Adapters and Reinforcement Learning for Data-Efficient Machine Translation

Shree Vignesh S, Abigail Kufeldt, Malini Kar, Parikshith Honnegowda and Pranjali Basmatkar

Abstract

Significant progress has been made in the field of machine translation, and the recent advancements in large language models have largely complemented this progress. Though translation systems have made significant progress on high-resource languages their attention has been wanting on low-resource languages. This work presents a new approach for enhancing the performance of low-resource machine translation using "Monolingual Adapters" and "Reinforced Self-Training". Our findings indicate that this approach - of decoupling adapters into their respective language components along with self-referential training using reinforcement learning - achieves promising results for low-resource machine translation, which is both compute and data-efficient. On the WMT22 dataset of four low-resource African languages—Yoruba, Igbo, Hausa, and Swahili, our monolingual adapter-based approach improves the translation BLEU of English-Igbo from 5.4 to 6.5 ($\approx$ 20% increase) and Hausa-Igbo from 1.6 to 2.4 ($\approx$ 50% increase). Furthermore, our reinforced self-training approach resulted in a notable performance boost of 1.8 BLEU points ($\approx$ 12% increase) over the baseline translation model for Hausa-English using just monolingual data from paracrawl corpus Bañón et al. (2020). Our strategy, while boosting the translation performance, is language agnostic and can be scaled to boost translation between other low-resource languages.

Contents

1. Introduction

The internet serves as a gateway to the world, providing opportunities for education, communication, commerce, and much more. Yet, a large chunk of the African population lacks access to the internet, and for the population with access, there is a major challenge — much of the content is unavailable in their local languages (Union, 2020). This lack of local language support hinders these communities from fully harnessing the potential of the Internet.

Contemporary methodologies in the realm of machine learning have demonstrated astounding triumphs across a multitude of natural language processing undertakings. These endeavors span across the fields of machine translation (MT), sentiment analysis, and speech recognition, to name a few. However, when one attempts to employ these potent techniques in Resource Languages (LRL), they face a unique conundrum due to insufficient data and resources.

Low Resource Languages are languages for which limited or even nonexistent data is available. This lack of data makes building robust language models or translation systems that can support these languages challenging. Adding to the complexity, data in these languages often lack labels — meaning they are unaligned with parallel sentences in the source language. The absence of labels necessitates using clever techniques to produce parallel data to train these models, thereby widening the language gap in digital content.

In light of these challenges, our project, "Adapters and Reinforcement Learning for Data-Efficient Machine Translation," aims to address and rectify this translation disparity by designing a reliable and efficient approach to LRL MT and demonstrating its success on our low-resource languages of interest. Our goal is to develop systems that make the Internet more accessible and inclusive to all, regardless of the language they speak.

The languages of interest in this work are four languages spoken in Africa: Swahili, Yoruba, Igbo, and Hausa. These languages alone each cover roughly 50 million speakers (native or second language) across central and southeast Africa, and yet common translation tools like Google Translate or even ChatGPT have no ability to effectively translate between them (Wikipedia, 2023b). A lack of training data necessitates using specialized techniques to leverage available data and benefit from the cross-lingual transfer of similar languages.

As for the typology of these languages, Hausa belongs to the Afro-Asiatic language family (Wikipedia, 2023a). Igbo and Yoruba are categorized as Niger-Congo languages (Wikipedia, 2023b). Crucially, Swahili is primarily described as a Bantu language, but some linguists will use Niger-Congo as an umbrella term above the designation Bantu (Wikipedia, 2023c). Thus, it's not particularly clear, linguistically, whether Swahili is best considered as a typologically distinct language (being Bantu) from Igbo and Yoruba or whether they share enough features to be treated as the same family (Niger-Congo). This gives us multiple different typology configurations to investigate cross-lingual transfer in our experiments down the line.

To mitigate the data constraints in LRL, we will explore current state-of-the-art techniques in Multilingual Neural Machine Translation (MNMT). Multilingual translation systems perform better than their bilingual counterparts because they inherently benefit from cross-lingual transfer from the High Resource Languages (HRL) they have been exposed to. Leveraging linguistic information from a data-rich language reduces the MNMT system's reliance on limited LRL data.

To further enhance the data pool, we aim to investigate the use of unsupervised learning techniques. We are also planning to use language adapters, which can serve as a powerful tool that allows models trained on languages rich in resources to be proficiently adjusted and made applicable to languages deficient in resources.

This work would essentially strive to break down digital language barriers and foster an inclusive digital ecosystem. By developing a reliable method for LRL MT, we can help bridge the language divide, providing everyone equal access to the wealth of knowledge, tools, and opportunities the internet offers.

2. Related Work

To align our research direction and construct our approach, we scrutinized extensive prior work in three main areas: Automatic Speech Recognition (ASR), Machine Translation (MT), and Text-to-Speech (TTS) synthesis.

The paper (Gibadullin et al., 2021) focuses on applying deep recurrent neural networks for English-Russian translation from speech to text. The authors compiled a neural network learning dataset by splitting an English audiobook. They used modern approaches of deep learning, such as a long short-term memory (LSTM) network, an encoder-decoder architecture, an attention mechanism, and ray search. The calculations were carried out on GPUs using the TensorFlow software environment.

Regarding low-resource language translation, automatic speech recognition techniques can be used to transcribe speech in the source language into text and then use machine translation techniques to translate the text into the target language. This approach can be useful when limited data is available for training machine translation models. However, it is important to note that the accuracy of this approach is dependent on the quality of the ASR system and the availability of parallel data for training machine translation models.

Another paper relevant to ASR that we explored was Massively Multilingual Speech (MMS), which aims to expand speech technology to support up to 1,107 different languages (Pratap et al., 2023). The authors use a labeled dataset to finetune pretrained ASR models, and then scale the number of languages for multilingual ASR from 61 to 1,107. They compare their models to existing multilingual work and build robust multilingual models supporting 1,162 languages trained on several existing corpora and the MMS-lab data. Finally, they evaluate their multilingual models on all languages they support. Crucially, these multilingual ASR models have been exposed to all four of our target languages — Igbo, Hausa, Swahili, and Yoruba.

One potential application of this technology is LRL translation. By using ASR models to transcribe speech in a low-resource language and then translating that text into a more widely spoken language, we can improve accessibility for people who speak those LRLs. This can be especially useful in areas where many different local languages are spoken but few resources are available for translation.

In the realm of MT, we were particularly intrigued by prior work in Active Learning and Language Adapters, which comprise the centerpiece of our research interests in this project. As for Active Learning, Zhao et al. (2020)'s work on Iterative Back Translation proposes a novel strategy to improve translation accuracy by iteratively translating monolingual data and training models without human annotators. This approach offers a promising way to leverage limited data, but its real-world applicability, particularly for low-resource machine translation is uncertain. Also this would necessitate the need for training a back translation model which would be computationally costly. Instead we focus our attention on reinforcement learning and specifically on reinforced self-training Gulcehre et al. (2023). This approach leverages a trained translation model and iteratively improve its performance by sampling translations from a policy (here the trained translation model is the policy) in conjunction with a suitable reward function.

Simultaneously, we dove into the concept of language adapters. We analyzed various studies that proposed different variations of adapters, their underlying mechanisms, and impacts on cross-lingual transfer in multilingual translation models. We were particularly interested in the idea of comparing language adapters alone versus their combination with family adapters, keeping in mind the linguistic typologies of our target languages.

The approach described in the paper Adelani et al. (2022a) uses pre-trained multilingual encoder-decoder models, specifically DeltaLM, which are fine-tuned with the allowed data sources for the WMT22 shared task on large-scale machine translation evaluation for African languages. The authors also incorporate language family and language-specific adapter units to enhance the performance of their translation model. Adelani et al. report that their approach, which incorporates language adapters and fine-tuning with allowed data sources, outperforms the benchmark for the constrained translation track of the WMT22 shared task on large-scale machine translation evaluation for African languages. Specifically, their best submission ranked second under the constrained setting.

As evidenced by Adelani et al. (2022a)'s work, using language adapters in MNMT facilitates better acquisition of the target language's specific linguistic features, which can improve translation quality even with limited training data. Additionally, fine-tuning with allowed data sources can help to further improve performance by incorporating additional relevant data.

Thus, after investigating prior work in LRL MT, we identified a handful of techniques useful for the given task — namely, reinforced self-training, and language adapters. These methods collectively demonstrate significant success in leveraging limited data and cross lingual transfer to improve translation quality, making them ideal for our task. To the best of our knowledge, these techniques have not been integrated into one coherent system for LRL MNMT, despite their individual success. Thus, we seek to consolidate these high-performing LRL techniques to build a language-agnostic approach to LRL NMT that can be used on any language pair in the future, satisfying our original goal of combatting the digital divide.

3. Dataset

For our text-to-text translation task, we chiefly explored the datasets involved in the WMT'22 data track. For the evaluation of our translation model, we intend to use Flores-200. For our ASR component, we will use FLEURS (Conneau et al., 2022).

3.1. MAFAND-MT African News Corpus

MAFAND-MT, or Masakhane Anglo & Franco Africa News Dataset for Machine Translation, involves eleven African languages, including six languages spoken predominantly in Francophone Africa and five languages spoken predominantly in Anglophone Africa (Adelani et al., 2022b). The data has been crawled and preprocessed from news websites from local newspapers. The articles were made certain to contain diverse topics (e.g., politics, sports, culture, technology, society, religion, and education). Native speakers of the target language did this with source language proficiency.

Language Pair	Train	Dev	Test	Total
English-Igbo	6998	1500	1500	9998
English-Swahili	30782	1791	1835	34408
English-Yoruba	6644	1544	1558	9746
English-Hausa	3098	1300	1500	5898

Table 1. *Data Splits for Mafand-MT.*

3.2. WebCrawl African

WebCrawl African is a collection of African Multilingual parallel corpora comprising 695,000 (approx) sentence pairs, covering 15 African languages plus English and 73 language pairs (Vegi et al., 2022). African languages covered include Afrikaans - afr, Lingala - lin, Swati - ssw, Amharic - amh, Luganda - lug, Tswana - tsn, Chichewa - nya, Hausa - hau, Oroma - orm, Xhosa - xho, Igbo - ibo, Xitsonga - tso, Yoruba - yor, Swahili - swh, Zulu - zul. This corpus is submitted as part of the large-scale multilingual African shared task organized by WMT 2022.

Language Pair	Total Sentence Pairs	Language Pair	Total Sentence Pairs
English-Igbo	1124	Hausa-Igbo	123
English-Swahili	64506	Hausa-Swahili	3748
English-Yoruba	6308	Hausa-Yoruba	4153
English-Hausa	5650	Igbo-Swahili	185
Igbo-Yoruba	143	Swahili-Yoruba	4135

Table 2. *Data Splits for WebCrawl African.*

3.3. ParaCrawl (Monolingual Data)

ParaCrawl contains derived corpora built from language-classified extracts of the ParaCrawl project (Bañón et al., 2020). It involves data from the Internet Archives, and targets crawls performed in the ParaCrawl project with document-level language classification. It covers 22 of the 24 African languages required for WMT22.

Language Pair	Total Sentences	Average Length	Unique Tokens
Igbo	453,839	19	453,835
Swahili	12,648,480	22	7,480,015
Yoruba	510,412	21	504,264
Hausa	3,520,670	21	3,372,487

Table 3. *Data from ParaCrawl.*

3.4. Evaluation Dataset: Flores-200

FLORES-200 consists of translations from 842 distinct web articles, totaling 3001 sentences (Guzmán et al., 2019). These sentences are divided into three splits: dev, devtest, and test (hidden). On average, sentences are approximately 21 words long.

4. Methodology

Our project uses Meta's M2M100 model as our foundational architecture. M2M100 stands as a linguistically diverse model, covering a staggering array of 100 languages. This model enables direct translation without the need for intermediate pivoting through a common language, making it crucial for our project's multilingual translation needs.

4.1. Monolingual Language Adapters

Our first approach to the problem involves the usage of language adapters (Houlsby et al., 2019). *Adapter* refers to a set of newly introduced weights, typically within the layers of a transformer model. Adapters provide an alternative to fully fine-tuning the model for each downstream task while maintaining performance. They also have the added benefit of requiring as little as 1MB of storage space per task, making them extremely parameter-efficient.

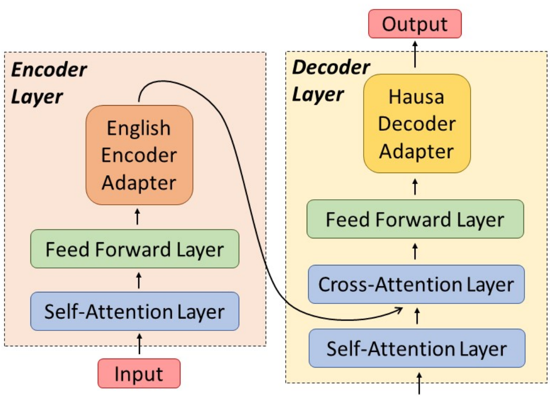

Figure 1. *Monolingual Language Adapter Setup.*

Adapters, applied to each layer of the Transformer, are inserted after the attention and feed-forward layers of the base model as illustrated in Fig. 1 above. A skip connection is applied across these layers, and their outputs undergo layer normalization. Two adapters are inserted after these layers, projecting the original features to a smaller dimension, applying a nonlinearity, and then projecting back to the original dimension. This helps limit parameter count per layer, typically using only 0.5% to 8% of the original model's parameters. Setting a smaller bottleneck dimension makes the trade-off between performance and parameter efficiency manageable. The adapter module, featuring an internal skip connection, initializes itself close to an identity function when projection layer parameters start near zero. With the reduction of parameters, there is an obvious trade-off between performance and parameter efficiency, which we try to balance. The effect of the added adapter layers is that they allow the model to focus on the higher layers of the network, which has generally been found effective in transfer learning.

Monolingual adapters refer to a specific adapter setup proposed in Philip et al. (2020), in which the source and target languages are associated with their corresponding encoder or decoder adapters. In this way, one can train a decoder adapter for the target language of interest and modularly mix and match it with various source language adapters at inference time, for example. This approach is powerful for our work because of the different target languages we're interested in exploring and the need of our final translation system to be flexible and able to perform translation on any language direction.

For this project, we use the publicly available AdapterHub library (Pfeiffer et al., 2020), a central repository for pre-trained adapter modules. AdapterHub offers a standardized interface for adding adapters to existing models and simplifies the process of adapting pre-trained models to new tasks by providing a flexible and modular framework. We first extended the support of AdapterHub to encompass our chosen baseline model, i.e, Meta's M2M100 model (Fan et al., 2020).

Next, we defined a new adapter config type to enable the use of monolingual adapters; the current version of AdapterHub does not allow you to separate the encoder and decoder side adapters, so it proved necessary to edit the original code to allow this particular configuration. In addition to this, we defined a new adapter composition (method of combining multiple adapters), which we called Adapter Pair. This allows two separate adapters to be active simultaneously – in this case, those being the encoder and decoder side adapters.

4.2. ReST

The second approach involves experimenting with the Reinforced self-training methodology Gulcehre et al. (2023). The methodology is grounded in the concept of Reinforced Self-Training (ReST), which is a self-training approach that generates new data based on the model's current understanding, filters these samples based on a reward function, and uses the filtered data to improve the model's performance iteratively. This creates a bootstrapping effect, leading to progressive enhancements in the model. The procedure has three major sections:

- **Grow**: In this phase, the language model policy generates multiple outputs, creating an augmented dataset. This step involves using the initial language model policy, typically a supervised policy, to produce various output predictions for each context. The objective is to expand the training dataset with diverse and rich data samples.
- **Filter**: The augmented dataset generated in the Grow step is then subjected to a ranking and filtering process using a reward function. This reward function is essentially a model trained on human preferences, which evaluates the quality and relevance of the generated samples. Samples that meet a certain quality threshold, as determined by the reward function, are retained for further processing.
- **Improve**: The final phase involves the integration of the filtered augmented dataset with the original supervised dataset. The combined dataset is then used to fine-tune the original policy. With an increasing filtering threshold, this step can be iteratively repeated to progressively refine the policy. The refined policy from each iteration can be employed in subsequent Grow steps.

The ReST method provides several advantages over typical RLHF (Reinforcement Learning from Human Feedback) methods, such as reduced computational burden and the ability to generate new training data from an improved policy, thus ensuring continuous improvement and adaptation.

The authors tested this approach on machine translation tasks using the IWSLT 2014 and WMT 2020 benchmarks and high-fidelity internal benchmarks on the Web Domain. The results showed that ReST

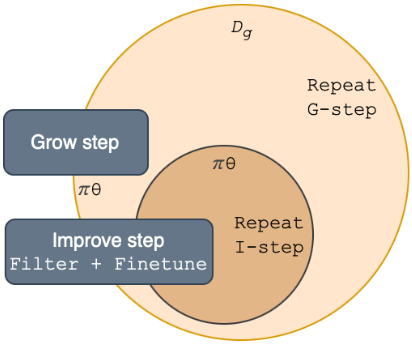

Figure 2. *Reinforced self-training.*

significantly improves reward model scores on test and validation sets and generates higher quality translations according to human raters compared to a supervised learning baseline.

Overall, the ReST approach represents a significant step forward in developing language models that are more aligned with human preferences, particularly in the context of machine translation. Its iterative process of generation and refinement offers a dynamic and effective way to enhance model performance while ensuring that the outputs remain closely aligned with human expectations and linguistic norms.

5. Experiments

5.1. Baselines

To get started with the experimentation phase, we set up a few baselines to understand the problem statement better.

Weak baseline: Here, we conduct a zero-shot evaluation of M2M-100 (Fan et al., 2020) on the English-Hausa (en-ha) and Hausa-English (ha-en) language direction.

Strong Baseline: Here, we fine-tune the M2M-100 on the English-Hausa (en-ha) and Hausa-English (ha-en) language direction and evaluate the results.

The baseline results for the given language directions can be seen in below.

Model	Language direction	Bleu
M2M-100 zero-shot	English - Hausa	2.39
M2M-100 zero-shot	Hausa - English	6.76
M2M-100 fine-tuned	English - Hausa	11.53
M2M-100 fine-tuned	Hausa - English	15.34

Table 4. *Baseline results.*

5.2. Large Language Models

We also compare how well large language models (LLMs) perform on low resource machine translation. We experiment with two LLMs GPT-3.5-turbo and GPT-4 by directly prompting them in a zero-shot

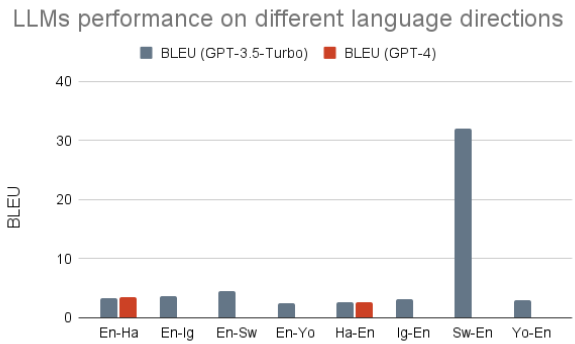

Figure 3. *Large Language Models performance on different language directions.*

setting. We tried multiple different prompts and show the results of the best performing prompt in figure 3.

5.3. *Language Adapters*

We show that adapters achieve parameter-efficient results for low-resource language translation, comparable to fine-tuning the base model. We train our language adapters in two ways: (i) Training individual source and target language adapters and (ii) Training adapters from all given languages to a particular target language adapter.

5.3.1. Monolingual Adapters

The experiments follow a structured approach to evaluate the impact of language adapters within the M2M-100 model for multilingual translation. As per the setup of monolingual adapters, we train two adapters in one run: one associated with the encoder, aligning with the source language, and another linked to the decoder, corresponding to the target language.

The initial experiment focuses on training adapters directly on the base M2M-100 model to assess their impact. This phase aims to understand how adapters influence the model's multilingual translation capabilities without fine-tuning in specific language directions. Training adapters on the base model is a foundational step to evaluate their effectiveness in enhancing translation quality across various language pairs. However, the learning capability of adapters on the baseline model is extremely limited. We think this is due to the diluted nature of the M2M-100 model, which houses a wide spectrum of languages and is likely heavily biased towards the high-resource languages it was exposed to during its pre-training. This situation highlights a probable drawback of a highly multilingual model like M2M-100, where its wide linguistic diversity might pose challenges in optimizing performance for specific, low-resource language pairs.

The graph below illustrates the loss convergence and drop in BLEU score observed when training en-ha adapters on the base M2M-100 model, using a learning rate of 5e-5 with the AdamW optimizer and 1000 warm-up steps over 15 epochs and a weight decay of 0.01.

To confirm our theory, we repeat this training process with many epochs (n=150) and a lower learning rate (5e-6) to ensure that the adapters' poor performance is not due to a need for longer, slower training.

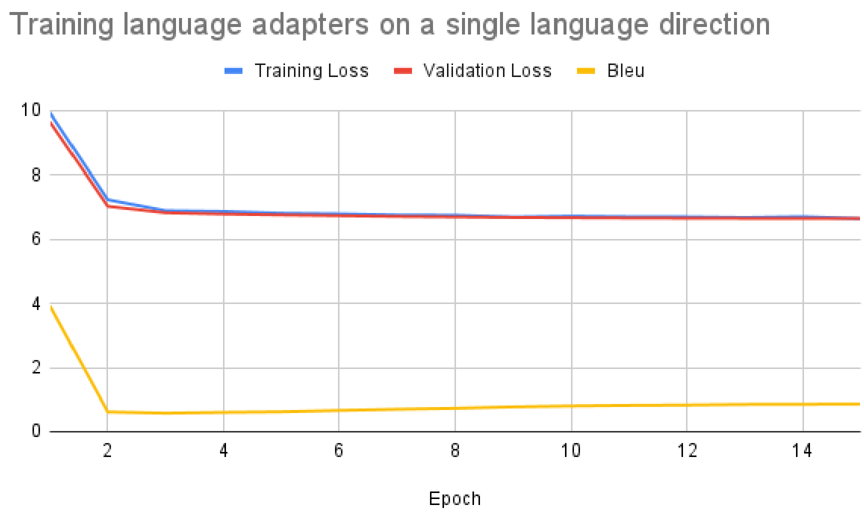

Figure 4. *Training en-ha adapters on base M2M-100 model.*

Following the adapter training on the base model, the subsequent phase involves fine-tuning the M2M-100 model (Fan et al. (2020)) specifically in the English-to-Hausa direction. Post-fine-tuning, our language adapters are trained in multiple language directions, encompassing English, Hausa, Igbo, Swahili, Yoruba, and various paired translations. Interestingly, observations indicate notable improvements when the adapter's target language differs from the fine-tuned model's target language (Hausa). This suggests potential limitations or saturation within the Hausa decoder due to fine-tuning. However, using this fine-tuning process allows for enhancing performance in other language directions using adapters, requiring significantly fewer parameters. Notably, limitations in the quantity of Ha-Ig direction's data result in reduced performance despite the different target language, emphasizing the impact of data availability on the model's proficiency across diverse language pairs.

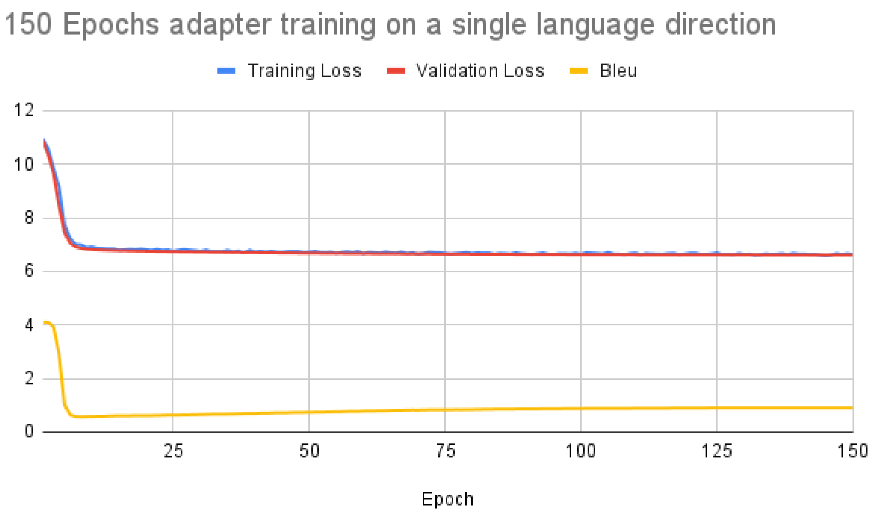

Figure 5. *Training en-ha adapters for 150 epochs on base M2M-100 model.*

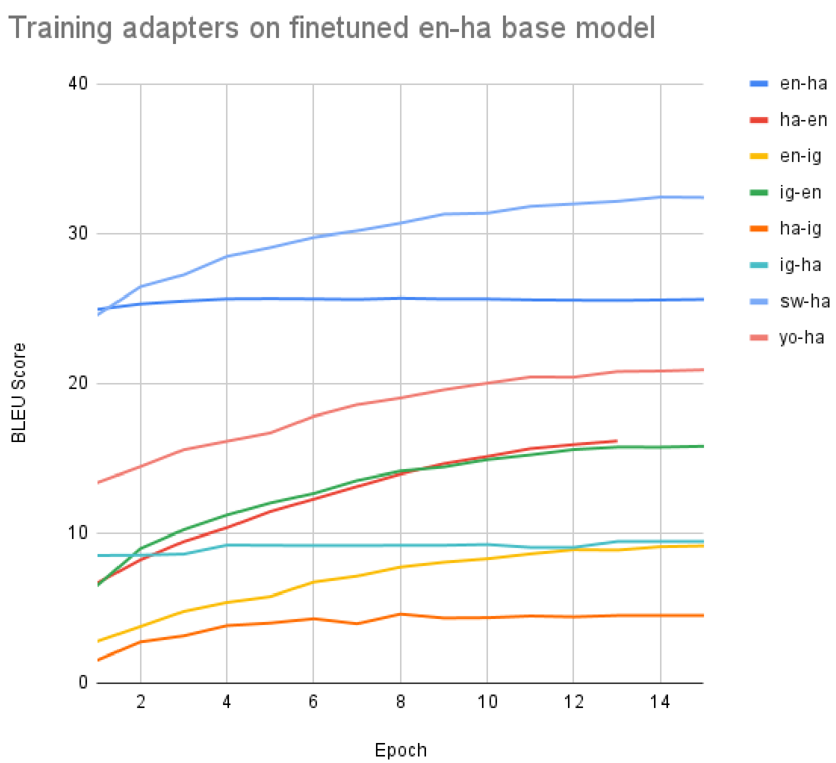

Figure 6. *Training adapters of various language directions on finetuned en-ha model.*

Collectively, these experiments offer insights into the adaptability and efficiency of language adapters in augmenting multilingual translation while considering varied language pairings and data constraints.

5.3.2. Mixed Language Training of Monolingual Adapters

The subsequent phase of experiments involved a mixed language training approach to explore the efficacy of training various source language adapters toward a single target language (X-Target). This methodology, termed mixed language training, entailed training multiple source language adapters targeting a single language, such as X-Hausa, X-Igbo, and X-Yoruba, derived from baseline models and those fine-tuned in English-Hausa (en-ha) and English-Igbo (en-ig) directions. Notably, except for fine-tuned (en-ha) + X-Hausa and fine-tuned (en-ig) + X-Igbo, all configurations exhibited discernible learning curves and enhanced performance compared to standard monolingual adapter training.

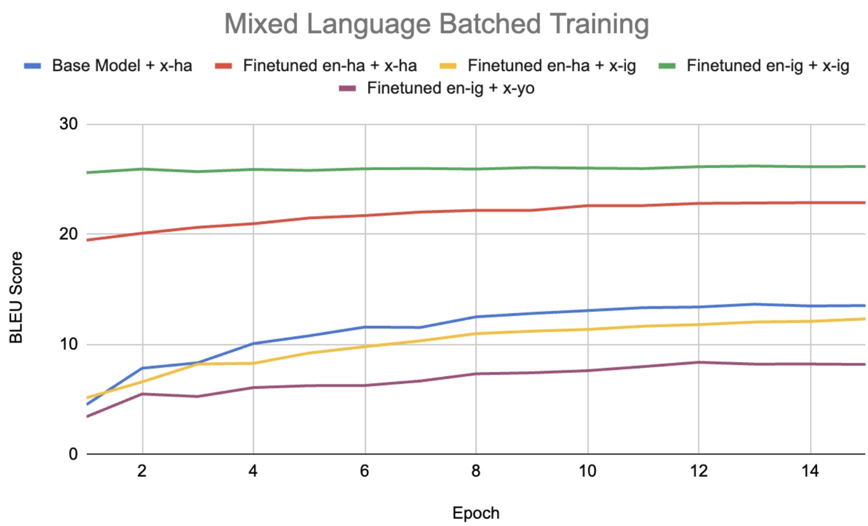

Figure 7. *Mixed Language Batched Training.*

These outcomes further corroborate the hypothesis of potential saturation within the decoder, indicating that the decoder's optimization might have reached its limits during fine-tuning. Additionally, the observed slightly higher performance likely arises from the increased exposure to data from various source languages during training, supporting the theory that more diverse data inputs benefit adapter performance. This demonstrates the adaptability and effectiveness of adapters, particularly in scenarios where data for multiple languages targeting a specific language is available. This approach potentially reduces the necessity for an extensive corpus of data exclusively for a single language direction, thereby showcasing the utility of adapters in leveraging multilingual data for enhanced translation capabilities.

5.4. ReST

In our Reinforced Self-Training (ReST) experimentation for low-resource language translation, we employed the m2m100 model, initially fine-tuning it on English-Hausa translation data. Our methodology involved two distinct approaches rooted in iterative training to enhance model performance over successive iterations.

In the first approach, we utilized our initial training data, applying the filter and improving steps iteratively. We incorporated the filtered data back into our training dataset during these iterations. This process involved progressively tightening the filtering threshold, allowing us to observe and measure improvements in model performance.

The second approach extended this methodology by incorporating a growth step applied to monolingual data. Here, we generated translations from the monolingual data, filtering these translations and adding the refined output to our training dataset. This approach allowed us to retrain the model with this enhanced dataset, observing the impact on model performance.

Our experimentation aimed to explore the efficacy of ReST in improving translation quality for low-resource languages, particularly in cases where traditional training datasets are limited. The iterative nature of the ReST methodology allows for continual refinement and improvement of the model's translation capabilities, making it a suitable approach for low-resource language translation tasks.

5.4.1. Grow

For the grow step, we chose the fine-tuned ha-en model as our policy to generate an augmented dataset. For the first experiment, we considered the Ha-En parallel corpus consisting of 10,000 data points as our source and sampled three translations per data point using our ha-en finetuned model as our policy. In the second setup, we considered monolingual Hausa data (Bañón et al. (2020)) consisting of 1.1M data points; we sampled one translation per datapoint using our policy, leading to 1.1M translated Ha-En pairs.

5.4.2. Filter & Improve

In an ideal scenario to filter data we would receive human feedback to weed out translations which are not good. Lacking such feedback we explored two different reward functions based on comet Rei et al. (2020):

- COMET-DA: A comet reward function which scores the translations given the source and reference translation.
- COMET-QE: A reference variant of comet called comet quality estimator which scores a translation given just the source.

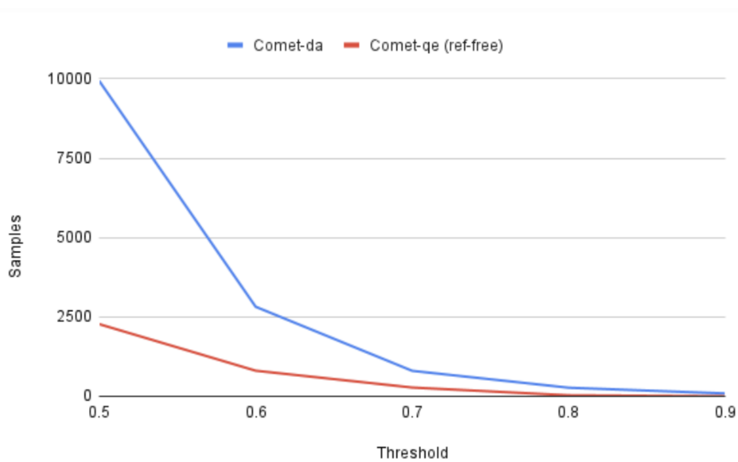

Figure 8. *Reward Function Comparison on Training Data.*

We used comet-da when we filtered Ha-En parallel train data as we had the reference English translations. When filtering the translations generated using the monolingual data we used comet-qe as we didn't have any reference translations.

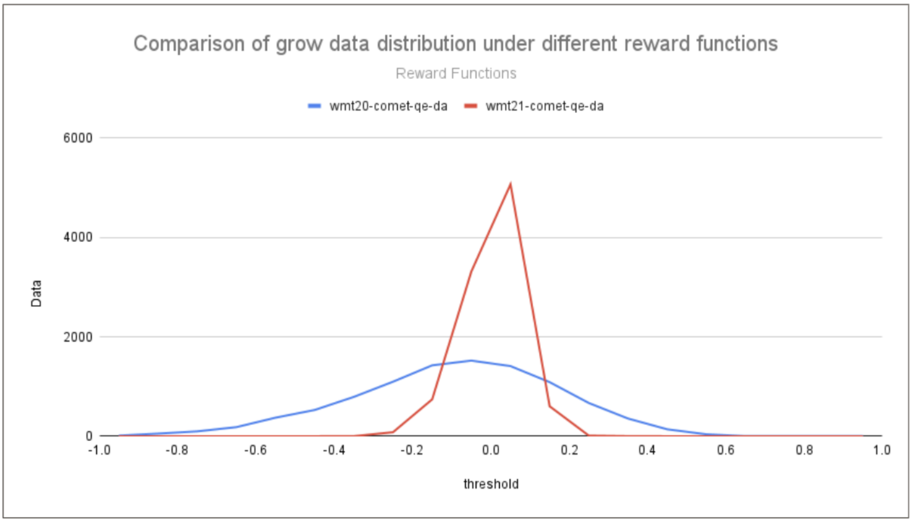

Figure 9. *Reward Function Comparison on Monolingual Data.*

5.4.3. Experimentation

In the experimental setup, we use the M2M100 model fine-tuned on the English-Hausa language data.

The preliminary phase of our experimentation focused on the filtration of translations using the COMET-Quality Estimation (COMET-DA) model, which served as a substitute for the typically human-centric feedback mechanism in Reinforcement Learning from Human Feedback (RLHF). This reward function was used to discern the quality of translations generated by the model. We instituted four COMET-DA threshold levels—0.6, 0.7, 0.8, and 0.9—and revised the training dataset by appending translations that exceeded these threshold values. Contrary to our expectations, this strategy did not result in a substantive enhancement of the model's translation quality, with BLEU scores marginally approaching but not surpassing our baseline metrics, as elaborated in Table 5.

In response to this, we redirected our strategy towards leveraging monolingual data sourced from the extensive Paracrawl dataset, which holds more than 1.1 million sentences. The rationale behind this pivot was the hypothesis that a broader spectrum of training data could potentially enrich the model with a richer understanding of language nuances and, consequently, better translation proficiency. Given the lack of references, we used the reference-free COMET-QE (WMT20 and WMT21) metric to visualize and filter the data distribution given various thresholds, as seen in Figure 9.

With these refined thresholds in place, we undertook a subsequent model fine-tuning. The results were encouraging, as evidenced by a marked improvement in BLEU scores, illustrated in Figure 14. This bar graph elucidates the progressive improvement of the model's performance across the spectrum of COMET-QE thresholds applied to the monolingual data. The fine-tuned model, now enriched with a diversity of high-quality, filtered data, exhibited an advancement in translation accuracy over the initial baseline, thereby validating our experimental setup.

Table 5. Experimentation Results using Training Data.

Sr. No.	Threshold (x)	Bleu (Test)			
		(x - 0.6)	**(x - 0.7)**	**(x - 0.8)**	**(x - 0.9)**
1	Grow-1 Improve-1 (0.6)	14.93	14.4966	14.6034	13.9324
2	Grow-1 Improve-1 (0.7)	-	**15.323**	14.6843	14.4296
3	Grow-1 Improve-1 (0.8)	-	-	14.87	14.6792
4	Grow-1 Improve-1 (0.9)	-	-	-	15.01

6. Results

6.1. Large Language Models

As shown in figure 3 LLMs don't tend to perform well typically for low-resource language translations. Also there doesn't seem to be a significant difference in performance between GPT-3.5-turbo and GPT-4. This tells us that the performance doesn't improve as the model scales.

Two crucial criteria must be met to achieve a high-quality translations:

- Having a good language model for the target language of the translation.
- Access to extensive and high-quality source-target parallel translation data.

All language pairs except Swahili-English have low BLEU scores ranging from ≈ 2.3 to 3.7. The reason behind the exceptional performance of Swahili-English translation (≈ 33 BLEU) is because it satisfies both the criteria. First, GPTs are trained on web corpora, which tend to be heavily biased towards high-resource languages, and English is one of them. Second, there is abundant, vast, high-quality Swahili-English parallel data available on the web.you

Conversely, English-Swahili translation doesn't perform as well (≈ 4.5) because even though there is an abundance of Swahili-English parallel data (conversely vast English-Swahili parallel data), language models like GPTs are not proficient in Swahili. This is due to the scarcity of Swahili data in web corpora for language modelling compared to high-resource languages like English.

6.2. Language Adapters

Language adapters offer a targeted approach to address language-specific nuances within a multilingual model like M2M100. They enable the model to focus on individual languages or language pairs, significantly improving translation efficiency in diverse linguistic contexts. Another key advantage of adapters is their parameter efficiency. While the M2M100 model has a vast architecture with 418 million hyperparameters, the adapters streamline this parameter count to a mere 4.76 million. This balance between performance enhancement and parameter optimization makes adapters a promising and pragmatic solution in multilingual translation tasks.

As discussed in the Experiments section, adapters train very poorly on the base M2M-100 model, even if trained with a low learning rate and many epochs.

However, when the base model has been finetuned in a related language direction, we see significant performance increases, matching our intuitions from the trends during training. Again, it seems to be the case that using the same target language as fine-tuning (Hausa) doesn't improve performance very much. Conversely, when training adapters of a different target language, we see adapters resulting in large jumps in evaluation set BLEU score. Extrapolating from this, we can conclude that when using adapters to improve the performance of a particular language direction, it would be prudent to finetune the base model in a related language direction to leverage knowledge transfer from that language pair to the target language of interest.

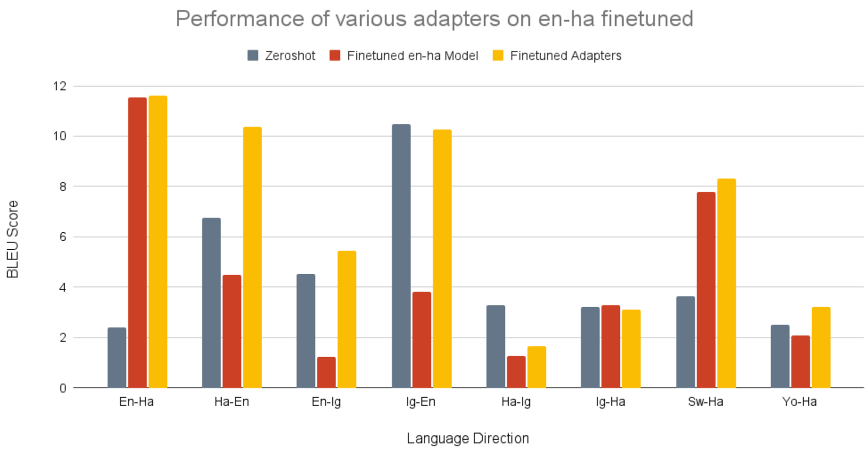

Figure 10. *Test performance of various adapters on finetuned en-ha model.*

The mixed language training approach shows advantages by optimizing adapter performance through diverse multilingual data. This method reduces dependency on large single-direction language datasets while enhancing translation outcomes across various language pairs. Its adoption enhances adaptability and resource efficiency, using multilingual data effectively to improve translation capabilities.

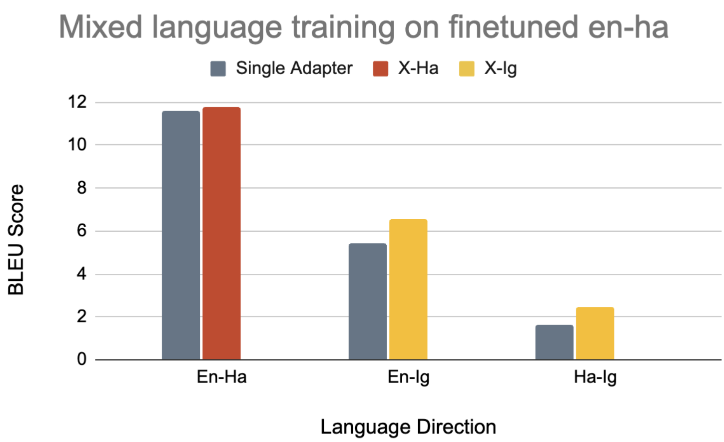

Figure 11. *Test performance of the same in a mixed language setup.*

Observations reveal varying levels of adapter performance across different language directions. While some language pairs exhibit significant improvements with adapters, others encounter limitations. These disparities indicate the impact of language specificity in source and target languages. The experiment outcomes support the hypothesis of potential decoder limitations or saturation during fine-tuning. This phenomenon significantly impacts adapter performance, constraining the decoder's ability to optimize further or enhance translation quality. Exposure to diverse, multilingual data during adapter training also positively influences the efficiency of language adapters, as seen in a mixed-language setting.

6.3. ReST

The Reinforced Self-Training (ReST) method demonstrates several important advantages for aligning conditional language models. It is computationally efficient, requires minimal hyperparameter tuning and leads to measurable gains. In the first approach, we used the available parallel (Ha-En) corpus to experiment with the ReST. As shown in Figure 12, we can observe that it didn't improve over the base Ha-En fine-tuned model. The main reason for this being less diverse data and multiple competing translations for the same source sentence, leading to slower model convergence.

Figure 12. *Test performance of finetuned en-ha model on parallel corpus using ReST.*

To overcome the limitations of the previous approach, we used 1.1 million samples of monolingual Hausa data (Bañón et al. (2020)). In the grow step, we generated the English translations for the 1.1 million samples using the fine-tuned Ha-En model. As this was a monolingual corpus with no reference, we used a reference-free reward function COMET-QE to score and sampled the data using different thresholds. Figure 13 shows the distribution of samples for different thresholds.

By keeping the graph in Figure 13 as a reference, we choose to experiment with thresholds from -0.2 to 0.5. We found that the model performed well at 0.2 thresholds with an approximate improvement of 12% compared to the base fine-tuned model. We can see that the model performance decreases with the decrease in the threshold; this is because there are more noisy data samples as the threshold decreases. Also, when we increase the threshold, we get good quality data, but the quantity of data samples is less, leading to a decrease in performance compared to the best model. Figure 14 depicts the performance of ReST for different thresholds over the monolingual data.

7. Conclusion

In this paper, we presented and explored "Monolingual adapters" and "Reinforced self-training" as approaches for low-resource non-English-centric translations (X-X) using English-centric translation (En-X / X-En) data.

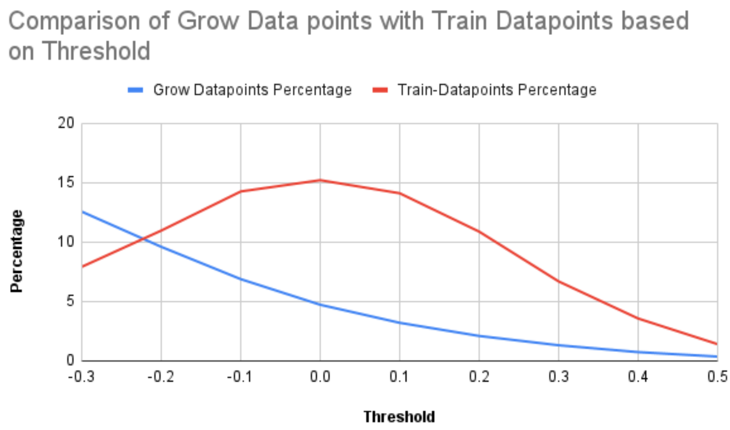

Figure 13. *BLEU Scores with the model fine-tuned on Monolingual Data.*

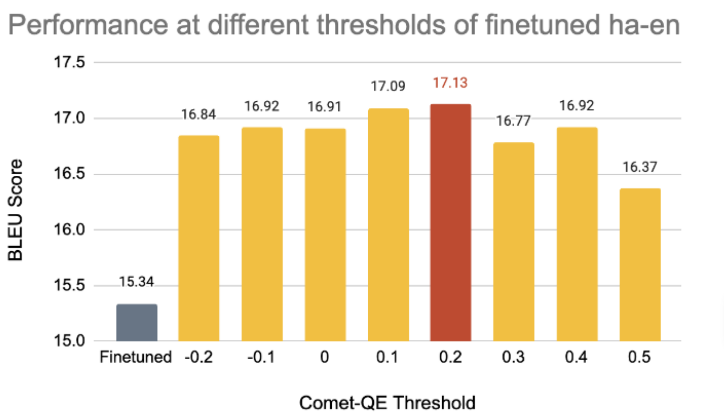

Figure 14. *BLEU Scores with the model fine-tuned on Monolingual Data.*

By splitting the adapters and decoupling them into two halves corresponding to the encoder and decoder languages of the translation pair, our proposed monolingual adapters show promising results by (i) enabling effective zero-shot transfer to unseen language pairs and (ii) incorporating additional parallel data of similar language directions to strengthen the target language decoder adapter. The most significant gains in performance are observed when adapters are exposed to many different language directions. Our experiments also suggest that adapters need some sufficient quantity of knowledge stored in the base model to build on; those interested in using this approach for LRL machine translation may need to fine-tune their base model before attempting to train their adapters.

Carefully selecting the reward function (comet-quality-estimator) and threshold, our reinforced self-training approach yields a performance boost of ≈ 1.8 BLEU points absolute ($\approx 12\%$ increase) over base fine-tuned Hausa-English translation model. Further, on analysis, we find that additional grow and subsequent improve steps yield progressively diminishing gains.

8. Future Work

We have two directions of exploration for future work.

8.1. Adapters

In our current work, we used M2M100 as our base multilingual translation model. It would be prudent to experiment with newer and more powerful multilingual models like DeltaLM Ma et al. (2021) and SeamlessM4T Communication et al. (2023) to explore the extent of the impact the base model has on adapter performance and training.

Another direction of exploration is to add a hierarchy of adapters (language and family adapters) to see if we can leverage linguistic families for better cross-lingual learning between adapters. It would be especially fascinating to experiment with different family designations for Swahili, as there is much debate in the linguistics community over the best classification for that particular language. Some researchers consider Swahili a Niger-Congo language like Igbo and Yoruba. In contrast, others argue it is better understood as a separate Bantu language under the Niger-Congo umbrella term, distinct from Igbo, Yoruba, and other non-Bantinoid Niger-Congo languages. Thus, it would be interesting to explore what association, if any, the different classifications for this language may have on family adapter training.

8.2. ReST

In our current work, we use one reward function for our reinforced self-training. We plan to explore how to use a combination of multiple rewards and learn from learning from weak supervision. We would also want to see how the hyper-parameters impact the ReST training. In each subsequent improve step we tighten the threshold and consequently add in less data but of a higher quality. We want to perform ablation studies to explore and quantitatively find any underlying trends of hyper-parameters and how they impact performance. Additionally, our current approach uses one source of monolingual data (Bañón et al. (2020)). We plan to explore the impact of using diverse data sources from different domains. Finally, we also plan to investigate the benefits of using ReST to fine-tune monolingual adapters.

Acknowledgments

We are grateful for the assistance of our project mentors, UCSC Professor Ian Lane and Yinka Iyinolakan of CDIAL, and Kaleen Shrestha for assisting us in this work.

References

David Adelani, Md Mahfuz Ibn Alam, Antonios Anastasopoulos, Akshita Bhagia, Marta R. Costa-jussÃ, Jesse Dodge, Fahim Faisal, Christian Federmann, Natalia Fedorova, Francisco GuzmÃ¡n, Sergey Koshelev, Jean Maillard, Vukosi Marivate, Jonathan Mbuya, Alexandre Mourachko, Safiyyah Saleem, Holger Schwenk, and Guillaume Wenzek. 2022a. 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*, pages 773–800, Abu Dhabi. Association for Computational Linguistics.

David Ifeoluwa Adelani, Jesujoba Oluwadara 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 Chinenye Emezue, Colin Leong, Michael Beukman, Shamsuddeen Hassan Muhammad, Guyo Dub Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ayoade Ajibade, Tunde Oluwaseyi Ajayi, Yvonne Wambui Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Koffi 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. 2022b. A few thousand translations go a long way! leveraging pre-trained models for african news translation.

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.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady ElSahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Peng Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Shang-Wen Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, M.L. Ramadan, Abinesh Ramakrishnan, Anna Sun, Ke M. Tran, Tuan Tran, I. A. Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bo Yu, Pierre Yves Andrews, Can Balioglu, Marta Ruiz Costa-jussà, Onur Çelebi, Maha Elbayad, Cynthia Gao, Francisco Guzm'an, Justine T. Kao, Ann Lee, Alexandre Mourachko, Juan Miguel Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. 2023. Seamlessm4t: Massively multilingual&multimodal machine translation.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. Fleurs: Few-shot learning evaluation of universal representations of speech.

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

R. F. Gibadullin, M. Yu. Perukhin, and A V Ilin. 2021. Speech recognition and machine translation using neural networks. In *2021 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM)*, pages 398–403.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. 2023. Reinforced self-training (rest) for language modeling.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. 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 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp.

Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, and Furu Wei. 2021. Deltalm: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders. *ArXiv*, abs/2106.13736.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. Adapterhub: A framework for adapting transformers.

Jerin Philip, Alexandre Berard, Matthias Gallé, and Laurent Besacier. 2020. Monolingual adapters for zero-shot neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4465–4470, Online. Association for Computational Linguistics.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1,000+ languages.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation.

International Telecommunication Union. 2020. Measuring digital development. facts and figures 2020. https://www.itu.int/en/ITU-D/Statistics/Documents/facts/FactsFigures2020.pdf. [Online; accessed 05-June-2023].

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.

Wikipedia. 2023a. Hausa language — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Hausa%20language&oldid=1157191964. [Online; accessed 02-June-2023].

Wikipedia. 2023b. Languages of Nigeria — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Languages%20of%20Nigeria&oldid=1156950980. [Online].

Wikipedia. 2023c. Swahili language — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Swahili%20language&oldid=1158074348. [Online; accessed 02-June-2023].

Yuekai Zhao, Haoran Zhang, Shuchang Zhou, and Zhihua Zhang. 2020. Active learning approaches to enhancing neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1796–1806, Online. Association for Computational Linguistics.