Back to the Stats: Rescuing Low Resource Neural Machine Translation with Statistical Methods

Velayuthan Menan[1], Dilith Jayakody[1], Nisansa de Silva[1], Aloka Fernando[1], Surangika Ranathunga[2]
[1]Dept. of Computer Science & Engineering, University of Moratuwa,10400, Sri Lanka
[2]Massey University, Palmerston North, 4443, New Zealand
{velayuthan.22,dilith.18,NisansaDdS,alokaf}@cse.mrt.ac.lk
s.ranathunga@massey.ac.nz

Abstract

This paper describes our submission to the *WMT24 shared task for Low-Resource Languages of Spain* in the Constrained task category. Due to the lack of deep learning-based data filtration methods for these languages, we propose a purely statistical-based, two-stage pipeline for data filtration. In the primary stage, we begin by removing spaces and punctuation from the source sentences (Spanish) and deduplicating them. We then filter out sentence pairs with inconsistent language predictions by the language identification model, followed by the removal of pairs with anomalous sentence length and word count ratios, using the development set statistics as the threshold. In the secondary stage, for corpora of significant size, we employ a Jensen-Shannon divergence-based method to curate training data of the desired size. Our filtered data allowed us to complete a two-step training process in under 3 hours, with GPU power consumption kept below 1 kWh, making our system both economical and eco-friendly. The source code, training data, and best models are available on the project's GitHub page[1].

1 Introduction

We[2] participated in the **Constrained submission** category of the WMT24 shared task for Low-Resource Languages of Spain (Sánchez-Martínez et al., 2024), focusing on the **Spanish-Asturian** language pair. For the **Constrained submission** category, we are limited to using only the resources provided on the official shared task site[3], and all models utilized must not exceed 1 billion parameters.

Previous shared tasks on data filtering have used deep learning-based scoring methods like LASER (Heffernan et al., 2022) and LaBSE (Feng et al., 2022), as well as sentence alignment methods such as SentAlign (Steingrimsson, 2023) and Vecalign (Thompson and Koehn, 2019). However, these methods often fail with low-resource languages (LRL) due to a lack language of support.

Following prior work (Cruz and Sutawika, 2022; Vegi et al., 2022; Zhang, 2023), we focus on statistical data filtration and sampling techniques to curate our datasets, ensuring our method is not limited to specific languages. Given our compute resource constraints, we design our pipeline to utilize small dataset sizes, enabling a larger volume of experiments. As noted in Ranathunga et al. (2024), randomly sampling a large corpus and training on that sample yields sub-optimal results. Therefore, we use the Jensen-Shannon Divergence (JSD) (Lu et al., 2020) to filter subsets from large corpora (see Section 3.1.1). We favor JSD over Kullback-Leibler (KL) divergence and higher-order domain discriminators due to its symmetric property and relatively simple implementation.

In addition to data filtration, we experiment with a two-step training schema. First, we train the entire model on a larger filtered dataset. Then, in the second step, we freeze the encoder layers and fine-tune the model on the filtered dataset for fine-tuning. This approach proved effective for the task. We select models with under 1 billion parameters for all experiments to adhere to the rules of the *Constrained task*.

Our key contributions are:

- We propose a two-stage data filtration system that can be applied to any language. This system includes statistical data filtration methods for bilingual and monolingual data, along with a Jensen-Shannon divergence-based filtration method.

[1]https://github.com/vmenan/wmt24-lowres-spain
[2]Team Mora-translate, Primary submission id 547
[3]https://www2.statmt.org/wmt24/romance-task.html

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 901–907
November 15–16, 2024.

- We achieve competitive results in a compute resource-constrained environment (Table 4).
- Based on our experiments, we show that fine-tuning a multilingual translation model for a high-resource source language and a low-resource target language is most effective when high-quality monolingual target data is leveraged, and the encoder is frozen to preserve the source language knowledge while training the model.
- We maintain an effective training time well under 3 hours and keep the total GPU power consumption of our best model (training + fine-tuning) under 1 kWh, resulting in a minimal carbon footprint and enhancing the eco-friendliness of our training schema.

2 Related Work

Bilingual parallel data curated from web-mined corpora are prone to various types of noise. Kreutzer et al. (2022) investigated the issue of noise in web-mined corpora by analyzing a sample of 100 sentence pairs, providing evidence of the problem. In a similar vein, Khayrallah and Koehn (2018) examined how different types of noise in parallel training data impact the quality of neural machine translation systems. Building on this body of work, Ranathunga et al. (2024) further established that data quality is more critical than data quantity, particularly for low-resource languages. These studies collectively highlight the importance of addressing *bad* data in web-mined corpora.

Handling noisy parallel data in machine translation had been extensively studied, with various methods proposed in the literature. Below, we present these methods, grouping them into two main categories: (1) deep learning-based approaches and (2) statistical-based approaches. Additionally, we include methods that do not fit broadly into these categories under the section *Other Approaches*.

Deep Learning-Based Approaches. Recent advances in deep learning have introduced several methods to handle unclean data. Zhang (2023) proposed a denoising approach by pretraining on corrupted data and regenerating the original content. They utilized text corruption techniques as proposed by Lewis et al. (2020), including token masking, sentence permutation, document rotation, token deletion, and text infilling. They pretrained models on synthetic and monolingual data and fine-tuned them on clean parallel corpora, achieving translation perplexity scores by training two models and analyzing prediction difficulties. Ensembling methods were also employed to enhance performance. Chaudhary et al. (2019) utilized LASER and an ensemble of scoring methods to check the similarity between embeddings and cross-entropy scores for both directions, penalizing significant differences. Abdulmumin et al. (2022) developed a binary classifier to predict translation accuracy, collecting positive data from a gold standard dataset and negative data from the worst LASER alignment scores. Unfortunately, majority of the methods and models mentioned above do not support Asturian, Aragonese, or Aranese.

Statistical Approaches. Steingrímsson et al. (2023) applied rules such as filtering sentences with three tokens or less, ensuring 60% or more token overlap between languages, and requiring at least 70% alphabetical characters in both sentences. Vegi et al. (2022) introduced constraints such as filtering sentences where the source or target exceeds 800 characters, where the length ratio is greater than 2.5 or less than 0.4, or where words exceed 10 characters. Cruz and Sutawika (2022) extended these rules to include sentences with too many contiguous punctuations (three or more), a large percentage of numbers or punctuations, and additional filtering criteria. Minh-Cong et al. (2023) proposed building a dictionary using MGiza++ and clean parallel corpora, translating source sentences, calculating edit distances, and iteratively training NMT models, assuming the availability of clean parallel corpora.

Other Approaches. Other approaches for handling unclean data include additional filtering rules proposed by Vegi et al. (2022), such as removing sentences that are empty or identical between source and target. Cruz and Sutawika (2022) suggested removing sentences with missing punctuations in one language, sentences containing HTML or URLs, ensuring numbers appear in both source and target, and deduplicating data after preprocessing. Steingrímsson et al. (2023) recommended using a language filter to ensure both languages are in the top two predictions and removing near-duplicate pairs.

3 Methodology

Training Datasets: For training, we use bilingual datasets from OPUS[4] CCMatrix and WikiMedia (Spanish-Asturian), subjected to the filtration outlined in section 3.1.1. We also use monolingual datasets, specifically the PILAR (Galiano-Jiménez et al., 2024b) Asturian monolingual dataset and the Spanish side of the English-Spanish Wikimedia dataset from OPUS, with filtration procedures detailed in section 3.1.2.

Development Set: We evaluate the trained models using the FLORES+[5] Spanish-Asturian development set, referred to as the development set throughout the paper.

Hardware Specifications: All experiments were conducted on a single machine with an Intel i9-9900K CPU, 64GB of RAM, and an Nvidia Quadro RTX 6000 (24GB VRAM).

Software Specifications: All models and training code were developed using the HuggingFace (HF) Transformers (Wolf et al., 2020) library. For evaluation, we use chrF and BLEU scores from the `evaluate`[6] library of HF. We utilized the work done by Nayak et al. (2023) to obtain the Jensen-Shannon divergence scores.

Models: We use NLLB-200-600M (NLLB Team et al., 2022) (we will address it as NLLB-600M throughout the paper), M2M100-418M (Fan et al., 2020) (we will address it as M2M100 throughout the paper), and SMaLL-100 (Mohammadshahi et al., 2022) to conduct experiments. All training and fine-tuning was performed on SMaLL-100 (see section 3.2)

Training Details: We use the HF Transformers Trainer API with the AdamW optimizer, a learning rate of 1×10^{-5}, and a batch size of 16. Gradient accumulation steps of 16 were used to increase the effective batch size to 256. Training is conducted in two steps: first, training the entire model using the filtered Spanish-Asturian CCMatrix dataset (Table 3); then, fine-tuning the best model from the training phase by freezing the encoder layers (Table 4).

Dataset used in fine-tuning step: For fine-tuning, we combined multiple datasets as follows (see Table 4 for results): Dataset A is the filtered Spanish-Asturian Wikimedia; Dataset B includes A + the filtered PILAR crawled monolingual data; Dataset C includes B + the filtered PILAR literary monolingual data; Dataset D includes C + the Spanish monolingual data from English-Spanish Wikimedia. All monolingual data were translated using NLLB-600M.

3.1 Data Filtration and Curation

3.1.1 Bilingual Data Filtration

Our Bilingual Data Filtration pipeline consists of two stages: Primary and Secondary.

Primary Filtration: We start by removing punctuation and whitespace from the Spanish text, then deduplicate it. Using Idiomata Cognitor (Galiano-Jiménez et al., 2024a), we classify the language of each sentence, removing those with inconsistent predictions. We analyze sentence length ratios and word counts, using the development set as a benchmark to remove anomalies.

Dataset	Before (M)	After (M)	% drop
Wikimedia - es_ast	0.04	0.03	38.99
CCMatrix - es_ast	5.39	2.01	62.72
Wikipedia - es_en	2.80	1.26	55.19
Wikimedia - es_en	1.81	1.31	27.47

Table 1: The table presents the sample count (in millions) before and after primary filtration, along with the percentage of samples dropped during this phase.

Secondary Filtration: For datasets over 300K sentences, we limit the size to 50K-300K sentences. We use the Jensen-Shannon divergence to refine the data. We sample sets of 2000 sentence pairs (with replacement) to form 1000 sets, remove Spanish stop words and punctuation with the *NLTK library*[7], and calculate word frequency distributions for each sample and the development set. We sort the samples by their divergence scores (Nayak et al., 2023) against the development set, and iteratively merge and deduplicate low-divergence samples until we have the target dataset size (as specified in Table 3). The implementation of the secondary filtration method is detailed in Algorithm 1.

[4] https://opus.nlpl.eu/
[5] https://github.com/transducens/PILAR
[6] https://github.com/huggingface/evaluate
[7] https://www.nltk.org/

Algorithm 1: Secondary Filtration

Input: Dataset D with $|D| \geq 500K$, development set E, number of batches N, batch size L, desired size S

Output: Deduplicated batch set B_{final} of size S

```
// Initialize empty batch list
```
$B \leftarrow \{\}$

for $i \leftarrow 1$ **to** N **do**

```
    // Randomly sample L rows with
       replacement
```
    $B_i \leftarrow RandomSample(D, L)$

    $B \leftarrow B \cup \{B_i\}$

end

```
// Initialize empty scores list
```
$scores \leftarrow \{\}$

for *each batch* $B_i \in B$ **do**

```
    // Jensen-Shannon Divergence
       between B_i and E
```
    $JS_i \leftarrow JS\_div(B_i, E)$

    $scores \leftarrow scores \cup \{(B_i, JS_i)\}$

end

Sort($scores$) by JS_i in ascending order

```
// Initialize final batch List
```
$B_{final} \leftarrow \{\}$

$current\_size \leftarrow 0$

while $current\_size < S$ **do**

    $B_{candidate} \leftarrow scores[i].batch$

    $B_{final} \leftarrow B_{final} \cup \{B_{candidate}\}$

    De-duplicate(B_{final})

    $current\_size \leftarrow Length(B_{final})$

end

return B_{final}

3.1.2 Monolingual Data Filtration

We extract monolingual data from PILAR crawled and literacy datasets for Asturian, and Wikimedia OPUS datasets for Spanish using the English-Spanish direction. Using *sentence-splitter*[8], we segment the PILAR text into sentences using the Spanish setting, achieving good performance despite the library's lack of support for Asturian. We removed URL links and retained sentences with a word count between four and sixty (the maximum in the development set). Sentences were then classified using the language identifier model Idiomata Cognitor (Galiano-Jiménez et al., 2024a), and those not identified as Spanish or Asturian were removed.

[8] https://github.com/mediacloud/sentence-splitter

3.2 Model Selection

Translation Model Selection: Based on zero-shot performance scores, NLLB-600M was selected as the best model for translating the filtered monolingual dataset, outperforming both M2M-100 and SMaLL-100 in chrF and BLEU scores (Table 2).

Training Model Selection: Given the limited GPU resources in our training environment, we selected SMaLL-100 as the model for training and experimentation due to its smaller size and superior performance compared to M2M-100 (Table 2).

4 Results and Discussion

In this section, we present the results of our experiments using the proposed methods. The results from our primary data filtration step (Section 3.1.1) demonstrate the row counts of each dataset before and after filtration. Notably, our filtration method had the most significant impact on the CCMatrix (es-ast) and Wikipedia (es-en) datasets, with data reduction percentages exceeding 50%. Further investigation could be conducted to understand the factors contributing to this substantial data drop in these sources and to determine whether this observation is consistent across other language pairs in these datasets.

Table 2 displays the zero-shot performance of three state-of-the-art open-source models: NLLB-600M, M2M100, and SMaLL-100. NLLB-600M was selected as the model for generating translations for monolingual sentences due to its significantly better performance compared to the other two models. Given its smaller size and superior performance compared to M2M100, SMaLL-100 was chosen as the model for training and experimentation. By choosing the SMaLL-100 model, we gained the added advantage of using larger batch sizes due to the model's small size. This proved crucial in our low-compute resource environment.

Table 3 presents the results from the first step of the two-step training regime described in Section 3, applied to the CCMatrix filtered dataset. The data was incrementally filtered based on increasing Jensen-Shannon divergence scores in steps of 50k. We observe that model performance improves up to a subset size of 100K, after which it gradually declines. This observation aligns with the findings of Ranathunga et al. (2024), emphasizing that data

Model	chrF	BLEU	# of Trainable Params (M)
NLLB-600M	**49.77**	**17.16**	615.07
M2M100	46.27	14.71	483.91
SMaLL-100	48.47	14.85	**332.74**

Table 2: Scores for the zero-shot performance of the models evaluated on FLORES+ Spanish-Asturian dev set and number of trainable parameters for each model.

quality is more important than the sheer size of the dataset. Selecting smaller, higher-quality datasets not only enhances performance but also offers the additional benefits of reduced training time and lower computational requirements, which in turn minimizes the carbon footprint. As shown in Table 3, the best-performing subset required only 1.48 hours of training and consumed just 0.44 kWh of GPU power.

Size (k)	chrF	BLEU	Time (hrs)	Power (kWh)
50	49.63	17.25	0.72	0.21
100	**50.04**	**17.52**	1.48	0.44
150	49.84	17.26	2.25	0.66
200	49.95	17.34	2.92	0.86
250	49.81	17.36	3.68	1.09
300	49.70	17.02	4.42	1.30

Table 3: Scores, training durations, and GPU power consumption for training CC-Matrix Spanish-Asturian filtered datasets at intervals of 50K jumps. This is the first step in the training phase.

The results of our second step of model training (fine-tuning), where the encoder layers were frozen, are presented in Table 4. Among the various combinations, Dataset C demonstrated the best performance. This dataset includes Spanish-Asturian Wikimedia data as well as Spanish-Asturian PILAR crawled and literary datasets. Notably, the Spanish side of this dataset was generated using the translation model (NLLB-600M) for the PILAR Asturian monolingual crawled and literary datasets.

Interestingly, the performance drops when using Dataset D, which consists of Dataset C combined with Spanish data (Spanish-English Wikimedia) translated into Asturian using the translation model. This observation underscores the importance of high-quality target-side sentences, as the monolingual PILAR dataset comprises carefully curated, high-quality Asturian data. We hypothesize that since Spanish is a high-resource language, the model's encoder has likely been exposed to extensive Spanish data. By freezing the encoder and allowing the model to learn during this fine-tuning step, the model was better able to focus on the target language. Based on these observations, we conclude that when fine-tuning a pre-trained multilingual translation model for a high-resource source language and a low-resource target language, it is essential to leverage high-quality monolingual data for the target language and freeze the encoder to retain the learned knowledge of the source language while making the other layers trainable.

	Size (k)	chrF	BLEU	Time (hrs)	Power (kWh)
A	24.8	51.14	17.80	0.28	0.08
B	38.5	51.38	18.02	0.60	0.18
C	60.2	**51.47**	**18.17**	0.95	0.28
D	84.9	50.92	17.97	1.35	0.40

Table 4: Dataset name, Size, Scores, Time duration and the GPU Power consumption of fine-tuning the best model from Table 3. A = filtered Spanish-Asturian Wikimedia; B = A + filtered and translated PILAR crawled data; C = B + filtered and translated PILAR literary data; D = C + Spanish-Asturian data from English-Spanish Wikimedia.

The training time and GPU power consumption for the best datasets from our two-step training procedure (as shown in Table 3 and Table 4) remains well within 3 hours and consumes less than 1 kWh of GPU power. This makes our proposed method highly suitable for low-compute environments.

5 Conclusion

We presented a purely statistical-based pipeline for data filtering, demonstrating that simple statistical methods should not be overlooked, particularly for low-resource languages where deep learning-based methods may fail to provide adequate support. Our proposed pipeline achieved competitive performance in a low-compute environment for the constrained task, proving to be both economical, with training times well under 3 hours as well as eco-friendly, with GPU power consumption kept under 1 kWh. This work reinforces the findings of previous studies that emphasize the importance of data quality over quantity. We hope that our methodology will encourage and empower researchers in low-compute environments to contribute to an egalitarian representation of lan-

guages.

Acknowledgements

We would like to thank the National Languages Processing (NLP) Center, at the University of Moratuwa for providing the GPUs to execute the experiments related to the research.

We would like to thank Emojot Private Limited for allowing us to utilize their compute resources for this paper.

References

Idris Abdulmumin, Michael Beukman, Jesujoba Alabi, Chris Chinenye Emezue, Everlyn Chimoto, Tosin Adewumi, Shamsuddeen Muhammad, Mofetoluwa Adeyemi, Oreen Yousuf, Sahib Singh, and Tajuddeen Gwadabe. 2022. Separating grains from the chaff: Using data filtering to improve multilingual translation for low-resourced African languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1001–1014, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Vishrav Chaudhary, Yuqing Tang, Francisco Guzmán, Holger Schwenk, and Philipp Koehn. 2019. Low-resource corpus filtering using multilingual sentence embeddings. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 261–266, Florence, Italy. Association for Computational Linguistics.

Jan Christian Blaise Cruz and Lintang Sutawika. 2022. Samsung research Philippines - datasaur AI's submission for the WMT22 large scale multilingual translation task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1034–1038, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding.

Aarón Galiano-Jiménez, Felipe Sánchez-Martínez, and Juan Antonio Pérez-Ortiz. 2024a. Idiomata cognitor.

Aarón Galiano-Jiménez, Felipe Sánchez-Martínez, and Juan Antonio Pérez-Ortiz. 2024b. Pan-iberian language archival resource.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2101–2112, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jinghui Lu, Maeve Henchion, and Brian Mac Namee. 2020. Diverging divergences: Examining variants of Jensen Shannon divergence for corpus comparison tasks. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6740–6744, Marseille, France. European Language Resources Association.

Nguyen-Hoang Minh-Cong, Nguyen Van Vinh, and Nguyen Le-Minh. 2023. A fast method to filter noisy parallel data WMT2023 shared task on parallel data curation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 359–365, Singapore. Association for Computational Linguistics.

Alireza Mohammadshahi, Vassilina Nikoulina, Alexandre Berard, Caroline Brun, James Henderson,

and Laurent Besacier. 2022. Small-100: Introducing shallow multilingual machine translation model for low-resource languages. *arXiv preprint arXiv:2210.11621*.

Shravan Nayak, Surangika Ranathunga, Sarubi Thillainathan, Rikki Hung, Anthony Rinaldi, Yining Wang, Jonah Mackey, Andrew Ho, and En-Shiun Annie Lee. 2023. Leveraging auxiliary domain parallel data in intermediate task fine-tuning for low-resource translation.

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, 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, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Surangika Ranathunga, Nisansa de Silva, Velayuthan Menan, Aloka Fernando, and Charitha Rathnayake. 2024. Quality does matter: A detailed look at the quality and utility of web-mined parallel corpora. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 860–880, St. Julian's, Malta. Association for Computational Linguistics.

Steinþór Steingrímsson, Hrafn Loftsson, and Andy Way. 2023. Filtering matters: Experiments in filtering training sets for machine translation. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 588–600, Tórshavn, Faroe Islands. University of Tartu Library.

Steinthor Steingrimsson. 2023. A sentence alignment approach to document alignment and multi-faceted filtering for curating parallel sentence pairs from web-crawled data. In *Proceedings of the Eighth Conference on Machine Translation*, pages 366–374, Singapore. Association for Computational Linguistics.

Felipe Sánchez-Martínez, Juan Antonio Pérez-Ortiz, Aarón Galiano-Jiménez, and Antoni Oliver. 2024. Findings of the WMT 2024 shared task on translating into low-resource languages of spain: Blending rule-based and neural systems. In *Proceedings of the Ninth Conference on Machine Translation*, Miami, USA. Association for Computational Linguistics.

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

Pavanpankaj Vegi, Sivabhavani J, Biswajit Paul, Prasanna K R, and Chitra Viswanathan. 2022. ANVITA-African: A multilingual neural machine translation system for African languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1090–1097, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wenbo Zhang. 2023. IOL research machine translation systems for WMT23 low-resource Indic language translation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 978–982, Singapore. Association for Computational Linguistics.