Separating Grains from the Chaff: Using Data Filtering to Improve Multilingual Translation for Low-Resourced African Languages

Idris Abdulmumin[1,2*], Michael Beukman[3*], Jesujoba O. Alabi[4*], Chris Emezue[5,6], Everlyn Asiko[7,8], Tosin Adewumi[9], Shamsuddeen Hassan Muhammad[2,10], Mofetoluwa Adeyemi, Oreen Yousuf[11], Sahib Singh[12], Tajuddeen Rabiu Gwadabe[2,13]

All: MasakhaneNLP, [1]Ahmadu Bello University, Zaria, Nigeria, [2]HausaNLP, [3]University of the Witwatersrand, South Africa, [4]Saarland University, Germany, [5]TUM, Germany, [6]Mila - Quebec AI Institute, [7]University of Cape Town, South Africa, [8]African Institute for Mathematical Sciences, [9]Luleå University of Technology, Sweden, [10]LIAAD-INESC TEC, Porto, Portugal, [11]Uppsala University, Sweden, [12]Ford Motor Company, [13]University of Chinese Academy of Sciences, China

iabdulmumin@abu.edu.ng

Abstract

We participated in the WMT 2022 Large-Scale Machine Translation Evaluation for the African Languages Shared Task. This work describes our approach, which is based on filtering the given noisy data using a sentence-pair classifier that was built by fine-tuning a pre-trained language model. To train the classifier, we obtain positive samples (i.e. high-quality parallel sentences) from a gold-standard curated dataset and extract negative samples (i.e. low-quality parallel sentences) from automatically aligned parallel data by choosing sentences with low alignment scores. Our final machine translation model was then trained on filtered data, instead of the entire noisy dataset. We empirically validate our approach by evaluating on two common datasets and show that data filtering generally improves overall translation quality, in some cases even significantly.

1 Introduction

This paper presents Masakhane NLP's submission to the WMT 2022 large-scale machine translation evaluation for African languages. We participated in the constrained translation task and chose to focus on a subset of all the language pairs considered for this task due to resource constraints. We specifically explore the language directions **{hau, ibo, lug, swa, tsn, yor, zul}↔eng** and **wol↔fra**, and submitted our primary and secondary systems which were competitive with other submissions for this task.

Machine translation has received much attention recently, especially for low-resourced languages (Adelani et al., 2022a; Fan et al., 2021; Haddow et al., 2022; Hoang et al., 2018; Nekoto et al., 2020). A promising approach for such setups is to fine-tune large pre-trained language models on the available small amount of translation data (Neubig and Hu, 2018; Adelani et al., 2021a, 2022a). While most of these language models are trained on predominantly high-resourced language datasets (Conneau et al., 2020; Devlin et al., 2019; Radford et al., 2018), there have been a few models that were pre-trained (Ogueji et al., 2021) or adaptively fine-tuned (Alabi et al., 2022) only on low-resourced languages.

Recent works have tried, successfully, to supplement the existing small amounts of natural data in low-resource languages with artificially generated parallel data. For instance, in machine translation, Sennrich et al. (2016) and Ueffing (2006) padded the true parallel data with automatic translations of monolingual sentences through back-translation and self-learning respectively. Others, such as Bañón et al. (2020); El-Kishky et al. (2020); and Schwenk et al. (2021), have used different approaches for detecting potentially aligned sentences within web datasets. While significant improvements have been achieved with these synthetic datasets, an in-depth investigation by Kreutzer et al. (2022) has found them to be fraught with many issues, such as misalignment, wrongful language codes, etc.

Similarly, research has shown that data quality plays an important role in the performance of natural language processing (NLP) models, in machine translation specifically (Arora et al., 2021; Dutta et al., 2020; Hasan et al., 2020; Tchistiakova et al., 2021), but also more generally in other NLP tasks (Abdul-Rauf et al., 2012; Alabi et al., 2020). It was found that often times, models that were trained on smaller amounts of high-quality data outperform their counterparts that are trained on larger amounts of noisy datasets (Gascó et al., 2012; Przystupa and Abdul-Mageed, 2019; Abdulmumin et al., 2022; de Gibert et al., 2022). This has led to many studies (Eetemadi et al., 2015) and prior WMT tasks (Koehn et al., 2018, 2019, 2020) that

* Equal contribution.

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 1001–1014
Abu Dhabi, December 7–8, 2022.

attempt to find ways to improve the quality of existing data, which, as mentioned before, is often rife with errors.

Therefore, in our submission to the shared task, we experimented with filtering web-mined data for African languages using pre-trained language models and evaluated the effect of using this filtered data on machine translation performance. We defined our filtering approach as a sentence-pair binary classification task and fine-tuned a pre-trained language model using positive and negative samples. We used sentences from the high-quality MAFAND-MT (Adelani et al., 2022a) dataset (which was included in the training data for the constrained task) as positive examples and created negative examples by extracting sentences with low language-agnostic sentence representations (LASER) (Artetxe and Schwenk, 2019b) alignment scores from the `wmt22-african` (NLLB Team et al., 2022) corpus that was provided for this task. Our results highlight the importance of filtering on the quality of the final machine translation system. We also detail how to create a high-quality filter for African languages using a few gold-standard parallel sentences. We release our codes on GitHub.[1]

The rest of the paper is organized as follows: in Section 2, we review related work, and in Section 3, we present the dataset we used. Section 4 provides an overview of the bitext filtering approach, while Section 5 details experimental settings and the translation model architecture. In Section 6, we evaluate the model's performance, and lastly in Section 7, we conclude and highlight some future research directions.

2 Related Work

One of the difficulties when dealing with low-resourced settings, as we do here, is that high-quality parallel texts are particularly scarce (Koehn and Knowles, 2017). To curate data for such language pairs, methods for automatically mining parallel text from the web using heuristics (Resnik, 1999) or latent space and similarity-based filters (Artetxe and Schwenk, 2019a; Schwenk et al., 2021) have been proposed. These have led to the curation of publicly available web-mined datasets such as CCAligned (El-Kishky et al., 2020), CC-Matrix (Fan et al., 2021; Schwenk et al., 2021), ParaCrawl (Esplà et al., 2019), and WikiMatrix (Schwenk et al., 2019) to mention just a few.

However, the recent research work by Kreutzer et al. (2022) shows that the automatically aligned and mined parallel bitexts, especially for low-resource language pairs, contain various degrees of errors and less than half of the data are of good quality. Additionally, many approaches generate large amounts of synthetic data, often through back-translation, where synthetic parallel data is generated by automatically translating monolingual data (Bojar and Tamchyna, 2011; Lambert et al., 2011; Sennrich et al., 2016). While additional data has the potential to improve the trained models, these synthetic datasets are often of low quality (Xu et al., 2019). These observations have led to an increased interest in the automatic filtering of noisy bitexts as a key research topic in machine translation (MT).

One approach to improve data quality is to filter out the noisy or invalid parts of a large corpus, keeping only a high-quality subset thereof (Abdulmumin et al., 2021). In this vein, numerous filtering methods have been developed (Axelrod et al., 2011; Eetemadi and Toutanova, 2015; Junczys-Dowmunt, 2018). For instance, Xu et al. (2019) use the cosine similarity between sentence embeddings as a measure of how closely aligned two sentences are. Imankulova et al. (2017) perform back-translation and then filter based on the sentence-level BLEU score, keeping only those sentences with a high BLEU. Similarly, Adjeisah et al. (2021) perform a round-trip translation and only use the sentence pair if it is sufficiently close to the original sentence, according to a chosen similarity measure. There has also been work on alignment between two parallel corpora, and Hasan et al. (2020) uses the LASER score[2] to evaluate alignment, and filter out all sentences below a specific threshold.

3 Datasets

We participated in the constrained translation track and used only the provided dataset. We present the various dataset used, their sizes and corresponding sources in Table 9 in Appendix A. For our experiment, we selected 8 language pairs and developed different multilingual machine translation systems for them. These language pairs are **{hau, ibo, lug, swa, tsn, yor, zul}**↔**eng** and **wol**↔**fra**. According to the recommendation

[1] https://github.com/abumafrim/WMT22-MasaKhane

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

Direction	Parallel sentences	Problem
eng → hau	**src:** I booked the house for my husband's family as we were getting married in Ericeira. **tgt:** na tsarr da aba a ka kasarr ni ila ure imbarr yi ngbangbamu.	**tgt** is not a Hausa sentence
eng → hau	**src:** "Go hunt, and may the light be with you.""" **tgt:** """Zo, zo muje, ke kika hada fitinar ke za ki warware ta."""	**tgt** is not a translation of the **src**
eng → hau	**src:** The Moslem creed. **tgt:** Musa Aminta	mismatched named entities
eng → hau	**src:** Israel **tgt:** oooooooooooooooooooooooooooooooooooooooooooooooo ooooooooBBBBBBBBBBBBBBBBBBBBBBBB	mistranslation; foreign characters

Table 1: Examples of noise in the auto-aligned bitext

of Kreutzer et al. (2022), we carefully examined the training dataset provided by manual inspection and divided it into two categories based on the source of the data and the amount of noise included therein. In the following subsections, we describe these two categories of data.

3.1 Clean Bitext

This category of training data comprises all the datasets that are considered to be manually curated. The datasets in this category include: bible-uedin (Christodouloupoulos and Steedman, 2015), MAFAND-MT,[3] QED (Abdelali et al., 2014), Mozilla-I10n,[4] Tanzil,[5] and several others listed in Table 9. The clean bitext consists of sentences mostly in the news and religious domains, with a few in the health, education, and technology domains. We also refer to the clean bitext as True Parallel in this paper.

3.2 Noisy Bitext

We categorized all the automatically aligned bitext as noisy bitext. This also includes the LASER filtered data. The sentences in this category make up the majority of the training dataset, making up 99.2% of the total training data. The datasets in this category include: CCAligned, CCMatrix, LASER `wmt22_african`,[6] WebCrawl African,[7] and the following datasets from OPUS (Tiedemann, 2012): MultiCCAligned (El-Kishky et al., 2020), TED2020 (Reimers and Gurevych, 2020), WikiMatrix (Schwenk et al., 2019), XLEnt (El-Kishky et al., 2021) and others highlighted in Table 9.

Language pair		Data size	% of original
eng	hau	9,122,559	99.9
	ibo	520,544	99.6
	lug	3,511,275	99.8
	swa	32,898,533	99.6
	tsn	6,036,656	99.1
	yor	1,718,105	99.3
	zul	4,142,146	97.6
fra	wol	237,348	100.0

Table 2: Training data after filtering using heuristics

On manual inspection, however, we found numerous issues with the data, including non-parallel sentences, sentences that consist of only numbers and/or punctuation, sentences in different languages, etc. Examples of noise in the auto-aligned data can be seen in Table 1.

3.3 Validation and Test Data

For the optimization of our translation systems, we combined the FLORES-101 (Goyal et al., 2022) and MAFAND-MT (Adelani et al., 2022a) development sets for each of the 8 language pairs. To compare the performance of the developed MT engines, we evaluated on the FLORES-101 devtest set and the MAFAND-MT test set.

4 Parallel Data Filtering

To attempt to deal with the highly noisy data, we opted to use filtering techniques to remove many invalid or incorrectly aligned sentences, similar to prior work (Arora et al., 2021; Hasan et al., 2020; Xu et al., 2019). We first used some simple heuristic approaches, described in Section 4.1, and then progress to an automatic filtering method, detailed in Section 4.2.

4.1 Heuristics

We filtered sentences that consist of only numbers and/or punctuation marks. After filtering, the statis-

[3] https://github.com/masakhane-io/lafand-mt.git
[4] https://github.com/mozilla-l10n/mt-training-data
[5] https://tanzil.net/trans/
[6] https://huggingface.co/datasets/allenai/wmt22_african
[7] https://github.com/pavanpankaj/Web-Crawl-African

Data	eng							fra
	hau†	ibo†	lug†	swa†	tsn†	yor†	zul†	wol†
Train	6,198	13,998	8,152	61,566	4,202	13,290	7,002	6,722
Dev	2,602	3,002	3,002	3,584	2,686	3,090	2,480	3,014
Test	3,002	3,002	3,002	3,672	3,002	3,118	1,998	3,002

Table 3: Sentence-pair classification training data: a mixture of MAFAND-MT† sentence pairs, taken as positive samples, and `wmt22_african` (worst pairs based on LASER scores), taken as negative samples.

tics of the resulting training dataset are shown in Table 2. The table shows that 2.4% of the original Zulu (`zu`) data consisted of just numbers or punctuation, while other languages had smaller invalid portions, between 0.0% and 0.1%.

4.2 Automatic Filtering

Due to the large size of the automatically aligned dataset, we adopted an automatic approach to determine the quality of parallel sentences to train our translation models. The approach we adopted is sentence-pair binary classification (Nguyen et al., 2021), where we used a transformer-based model to predict the probability that two aligned sentences are actual translations of each other. We explain the process of training data generation and the experimental choices for building the filtering model.

4.2.1 Positive and negative samples

To create the training and evaluation data for the sentence-pair classification-based filtering, we generated positive and negative samples from the training data available for this task. We used the train, dev and test sets from the MAFAND-MT dataset, which is a gold-standard parallel dataset, as positive examples. For the negative examples, however, we sorted the sentences in `wmt22_african` dataset that was provided for this task based on their LASER alignment scores, and selected the least scored sentences in equal amounts to each of the positive examples. The distribution of the train, dev and test samples is presented in Table 3.

4.2.2 Model

We fine-tuned two pre-trained language models, ALBERT (Lan et al., 2020) and AfroXLMR (Alabi et al., 2022) for the sentence pair binary classification task. ALBERT was selected based on its performance on downstream NLP tasks (Lan et al., 2020), even though it has fewer parameters than other BERT-based models (Nguyen et al., 2021). AfroXLMR, on the other hand, was chosen because it was trained on African languages (Alabi et al., 2022), and such a setup has been shown to improve performance on downstream tasks for these languages (Adelani et al., 2022a).

4.3 Filter Training Setup

The filtering models were trained to accept a pair of sentences from the source and target languages. During training, the `[CLS]` token hidden representation of the input sentence pairs is fed into a linear Layer and the model is optimized using binary cross entropy loss. However, at inference time, we add a sigmoid layer to the output to predict a number between 0.0 and 1.0 indicating the likelihood of the bitexts being translations of each other. We fine-tuned these models using each language's train split of positive and negative samples, then evaluated performance on the test set while optimizing on the development set.

The performance of the various automatic filtering models and the subsequent sizes of the filtered datasets for the 8 language pairs are shown in Table 4. This table shows the number of sentence pairs the models classified as actual translation pairs using a threshold of 0.5 and 0.7 as well as the F1 score when using the 0.5 threshold. Additionally, in Table 5, we show the number of sentences that were classified by two or all three of the models as being high-quality.

5 MT Experiments

To evaluate the effect of our filtering techniques, we trained some multilingual NMT models for the 8 language pairs that we have selected for this task. In the following subsections, we highlight the model architectures, training setups, and different multilingual models that were trained.

5.1 Model Architecture

For our experiments, we fine-tune M2M-100 (Fan et al., 2021) on different subsets of the provided data. M2M-100 is a pretrained translation model trained on several languages including African languages, as such it has seen all the languages we have chosen for this task during pre-training. We use the model with $418M$ parameters.

5.2 Training Setup

We fine-tuned the M2M-100 model based on the implementation within the Fairseq[8] toolkit (Ott

[8] https://github.com/facebookresearch/fairseq

Model		en							fr	$F1_{avg.}$
		hau	ibo	lug	swa	tsn	yor	zul	wol	
ALBERT-base	F1	95.6	94.2	94.7	89.6	95.7	91.1	87.4	95.1	92.9
	t=0.5	278,930	78,056	119,516	5,832,820	346,329	151,886	363,739	6,552	
	t=0.7	197,232	63,207	82,243	3,921,959	252,499	91,366	213,991	4,365	
ALBERT-xlarge	F1	93.2	92.8	96.3	63.7	95.3	90.7	89.1	84.4	88.2
	t=0.5	115,987	129,304	146,948	3,263,429	273,154	113,860	613,483	49,926	
	t=0.7	81,641	111,562	102,354	1,638,528	217,200	86,558	302,951	41,283	
AfroXLMR-base	F1	96.9	94.4	95.4	94.6	96.1	98.4	88.0	97.1	95.1
	t=0.5	296,881	75,102	149,051	6,139,327	363,155	81,902	281,803	6,997	
	t=0.7	226,666	59,995	84,499	5,064,365	276,490	73,786	171,778	5,189	

Table 4: Training data after filtering using sentence-pair classifier — t=Threshold; F1 was computed at t=0.5

t=0.5	Albert-base	Albert-xlarge	AfroXLMR
Albert-base	2,984,862	1,750,707	2,575,408
Albert-xlarge	-	2,107,204	1,058,711
AfroXLMR	-	-	3,925,612
sents. in ALL			**668,633**
t=0.7			
Albert-base	1,977,486	909,203	1,884,922
Albert-xlarge	-	1,206,493	547,925
AfroXLMR	-	-	3,420,147
sents. in ALL			**331,208**

Table 5: Data overlap after filtering using the sentence-pair classifier models

et al., 2019). We used batch sizes of 2,048 tokens, a maximum sentence length of 1,024, and a dropout of 0.3. For optimization, we used Adam (Kingma and Ba, 2015) with $\beta_1 = 0.9$ and $\beta_2 = 0.998$, a learning rate of $5e-5$ and a warmup of 2,500 updates. The optimizer uses a label-smoothed cross-entropy loss function with a label-smoothing value of 0.2. All models were trained for a maximum of 1,000,000 update steps. We tokenized all data using the model's SentencePiece (Kudo and Richardson, 2018) tokenizer.

To evaluate our models and to choose the best checkpoints, we used the BLEU score (Papineni et al., 2002) calculated with the SacreBLEU (Post, 2018) implementation. In addition, we also evaluated the models using CHRF (Popović, 2015).

5.2.1 Baseline models

We train many-to-many (M2M) translation models by fine-tuning M2M-100 on the following subsets of the datasets described in Section 3. These include, the clean bitexts described in Section 3.1, noisy bitext described in Section 3.2, and a mixture of the clean and noisy bitexts. The noisy bitext was only partially cleaned, as evidenced in Table 2, using the heuristic rules mentioned in Section 4.1 without applying the proposed automatic filtering on data.

We trained these baseline models to compare and measure the efficacy of our filtering technique on the quality of the translation models. We submitted the model in (i) as our secondary system for this task.

5.2.2 Models on filtered data only

To evaluate the effect of the filtered data on the quality of the translation output, we train M2M models on the filtered data from the different models using a threshold of 0.5 and 0.7.

5.2.3 Models on filtered and clean data

We went further to train multilingual models on the concatenation of the noisy and clean text, and on the filtered and clean data for easier comparison. With this system, we were able to measure the amount of improvement we can obtain by including the clean bitext compared to training models only on the filtered bitext.

6 Results and Discussion

In Tables 6 and 7, we report the BLEU and CHRF scores obtained by the different models that we trained, as evaluated on the FLORES-101 devtest and MAFAND-MT test datasets, respectively.

6.1 Baseline Models

On average, the baseline model trained on the clean bitext performed impressively on the two evaluation datasets, despite the limited dataset size. On MAFAND-MT, the model trained on the clean bitext obtained a higher BLEU score than the model trained on the noisy bitext, and on FLORES-101, the reverse was true. This is likely due to the fact that the MAFAND-MT data is present in the clean

Models	eng→x							fra→x	x→*eng*							x→*fra*	Avg.
	hau	ibo	lug	swa	tsn	yor	zul	wol	hau	ibo	lug	swa	tsn	yor	zul	wol	
BLEU																	
Baselines																	
Clean bitext	9.30	13.19	**4.00**	23.17	8.56	**3.60**	9.43	**3.56**	14.24	12.56	11.24	26.86	8.78	8.90	18.51	6.03	11.37
Noisy bitext	15.32	10.77	2.14	30.64	12.87	2.57	12.35	0.69	20.58	14.69	13.19	31.80	16.29	11.40	24.68	3.22	13.95
Clean + Noisy bitext	15.34	11.37	2.40	30.48	13.31	2.48	12.61	0.73	20.53	15.07	13.34	31.61	16.50	11.75	24.29	3.88	14.11
Filtered only																	
albert-xlarge-0-7	16.43	15.38	2.54	29.89	**16.31**	3.00	15.18	0.65	20.05	17.32	12.51	34.24	18.55	12.62	27.31	5.14	15.45
Filtered + Clean bitext																	
albert-xlarge-0-5	16.05	15.01	3.22	**33.31**	15.96	3.08	14.97	1.99	**20.92**	17.45	13.93	**34.99**	18.24	13.24	27.65	6.43	16.03
albert-xlarge-0-7	16.55	**15.70**	3.45	31.97	**16.31**	3.16	**15.50**	2.12	20.85	**17.88**	**13.97**	34.40	**18.29**	**13.38**	**27.35**	**7.20**	**16.13**
CHRF																	
Baselines																	
Clean bitext	34.01	**45.31**	**42.14**	55.14	45.58	**30.56**	43.62	**30.55**	34.70	**45.20**	**46.21**	54.53	45.37	39.04	46.50	**30.77**	41.83
Noisy bitext	30.04	34.18	33.04	54.34	43.51	16.23	46.30	8.92	35.15	37.46	35.96	54.38	45.39	33.84	49.85	15.72	35.89
Clean + Noisy bitext	30.53	35.75	33.69	54.66	44.23	15.90	46.37	10.91	35.70	38.82	37.50	54.78	45.62	35.35	49.71	19.21	36.79
Filtered only																	
albert-xlarge-0-7	36.18	41.71	36.94	54.79	51.64	18.85	51.14	10.86	37.18	41.38	39.07	56.81	**56.81**	38.27	52.71	22.20	40.41
Filtered + Clean bitext																	
albert-xlarge-0-5	36.56	43.19	40.44	56.65	51.25	20.33	50.77	23.44	37.79	43.72	44.34	57.70	51.98	40.01	52.75	27.76	42.42
albert-xlarge-0-7	36.64	44.32	41.44	56.60	**52.98**	21.88	**51.43**	25.22	38.11	44.23	45.14	**57.73**	52.22	**40.68**	**53.02**	29.29	**43.18**

Table 6: Performance of the multilingual model on the **FLORES-101** devtest set, with the maximum BLEU per column in **bold**. x represents African languages.

bitext, and that the noisy bitext contains sentences that were taken from the web, including Wikipedia, which is the source of the FLORES-101 dataset. When we compared the model trained on the clean bitext to the model trained on the noisy bitext, we saw between a +1 and +2 improvement on FLORES-101 and between +5 and +8 improvement on MAFAND-MT for lug, wol, and yor. This confirms not only the importance of the data domain, but also the importance of data quality on the quality of the machine translation output.

After mixing the two datasets, the performance improved over using only the clean bitext by more than 6 BLEU on hau↔eng, and almost 3 BLEU on average across all languages on FLORES. The performance, though, was similar to using only the noisy bitext. On the MAFAND-MT test set, however, the performance deteriorated by almost 2 BLEU when compared to training on the clean bitext only. At language-pair level, eng→ibo was affected more (−9.14 BLEU), followed by eng→wol, whereas yor→eng benefited tremendously (+17.83 BLEU). On average, training on the two bitexts marginally improves over using only the noisy bitext, and this is consistent on all the test sets.

Investigating the results in more depth, we found that the BLEU scores of the models are lower when translating into an African language, similar to the findings of Adelani et al. (2022a). This effect is exacerbated for the languages with the fewest parallel sentences, such as lug, wol, and yor, except for ibo, which overall has the second-fewest parallel sentences, as shown in Table 9.

6.2 Data Filtering Analysis

We generally see that more filtering results in improved performance, corresponding to removing more noisy sentences from the data. Using less filtering, with a threshold of 0.5, generally performed slightly worse than using a threshold of 0.7. Both of these settings outperformed (a) using no filtering and (b) using no additional data.

We can also see the effect of the filtering steps on the training data in Tables 2 and 4. Filtering the data using heuristics resulted in only a small portion of the data being filtered out. Using the classifier, however, caused a large amount of noisy data to be removed. When looking at the F1 scores of the classification models, we can see that ALBERT-xlarge has the lowest F1, followed by ALBERT-base and AfroXLMR-base. Looking at Table 5, we can see that ALBERT-xlarge is also the most strict filter, removing the most data, whereas AfroXLMR-base removes the least amount of data. Interestingly, the number of sentences marked as high-quality by all three models is surprisingly low, possibly indicating that these different models (particularly ALBERT-xlarge and AfroXLMR-base) focus on different features of the data.

Finally, we saw that a higher threshold resulted in improved translation performance, but ALBERT-xlarge (which is quite strict) had a lower F1 than the

Models	eng→x							fra→x	x→eng							x→fra	Avg.
	hau	ibo	lug	swa	tsn	yor	zul	wol	hau	ibo	lug	swa	tsn	yor	zul	wol	
BLEU																	
Baselines																	
Clean bitext	**9.00**	**20.83**	**11.67**	25.81	18.64	**9.86**	14.50	**8.91**	**12.49**	**19.24**	20.00	29.28	20.44	16.98	23.20	**7.77**	16.79
Noisy bitext	5.24	10.37	6.12	25.35	16.61	3.61	15.23	0.98	8.52	12.83	14.35	28.37	21.34	13.14	26.74	1.57	13.15
Clean + Noisy bitext	5.59	11.69	6.54	25.55	17.25	3.42	15.10	1.99	8.80	13.64	15.67	28.67	21.74	**34.81**	26.68	2.33	14.97
Filtered only																	
albert-xlarge-0-7	7.75	16.33	7.56	26.45	23.01	4.59	17.63	0.86	9.93	15.59	16.77	30.92	30.92	16.46	29.47	3.09	16.08
Filtered + Clean bitext																	
albert-xlarge-0-5	8.49	18.16	10.11	**27.89**	22.99	5.37	17.68	5.46	11.73	17.53	20.63	32.38	27.07	17.84	29.88	5.52	17.42
albert-xlarge-0-7	8.74	19.08	10.26	27.80	**24.25**	6.09	**18.25**	6.05	12.32	17.58	**21.15**	**32.60**	**27.40**	18.54	**30.02**	6.77	**17.93**
CHRF																	
Baselines																	
Clean bitext	36.23	34.10	**31.59**	54.59	33.85	**21.97**	41.70	**26.22**	37.74	37.32	33.85	51.39	32.43	30.51	43.20	29.31	36.00
Noisy bitext	40.24	31.27	25.84	59.14	38.88	19.18	46.98	8.90	44.80	38.71	34.58	56.25	40.57	33.28	49.26	19.15	36.69
Clean + Noisy bitext	40.91	31.67	26.04	59.13	39.60	19.06	47.14	9.66	44.63	39.18	34.76	56.20	40.65	33.71	49.23	21.86	37.09
Filtered only																	
albert-xlarge-0-7	**44.19**	38.13	27.37	59.40	43.97	20.85	**51.96**	11.11	44.98	42.95	33.98	58.60	**43.12**	35.55	**52.09**	24.51	39.55
Filtered + Clean bitext																	
albert-xlarge-0-5	43.38	37.88	29.70	**61.47**	43.30	20.57	51.06	18.73	**45.53**	42.77	36.14	**58.93**	43.11	36.61	52.06	28.26	40.59
albert-xlarge-0-7	44.15	**38.72**	30.78	60.63	**44.11**	21.01	51.85	19.82	45.40	**43.31**	**36.15**	58.45	42.90	**36.81**	52.06	**29.52**	**40.98**

Table 7: Performance of the multilingual model on the **MAFAND-MT** test set, with the maximum BLEU per column in **bold**. x represents African languages.

other models, possibly suggesting that F1 performance does not fully indicate the expected downstream performance on the actual translation task.

6.2.1 The effect of filtering on translation models

We fine-tune M2M-100 for multilingual translation on the filtered data, and as expected, our results (on average) demonstrate a considerable improvement when the translation model is trained on the filtered data rather than the original noisy texts. In particular, for many languages, training on the filtered data from ALBERT-xlarge with a threshold of 0.7 outperformed the model trained on just the noisy bitext with at least a BLEU point.

Furthermore, we compared the performance of the model trained on only the clean data and on only the filtered data. Just as we saw with the baseline system, on MAFAND-MT, the model trained on the clean bitext performed better than the model trained on the filtered bitext, and on FLORES-101, the reverse was true. These results again confirm the importance of the filtering approach and further supports the observation that NMT engines are less robust to noise as found by Khayrallah and Koehn (2018), especially for low-resource settings.

6.3 Clean vs. filtered data

We find that on FLORES-101, adding in noisy, unfiltered data improves the results over just using the true parallel data. On MAFAND-MT, however, it generally reduces the BLEU score significantly. For both datasets, adding appropriately filtered data results in the highest performance averaged over all the languages, although for some specific languages, just using true parallel data resulted in the best performance.

Our performance on the test set provided by the organizers (Adelani et al., 2022b) is shown in Table 8. Here we can see that our primary model, which was trained on the clean bitext as well as the filtered data (filtered using ALBERT-xlarge, $t = 0.7$), significantly outperforms the model trained only on the clean bitext. We also see that our approach seems to have a larger performance gain when translating *from* African languages compared to translating *to* them.

7 Conclusion and Future Work

In this work, we used a sentence-pair classifier to classify parallel data as being aligned, or not. Using this approach, we filtered out a large portion of the original, noisy, data and fine-tuned existing large language models on this new data. Our results show that training on the filtered data significantly increases the performance of the models, resulting in improved translations. In particular, our approach outperforms (i) training only on clean data, (ii) training only on filtered data, and (iii) training on the original dataset, consisting of clean and noisy data. This provides additional evidence in favor of prioritizing data quality over quantity, as well as the need for more advanced noise detection

Models	eng→x							fra→x	x→*eng*							x→*fra*	$x \rightarrow$ afr	afr $\rightarrow x$	Avg.
	hau	ibo	lug	swa	tsn	yor	zul	wol	hau	ibo	lug	swa	tsn	yor	zul	wol			
BLEU																			
Clean only	10.7	11.9	4.5	24.3	10.1	**4.2**	6.0	**4.4**	15.7	15.0	12.2	27.5	9.7	8.8	18.5	7.1	9.5	14.3	11.9
Filtered + Clean	**17.7**	**15.3**	**4.6**	**31.5**	**17.8**	3.2	**11.1**	1.5	**22.7**	**20.9**	**15**	**35.2**	**21.2**	**14.2**	**26.8**	**7.6**	**12.8**	**20.4**	**16.6**
chrF2++																			
Clean only	36.0	34.6	**29.0**	52.2	33.8	**21.8**	36.3	**25.4**	38.0	38.2	33.4	50.4	31.6	29.4	41.6	**28.0**	33.6	36.3	35.0
Filtered + Clean	**43.4**	**38.6**	27.2	**57.7**	**41.9**	19.4	**44.8**	17.9	**45.2**	**44.6**	**35.4**	**57.1**	**43.6**	**35.3**	**49.1**	27.7	**36.4**	**42.2**	**39.4**

Table 8: Performance of the submitted models on the wmt22 test sets as provided by the organizers. We submitted two models. The primary one, denoted *Filtered + Clean*, was trained on the clean bitext as well as the data filtered by ALBERT-xlarge with a threshold of 0.7. The secondary (or contrastive) approach, denoted *Clean only*, was trained only on the clean bitext. The $x \rightarrow$ afr and afr $\rightarrow x$ columns contain the average performance for translations to and from African languages, respectively. *avg* contains the average over all language pairs.

and filtering tools. There are numerous potential avenues for future work; one option is to use a multilingual model as the sentence classifier instead of using a separate model per language, to leverage commonalities between different languages (Adelani et al., 2021b; Conneau et al., 2020). Secondly, a more in-depth study of the effect of the threshold parameter on the final BLEU score would be useful. We would also like to understand the reasons behind the performance by analyzing the filtered data more in depth. Finally, given more computational resources, we will (i) train the classifier for more epochs, using other language models and/or using different quality thresholds, (ii) use longer sentence length than the current 128, (iii) train the translation models on AfroXLMR and ALBERT-base filtered data, and (iv) use the filtering approach on more languages, to evaluate its generalizability. Ultimately, we hope that this filtering approach could lead to the use of cleaner data to train translation models, improving the overall translation quality for low-resourced languages.

Acknowledgements

Idris Abdulmumin acknowledges funding from Google through the Africa Travel Grants. Computations were performed using High Performance Computing infrastructure provided by the Mathematical Sciences Support unit at the University of the Witwatersrand, LIAAD INESC TEC, Portugal, and DFKI GmbH. We acknowledge support from the Machine Learning Group, EISLAB, Luleå University of Science and Technology, Sweden. In addition, we thank the organizers for the compute grant, which allowed us to run some of our experiments. Finally, we thank Dr. Rachel Bawden, and the anonymous reviewers of the 2022 WMT shared task for their helpful feedback. We also acknowledge the useful inputs of Julia Kreutzer and Angela Fan.

References

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. The AMARA corpus: Building parallel language resources for the educational domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 1856–1862, Reykjavik, Iceland. European Languages Resources Association (ELRA).

Sadaf Abdul-Rauf, Mark Fishel, Patrik Lambert, Sandra Noubours, and Rico Sennrich. 2012. Extrinsic evaluation of sentence alignment systems. In *Workshop on Creating Cross-language Resources for Disconnected Languages and Styles*, pages 6–10.

Idris Abdulmumin, Bashir Shehu Galadanci, Abubakar Isa, Habeebah Adamu Kakudi, and Ismaila Idris Sinan. 2021. A Hybrid Approach for Improved Low Resource Neural Machine Translation using Monolingual Data. *Engineering Letters*, 29(4):339–350.

Idris Abdulmumin, Bashir Shehu Galadanci, Shamsuddeen Hassan Muhammad, and Garba Aliyu. 2022. Quantity vs. quality of monolingual source data in automatic text translation: Can it be too little if it is too good? In *2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON)*, pages 1–5. IEEE.

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin

Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022a. A few thousand translations go a long way! Leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

David Adelani, Dana Ruiter, Jesujoba Alabi, Damilola Adebonojo, Adesina Ayeni, Mofe Adeyemi, Ayodele Esther Awokoya, and Cristina España-Bonet. 2021a. The effect of domain and diacritics in Yoruba–English neural machine translation. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 61–75, Virtual. Association for Machine Translation in the Americas.

David Ifeoluwa Adelani, Jade Z. Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin P. Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane Mboup, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima Diop, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021b. Masakhaner: Named entity recognition for african languages. *Trans. Assoc. Comput. Linguistics*, 9:1116–1131.

David Ifeoluwa Adelani, Md Mahfuz Ibn Alam, Antonios Anastasopoulos, Akshita Bhagia, Marta Costa-Jussá, Jesse Dodge, Fahim Faisal, Christian Federmann, Natalia Fedorova, Francisco Guzmán, Sergey Koshelev, Jean Maillard, Vukosi Marivate, Jonathan Mbuya, Safiyyah Saleem, and Holger Schwenk. 2022b. Findings of the WMT 2022 shared task on large-scale machine translation evaluation for african languages. In *Proceedings of the Seventh Conference on Machine Translation*, Online. Association for Computational Linguistics.

Michael Adjeisah, Guohua Liu, Douglas Omwenga Nyabuga, Richard Nuetey Nortey, and Jinling Song. 2021. Pseudotext injection and advance filtering of low-resource corpus for neural machine translation. *Comput. Intell. Neurosci.*, 2021:6682385:1–6682385:10.

Jesujoba Alabi, Kwabena Amponsah-Kaakyire, David Adelani, and Cristina España-Bonet. 2020. Massive vs. curated embeddings for low-resourced languages: the case of Yorùbá and Twi. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2754–2762, Marseille, France. European Language Resources Association.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Karunesh Kumar Arora, Geetam S Tomar, and Shyam S Agrawal. 2021. Studying the role of data quality on statistical and neural machine translation. In *2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT)*, pages 199–204. IEEE.

Mikel Artetxe and Holger Schwenk. 2019a. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019b. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Trans. Assoc. Comput. Linguistics*, 7:597–610.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 355–362. ACL.

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.

Ondrej Bojar and Ales Tamchyna. 2011. Improving translation model by monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine*

Translation, WMT@EMNLP 2011, Edinburgh, Scotland, UK, July 30-31, 2011, pages 330–336. Association for Computational Linguistics.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the Bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395.

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, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sourav Dutta, Jesujoba Alabi, Saptarashmi Bandyopadhyay, Dana Ruiter, and Josef van Genabith. 2020. UdS-DFKI@WMT20: Unsupervised MT and very low resource supervised MT for German-Upper Sorbian. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1092–1098, Online. Association for Computational Linguistics.

Sauleh Eetemadi, William Lewis, Kristina Toutanova, and Hayder Radha. 2015. Survey of data-selection methods in statistical machine translation. *Mach. Transl.*, 29(3-4):189–223.

Sauleh Eetemadi and Kristina Toutanova. 2015. Detecting translation direction: A cross-domain study. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 103–109, Denver, Colorado. Association for Computational Linguistics.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pages 5960–5969, Online. Association for Computational Linguistics.

Ahmed El-Kishky, Adi Renduchintala, James Cross, Francisco Guzmán, and Philipp Koehn. 2021. XLEnt: Mining cross-lingual entities with lexical-semantic-phonetic word alignment. In *Preprint*, Online.

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 Volume 2: Translator, Project and User Tracks*, pages 118–119, Dublin, Ireland. European Association for Machine Translation.

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. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

Guillem Gascó, Martha Alicia Rocha, Germán Sanchis-Trilles, Jesús Andrés-Ferrer, and Francisco Casacuberta. 2012. Does more data always yield better translations? In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics - EACL 2012*, pages 152–161. Association for Computational Linguistics.

Ona de Gibert, Ksenia Kharitonova, Blanca Calvo Figueras, Jordi Armengol-Estap'e, and Maite Melero. 2022. Quality versus quantity: Building catalan-english mt resources. In *Proceedings of SIGUL2022 @LREC2022*, pages 59–69, Marseille. European Language Resources Association (ELRA).

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.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindrich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Comput. Linguistics*, 48(3):673–732.

Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M. Sohel Rahman, and Rifat Shahriyar. 2020. Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for bengali-english machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2612–2623. Association for Computational Linguistics.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

Aizhan Imankulova, Takayuki Sato, and Mamoru Komachi. 2017. Improving low-resource neural machine translation with filtered pseudo-parallel corpus. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 70–78, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 888–895. 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.

Diederick P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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

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

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

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

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.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Patrik Lambert, Holger Schwenk, Christophe Servan, and Sadaf Abdul-Rauf. 2011. Investigations on translation model adaptation using monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation, WMT@EMNLP 2011, Edinburgh, Scotland, UK, July 30-31, 2011*, pages 284–293. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi E. Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Hassan Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Z. Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkabir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Espoir Murhabazi, Elan Van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Itoro Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in african languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 2144–2160. Association for Computational Linguistics.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels, Belgium. Association for Computational Linguistics.

Vincent Nguyen, Sarvnaz Karimi, and Zhenchang Xing. 2021. Combining shallow and deep representations for text-pair classification. In *Proceedings of the The 19th Annual Workshop of the Australasian Language Technology Association*, pages 68–78, Online. Australasian Language Technology Association.

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.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

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, Philadelphia, Pennsylvania, USA. 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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Michael Przystupa and Muhammad Abdul-Mageed. 2019. Neural machine translation of low-resource and similar languages with backtranslation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 224–235, Florence, Italy. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Philip Resnik. 1999. Mining the web for bilingual text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 527–534, College Park, Maryland, USA. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Svetlana Tchistiakova, Jesujoba Alabi, Koel Dutta Chowdhury, Sourav Dutta, and Dana Ruiter. 2021. EdinSaar@WMT21: North-germanic low-resource multilingual NMT. In *Proceedings of the Sixth Conference on Machine Translation*, pages 368–375, Online. Association for Computational Linguistics.

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

Nicola Ueffing. 2006. Using Monolingual Source-Language Data to Improve MT Performance. In *International Workshop on Spoken Language Translation*, pages 174–181, Kyoto, Japan.

Guanghao Xu, Youngjoong Ko, and Jungyun Seo. 2019. Improving neural machine translation by filtering synthetic parallel data. *Entropy*, 21(12):1213.

A Appendix - Data Sources

Datasets used in this project and their sources, as listed in Table 9: MAFAND-MT, wmt22_african, LAVA Corpus,[9] XLEnt, Tanzil, WikiMatrix, CCAligned, CCMatrix, GlobalVoices,[10,11] ParaCrawl,[12] GNOME,[13] tico-19,[14] ELRC_2922,[15] EUbookshop,[16] KDE4,[17] TED2020, Tatoeba,[18] Ubuntu,[19] bible-uedin, wikimedia,[20] QED, MultiCCAligned and Mozilla-I10n.

[9] https://drive.google.com/drive/folders/179AkJ0P3fZMFS0rIyEBBDZ-WICs2wpWU
[10] https://casmacat.eu/corpus/global-voices.html
[11] https://globalvoices.org/
[12] https://paracrawl.eu/
[13] https://l10n.gnome.org/
[14] https://tico-19.github.io/index.html
[15] https://elrc-share.eu/repository/browse/covid-19-health-wikipedia-dataset-multilingual-53-en-x-language-pairs/fe23e2c28c8311ea913100155d0267066f62c6b30ac0429f8d497df0abd2ef72/
[16] http://bookshop.europa.eu
[17] http://www.lt-innovate.org/lt-observe/resources/kde4-kde4-localization-files-v2
[18] https://tatoeba.org/en/
[19] https://translations.launchpad.net/
[20] https://dumps.wikimedia.org/other/contenttranslation/

Data	en							fr
	hau	ibo	lug	swa	tsn	yor	zul	wol
True Parallel								
MAFAND-MT	3,098	6,998	4,075	30,782	2,100	6,644	3,500	3,360
Tanzil	128,376	-	-	138,253	-	-	-	-
GlobalVoices	-	-	-	32,307	-	137	-	-
tico-19	3,071	-	3,071	3,071	-	-	3,071	-
ELRC_2922	-	-	-	607	-	-	-	-
EUbookshop	-	-	-	18	-	-	-	-
Tatoeba	57	22	3	395	31	37	70	67
bible-uedin	-	-	-	-	-	-	15,907	7,918
QED	124	12	740	18,192	-	52	1,624	66
Mozilla-I10n	4,952	4,172	5,931	7,798	-	4,095	-	7,041
Total (TP)	139,678	11,204	13,820	231,423	2,131	10,965	24,172	18,452
Automatcally Aligned								
WMT22 African	2,309,758	172,973	3,450,573	23,358,739	5,931,529	1,455,571	3,862,020	189,659
WebCrawl Afr.	16,950	3,372	10,809	193,518	77,976	18,924	152,724	-
LAVA Corpus	-	-	20,993	371,864	-	-	-	-
WikiMatrix	-	-	-	51,387	-	-	-	-
CCAligned	339,178	148,147	14,702	2,044,993	71,254	175,193	126,103	–
CCMatrix	5,861,080	80,385	-	5,756,664	-	-	-	-
ParaCrawl	-	-	-	132,521	-	-	-	-
GNOME	5,466	23,767	4,578	40	-	10,234	44,605	-
KDE4	1,493	-	-	-	-	-	-	-
TED2020	27	210	-	9,745	-	-	-	-
XLEnt	436,602	69,820	1,054	871,902	4,781	51,173	28,394	4,082
Ubuntu	242	635	637	986	-	141	4,718	220
wikimedia	23,385	12,279	1,315	3,765	969	8,521	1,226	679
MultiCCAligned	-	-	-	-	-	-	-	24,256
Total (AA)	8,994,181	511,588	3,504,661	32,796,124	6,086,509	1,719,757	4,219,790	218,896
Total (ALL)	9,133,859	522,792	3,518,481	33,027,547	6,088,640	1,730,722	4,243,962	237,348

Table 9: Training Data Used — TP=True Parallel; AA=Automatically Aligned