaknesses in machine translation metrics through minimum Bayes risk decoding: A case study for COMET. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1125–1141, Online only. Association for Computational Linguistics.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Francisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the Translation Initiative for COvid-19.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Miquel Esplà, Mikel Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 118–119, Dublin, Ireland. European Association for Machine Translation.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, and 2 others. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. 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.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Anton Lozhkov, Loubna Ben Allal, Elie Bakouch, Leandro von Werra, and Thomas Wolf. 2024. Finemath: the finest collection of mathematical content.

James Cross Onur Çelebi Maha Elbayad Kenneth Heafield Kevin Heffernan Elahe Kalbassi Janice Lam Daniel Licht Jean Maillard Anna Sun Skyler Wang Guillaume Wenzek Al Youngblood Bapi Akula Loic Barrault Gabriel Mejia Gonzalez Prangthip Hansanti John Hoffman Semarley Jarrett Kaushik Ram Sadagopan Dirk Rowe Shannon Spruit Chau Tran Pierre Andrews Necip Fazil Ayan Shruti Bhosale Sergey Edunov Angela Fan Cynthia Gao Vedanuj Goswami Francisco Guzmán Philipp Koehn Alexandre Mourachko Christophe Ropers Safiyyah Saleem Holger Schwenk Jeff Wang NLLB Team, Marta R. Costa-jussà. 2022. No language left behind: Scaling human-centered machine translation.

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

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024a. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. *arXiv preprint*. ArXiv:2406.17557 [cs].

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Martin Jaggi, Leandro von Werra, and Thomas Wolf. 2024b. Fineweb2: A sparkling update with 1000s of languages.

Jan-Thorsten Peter, David Vilar, Daniel Deutsch, Mara Finkelstein, Juraj Juraska, and Markus Freitag. 2023. There's no data like better data: Using QE metrics for MT data filtering. In *Proceedings of the Eighth Conference on Machine Translation*, pages 561–577, Singapore. Association for Computational Linguistics.

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.

Shenbin Qian, Archchana Sindhujan, Minnie Kabra, Diptesh Kanojia, Constantin Orasan, Tharindu Ranasinghe, and Fred Blain. 2024. What do large language models need for machine translation evaluation? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3660–3674, Miami, Florida, USA. Association for Computational Linguistics.

Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón, and Sergio Ortiz Rojas. 2020. Bifixer and bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal. European Association for Machine Translation.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

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.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. 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)*, pages 6490–6500, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

01.AI Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, and 11 others. 2024. Yi: Open foundation models by 01.ai. *ArXiv*, abs/2403.04652.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *Preprint*, arXiv:2401.10020.

Jaume Zaragoza-Bernabeu, Gema Ramírez-Sánchez, Marta Bañón, and Sergio Ortiz Rojas. 2022. Bicleaner AI: Bicleaner goes neural. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 824–831, Marseille, France. European Language Resources Association.

Vilém Zouhar, Peng Cui, and Mrinmaya Sachan. 2025. How to select datapoints for efficient human evaluation of nlg models? *Preprint*, arXiv:2501.18251.

A LLM Prompts

```
Evaluate the quality of the translation using the additive 5-point scoring system described below.
Points are accumulated based on the satisfaction of each criterion:
- Add 1 point if the both the source sentence and the translation are fluent well formed sentences.
- Add 1 point if the translation is a feasible translation of the sentence. The translation may be
suboptimal but should still convey the basic sense of the orginal sentence.
- Add 1 point if the translation adequately conveys the entire meaning of the original sentence. Such
a translation should not have any errors, but does not need to be completely unambigous or natural.
- Add 1 point if the translation contains the exact same information as the original sentence. Such
translations should be of professional standard and entail the same information as the original
sentence.
- Add 1 point if the translation quality is extremly high, the translation accuralety conveys the tone
of the original sentence or the translation accounts for cultural differences between the languages.
Below is an {SRC_LANGUAGE} sentence and a translation into {TGT_LANGUAGE}.
The sentence: {SRC}
The translation: {TGT}
After examining the extract:
- Briefly justify each point on the 5-point scoring system, up to 100 words.
- Conclude with the score using the format: “Translation score: <total points>”
```

Figure 5: Template prompt used for scoring data with `Llama-3.1-70B-Instruct` for translation quality.

```
Evaluate whether the sentence below is from the medical domain and could be helpful in a medical,
biological or public health context using the additive 5-point scoring system described below. Points
are accumulated based on the satisfaction of each criterion:
- Add 1 point if the sentence contains any information related to the medical domain.
- Add 1 point if the medical content is clear and presented in an organised manner.
- Add 1 point if the sentences only contain medical, biological or public health content.
- Add 1 point if the sentence is highly relevant and beneficial for medical, biological or public
health purposes whilst exhibiting a clear and consistent writing style.
- Add 1 point if the sentence is an outstanding example of scientific medical or biological content.
Below is an {SRC_LANGUAGE} sentence.
The sentence: {SRC}
After examining the sentence:
- Briefly justify each point on the 5-point scoring system, up to 100 words.
- Conclude with the score using the format: Medical score: <total points>”
```

Figure 6: Template prompt used for scoring data with `Llama-3.1-70B-Instruct` for medical content.

B Keywords

Table 5 gives the 30 keywords that are used to filter for medical sentences on the English side of the parallel data as described in Section 3.3. They were manually selected to be as unambiguous as possible.

vaccine	drug	health	infect	doctor	patient
disease	innoculate	liver	bone	illness	injury
treatment	injection	medicine	symptom	tissue	infection
surgery	aorta	therapy	hospital	pancreas	blood
cancer	influenza	protein	dental	pregnant	virus

Table 5: List of the English medical keywords used to filter for medical sentences for the KEYWORD baseline.

C Domain-specific terminology evaluation details

We present the evaluation prompt in Figure 7. Following the findings of Qian et al. (2024), we include a chain of thought to the prompt to improve the LLM evaluation. The experiment was done using `gpt-4o-mini` as a judge.

The terminology evaluation experiment uses the 2.5 million threshold systems from the experiment depicted in Figure 3 and described in Section 5.2. As a representative of MDFS, we employ MDFS-CLASS (English).

```
Please find the medical word pairs in the source and
target language sentences. Refer to the above word
pairs to count the disambiguation accuracy in the
generated sentences of System A and System B.
Think step by step and produce a final score: 0 if
System A produced a better translation, 1 if it is a
tie, 2 if System B produced a better translation.

Source: "{source}"
Target: "{target}"
System A: "{system_a}"
System B: "{system_b}"
```

Figure 7: Template prompt used for medical terminology LLM-based evaluation.

D NMT Results

Further to the chrf++ scores given in 5 we report spBLEU, chrf++ and COMET in the tables below. Tables 6 and 7 give the results for the translation quality experiments and Tables 8 and 9 give the results for the medical domain experiments.

Threshold	Method	spBLEU	chrf++	COMET
1%	RANDOM	32.1 ± 0.2	54.7 ± 0.2	0.783 ± 0.001
	COMET-KIWI	32.0 ± 0.3	54.8 ± 0.2	0.763 ± 0.003
	MDFS-LINEAR	34.6 ± 0.1	56.5 ± 0.1	0.800 ± 0.001
	MDFS-CLASS	34.7 ± 0.1	56.7 ± 0.1	0.801 ± 0.001
10%	RANDOM	37.8 ± 0.2	59.0 ± 0.1	0.833 ± 0.001
	COMET-KIWI	40.8 ± 0.2	61.0 ± 0.1	0.848 ± 0.000
	MDFS-LINEAR	40.3 ± 0.2	60.6 ± 0.1	0.845 ± 0.001
	MDFS-CLASS	39.6 ± 0.3	60.2 ± 0.2	0.844 ± 0.000
25%	RANDOM	38.0 ± 0.2	59.2 ± 0.1	0.835 ± 0.001
	COMET-KIWI	41.0 ± 0.1	61.1 ± 0.1	0.852 ± 0.000
	MDFS-LINEAR	40.4 ± 0.2	60.7 ± 0.1	0.847 ± 0.001
	MDFS-CLASS	39.8 ± 0.1	60.4 ± 0.1	0.847 ± 0.001
50%	RANDOM	38.3 ± 0.2	59.3 ± 0.1	0.836 ± 0.001
	COMET-KIWI	40.3 ± 0.2	60.6 ± 0.1	0.849 ± 0.001
	MDFS-LINEAR	39.7 ± 0.2	60.3 ± 0.1	0.847 ± 0.000
	MDFS-CLASS	39.7 ± 0.2	60.3 ± 0.1	0.846 ± 0.001

Table 6: Mean spBLEU, chrf++ and COMET scores for the En→De translation quality experiments. The mean is take from three runs with different random seeds and the errors are the Standard Error of the Mean.

Threshold	Method	spBLEU	chrf++	COMET
1%	RANDOM	28.8 ± 0.2	49.7 ± 0.1	0.803 ± 0.001
	COMET-KIWI	24.4 ± 0.1	45.2 ± 0.0	0.749 ± 0.001
	MDFS-LINEAR	28.6 ± 0.1	48.3 ± 0.0	0.789 ± 0.000
	MDFS-CLASS	28.5 ± 0.1	48.2 ± 0.0	0.792 ± 0.001
10%	RANDOM	35.2 ± 0.1	55.0 ± 0.1	0.846 ± 0.001
	COMET-KIWI	35.0 ± 0.1	54.8 ± 0.1	0.846 ± 0.000
	MDFS-LINEAR	36.6 ± 0.2	55.5 ± 0.1	0.852 ± 0.001
	MDFS-CLASS	36.8 ± 0.1	55.5 ± 0.0	0.854 ± 0.000
25%	RANDOM	35.5 ± 0.1	55.5 ± 0.1	0.849 ± 0.000
	COMET-KIWI	36.1 ± 0.1	55.9 ± 0.1	0.856 ± 0.000
	MDFS-LINEAR	36.8 ± 0.3	56.2 ± 0.2	0.856 ± 0.000
	MDFS-CLASS	36.6 ± 0.1	56.0 ± 0.1	0.856 ± 0.001
50%	RANDOM	35.5 ± 0.1	55.3 ± 0.1	0.849 ± 0.001
	COMET-KIWI	36.3 ± 0.1	56.1 ± 0.1	0.858 ± 0.000
	MDFS-LINEAR	36.2 ± 0.2	56.0 ± 0.2	0.856 ± 0.001
	MDFS-CLASS	36.3 ± 0.1	56.2 ± 0.1	0.858 ± 0.001

Table 7: Mean spBLEU, chrf++ and COMET scores for the En→Ar translation quality experiments. The mean is take from three runs with different random seeds and the errors are the Standard Error of the Mean.

Threshold	Method	spBLEU	chrf++	COMET
1.0	RANDOM	34.5 ± 0.2	54.6 ± 0.2	0.832 ± 0.001
	KEYWORD	37.4 ± 0.1	56.8 ± 0.2	0.846 ± 0.001
	MDFS-LINEAR (EN)	35.8 ± 0.1	55.7 ± 0.1	0.835 ± 0.001
	MDFS-CLASS (EN)	35.9 ± 0.2	55.7 ± 0.2	0.836 ± 0.001
	MDFS-LINEAR	36.8 ± 0.2	56.1 ± 0.1	0.840 ± 0.001
	MDFS-CLASS	37.4 ± 0.1	56.8 ± 0.1	0.844 ± 0.000
2.5	RANDOM	35.9 ± 0.2	55.5 ± 0.2	0.840 ± 0.001
	KEYWORD	38.1 ± 0.2	57.4 ± 0.1	0.848 ± 0.001
	MDFS-LINEAR (EN)	37.3 ± 0.3	57.0 ± 0.1	0.845 ± 0.001
	MDFS-CLASS (EN)	37.6 ± 0.1	57.1 ± 0.1	0.847 ± 0.001
	MDFS-LINEAR	37.9 ± 0.2	57.3 ± 0.1	0.849 ± 0.000
	MDFS-CLASS	38.1 ± 0.3	57.5 ± 0.2	0.850 ± 0.001
5.0	RANDOM	36.3 ± 0.2	56.0 ± 0.2	0.840 ± 0.001
	KEYWORD	37.9 ± 0.3	57.2 ± 0.1	0.848 ± 0.001
	MDFS-LINEAR (EN)	37.9 ± 0.1	57.5 ± 0.1	0.849 ± 0.001
	MDFS-CLASS (EN)	38.0 ± 0.1	57.5 ± 0.1	0.848 ± 0.001
	MDFS-LINEAR	38.2 ± 0.1	57.6 ± 0.0	0.850 ± 0.000
	MDFS-CLASS	38.0 ± 0.2	57.6 ± 0.1	0.850 ± 0.001
10.0	RANDOM	36.0 ± 0.0	56.0 ± 0.2	0.841 ± 0.000
	KEYWORD	37.5 ± 0.2	56.9 ± 0.2	0.846 ± 0.001
	MDFS-LINEAR (EN)	37.9 ± 0.1	57.5 ± 0.2	0.850 ± 0.001
	MDFS-CLASS (EN)	38.3 ± 0.2	57.6 ± 0.1	0.850 ± 0.000
	MDFS-LINEAR	38.0 ± 0.0	57.6 ± 0.0	0.849 ± 0.000
	MDFS-CLASS	37.9 ± 0.1	57.4 ± 0.0	0.849 ± 0.001

Table 8: Mean spBLEU, chrf++ and COMET scores for the En→De medical domain experiments. The mean is take from three runs with different random seeds and the errors are the Standard Error of the Mean. The threshold is millions of sentences.

Threshold	Method	spBLEU	chrf++	COMET
1.0	RANDOM	39.2 ± 0.2	58.3 ± 0.1	0.864 ± 0.001
	KEYWORD	41.8 ± 0.3	60.0 ± 0.1	0.878 ± 0.001
	MDFS-LINEAR (EN)	42.8 ± 0.1	60.6 ± 0.1	0.877 ± 0.000
	MDFS-CLASS (EN)	42.7 ± 0.3	60.6 ± 0.2	0.878 ± 0.000
	MDFS-LINEAR	39.9 ± 1.2	58.2 ± 0.5	0.837 ± 0.007
	MDFS-CLASS	42.0 ± 0.2	60.0 ± 0.1	0.873 ± 0.002
2.5	RANDOM	40.0 ± 0.2	58.9 ± 0.2	0.870 ± 0.001
	KEYWORD	42.5 ± 0.2	60.5 ± 0.1	0.880 ± 0.000
	MDFS-LINEAR (EN)	43.0 ± 0.2	60.7 ± 0.1	0.880 ± 0.001
	MDFS-CLASS (EN)	42.9 ± 0.3	60.7 ± 0.2	0.881 ± 0.001
	MDFS-LINEAR	42.1 ± 0.4	60.1 ± 0.2	0.874 ± 0.001
	MDFS-CLASS	42.7 ± 0.1	60.5 ± 0.1	0.875 ± 0.002
5.0	RANDOM	40.5 ± 0.2	59.1 ± 0.2	0.872 ± 0.000
	KEYWORD	42.1 ± 0.1	60.3 ± 0.1	0.878 ± 0.000
	MDFS-LINEAR (EN)	42.7 ± 0.2	60.6 ± 0.1	0.881 ± 0.000
	MDFS-CLASS (EN)	42.8 ± 0.2	60.6 ± 0.1	0.881 ± 0.000
	MDFS-LINEAR	42.8 ± 0.3	60.6 ± 0.2	0.881 ± 0.001
	MDFS-CLASS	42.3 ± 0.1	60.4 ± 0.1	0.881 ± 0.000
10.0	RANDOM	40.8 ± 0.1	59.4 ± 0.1	0.872 ± 0.001
	KEYWORD	41.8 ± 0.1	60.1 ± 0.1	0.877 ± 0.001
	MDFS-LINEAR (EN)	41.8 ± 0.1	60.0 ± 0.0	0.879 ± 0.000
	MDFS-CLASS (EN)	42.1 ± 0.0	60.3 ± 0.0	0.879 ± 0.000
	MDFS-LINEAR	42.2 ± 0.2	60.3 ± 0.1	0.879 ± 0.000
	MDFS-CLASS	42.1 ± 0.2	60.3 ± 0.1	0.880 ± 0.001

Table 9: Mean spBLEU, chrf++ and COMET scores for the En→Ro medical domain experiments. The mean is take from three runs with different random seeds and the errors are the Standard Error of the Mean. The threshold is millions of sentences.

E NLLB Finetuning

In order to evaluate how filtered data performs at finetuning, we train `nllb-600` and `nllb-1.3B` for one epoch using 1M sentences of En→Ro data. The `nllb-1.3B` is finetuned using LoRA (Hu et al., 2022), whereas for `nllb-600`, we update all parameters. The results demonstrate that MDFS improves medical domain translation over the RANDOM baseline. When using LoRA MDFS-LINEAR, Romanian results in comparable chrf++ to other models, whereas the full finetuning using the `nllb-600` model results in reduced scores compared to the other models.

	nllb-600M	**nllb-1.3B**
BASELINE	55.8	58.1
RANDOM	58.2	59.0
KEYWORD	59.3	59.6
MDFS-CLASS (En)	59.5	59.7
MDFS-LINEAR (En)	59.6	59.7
MDFS-CLASS	59.5	59.7
MDFS-LINEAR	58.5	59.6

Table 10: chrf++ scores for finetuning `nllb-600` and `nllb-1.3B` using 1M sentences of En→Ro data. Both models are trained for 1 epoch, `nllb-1.3B` is trained using LoRA.

F GPU Hours

Labelling datasets with `Llama-3.1-70B-Instruct` was run on a single `A100-80GB` GPU. We labelled four datasets, each running taking around $\sim$ 70 hours.

Training MDFS models took $\sim$ 10 hours on either one `A100-40GB` GPU or two `RTX 3900` GPUs. Labelling the NMT data takes $\sim$ 24 hours, again run on either `A100-40GB` GPU or two `RTX 3900` GPUs. We train and predict twice for each language pair and task for a total of 8 runs.

NMT training and evaluation is run on either one `A100-40GB` GPU or one `RTX 3900`, with each run and evaluation taking $\sim$ 4 hours; we train 240 NMT models.