Critical Learning Periods: Leveraging Early Training Dynamics for Efficient Data Pruning

Everlyn Asiko Chimoto[1,2,3] **Jay Gala**[1,5] **Orevaoghene Ahia**[1,6]
Julia Kreutzer[7] **Bruce A. Bassett**[2,4] **Sara Hooker**[7]

[1]Cohere For AI Community [2]University of Cape Town, South Africa
[3]African Institute for Mathematical Sciences [4]South African Astronomical Observatory
[5]Mohamed bin Zayed University of Artificial Intelligence
[6]University of Washington [7]Cohere For AI

Abstract

Neural Machine Translation models are extremely data and compute-hungry. However, not all data points contribute equally to model training and generalization. Data pruning to remove the low-value data points has the benefit of drastically reducing the compute budget without a significant drop in model performance. In this paper, we propose a new data pruning technique: *Checkpoints Across Time* (*CAT*), that leverages early model training dynamics to identify the most relevant data points for model performance. We benchmark *CAT* against several data pruning techniques including COMET-QE, LASER and LaBSE. We find that *CAT* outperforms the benchmarks on Indo-European languages on multiple test sets. When applied to English-German, English-French and English-Swahili translation tasks, *CAT* achieves comparable performance to using the full dataset, while pruning up to 50% of training data. We inspect the data points that *CAT* selects and find that it tends to favor longer sentences and sentences with unique or rare words.

1 Introduction

In recent years, there has been significant improvement in the quality of Neural Machine Translation (NMT) models (Johnson et al., 2017; Arivazhagan et al., 2019; Liu et al., 2020; Fan et al., 2021; Reid et al., 2021; Ramesh et al., 2022; Costa-jussà et al., 2022; Bapna et al., 2022; Gala et al., 2023). The success of these models is primarily due to many factors including pretraining data size, compute abundance and the ever-increasing model size. Despite the gains due to growing datasets, large dataset sizes have also posed significant hurdles for maintaining data quality. Significant portions of web-scraped data used for language model pretraining have been shown to be of low quality, machine-generated spam, pornographic content (Kreutzer et al., 2022). Given the ever-growing size of parallel corpora, it often becomes laborious for humans to assess the quality at scale (Freitag et al., 2021; Agarwal et al., 2022; Longpre et al., 2023). Quality of pretraining data for large NMT models often faces amplified quality issues because of extensive bitext mining approaches to automatically extracting translation pairs from either monolingual corpora (Guo et al., 2018; Schwenk et al., 2021a,b; Costa-jussà et al., 2022; Ramesh et al., 2022; Vegi et al., 2022; Gala et al., 2023) or document-aligned corpora (Bañón et al., 2020; Steingrímsson et al., 2021; Ramesh et al., 2022; Gala et al., 2023). Bitext mining approaches, although far more cost-effective than human curation, have been shown to introduce significant noise into datasets by relying on sub-optimal sentence embedding models (Thompson and Koehn, 2019; Feng et al., 2022; Heffernan et al., 2022) to match sub-optimal translation pairs (Kreutzer et al., 2022).

Recent work has shown that quality matters. Large neural models trained on high-quality data outperform models trained on noisy data (Khayrallah and Koehn, 2018; Arora et al., 2021; Abdulmumin et al., 2022; Gala et al., 2023) or match performance using far fewer data points (Sorscher et al., 2022; Siddiqui et al., 2023; Koneru et al., 2022; Li et al., 2018; Xu et al., 2023). In this work, we ask if we can arrive at a rigorous estimator of data quality through *data pruning*. Our goal is to focus on metrics which are scalable across large datasets, and which identify a subset that preserves performance.

We leverage distinctive periods of training to indicate examples most critical to model generalization. Specifically, we propose *Checkpoints Across Time* (**CAT**): a metric that leverages the variability in perplexity across early checkpoints to identify such critical data points when training MT models. We use this to rank data points, selecting the instances with the highest variability and continuing

Findings of the Association for Computational Linguistics ACL 2024, pages 9407–9426

to train the model solely on this subset. This scoring is used to remove the data points estimated to be least important, thus creating gains in efficiency for training.

CAT is motivated by fundamental work in machine learning which has shown that there are distinctive periods to learning in deep neural network training. Prior work shows that more common, easier features are learned earlier in training, with the most challenging features learned in the last stages of training (Jiang et al., 2020; Achille et al., 2017; Siddiqui et al., 2023). Prior work has also shown that variance in gradients with respect to inputs (Agarwal et al., 2022) or gradient patterns themselves vary depending on the period of learning (Faghri et al., 2020). However, gradients remain expensive to leverage as a learning signal to identify relevant instances. Furthermore, almost the entirety of this work has focused on a computer vision setting, with the exception of Swayamdipta et al. (2020) which uses differences in the model's confidence in the true class, and the variability of this confidence across epoch to cluster data. Here – we explore applying the insights that there are distinctive periods of training to an NLP setting and leveraging for the purpose of pruning. We propose a more efficient method to compute signal relative to prior methods which have required computing gradients instead. Our hypothesis is that those with the largest variability indicate the easy examples which show the quickest learning and gains in certainty early in training.

Compared to existing data pruning strategies in NMT, e.g. the popular (dual) cross-entropy method (Axelrod et al., 2011; Junczys-Dowmunt, 2018) and contrastive data selection (Wang et al., 2018), our approach doesn't require an initial clean dataset to train on as this is one limitation that often prevents the adoption of such methods to low-resourced languages. Additionally, it does not require significant time to run as we rely only on two early checkpoints rather than a training run over multiple epochs. This makes **CAT** methods suitable for low-compute and data-scarce environments which are usually co-occurring (Ahia et al., 2021). Additionally, it eliminates dependency on other models that may not capture information from various languages and domains.

We study the effectiveness of **CAT** across diverse linguistic contexts. Specifically, we utilize English-German and English-French pairs from the Indo-European family using WMT datasets, and the English-Swahili pair from the Bantu family with automatically aligned datasets. Through a series of extensive experiments, we show that **CAT** results in significant improvements in translation quality over random pruning (randomly selecting $X\%$ subset of the training data) and other embedding-based quality estimation methods such as LaBSE (Feng et al., 2022). We summarize our key contributions below:

1. Focusing on Machine Translation, we perform an empirical evaluation of existing quality estimation techniques and embedding models as a means of data pruning. These techniques were not designed for data pruning however we observe that quality estimation techniques are less effective for pruning high-quality MT datasets compared to embedding models used for automatic alignment.

2. We propose a novel method, *Checkpoints Across Time* (**CAT**), that leverages perplexity variation to prune MT datasets with significant improvements over existing methods for Indo-European languages. We show that **CAT** results in significant improvements in translation quality over random pruning (randomly selecting $X\%$ subset of the training data) and other embedding-based quality estimation methods such as LaBSE (Feng et al., 2022). Using our best CAT-based method on English-German, English-French and English-Swahili translation tasks, we are able to achieve comparable performance while pruning up to 50% of the training data.

3. We complement our findings with extensive analysis, studying the relationships between performance metrics and sentence characteristics such as sentence length in MT.

2 Background

2.1 Data Pruning in ML

Data pruning has been extensively studied in the work of (Marion et al., 2023; Li et al., 2018; Raju et al., 2021; Swayamdipta et al., 2020; Boubdir et al., 2023). Li et al. (2018); Sorscher et al. (2022) and Raju et al. (2021) focus on data pruning for computer vision whereas Swayamdipta et al. (2020) focuses on data pruning for language tasks. Various techniques have been proposed for data pruning, such as utilization of loss profile (Siddiqui

et al., 2023), gradients confidence score (Agarwal et al., 2022), and data classification using a model (Swayamdipta et al., 2020). Increasingly techniques have been extended from primarily computer vision settings (Sorscher et al., 2022) to more recently expanding the treatment of data pruning language settings (Marion et al., 2023). These methods quantify the importance of data points to model generalization. Our work follows along these lines of identifying critical data points by leveraging the difference in perplexity between early checkpoints.

2.2 Model-based Data Pruning for NMT

Specifically for the task of NMT, model-based data pruning approaches have been most popular and originate from the cross-entropy scoring method, proposed by Moore and Lewis (2010). Data points are ranked either based on their perplexity under an in-domain language model (LM), or based on the difference of this in-domain cross-entropy and the cross-entropy of a general-domain LM. Axelrod et al. (2011) first apply this technique for MT data selection with a bilingual extension with LM cross-entropy scores for source and target sides. Duh et al. (2013) use early neural LMs instead of n-gram LMs, and Junczys-Dowmunt (2018) replace LMs with NMT models trained in both directions. Zhang et al. (2020) propose to use scores from pre-trained large language models like GPT for domain filtering instead of custom-trained models. van der Wees et al. (2017) make the data selection dynamic, by gradually fine-tuning an NMT model on newly selected data points every epoch. While the focus of model-based data selection was originally on in-domain data selection, it was later extended to generally filtering noisy data sources. Wang et al. (2018) propose an online data denoising approach, measuring noise through differences in log probabilities between noisy and denoised NMT models. Most recently, MT quality estimation (QE) models have also been leveraged for corpus mining and filtering (Kocyigit et al., 2022; Batheja and Bhattacharyya, 2023). Chimoto and Bassett (2022) show that COMET-QE (Rei et al., 2020) improve over random data selection in active learning MT settings.

All these approaches *rely on the existence of a small trusted data set or trusted models* to extrinsically define what "good data" looks like. Our approach, while also using model perplexity as a metric, however, does not require such pre-selection, and focuses on model training dynamics instead. Thereby, it can be applied more broadly. Our experiments aim at this less constrained setting without any given high-quality data, so we compare against model-based selection with an MT-QE model and a multilingual LLM.

2.3 Embedding-based Data Pruning for NMT

Embedding-based similarity metrics have recently become popular for NMT data selection, as they rely on unsupervised learning and *intrinsic notions of quality*, assuming that high-quality parallel sentences have high similarity in a cross-lingual representation space (Schwenk, 2018). Few examples of language-agnostic sentence embedders include MUSE (Lample et al., 2018), XLM-R (Conneau et al., 2020), LaBSE (Feng et al., 2022) and LASER (Heffernan et al., 2022). Schwenk et al. (2021a,b); Costa-jussà et al. (2022) utilize the LASER sentence embedder for bitext mining, whereas Ramesh et al. (2022); Gala et al. (2023) use the LaBSE sentence embedder for the same purpose. In our experiments, we compare against data pruning with LASER and LaBSE embeddings.

2.4 Comparison Studies

With this plethora of filtering methods and application scenarios, there have been few works to systematically compare them. The WMT shared tasks for corpus filtering (Koehn et al., 2018, 2019, 2020) have shed light on how different languages and resourcefulness conditions pose different challenges for above-described filtering methods in practice, and how they affect downstream NMT performance. Bane and Zaretskaya (2021) compare data filtering techniques with a more *qualitative* angle: They find that NMT scores and MUSE embeddings have the highest correlation with human quality judgments, NMT scores work best for in-domain selection, and MUSE/XLM-R for out-of-domain generalization. Herold et al. (2022) compare multiple data noise detection techniques according to their effectiveness for filtering out specific types of data noise. They find that LASER does not detect incorrect languages, and both LASER and cross-entropy perform weakly on detecting misaligned sentences and over/under translation. Similarly, Bane et al. (2022) find cross-entropy filtering empirically superior to other methods (XLM-R, MUSE, LASER, COMET) for filtering out various synthetic noise types (Khayrallah and Koehn, 2018). In their setup, COMET fails at

filtering out misaligned sentences but shows particular sensitivity for target-side omissions, NMT scoring for source-side omissions, while LASER and COMET do not filter mismatching numbers well. Dakwale et al. (2022) compare LASER and LaBSE on low-resource corpus filtering tasks and find them competitive if languages are included in the pretraining of the underlying embedding models.

In this work, we provide both a systematic empirical comparison of different types of data pruning techniques on downstream NMT performance for three target languages, as well as qualitative analyses of what kind of data is getting selected.

3 CAT: *Checkpoints Across Time*

Our goal is to extract the most valuable data points that would contribute to model generalization and train the model solely on this. Our approach, while simple, leverages the difference in learning stages in deep neural network optimization. Firstly, we train an NMT model on the full dataset for only a few epochs, ideally a small fraction of the total time needed for the model to converge. Subsequently, we use the checkpoints from this initial stage to compute the perplexity of each data point. These perplexity scores serve as the basis for selecting candidate data points for subsequent pruning using different perplexity profiles. This approach draws upon the findings of Swayamdipta et al. (2020) that categorize training data relevance based on the learning dynamics during training, and Agarwal et al. (2022) which leverage the difference in gradients across training to identify easy versus challenging examples. Unlike Agarwal et al. (2022) which uses computationally expensive variance of gradients, we use perplexity. For CAT we can utilize the perplexity scores in two ways:

3.1 CAT-DIFF

Let $X = \{x_1, x_2, x_3, ..., x_n\}$ be all instances in the training data. To select $s \subseteq X$ instances to be pruned, we compute the perplexity of each instance at epoch j of all data points in X using the model trained on X with weight parameters θ_j:

$$\text{PPL}_j(x_i) = \exp\left(-\frac{1}{T}\sum_{t=1}^{T}\log P_{\theta_j}(x_{i,t})\right) \quad (1)$$

where T is the number of words in each example x_i.

We then rank all data points based on differences in perplexity between the two epochs that form our checkpoints:

$$\Delta\text{PPL}(x_i) = \text{PPL}_1(x_i) - \text{PPL}_k(x_i) \quad (2)$$

From the ranked data points, we select those with the highest differences between early epochs,[1] according to a threshold at a chosen percentile. For example, for pruning of 50% of data points, we keep all data points that rank higher than the 50th percentile.

3.2 CAT-VAR

An alternative is to compute the perplexity at $N > 1$ epoch checkpoints, but to rank data points according to the variance of the perplexity across the checkpoints:

$$\text{Var}(\text{PPL}(x_i)) = \frac{1}{N}\sum_{j=1}^{N}\left(\text{PPL}_j(x_i) - \frac{1}{N}\sum_{l=1}^{N}\text{PPL}_l(x_l)\right)^2 \quad (3)$$

In our experiments, we use $N = 3$ corresponding to epochs $1, 3$ and 5.

In contrast to CAT-DIFF, we then select data points around the median with a range according to the pruning percentage that we want to achieve, e.g. data points between the 25th and 75th percentile for a pruning percentage of 50%. This aligns with several works which have found variance in model signal to be predictive of difficulty (Swayamdipta et al., 2020; Agarwal et al., 2022). Swayamdipta et al. (2020) show that data points that fall on the tails of the variance tend to be either too easy or too difficult and thus the model may not benefit most from being trained on such examples.

4 Experimental Setup

To fully test the CAT methods for pruning we explore the following variations: (1) different target languages, (2) different pruning levels, (3) different test domains, and (4) configurations such as model sizes and choice of checkpoints. We compare the effectiveness of the pruning against embedding-based, LLM-based, and random selection.

4.1 Training Data

We conduct our experiments on datasets that translate from English into three languages: French,

[1] Either 5th and 1st or 2nd and 1st.

German, and Swahili. For German, we use the English-German WMT19 dataset (Barrault et al., 2019), comprising 38M parallel sentences sourced from various origins, and for Swahili, the English-Swahili WMT22 dataset (Costa-jussà et al., 2022; Adelani et al., 2022b), consisting of 22M parallel sentences sourced from the web (Bañón et al., 2020) then automatically identified using LASER. We use the English-French WMT15 (Bojar et al., 2015) dataset for French. Due to compute constraints, the majority of our experiments and ablations were carried out on a randomly selected subset of 3.8M parallel sentences from all datasets, which constitutes 10% of the original German and French datasets and 17% of the Swahili dataset. Subsequently, we implement the best variant of our method on the entire dataset to evaluate the scalability of our approach.

4.2 Evaluation Data

Our evaluation sets consist of WMT18 (Bojar et al., 2018), WMT15 (Bojar et al., 2015), FLORES (Goyal et al., 2022; Costa-jussà et al., 2022) and MAFAND-MT (Adelani et al., 2022a) test sets. These test sets were curated using human annotators thereby mitigating common issues associated with automatically aligned datasets, such as erroneous alignments. The WMT18 and WMT15 test sets were released as part of the WMT machine translation task consisting of 2998 and 1500 parallel sentences respectively from various sources of the news domain. WMT18 was used to evaluate German models while WMT15 test set was used to evaluate French. The FLORES test set was used to evaluate all languages: German, French and Swahili. The FLORES test set consists of 1012 sentences from 3 sources: WikiNews, WikiJunior and WikiVoyage. We use the MAFAND-MT test set for Swahili model evaluation. MAFAND-MT test set consists of 1835 sentences from the news domain.

4.3 Training Details

We train all our NMT models using the fairseq library[2] (Ott et al., 2019). Our models follow the same architecture as the vanilla transformer (Vaswani et al., 2017) with GeLU activation (Hendrycks and Gimpel, 2016) instead of ReLU activation (Nair and Hinton, 2010). We report all the hyperparameters used for training the models in Table 1. We conduct all our experiments on 4 A100 40GB GPUs with the duration for runs at 90%, 70%, and 50% prune level, and full dataset utilization approximately taking 1, 2, 3, and 4.5 hours, respectively. Furthermore, we train separate SentencePiece tokenizers (Kudo and Richardson, 2018) for source and target languages with a vocab size of 16K using 3.8M translation pairs for both English-German and English-Swahili experiments. We use SacreBLEU library[3] (Post, 2018) to compute BLEU score (Papineni et al., 2002) and used it as our evaluation metric. We also calculate chrF++ and COMET scores and report them in the Appendix. All the metrics exhibit the same trend in our experiments. Our baseline models are trained on the entire 3.8M translation pairs and denoted as "`full`" for all language pairs; English-German, English-French and English-Swahili. We investigate the impact of data pruning with sparsity levels of 50%, 70% and 90% using various techniques briefly described below.

4.4 Baselines

As outlined in Section 2, we compare with the most common NMT data pruning techniques:

1. **Random Selection:** This involves choosing data instances without specific criteria purely by chance. Although this approach may seem intuitively suboptimal, Azeemi et al. (2023) shows that this also performs competitively with different pruning strategies for a few cases. Therefore, we include this approach as one of our baselines.

2. **COMET-QE:** Rei et al. (2020) proposed model-based MT evaluation metric that utilizes embeddings from XLM-RoBERTa (Conneau et al., 2020) to access translation quality when provided source, labels and target translations or just source and target translations.

3. **LaBSE & LASER:** Both are multilingual sentence embedding models trained on large aligned corpora covering a number of languages. LASER (Heffernan et al., 2022) covers 200 languages and was trained with an encoder-decoder LSTM architecture. LaBSE (Feng et al., 2022) on the other hand covers 93 languages and was trained on top of BERT (Devlin et al., 2019).

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

[3] With parameters: `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.2`

Hyperparameter	Value
Optimizer	Adam (Kingma and Ba, 2014)
Beta values (β_1, β_2)	$(0.9, 0.98)$
Learning rate	$5e-4$
Scheduler	Inverse sqrt
Criterion	Cross-entropy
Label smoothing (Szegedy et al., 2016)	0.1
Warmup learning rate	$1e$-7
Warmup steps	$4,000$
Gradient clipping	1.0
Dropout (Srivastava et al., 2014)	0.2
Effective batch size	16K
Mixed precision training	FP16
Maximum epochs	30
Maximum sequence length	256

Table 1: Model hyperparameter settings for all our experiments.

4. **BLOOM LLM:** Using the multilingual pretrained BLOOM LM (Scao et al., 2022), we explore the selection of relevant data points by computing the perplexity of the reference sentence.[4] This is the closest to the CAT method, as it also leverages insights from a trained Transformer model, where we expect easy and hard-to-learn examples to fall on the tails and thus will not enable the model to learn effectively. In contrast to CAT, its perplexity scores might be more expressive, because it was trained on more data and with more parameters. On the other hand, we do not get any insights on training dynamics. We use BLOOM variants with 560M, 1.1B and 1.7B parameters for our experiments and assess the impact of model size on perplexity computation.

In the following, we will summarize embedding-based selection and QE-based selection as *Translation Quality Estimators*.

5 Results and Discussion

We conducted experiments on English-German and English-Swahili datasets, employing various pruning techniques. We pruned 90%, 70% and 50% of the training data and compared this to training on the full set. To evaluate these models, we used WMT18 and FLORES test sets for German, and FLORES and MAFAND-MT as the test sets for Swahili. Furthermore, we applied the most effective pruning techniques to an English-French dataset to verify the consistency of performance across different language pairs.

5.1 CAT techniques outperforms random pruning in German and Swahili

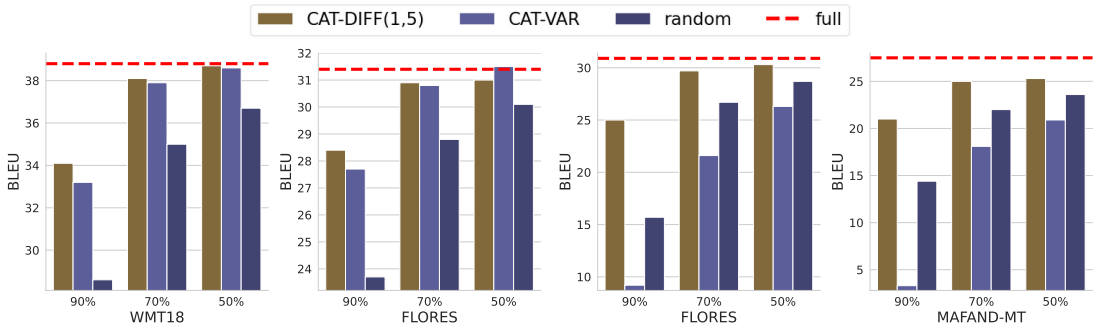
Figure 1a demonstrates that CAT-DIFF(1,5) and CAT-VAR outperform random selection in the German test sets. On the other hand, in Swahili, only CAT-DIFF(1,5) consistently outperforms random selection, while selecting data randomly yields better results than CAT-VAR (refer to Figure 1b). CAT techniques achieve efficient pruning while maintaining an upwards of 75% performance compared to training on the full dataset, even after pruning 90% of the data. For German, both CAT-DIFF(1,5) and CAT-VAR maintain over 80% of the performance achieved by training on 100% of the data. In the case of Swahili, CAT-DIFF(1,5) retains above 80% of the performance, whereas CAT-VAR tends to perform worse, retaining only around 30% of the performance. A noteworthy efficiency of CAT techniques lies in their ability to sidestep the inferencing with fully converged models, thus one doesn't need a lot of compute to utilize these techniques.

Table 2 presents the ablation study conducted for CAT-DIFF to compare the performance of CAT pruning techniques based on different checkpoints. We evaluate the difference between checkpoints 1 and 5, referred to as CAT-DIFF(1,5) and checkpoints 1 and 2, denoted as CAT-DIFF(1,2). We see that CAT-DIFF(1,5) yields better performance than CAT-DIFF(1,2). However, CAT-DIFF(1,2) offers the advantage of using fewer resources albeit at the cost of a slight dip in performance.

5.2 Translation quality estimators are less effective for German

We find translation quality estimators to be less effective in pruning German than Swahili. For instance, Figure 2a indicates that random data pruning outperforms translation quality estimators at 90% and 70% pruning levels for both test sets. However, with Swahili, we observe a different behavior, where LABSE outperforms random pruning on all pruning levels, as depicted in Figure 2b. Interestingly, pruning Swahili data with LaBSE and LASER yields superior performance compared to training on the full data set, particularly for the 50% pruning level for the FLORES test set. Further, translation quality estimators consistently outperform random pruning in Swahili and do not sacri-

[4]Past work has found that NMT is more susceptible to target-side noise (Khayrallah and Koehn, 2018), so we focus on target-side perplexity. In principle, this method could be extended to a combination of source and target perplexities as in the cross-entropy method (Moore and Lewis, 2010), but it would require twice the computation.



(a) German perplexity

(b) Swahili Perplexity

Figure 1: Performance of the CAT methods for both German and Swahili. The CAT-DIFF(1,5), referring to the difference between checkpoints 1 and 5, consistently performs better than random. CAT-VAR yields very poor performance for Swahili and also underperforms CAT-DIFF for German.

fice performance even at 90% pruning level with the exception of LASER on the MAFAND test set. We believe that these differences in behavior can be attributed to varying data quality between the two languages. The German dataset contains sentences with high variance in sentence length and included sentences with foreign language. Conversely, the Swahili dataset consists of processed automatically aligned sentences, where there is a low sentence length variance and the sentences do not contain any foreign language. See Figure 7 and Figure 8 in the Appendix. Therefore, selecting the best sentences ensures the inclusion of high-quality translation pairs.

5.3 LLM perplexity techniques are competitive for Swahili but sub-par for German

Figure 3 illustrates that utilizing LLMs for pruning provides competitive results for Swahili but shows sub-par performance for German. We further observe that the LLM perplexity technique across different BLOOM model sizes performs superior compared to random pruning across pruning levels for Swahili (see Figure 3b). Among the various BLOOM model sizes, the smallest model, BLOOM 560M consistently outperforms the others, followed by the BLOOM 1B model. It is important to note that LLM techniques perform worse than random for German. This suggests that in cases where limitations in data and compute double bind, utilizing pretrained LLMs can save on the limited resources available.

5.4 What pruning techniques work best?

Figure 4 shows the comparison of random, LaBSE and CAT-DIFF pruning techniques for German, French and Swahili. It is evident that CAT-DIFF, outperforms other methods for German, whereas LaBSE demonstrates superior performance for Swahili. Specifically, CAT-DIFF(1,5) surpasses the other techniques for German at each pruning level and retains 99% of the full data's performance even at 50% prune level (refer to Figure 4a). We observe similar trends for French in Figure 4b with CAT-DIFF(1,5) also generally outperforms LaBSE and random selection, except for FLORES at the 50% prune level. Conversely, for Swahili, LaBSE is consistently superior to other pruning techniques, followed by CAT-DIFF(1,5) (see Figure 4c). We also find that pruning using LaBSE for Swahili yields competitive performance as training on the full dataset.

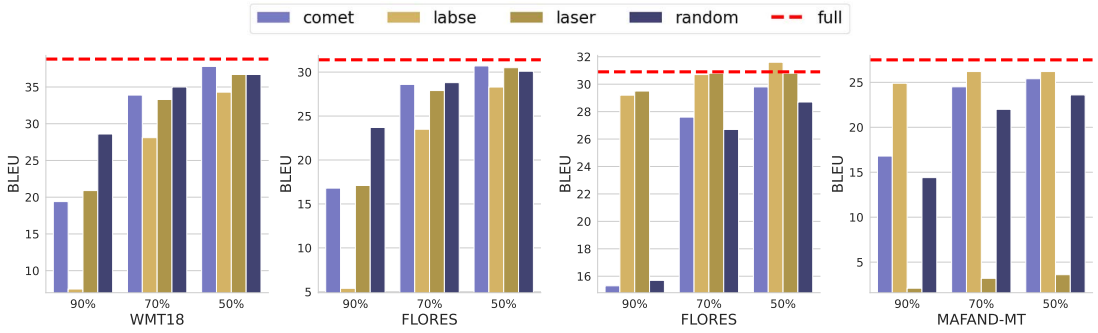
We also report the statistical significance results with paired bootstrap resampling(Koehn, 2004) for SacreBLEU and COMET in the Appendix, see Tables 4 to 6.

		WMT18			FLORES			MAFAND-MT		
		90%	*70%*	*50%*	*90%*	*70%*	*50%*	*90%*	*70%*	*50%*
German	CAT-DIFF(1,5)	**34.1**	**38.1**	**38.7**	**28.4**	**30.9**	**31.0**			
	CAT-DIFF(1,2)	33.4	37.7	38.1	27.7	30.1	30.9			
Swahili	CAT-DIFF(1,5)				**25.0**	29.7	**30.3**	**21.0**	**25.0**	25.3
	CAT-DIFF(1,2)				23.7	**30.1**	30.0	20.6	24.6	**25.6**

Table 2: Ablation study of the CAT-DIFF techniques for both German and Swahili. CAT-DIFF(1,5) outperforms CAT-DIFF(1,2) except for Swahili at 70% and 50% prune levels for FLORES and MAFAND-MT respectively. As a result, we use CAT-DIFF(1,5) in all experiments.

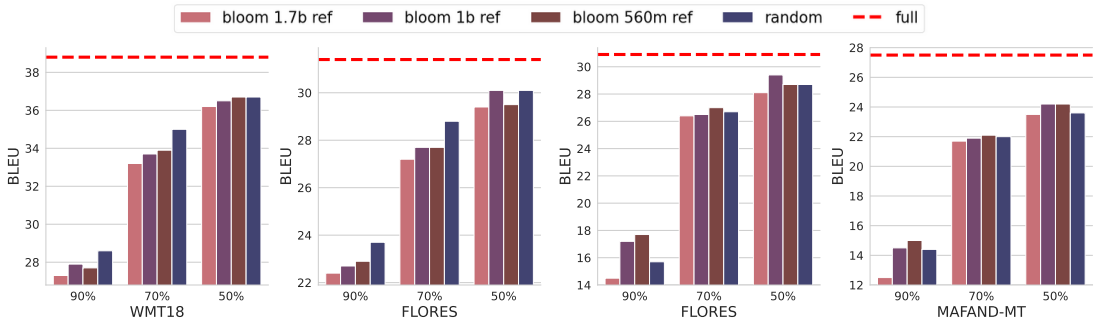
5.5 Scaling best pruning techniques

To verify the generalizability of CAT-DIFF along with baselines such as LaBSE and Random Selection, we trained NMT models on the entire dataset,



(a) German Translation Quality (b) Swahili Translation Quality

Figure 2: Performance of the translation quality metrics for both German and Swahili. Quality estimation metrics perform better than random in an automatically aligned setting (Swahili) than in the high-quality setting (German).



(a) German LLM (b) Swahili LLM

Figure 3: Performance of the LLM perplexity metrics for both German and Swahili. For German, the LLM perplexity metrics perform worse than random whereas, for Swahili, the LLM perplexity metrics are approximately the same as random.

∼23M parallel sentences for Swahili and ∼38M parallel sentences for German and French. We compare CAT-DIFF(1,5), LABSE and random data pruning at 90% prune level. We find the results to be similar to 3.8M scale experiments. Table 3 show that CAT-DIFF(1,5) outperforms random in German and French while pruning using LaBSE performs better in Swahili.

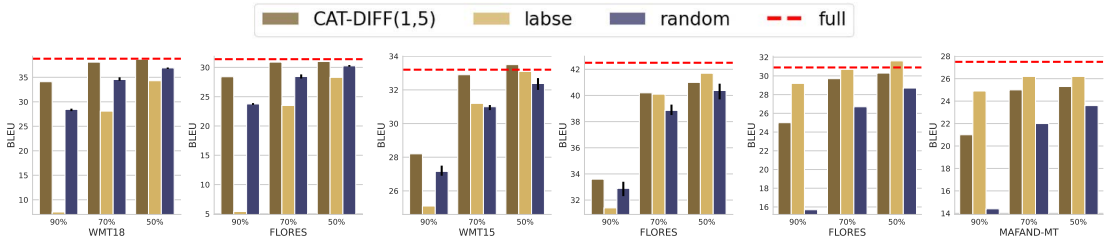
6 Analysis of Sentence Selection for German, French, and Swahili

To better understand the performance gap between pruning techniques for German, French(Indo-European languages) and Swahili(a Bantu language), we analyze the sentences selected randomly as well as those selected using CAT-DIFF(1,5) and LaBSE. Figure 5 illustrates that longer sentences correlate with higher BLEU scores for German and French when considering target sentence length. This observation of longer sequence lengths giving better scores is also seen in Junczys-Dowmunt (2019). Most of the points in the top right quadrant are chosen by CAT-DIFF(1,5), indicating a preference for longer sentences and performs better. However, we observe that picking the longest sentences for Swahili doesn't lead to better BLEU scores. We, therefore also look at the lexical diversity to distinguish the sentences selected.

Language	Test Set	Random	LaBSE	CAT-DIFF(1,5)	Full
German	WMT18	38.8	9.2	40.3	42.6
	FLORES	31.4	6.4	32.1	34.3
Swahili	FLORES	29.3	33.6	20.3	33.0
	MAFAND-MT	24.5	28.9	15.3	27.5
French	WMT15	33.2	29.2	34.7	35.8
	FLORES	42.5	36.2	43.6	45.1

Table 3: BLEU scores for German, French and Swahili on the entire datasets (i.e. ∼38M for German and French and ∼23M for Swahili). Random and CAT-DIFF are performed with 10% sparsity (i.e. 90% prune level).

Figure 6 shows that CAT-DIFF(1,5) selects sentences that have more unique and rare words as opposed to German which is less lexically diverse and thus results in lower mean frequency. Simi-



(a) German (b) French (c) Swahili

Figure 4: Best performing techniques against the performance of full set. For German and French, CAT-DIFF(1,5) achieves above 77% of the performance of training on the full set for the 90% prune level. At 50% pruning, CAT-DIFF(1,5) achieves above 92% of the full training performance. The error bars on the German and French plots represent the minimum and maximum BLEU score over 5 runs with different random seeds for data selection.

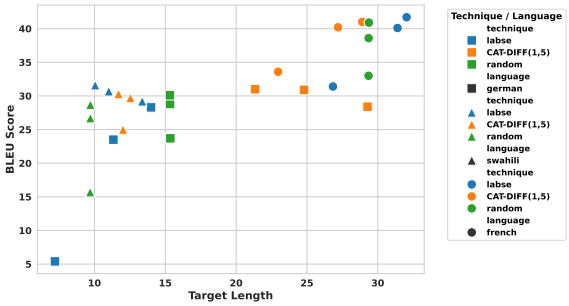


Figure 5: Target sentence length vs BLEU score for all languages and techniques. We see that longer sentences correlate on average with better performance. Notably, pruning techniques like CAT-DIFF(1,5), which favor longer sentences, achieve higher BLEU scores.

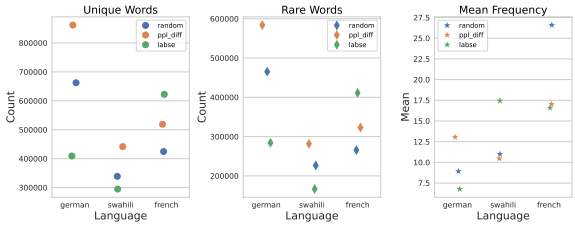


Figure 6: Analysis of word characteristics selected by different pruning techniques. CAT-DIFF(1,5) tends to select more unique and rare words for German and Swahili and the words selected are characterized by a lower mean word frequency. In contrast, for French, LaBSE favors unique and rare words, and its selection is less pronounced compared to CAT-DIFF(1,5)'s for German. These statistics illustrate that pruning techniques prioritize nuanced examples across languages without adhering to a uniform pattern.

larly, CAT-DIFF(1,5) also selects more unique and rare words for Swahili. Although LaBSE performs relatively better in selecting more unique and rare words for French, it still falls short compared selection of words using CAT-DIFF(1,5) for German. This implies that CAT-DIFF(1,5) selection is far more complex than selecting the longest sentences. We leave a deeper investigation of sentences selected using various techniques for future studies.

7 Conclusion

Identifying optimal subsets of a given dataset for training a machine translation model is a crucial problem that enables data efficiency and saves on compute. In this paper, we present *Checkpoints Across Time* (CAT) methods as efficient techniques to prune datasets. CAT uses the change in perplexity scores of individual examples over the first few training epochs to find good subsets. Using 3.8M sentences for English-German, English-French and English-Swahili, we show that CAT techniques offer highly computationally efficient pruning that achieves above 75% of full performance even when 90% pruning is applied to all target languages. We also show that LaBSE, when used to select training examples from an automatically aligned dataset provides strong signals on examples crucial for model generalization as the examples selected perform on par or surpass training on the full dataset. Interestingly, we find that utilizing LLMs for pruning does not offer benefits in selecting data for German but does for Swahili.

8 Limitations

In this work, we focus on data pruning in a bilingual setting for 3 languages and limit the data size to 3.8M sentences. This is due to computation constraints. This means that the results may not apply to other languages. Also, it would be interesting to investigate performance on larger scales as we only conducted experiments on German and English for CAT-DIFF and random selection. Moreover, fur-

ther research is needed to explore the multilingual setup where data pruning would be conducted simultaneously on several languages.

9 Acknowledgements

EAC acknowledges a grant from the Carnegie Corporation of New York (provided through the African Institute for Mathematical Sciences). The statements made and views expressed are solely the responsibility of the authors.

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.

Alessandro Achille, Matteo Rovere, and Stefano Soatto. 2017. Critical learning periods in deep neural networks. *CoRR*, abs/1711.08856.

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, Md Mahfuz Ibn Alam, Antonios Anastasopoulos, Akshita Bhagia, Marta R. Costa-jussà, Jesse Dodge, Fahim Faisal, Christian Federmann, Natalia Fedorova, Francisco Guzmán, Sergey Koshelev, Jean Maillard, Vukosi Marivate, Jonathan Mbuya, Alexandre Mourachko, Safiyyah Saleem, Holger Schwenk, and Guillaume Wenzek. 2022b. Findings of the WMT'22 shared task on large-scale machine translation evaluation for African languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 773–800, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Chirag Agarwal, Daniel D'souza, and Sara Hooker. 2022. Estimating example difficulty using variance of gradients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10368–10378.

Orevaoghene Ahia, Julia Kreutzer, and Sara Hooker. 2021. The low-resource double bind: An empirical study of pruning for low-resource machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3316–3333, Punta Cana, Dominican Republic. Association for Computational Linguistics.

N. Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George F. Foster, Colin Cherry, Wolfgang Macherey, Z. Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *ArXiv*, abs/1907.05019.

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.

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*, pages 355–362, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Abdul Azeemi, Ihsan Qazi, and Agha Raza. 2023. Data pruning for efficient model pruning in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 236–246, Singapore. Association for Computational Linguistics.

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

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

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.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi N. Baljekar, Xavier García, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Z. Chen, Yonghui Wu, and Macduff Hughes. 2022. Building machine translation systems for the next thousand languages. *ArXiv*, abs/2205.03983.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Akshay Batheja and Pushpak Bhattacharyya. 2023. "a little is enough": Few-shot quality estimation based corpus filtering improves machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14175–14185, Toronto, Canada. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. 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.

Meriem Boubdir, Edward Kim, Beyza Ermis, Marzieh Fadaee, and Sara Hooker. 2023. Which prompts make the difference? data prioritization for efficient human llm evaluation.

Everlyn Chimoto and Bruce Bassett. 2022. COMET-QE and active learning for low-resource machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4735–4740, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Marta Ruiz Costa-jussà, James Cross, Onur cCelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Alison Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon L. Spruit, C. Tran, Pierre Yves Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzm'an, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *ArXiv*, abs/2207.04672.

Praveen Dakwale, Talaat Khalil, and Brandon Denis. 2022. Empirical evaluation of language agnostic filtering of parallel data for low resource languages. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 346–355, Manila, Philippines. De La Salle University.

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.

Kevin Duh, Graham Neubig, Katsuhito Sudoh, and Hajime Tsukada. 2013. Adaptation data selection using neural language models: Experiments in machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 678–683, Sofia, Bulgaria. Association for Computational Linguistics.

Fartash Faghri, David Duvenaud, David J. Fleet, and Jimmy Ba. 2020. A study of gradient variance in deep learning. *CoRR*, abs/2007.04532.

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, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

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.

Mandy Guo, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernandez Abrego, Keith Stevens, Noah Constant, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Effective parallel corpus mining using bilingual sentence embeddings. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 165–176, Brussels, Belgium. Association for Computational Linguistics.

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.

Dan Hendrycks and Kevin Gimpel. 2016. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *CoRR*, abs/1606.08415.

Christian Herold, Jan Rosendahl, Joris Vanvinckenroye, and Hermann Ney. 2022. Detecting various types of noise for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2542–2551, Dublin, Ireland. Association for Computational Linguistics.

Ziheng Jiang, Chiyuan Zhang, Kunal Talwar, and Michael C. Mozer. 2020. Exploring the memorization-generalization continuum in deep learning. *CoRR*, abs/2002.03206.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

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*, pages 888–895, Belgium, Brussels. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2019. Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 225–233, Florence, Italy. 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.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *International Conference On Learning Representations*.

Muhammed Kocyigit, Jiho Lee, and Derry Wijaya. 2022. Better quality estimation for low resource corpus mining. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 533–543, Dublin, Ireland. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. 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.

Sai Koneru, Danni Liu, and Jan Niehues. 2022. Cost-effective training in low-resource neural machine translation. *CoRR*, abs/2201.05700.

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.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Junyu Li, Ligang He, Shenyuan Ren, and Rui Mao. 2018. Data fine-pruning: A simple way to accelerate neural network training. In *Network and Parallel Computing*, pages 114–125, Cham. Springer International Publishing.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, Xinyi Wu, Enrico Shippole, Kurt Bollacker, Tongshuang Wu, Luis Villa, Sandy Pentland, and Sara Hooker. 2023. The data provenance initiative: A large scale audit of dataset licensing attribution in ai. *arXiv preprint arXiv: 2310.16787*.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. 2023. When less is more: Investigating data pruning for pretraining llms at scale. In *NeurIPS Workshop on Attributing Model Behavior at Scale*.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden. Association for Computational Linguistics.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *International Conference on Machine Learning*.

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 NAACL-HLT 2019: Demonstrations*.

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.

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, Brussels, Belgium. Association for Computational Linguistics.

Ravi Raju, Kyle Daruwalla, and Mikko H. Lipasti. 2021. Accelerating deep learning with dynamic data pruning. *ArXiv*, abs/2111.12621.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online. Association for Computational Linguistics.

Machel Reid, Junjie Hu, Graham Neubig, and Yutaka Matsuo. 2021. AfroMT: Pretraining strategies and reproducible benchmarks for translation of 8 African languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1306–1320, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa

Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv: 2211.05100*.

Holger Schwenk. 2018. Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–234, Melbourne, Australia. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021a. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021b. 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.

Shoaib Ahmed Siddiqui, Nitarshan Rajkumar, Tegan Maharaj, David Krueger, and Sara Hooker. 2023. Metadata archaeology: Unearthing data subsets by leveraging training dynamics. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. 2022. Beyond neural scaling laws: beating power law scaling via data pruning. In *Advances in Neural Information Processing Systems*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

Steinþór Steingrímsson, Pintu Lohar, Hrafn Loftsson, and Andy Way. 2021. Effective bitext extraction from comparable corpora using a combination of three different approaches. In *Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021)*, pages 8–17, Online (Virtual Mode). INCOMA Ltd.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826.

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.

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2017. Dynamic data selection for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1410, Copenhagen, Denmark. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Pavanpankaj Vegi, Sivabhavani J, Biswajit Paul, Abhinav Mishra, Prashant Banjare, Prasanna K R, and Chitra Viswanathan. 2022. WebCrawl African : A multilingual parallel corpora for African languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1076–1089, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Wei Wang, Taro Watanabe, Macduff Hughes, Tetsuji Nakagawa, and Ciprian Chelba. 2018. Denoising neural machine translation training with trusted data and online data selection. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 133–143, Brussels, Belgium. Association for Computational Linguistics.

Jiarong Xu, Renhong Huang, XIN JIANG, Yuxuan Cao, Carl Yang, Chunping Wang, and Yang Yang. 2023. Better with less: A data-active perspective on pre-training graph neural networks. In *Thirty-seventh*

Conference on Neural Information Processing Systems.

Boliang Zhang, Ajay Nagesh, and Kevin Knight. 2020. Parallel corpus filtering via pre-trained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8545–8554, Online. Association for Computational Linguistics.

A Dataset Analysis

Figure 7 and Figure 8 provide analysis of the 3.8M dataset for German, Swahili and French. Both German and French exhibit outliers in the sentence length with some sentences being 500 words or longer while Swahili seems to be filtered by sentence length as no sentence has more than 250 words. See Figure 7. We also see that foreign languages were removed from the Swahili dataset as both source and target sentences do not contain any other language whereas both German and French contain foreign language.

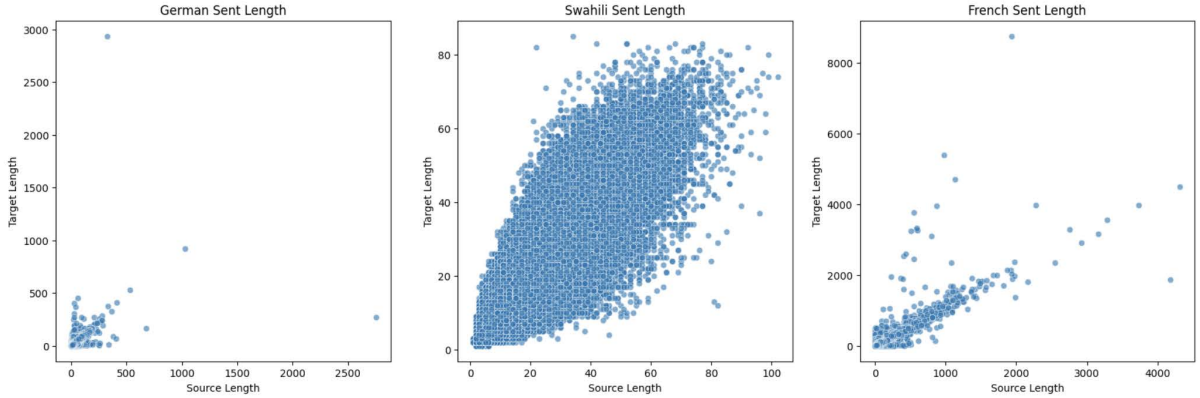


Figure 7: Source (English) sentence length vs Target (German, Swahili, French) sentence length. German and French show high variance while Swahili shows low variance.

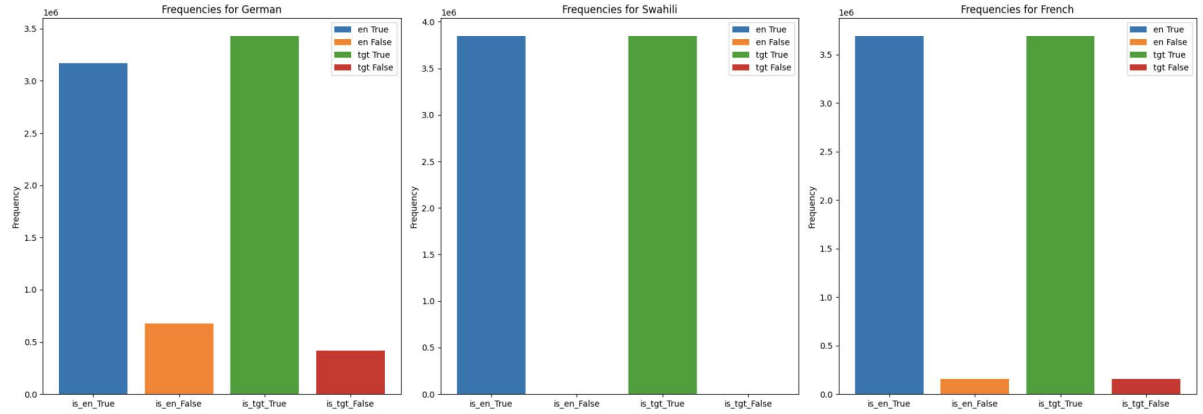


Figure 8: English sentences Language ID (en) vs Target (German, Swahili, French) sentence length (tgt). Swahili set does not contain any foreign languages while German and French seem to contain other languages.

B Other evaluation metrics and significance tests

We ran additional experiments to calculate chrF++ and COMET scores and report the statistical significance results with paired bootstrap resampling (Koehn, 2004) using sacreBLEU and COMET libraries with default configurations for the respective metrics below for all our experiments with on the 3.8M data scale. We find that CAT-DIFF demonstrates competitive or superior performance compared to other methods on en-de and en-fr pairs across most test sets. Bolded numbers in Tables 4 to 6 show that with 1000 resamples, CAT-DIFF outperforms the other technique 95% of the time thus CAT-DIFF is statistically significance at 95% confidence level. These results highlight the effectiveness of our proposed method.

Swahili	*FLORES*			*MAFAND-MT*		
	90%	*70%*	*50%*	*90%*	*70%*	*50%*
			BLEU			
CAT DIFF(1,5)	25.0 ± 1.0	29.7 ± 1.0	30.3 ± 1.1	21.0 ± 0.8	25.0 ± 0.8	25.3 ± 0.9
Random	**15.7 ± 0.8**	**26.7 ± 0.9**	28.7 ± 1.0	**14.4 ± 0.6**	**22.0 ± 0.8**	**23.6 ± 0.8**
LaBSE	29.2 ± 1.1	30.7 ± 1.0	31.6 ± 1.0	24.9 ± 0.9	26.2 ± 0.9	26.2 ± 0.9
LASER	**0.4 ± 0.1**	**1.4 ± 0.3**	**1.5 ± 0.3**	**2.1 ± 0.3**	**3.2 ± 0.4**	**3.6 ± 0.4**
COMET	**15.3 ± 0.9**	**27.6 ± 1.0**	29.8 ± 1.0	**16.8 ± 0.8**	**24.5 ± 0.8**	25.4 ± 0.9
BLOOM 560m	**17.7 ± 0.9**	**27.0 ± 0.9**	**28.7 ± 1.0**	**15.0 ± 0.7**	**22.1 ± 0.8**	**24.2 ± 0.8**
BLOOM 1b	**17.2 ± 0.9**	**26.5 ± 1.0**	**29.4 ± 1.1**	**14.5 ± 0.6**	**21.9 ± 0.8**	**24.2 ± 0.8**
BLOOM 1.7b	**14.5 ± 0.8**	**26.4 ± 1.0**	**28.1 ± 1.0**	**12.5 ± 0.6**	**21.7 ± 0.8**	**23.5 ± 0.8**
Full	30.9 ± 1.0	30.9 ± 1.0	30.9 ± 1.0	27.5 ± 1.0	27.5 ± 1.0	27.5 ± 1.0
			chrF++			
CAT DIFF(1,5)	52.2 ± 0.7	56.0 ± 0.7	56.4 ± 0.8	47.2 ± 0.7	50.9 ± 0.7	51.0 ± 0.7
Random	**42.7 ± 0.7**	**53.5 ± 0.7**	**55.0 ± 0.7**	**39.8 ± 0.6**	**48.5 ± 0.6**	**49.7 ± 0.7**
LaBSE	55.9 ± 0.8	57.0 ± 0.7	57.5 ± 0.7	51.1 ± 0.6	52.3 ± 0.7	52.3 ± 0.7
LASER	**8.2 ± 0.3**	**13.0 ± 0.3**	**13.2 ± 0.3**	**12.5 ± 0.3**	**14.6 ± 0.4**	**15.5 ± 0.4**
COMET	**42.2 ± 0.7**	**54.2 ± 0.7**	56.2 ± 0.7	**42.3 ± 0.7**	**50.6 ± 0.7**	**51.3 ± 0.7**
BLOOM 560m	**44.9 ± 0.7**	**53.6 ± 0.7**	**55.3 ± 0.8**	**41.2 ± 0.6**	**48.6 ± 0.6**	**50.1 ± 0.7**
BLOOM 1b	**44.1 ± 0.7**	**53.2 ± 0.7**	**55.5 ± 0.8**	**40.2 ± 0.6**	**48.2 ± 0.7**	**50.3 ± 0.7**
BLOOM 1.7b	**40.7 ± 0.7**	**53.0 ± 0.8**	**54.7 ± 0.8**	**37.4 ± 0.5**	**48.1 ± 0.7**	**49.6 ± 0.7**
Full	56.8 ± 0.7	56.8 ± 0.7	56.8 ± 0.7	52.9 ± 0.7	52.9 ± 0.7	52.9 ± 0.7
			COMET			
CAT DIFF(1,5)	75.81	79.25	79.49	76.38	80.18	80.45
Random	**68.97**	**77.66**	**79.05**	**71.30**	**78.39**	**79.66**
LaBSE	78.29	79.72	80.36	80.30	81.65	81.75
LASER	**30.82**	**41.91**	**44.56**	**40.43**	**45.88**	**48.23**
COMET	**71.42**	79.39	80.43	**75.70**	81.33	81.18
BLOOM 560m	**70.19**	**77.69**	79.10	**70.91**	**78.74**	80.27
BLOOM 1b	**69.31**	**77.12**	**78.95**	**69.84**	**78.43**	**79.90**
BLOOM 1.7b	**68.15**	**77.34**	**78.56**	**71.85**	**78.33**	**79.48**
Full	80.01	80.01	80.01	81.81	81.81	81.81

Table 4: Significance test for Swahili with the null hypothesis: mean translation score of CAT-DIFF is equal to the mean of the other technique. A bolded entry signifies that CAT-DIFF is significantly better (p-value < 0.05) than the method used for the corresponding value. For each dataset, we show results for different prune levels (90%, 70%, 50%) and metrics (BLEU, chrF++, COMET). Here "Full" means the value obtained when using all the training data.

German	*FLORES*			*WMT18*		
	90%	*70%*	*50%*	*90%*	*70%*	*50%*
			BLEU			
CAT DIFF(1,5)	28.3 ± 1.0	30.9 ± 1.1	31.0 ± 1.1	34.1 ± 0.7	38.1 ± 0.7	38.7 ± 0.8
Random	**23.7 ± 1.0**	**28.8 ± 1.1**	30.0 ± 1.1	**28.6 ± 0.7**	**34.9 ± 0.7**	**36.7 ± 0.8**
LaBSE	**5.4 ± 0.6**	**23.5 ± 1.0**	**28.3 ± 1.0**	**7.5 ± 0.5**	**28.1 ± 0.6**	**34.3 ± 0.7**
LASER	**17.1 ± 0.8**	**27.8 ± 1.0**	**30.5 ± 1.1**	**20.8 ± 0.6**	**33.3 ± 0.7**	**36.7 ± 0.8**
COMET	**16.7 ± 0.9**	**28.5 ± 1.1**	30.7 ± 1.1	**19.4 ± 0.6**	**33.9 ± 0.7**	**37.8 ± 0.7**
BLOOM 560m	**23.0 ± 1.0**	**27.7 ± 1.0**	**29.5 ± 1.1**	**28.0 ± 0.6**	**33.8 ± 0.7**	**36.6 ± 0.7**
BLOOM 1b	**22.6 ± 1.0**	**27.6 ± 1.0**	**29.5 ± 1.0**	**27.7 ± 0.6**	**33.9 ± 0.7**	**36.7 ± 0.7**
BLOOM 1.7b	**22.3 ± 0.9**	**26.9 ± 1.0**	**29.9 ± 1.1**	**27.1 ± 0.6**	**33.5 ± 0.7**	**35.9 ± 0.7**
Full	31.5 ± 1.2	31.5 ± 1.2	31.5 ± 1.2	38.8 ± 0.8	38.8 ± 0.8	38.8 ± 0.8
			chrF++			
CAT DIFF(1,5)	55.3 ± 0.8	57.5 ± 0.7	57.7 ± 0.8	59.2 ± 0.5	62.2 ± 0.5	62.6 ± 0.5
Random	**51.5 ± 0.7**	**55.5 ± 0.8**	**56.8 ± 0.8**	**54.6 ± 0.5**	**59.5 ± 0.5**	**61.2 ± 0.5**
LaBSE	**24.6 ± 0.7**	**51.4 ± 0.8**	**55.6 ± 0.8**	**27.9 ± 0.4**	**54.6 ± 0.5**	**59.4 ± 0.5**
LASER	**44.5 ± 0.8**	**55.3 ± 0.8**	**57.2 ± 0.8**	**47.4 ± 0.5**	**58.8 ± 0.5**	**61.2 ± 0.5**
COMET	**44.1 ± 0.7**	**55.2 ± 0.8**	**57.1 ± 0.8**	**46.2 ± 0.5**	**58.8 ± 0.5**	**61.6 ± 0.5**
BLOOM 560m	**50.9 ± 0.7**	**54.8 ± 0.8**	**56.2 ± 0.7**	**54.3 ± 0.5**	**58.8 ± 0.5**	**60.9 ± 0.5**
BLOOM 1b	**50.7 ± 0.7**	**54.8 ± 0.7**	**56.2 ± 0.7**	**54.0 ± 0.5**	**59.0 ± 0.5**	**60.9 ± 0.5**
BLOOM 1.7b	**50.3 ± 0.7**	**54.5 ± 0.7**	**56.5 ± 0.8**	**53.5 ± 0.5**	**58.6 ± 0.5**	**60.5 ± 0.5**
Full	57.7 ± 0.8	57.7 ± 0.8	57.7 ± 0.8	62.7 ± 0.5	62.7 ± 0.5	62.7 ± 0.5
			COMET			
CAT DIFF(1,5)	77.80	80.38	80.93	78.92	82.11	82.23
Random	**70.77**	**78.70**	**79.59**	**71.30**	**79.14**	**80.83**
LaBSE	**46.39**	**69.46**	**76.43**	**48.09**	**70.75**	**78.00**
LASER	**60.26**	**76.14**	**79.31**	**61.04**	**77.08**	**80.48**
COMET	**57.70**	**78.46**	81.69	**58.65**	**78.80**	82.59
BLOOM 560m	**69.70**	**77.34**	**79.00**	**71.18**	**78.15**	**80.93**
BLOOM 1b	**70.35**	**76.82**	**79.25**	**70.72**	**78.17**	**80.69**
BLOOM 1.7b	**69.58**	**76.83**	**78.38**	**70.14**	**77.77**	**80.06**
Full	80.86	80.86	80.86	82.81	82.81	82.81

Table 5: Significance test for German with the null hypothesis: mean translation score of CAT-DIFF is equal to the mean of the other technique. A bolded entry signifies that CAT-DIFF is significantly better (p-value < 0.05) than the method used for the corresponding value. For each dataset, we show results for different prune levels (90%, 70%, 50%) and metrics (BLEU, chrF++, COMET). Here "Full" means the value obtained when using all the training data.

French	*FLORES*			*WMT15*		
	90%	*70%*	*50%*	*90%*	*70%*	*50%*
			BLEU			
CAT DIFF(1,5)	33.5 ± 1.1	40.1 ± 1.3	41.0 ± 1.3	28.2 ± 1.0	32.8 ± 1.1	33.5 ± 1.1
Random	32.9 ± 1.2	**38.6 ± 1.3**	40.9 ± 1.3	**26.9 ± 1.0**	**31.0 ± 1.1**	**32.5 ± 1.1**
LaBSE	**31.4 ± 1.3**	40.1 ± 1.4	41.6 ± 1.4	**25.1 ± 1.1**	**31.2 ± 1.1**	33.1 ± 1.1
Full	42.4 ± 1.3	42.4 ± 1.3	42.4 ± 1.3	33.1 ± 1.1	33.1 ± 1.1	33.1 ± 1.1
			chrF++			
CAT DIFF(1,5)	57.5 ± 0.8	62.6 ± 0.9	63.0 ± 0.9	53.1 ± 0.8	56.7 ± 0.9	57.3 ± 0.9
Random	57.3 ± 0.9	**61.4 ± 0.9**	62.9 ± 1.0	**51.8 ± 0.8**	**55.5 ± 0.8**	**56.6 ± 0.8**
LaBSE	**55.3 ± 1.1**	62.3 ± 1.0	63.7 ± 1.0	**49.4 ± 1.1**	**55.9 ± 0.9**	57.4 ± 0.9
Full	63.9 ± 0.9	63.9 ± 0.9	63.9 ± 0.9	57.3 ± 0.9	57.3 ± 0.9	57.3 ± 0.9
			COMET			
CAT DIFF(1,5)	77.15	82.60	82.94	72.50	77.19	78.08
Random	**75.44**	**81.02**	**82.41**	**70.74**	**75.58**	**77.02**
LaBSE	**73.81**	**81.22**	82.86	**69.72**	**75.82**	**77.28**
Full	83.78	83.78	83.78	78.31	78.31	78.31

Table 6: Significance test for French with the null hypothesis: mean translation score of CAT-DIFF is equal to the mean of the other technique. A bolded entry signifies that CAT-DIFF is significantly better (p-value < 0.05) than the method used for the corresponding value. For each dataset, we show results for different prune levels (90%, 70%, 50%) and metrics (BLEU, chrF++, COMET). Here "Full" means the value obtained when using all the training data.