MMTAfrica: Multilingual Machine Translation for African Languages

Chris C. Emezue*
Technical University of Munich
Mila Quebec AI Institute†
Masakhane

Bonaventure F. P. Dossou*
Jacobs University Bremen
Mila Quebec AI Institute†
Masakhane

Abstract

In this paper, we focus on the task of multilingual machine translation for African languages and describe our contribution in the 2021 WMT Shared Task: Large-Scale Multilingual Machine Translation. We introduce MMTAfrica, the first many-to-many multilingual translation system for six African languages: Fon (fon), Igbo (ibo), Kinyarwanda (kin), Swahili/Kiswahili (swa), Xhosa (xho), and Yoruba (yor) and two non-African languages: English (eng) and French (fra). For multilingual translation concerning African languages, we introduce a novel backtranslation and reconstruction objective, BT&REC, inspired by the random online back translation and T5 modelling framework respectively, to effectively leverage monolingual data. Additionally, we report improvements from MMTAfrica over the FLORES 101 benchmarks (spBLEU gains ranging from +0.58 in Swahili to French to +19.46 in French to Xhosa).

In this paper, we make use of the following notations:

• ⊙ refers to any language in the set \{eng, fra, ibo, fon, swa, kin, xho, yor\}.
• ◇ refers to any language in the set \{eng, fra, ibo, fon\}.
• AL(s) refers to African language(s).
• X\to Y refers to neural machine translation from language X to language Y.

1 Introduction

Despite the progress of multilingual machine translation (MMT) and the many efforts towards improving its performance for low-resource languages, African languages suffer from under-representation. For example, of the 2000 known African languages (Eberhard et al., 2020) only 17 of them are available in the FLORES 101 Large-scale Multilingual Translation Task as at the time of this research. Furthermore, most research that look into transfer learning of multilingual models from high-resource to low-resource languages rarely work with ALs in the low-resource scenario. While the consensus is that the outcome of the research made using the low-resource non-African languages should be scalable to African languages, this cross-lingual generalization is not guaranteed(Orife et al., 2020) and the extent to which it actually works remains largely understudied. Transfer learning from African languages to African languages sharing the same language sub-class has been shown to give better translation quality than from high-resource Anglo-centric languages (Nyoni and Bassett, 2021) calling for the need to investigate AL\leftrightarrow AL multilingual translation.

This low representation of African languages goes beyond machine translation (Martinus and Abbott, 2019; Joshi et al., 2020; ∀ et al., 2020). The analysis conducted by ∀ et al. (2020) revealed that low-resourcedness of African languages can be traced to the poor incorporation of African languages in the NLP research community (Joshi et al., 2020). All these call for the inclusion of more African languages in multilingual NLP research and experiments.

With the high linguistic diversity in Africa, multilingual machine translation systems are very important for inter-cultural communication, which is in turn necessary for peace and progress. For example, one widely growing initiative to curb the large gap in scientific research in Africa is to translate educational content and scientific papers to various African languages in order to reach far more African native speakers (Abbott and Mart-
We take a step towards addressing the under-representation of African languages in MMT and improving experiments by participating in the 2021 WMT Shared Task: Large-Scale Multilingual Machine Translation with a major target of ALs←→ALs. In this paper, we focused on 6 African languages and 2 non-African languages (English and French). Table 1 gives an overview of our focus African languages in terms of their language family, number of speakers and the regions in Africa where they are spoken (Adelani et al., 2021b). We chose these languages in an effort to create some language diversity: the 6 African languages span the most widely and least spoken languages in Africa. Additionally, they have some similar, as well as contrasting, characteristics which offer interesting insights for future work in ALs:

- Igbo, Yorùbá and Fon use diacritics in their language structure while Kinyarwanda, Swahili and Xhosa do not. Various forms of code-mixing are prevalent in Igbo (Dossou and Emezue, 2021b).

- Fon was particularly chosen because there is only a minuscule amount of online (parallel or monolingual) corpora compared to the other 5 languages. We wanted to investigate and provide valuable insights on improving translation quality of very low-resourced African languages.

- Kinyarwanda and Fon are the only African languages in our work not covered in the FLORES Large-Scale Multilingual Machine Translation Task and also not included in the pretraining of the original model framework used for MMTAfrica. Based on this, we were able to understand the performance of multilingual translation finetuning involving languages not used in the original pretraining objective. We also offered a method to improve the translation quality of such languages.

Our main contributions are summarized below:

1. MMTAfrica – a many-to-many AL←→AL multilingual model for 6 African languages.

2. Our novel reconstruction objective (described in section 4.2) and the BT&REC finetuning setting, together with our proposals in section 5.1 offer a comprehensive strategy for effectively exploiting monolingual data of African languages in AL←→AL multilingual machine translation.

3. Evaluation of MMTAfrica on the FLORES Test Set reports significant gains in spBLEU over the M2M MMT (Fan et al., 2020) benchmark model provided by Goyal et al. (2021).

4. We further created a unique highly representative test set – MMTAfrica Test Set – and reported benchmark results and insights using MMTAfrica.

| Language   | Lang ID (ISO 639-3) | Family                       | Speakers | Region         |
|------------|---------------------|------------------------------|----------|----------------|
| Igbo       | ibo                 | Niger-Congo-Volta-Niger      | 27M      | West           |
| Fon        | fon                 | Niger-Congo-Volta            | 1.7M     | West           |
| Kinyarwanda| kin                 | Niger-Congo-Bantu            | 12M      | East           |
| Swahili    | swa                 | Niger-Congo-Bantu            | 98M      | Southern, Central & East |
| Xhosa      | xho                 | Niger-Congo-Nguni-Bantu      | 19M      | Southern       |
| Yorùbá     | yor                 | Niger-Congo-Volta-Niger      | 42M      | West           |

Table 1: Language, family, number of speakers (Eberhard et al., 2020), and regions in Africa. Adapted from (Adelani et al., 2021b)

2 Related Work

2.1 Multilingual Machine Translation (MMT)

The great success of the encoder-decoder (Sutskever et al., 2014; Cho et al., 2014) NMT on bilingual datasets (Bahdanau et al., 2015; Vaswani et al., 2017; Barrault et al., 2019, 2020) inspired the extension of the original bilingual framework to handle more language pairs simultaneously – leading to multilingual neural machine translation.

Works on multilingual NMT have progressed from sharing the encoder for one-to-many translation (Dong et al., 2015), many-to-one translation (Lee et al., 2017), sharing the attention mechanism across multiple language pairs (Firat et al., 2016a; Dong et al., 2015) to optimizing a single NMT model (with a universal encoder and decoder) for the translation of multiple language pairs (Ha et al., 2016; Johnson et al., 2017). The universal encoder-decoder approach constructs a shared vocabulary for all languages in the training set, and uses just one encoder and decoder for multilingual translation between language pairs. Johnson et al. (2017) proposed to use a single model and prepend
special symbols to the source text to indicate the target language. We adopt their model approach in this paper.

The current state of multilingual NMT, where a single NMT model is optimized for the translation of multiple language pairs (Firat et al., 2016a; Johnson et al., 2017; Lu et al., 2018; Aharoni et al., 2019; Arivazhagan et al., 2019b), has become very appealing for a number of reasons. It is scalable and easy to deploy or maintain (the ability of a single model to effectively handle all translation directions from \( N \) languages, if properly trained and designed, surpasses the scalability of \( O(N^2) \) individually trained models using the traditional bilingual framework). Multilingual NMT can encourage knowledge transfer among related language pairs (Lakew et al., 2018; Tan et al., 2019) as well as positive transfer from higher-resource languages (Zoph et al., 2016; Neubig and Hu, 2018; Arivazhagan et al., 2019a; Aharoni et al., 2019; Johnson et al., 2017) due to its shared representation, improve low-resource translation (Ha et al., 2016; Johnson et al., 2017; Arivazhagan et al., 2019b; Xue et al., 2021) and enable zero-shot translation (i.e. direct translation between a language pair never seen during training) (Firat et al., 2016b; Johnson et al., 2017).

Despite the many advantages of multilingual NMT it suffers from certain disadvantages. Firstly, the output vocabulary size is typically fixed regardless of the number of languages in the corpus and increasing the vocabulary size is costly in terms of computational resources because the training and inference time scales linearly with the size of the decoder’s output layer. For example, the training dataset for all the languages in our work gave a total vocabulary size of 1,683,884 tokens (1,519,918 with every sentence lowercased) but we were constrained to a decoder vocabulary size of 250,000.

Another pitfall of massively multilingual NMT is its poor zero-shot performance (Firat et al., 2016b; Arivazhagan et al., 2019a; Aharoni et al., 2019), particularly compared to pivot-based models (two bilingual models that translate from source to target language through an intermediate language). Neural machine translation is heavily reliant on parallel data and so without access to parallel training data for zero-shot language pairs, multilingual models face the spurious correlation issue (Gu et al., 2019) and off-target translation (Johnson et al., 2017) where the model ignores the given target information and translates into a wrong language.

Some approaches to improve the performance (including zero-shot translation) of multilingual models have relied on leveraging the plentiful source and target side monolingual data that are available. For example, generating artificial parallel data with various forms of backtranslation (Sennrich et al., 2015) has been shown to greatly improve the overall (and zero-shot) performance of multilingual models (Firat et al., 2016b; Gu et al., 2019; Lakew et al., 2018; Zhang et al., 2020) as well as bilingual models (Edunov et al., 2018). Zhang et al. (2020) proposed random online backtranslation to enhance multilingual translation of unseen training language pairs.

Additionally, leveraging monolingual data by jointly learning to reconstruct the input while translating has been shown to improve neural machine translation quality (Févry and Phang, 2018; Lample et al., 2017; Cheng et al., 2016; Zhang and Zong, 2016). Siddhant et al. (2020) leveraged monolingual data in a semi-supervised fashion and reported three major results:

1. Using monolingual data significantly boosts the translation quality of low resource languages in multilingual models.
2. Self-supervision improves zero-shot translation quality in multilingual models.
3. Leveraging monolingual data with self-supervision provides a viable path towards adding new languages to multilingual models.

3 Data Methodology

Table 2 presents the size of the gathered and cleaned parallel sentences for each language direction. We devised preprocessing guidelines for each of our focus languages taking their linguistic properties into consideration. We used a maximum sequence length of 50 (due to computational resources) and a minimum of 2. In the following sections we will describe the data sources for the the parallel and monolingual corpora.

Parallel Corpora: As NMT models are very reliant on parallel data, we sought to gather more parallel sentences for each language direction in an effort to increase the size and domain of each language direction. To this end, our first source was JW300 (Agić and Vulić, 2019), a parallel corpus of
over 300 languages with around 100 thousand biblical domain parallel sentences per language pair on average. Using OpusTools (Aulamo et al., 2020) we were able to get only very trustworthy translations by setting \( t = 1.5 \) (\( t \) is a threshold which indicates the confidence of the translations). We collected more parallel sentences from Tatoeba\(^1\), kde\(^2\) (Tiedemann, 2012), and some English-based bilingual samples from MultiParaCrawl\(^3\).

Finally, following pointers from the native speakers of these focus languages, we generated monolingual data in only these languages. The monolingual sources and volume are summarized in Table 3.

### 3.1 Data Set Types in our Work

Here we elaborate on the different categories of data set that we (generated and) used in our work for training and evaluation.

- **FLORES Test Set**: This refers to the dev test set of 1012 parallel sentences in all 101 language directions provided by Goyal

Table 2: Monolingual data sources and sizes (number of samples). We see for example that much more research on machine translation and data collation has been carried out on swa\(\rightleftharpoons\)eng than fon\(\rightleftharpoons\)fra, attesting to the under-representation of some African languages.

Table 2: Number of parallel samples for each language direction. We highlight the largest and smallest parallel samples. We see for example that much more research on machine translation and data collation has been carried out on swa\(\rightleftharpoons\)eng than fon\(\rightleftharpoons\)fra, attesting to the under-representation of some African languages.

| Language (ID) | Monolingual source | Size |
|--------------|--------------------|------|
| Xhosa (xho)  | The CC100-Xhosa Dataset created by Conneau et al. (2019), and OpenSLR (van Niekerk et al., 2017) | 158,660 |
| Yoruba (yor) | Yoruba Embeddings Corpus (Alabi et al., 2020) and MENYO20k (Adelani et al., 2021) | 45,218 |
| Fon/Fongbe (fon) | FFR Dataset (Dossou and Emezue, 2020), and Fon French Daily Dialogues Parallel Data (Dossou and Emezue, 2021a) | 42,057 |
| Swahili/Kiswahili (swa) | Shikali and Refuoe, 2019 | 23,170 |
| Kinyarwanda (kin) | KINNEWS-and-KIRNEWS (Nyongobo et al., 2020) | 7,586 |
| Igbo (ibo) | Ezeani et al., 2020 | 7,817 |
et al. (2021). We performed evaluation on this test set for all language directions except *←→* fon and *←→* kin.

- **MMTAfrica Test Set**: This is a substantial test set we created by taking out a small but equal number of sentences from each parallel source domain. As a result, we have a set from a wide range of domains, while encompassing samples from many existing test sets from previous research. Although this set is small to be considered as a test set, we open-source it because it contains sentences from many domains (making it useful for evaluation) and we hope that it can be built upon, by perhaps merging it with other benchmark test sets (Abate et al., 2018; Abbott and Martinus, 2019; Reid et al., 2021).

- **Baseline Train/Test Set**: We first conducted baseline experiments with Fon, Igbo, English and French as explained in section 4.4.1. For this we created a special data set by carefully selecting a small subset of the FFR Dataset (which already contained parallel sentences in French and Fon), first automatically translating the sentences to English and Igbo, using the Google Translate API, and finally re-translation with the help of Igbo (7) and English (7) native speakers (we recognized that it was easier for native speakers to edit/tweak an existing translation rather than writing the whole translation from scratch). In doing so, we created a data set of 13,878 translations in all 4 language directions.

We split the data set into 12,554 for training Baseline Train Set, 662 for dev and 662 for test Baseline Test Set.

### 4 Model and Experiments

#### 4.1 Model

For all our experiments, we used the mT5 model (Xue et al., 2021), a multilingual variant of the encoder-decoder, transformer-based (Vaswani et al., 2017) “Text-to-Text Transfer Transformer” (T5) model (Raffel et al., 2019). In T5 pre-training, the NLP tasks (including machine translation) were cast into a “text-to-text” format — that is, a task

where the model is fed some text prefix for context or conditioning and is then asked to produce some output text. This framework makes it straightforward to design a number of NLP tasks like machine translation, summarization, text classification, etc. Also, it provides a consistent training objective both for pre-training and fine-tuning. The mT5 model was pre-trained with a maximum likelihood objective using “teacher forcing” (Williams and Zipser, 1989). The mT5 model was also pretrained with a modification of the masked language modelling objective (Devlin et al., 2018).

We finetuned the **mt5-base** model on our many-to-many machine translation task. While Xue et al. (2021) suggest that higher versions of the mT5 model (Large, XL or XXL) give better performance on downstream multilingual translation tasks, we were constrained by computational resources to **mt5-base**, which has 580M parameters.

#### 4.2 Setup

For each language direction \( X \rightarrow Y \) we have its set of \( n \) parallel sentences \( D = \{(x_i, y_i)\}_{i=1}^{n} \) where \( x_i \) is the \( i \)th source sentence of language \( X \) and \( y_i \) is its translation in the target language \( Y \).

Following the approach of Johnson et al. (2017) and Xue et al. (2021), we model translation in a text-to-text format. More specifically, we create the input for the model by prepending the target language tag to the source sentence. Therefore for each source sentence \( x_i \) the input to the model is \( <Y_{tag}> x_i \) and the target is \( y_i \). Taking a real example, let’s say we wish to translate the Igbo sentence *Daalu maka ikwu eziokwu nke Chineke* to English. The input to the model becomes \( <eng> Daalu maka ikwu eziokwu nke Chineke \).

#### 4.3 Training

We have a set of language tags \( L \) for the languages we are working with in our multilingual many-to-many translation. In our baseline setup (section 4.4.1) \( L = \{eng, fra, ibo, fon\} \) and in our final experiment (section 4.4.2) \( L = \{eng, fra, ibo, fon, swa, kin, xho, yor\} \). We carried out many-to-many translation using all the possible directions from \( L \) except \( eng \leftarrow fra \). We skipped \( eng \leftarrow fra \) for this fundamental reason:

- our main focus is on African → African or \{eng, fra\} ←→ African. Due to the high-
resource nature of English and French, adding the training set for eng ←→ fra would overshadow the learning of the other language directions and greatly impede our analyses. Our intuition draws from the observation of Xue et al. (2021) as the reason for off-target translation in the mT5 model: as English-based finetuning proceeds, the model’s assigned likelihood of non-English tokens presumably decreases. Therefore since the mt5-base training set contained predominantly English (and after other European languages) tokens and our research is about AL←→AL translation, removing the eng ←→ fra direction was our way of ensuring the model designated more likelihood to AL tokens.

### 4.3.1 Our Contributions

In addition to the parallel data between the African languages, we leveraged monolingual data to improve translation quality in two ways:

1. **our backtranslation (BT):** We designed a modified form of the random online backtranslation (Zhang et al., 2020) where instead of randomly selecting a subset of languages to backtranslate, we selected for each language num_bt sentences at random from the monolingual data set. This means that the model gets to backtranslate different (monolingual) sentences every backtranslation time and in so doing, we believe, improve the model’s domain adaptation because it gets to learn from various samples from the whole monolingual data set. We initially tested different values of num_bt to find a compromise between backtranslation computation time and translation quality.

   Following research works which have shown the effectiveness of random beam-search over greedy decoding while generating backtranslations (Lample et al., 2017; Edunov et al., 2018; Hoang et al., 2018; Zhang et al., 2020), we generated num_sample prediction sentences from the model and randomly selected (with equal probability) one for our backtranslated sentence. Naturally the value of num_sample further affects the computation time (because the model has to produce num_sample different output sentences for each input sentence) and so we finally settled with num_sample = 2.

2. **our reconstruction:** Given a monolingual sentence \(x^m\) from language \(m\), we applied random swapping (2 times) and deletion (with a probability of 0.2) to get a noisy version \(\hat{x}\). Taking inspiration from Raffel et al. (2019) we integrated the reconstruction objective into our model finetuning by prepending the language tag \(<m>\) to \(\hat{x}\) and setting its target output to \(x^m\).

### 4.4 Experiments

In all our experiments we initialized the pretrained mt5-base model using Hugging Face’s AutoModelForSeq2SeqLM⁶ and tracked the training process with Weights&Biases (Biewald, 2020). We used the AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate (lr) of \(3 \times e^{-6}\) and transformer’s get_linear_schedule_with_warmup scheduler (where the learning rate decreases linearly from the initial lr set in the optimizer to 0, after a warmup period and then increases linearly from 0 to the initial lr set in the optimizer.)

#### 4.4.1 Baseline

The goal of our baseline was to understand the effect of jointly finetuning with backtranslation and reconstruction on the African←→African language translation quality in two scenarios: when the AL was initially pretrained on the multilingual model and contrariwise. Using Fon (which was not initially included in the pretraining) and Igbo (which was initially included in the pretraining) as the African languages for our baseline training, we finetuned our model on a many-to-many translation in all directions of \{eng, fra, ibo, fon\}/eng ←→ fra amounting to 10 directions. We used the Baseline Train Set for training and the Baseline Test Set for evaluation. We trained the model for only 3 epochs in three settings:

1. **BASE**: in this setup we finetune the model on only the many-to-many translation task: no backtranslation nor reconstruction.

2. **BT**: refers to finetuning with our backtranslation objective described in section 4.2. For our

---

⁶https://huggingface.co/transformers/model_doc/auto.html#transformers.AutoModelForSeq2SeqLM

⁷https://huggingface.co/transformers/main_classes/optimizer_schedules.html#transformers.get_linear_schedule_with_warmup
baseline, where we backtranslate using monolingual data in \{ibo, fon\}, we set $num_{bt} = 500$. For our final experiments, we first tried with 500 but finally reduced to 100 due to the great deal of computation required.

For our baseline experiment, we ran one epoch normally and the remaining two with backtranslation. For our final experiments, we first finetuned the model on 3 epochs before continuing with backtranslation.

3. **BT&REC** : refers to joint backtranslation and reconstruction (explained in section 4.2) while finetuning. Two important questions were addressed – 1) the ratio, backtranslation : reconstruction, of monolingual sentences to use and 2) whether to use the same or different sentences for backtranslation and reconstruction. Bearing computation time in mind, we resolved to go with $500 : 50$ for our baseline and $100 : 50$ for our final experiments. We leave ablation studies on the effect of the ratio on translation quality to future work. For the second question we decided to randomly sample (with replacement) different sentences each for our backtranslation and reconstruction.

For our baseline, we used a learning rate of $5e^{-4}$, a batch size of 32 sentences, with gradient accumulation up to a batch of 256 sentences and an early stopping patience of 100 evaluation steps. To further analyse the performance of our baseline setups we ran `comparemt`\(^8\) (Neubig et al., 2019) on the model’s predictions.

### 4.4.2 MMTAfrica

MMTAfrica refers to our final experimental setup where we finetuned our model on all language directions involving all eight languages $L = \{\text{eng}, \text{fra}, \text{ibo}, \text{fon}, \text{swa}, \text{kin}, \text{xho}, \text{yor}\}$ except eng $\leftarrow$ fra. Taking inspiration from our baseline results we ran our experiment with our proposed **BT&REC** setting and made some adjustments along the way.

The long computation time for backtranslating (with just 100 sentences per language the model was required to generate around 3,000 translations every backtranslation time) was a drawback. To mitigate the issue we parallelized the process using the multiprocessing package in Python.\(^9\) We further slowly reduced the number of sentences for backtranslation (to 50, and finally 10).

Gradient descent in large multilingual models has been shown to be more stable when updates are performed over large batch sizes are used (Xue et al., 2021). To cope with our computational resources, we used gradient accumulation to increase updates from an initial batch size of 64 sentences, up to a batch gradient computation size of 4,096 sentences. We further utilized PyTorch’s DataParallel package\(^10\) to parallelize the training across the GPUs. We used a learning rate (lr) of $3e^{-6}$

### 5 Results and Insights

All evaluations were made using spBLEU (sentencepiece (Kudo and Richardson, 2018) + sacreBLEU (Post, 2018)) as described in (Goyal et al., 2021). We further evaluated on the chrF (Popović, 2015) and TER metrics.

#### 5.1 Baseline Results and Insights

Figure 1 compares the spBLEU scores for the three setups used in our baseline experiments. As a reminder, we make use of the symbol ⋄ to refer to any language in the set \{eng, fra, ibo, fon\}.

**BT** gives strong improvement over **BASE** (except in eng $\rightarrow$ ibo where it’s relatively the same, and fra $\rightarrow$ ibo where it performs worse).

![Figure 1: spBLEU scores of the 3 setups explained in section 4.4.1](https://github.com/neulab/compare-mt)

When the target language is fon, we observe a considerable boost in the spBLEU of the **BT** setting, which also significantly outperformed **BASE**.

---

\(^8\)https://github.com/neulab/compare-mt
\(^9\)https://docs.python.org/3/library/multiprocessing.html
\(^10\)https://pytorch.org/docs/stable/generated/torch.nn.DataParallel.html
and BT&REC. BT&REC contributed very little when compared with BT and sometimes even performed poorly (in eng→fon). We attribute this poor performance from the reconstruction objective to the fact that the mt5-base model was not originally pretrained on Fon. Therefore, with only 3 epochs of finetuning (and 1 epoch before introducing the reconstruction and backtranslation objectives) the model was not able to meaningfully utilize both objectives.

Conversely, when the target language is ibo BT&REC gives best results – even in scenarios where BT underperforms BASE (as is the case of fra→ibo and eng→ibo). We believe that the decoder of the model, being originally pretrained on corpora containing Igbo, was able to better use our reconstruction to improve translation quality in ◊→ibo direction.

Drawing insights from fon←→ibo we offer the following propositions concerning AL←→AL multilingual translation:

- our backtranslation (section 4.2) from monolingual data improves the cross-lingual mapping of the model for low-resource African languages. While it is computationally expensive, our parallelization and decay of number of backtranslated sentences are some potential solutions towards effectively adopting backtranslation using monolingual data.

- Denoising objectives typically have been known to improve machine translation quality (Zhang and Zong, 2016; Cheng et al., 2016; Gu et al., 2019; Zhang et al., 2020; Xue et al., 2021) because they imbue the model with more generalizable knowledge (about that language) which is used by the decoder to predict better token likelihoods for that language during translation. This is a reasonable explanation for the improved quality with the BT&REC over BT in the ◊→ibo. As we learned from ◊→fon, using reconstruction could perform unsatisfactorily if not handled well. Some methods we propose are:

1. For African languages that were included in the original model pretraining (as was the case of Igbo, Swahili, Xhosa, and Yoruba in the mT5 model), using the BT&REC setting for finetuning produces best results. While we did not perform ablation studies on the data size ratio for backtranslation and reconstruction, we believe that our ratio of 2 : 1 (in our final experiments) gives the best compromise on both computation time and translation quality.

2. For African languages that were not originally included in the original model pretraining (as was the case of Kinyarwanda and Fon in the mT5 model), reconstruction together with backtranslation (especially at an early stage) only introduces more noise which could harm the cross-lingual learning. For these languages we propose:

(a) first finetuning the model on only our reconstruction (described in section 4.2) for fairly long training steps before using BT&REC. This way, the initial reconstruction will help the model learn that language representation space and increase its the likelihood of tokens.

5.2 MMTAfrica Results and Insights

In Table 4, we compared MMTAfrica with the M2M MMT(Fan et al., 2020) benchmark results of Goyal et al. (2021) using the same test set they used – FLORES Test Set. On all language pairs except swa→eng (which has a comparable −2.76 spBLEU difference), we report an improvement from MMTAfrica (spBLEU gains ranging from +0.58 in swa→fra to +19.46 in fra→xho). The lower score of swa→eng presents an intriguing anomaly, especially given the large availability of parallel corpora in our training set for this pair. We plan to investigate this in further work.

In Table 5 we introduce benchmark results of MMTAfrica on MMTAfrica Test Set. We also put the test size of each language pair.

Interesting analysis about Fon (fon) and Yoruba (yor): For each language, the lowest sp-BLEU scores in both tables come from the yor direction, except fon←→yor (from Table 5) which interestingly has the highest spBLEU score compared to the other fon→⊛ directions. We do not know the reason for the very low performance in the ⊛→yor direction, but we offer below a plausible explanation about fon←→yor.

The oral linguistic history of Fon ties it to the ancient Yoruba kingdom (Barnes, 1997). Furthermore, in present day Benin, where Fon is largely
spoken as a native language, Yoruba is one of the indigenuous languages commonly spoken. Therefore Fon and Yoruba share some linguistic characteristics and we believe this is one logic behind the fon→yor surpassing other fon→ directions.

This explanation could inspire transfer learning from Yoruba, which has received comparably more research and has more resources for machine translation, to Fon. We leave this for future work.

This experimental resource for running all experiments were provided by the FLORES compute grants as well as additional computational resources provided by Paco Guzman (Facebook AI) and Mila Quebec AI Institute. We express our profound gratitude to all who contributed in one way or the other towards the development of MMTAfrica including (in no order):

Mathias Müller (University of Zurich) for giving immense technical assistance in finetuning the model and advising us on best hyperparameter tuning practises.

Graham Neubig (Carnegie Mellon University) for explaining and setting up comparemert for us to better understand and evaluate the performance of our baseline models.

Angela Fan (FacebookAI) for guiding us during the shared task and taking the time to answer all our questions.

Julia Kreutzer (GoogleAI) for advising us on useful model comparison and evaluation techniques.

Colin Leong (University of Dayton) for painstakingly proof-reading the paper and helping us make it more reader-friendly.

Daria Yasafova (Technical University of Munich) and Yeno Gbenou (Drexel University) for additionally proof-reading the paper.

Finally the entire Masakhane community for, among many other things, 1) guiding us to existing parallel and monolingual data set for the focus African languages, 2) explaining their (the focus African languages) important linguistic characteristics which helped us work better with them (like in preprocessing) and 3) performing (in the future) human evaluation on MMTAfrica.

Indeed it took a village to raise MMTAfrica.

### 7 Acknowledgments

The computational resources for running all experiments were provided by the FLORES compute grants as well as additional computational resources provided by Paco Guzman (Facebook AI) and Mila Quebec AI Institute. We express our profound gratitude to all who contributed in one way or the other towards the development of MMTAfrica including (in no order):

Mathias Müller (University of Zurich) for giving immense technical assistance in finetuning the model and advising us on best hyperparameter tuning practices.

Graham Neubig (Carnegie Mellon University) for explaining and setting up comparemert for us to better understand and evaluate the performance of our baseline models.

Angela Fan (FacebookAI) for guiding us during the shared task and taking the time to answer all our questions.

Julia Kreutzer (GoogleAI) for advising us on useful model comparison and evaluation techniques.

Colin Leong (University of Dayton) for painstakingly proof-reading the paper and helping us make it more reader-friendly.

Daria Yasafova (Technical University of Munich) and Yeno Gbenou (Drexel University) for additionally proof-reading the paper.

Finally the entire Masakhane community for, among many other things, 1) guiding us to existing parallel and monolingual data set for the focus African languages, 2) explaining their (the focus African languages) important linguistic characteristics which helped us work better with them (like in preprocessing) and 3) performing (in the future) human evaluation on MMTAfrica.

Indeed it took a village to raise MMTAfrica.

### Table 4: Evaluation Scores of the Flores M2M MMT model and MMTAfrica on FLORES Test Set.

| Source | Target | spBLEU (FLORES) | spBLEU (Ours) | spCHRF↑ | spTER↑ |
|--------|--------|-----------------|---------------|---------|---------|
| swa    | ibo    | 19.80           | 21.84         | 31.95   | 68.22   |
| swa    | xho    | 21.71           | 21.71         | 39.86   | 73.16   |
| swa    | yor    | 11.68           | 27.44         | 75.23   |
| swa    | eng    | 27.67           | 30.43         | 56.12   | 55.91   |
| swa    | fra    | 16.46           | 26.69         | 46.20   | 63.47   |
| xho    | ibo    | 17.02           | 19.80         | 33.95   | 68.22   |
| xho    | owa    | 6.14            | 6.14          | 44.68   | 63.21   |
| xho    | yor    | 10.42           | 10.42         | 26.77   | 76.25   |
| xho    | eng    | 20.77           | 30.43         | 46.99   | 64.09   |
| xho    | fra    | 21.48           | 26.69         | 40.65   | 69.31   |
| yor    | ibo    | 11.45           | 1.85          | 25.26   | 74.99   |
| yor    | owa    | 14.99           | 1.93          | 30.49   | 79.90   |
| yor    | xho    | 9.31            | 1.94          | 26.34   | 86.08   |
| yor    | eng    | 8.15            | 4.18          | 30.65   | 86.94   |
| yor    | fra    | 10.59           | 3.57          | 27.60   | 81.32   |
| eng    | ibo    | 24.19           | 3.53          | 37.24   | 65.68   |
| eng    | owa    | 40.11           | 26.95         | 53.13   | 52.80   |
| eng    | xho    | 27.15           | 4.47          | 44.93   | 67.77   |
| eng    | yor    | 12.09           | 2.17          | 28.34   | 74.74   |
| fra    | ibo    | 19.48           | 1.69          | 34.47   | 68.50   |
| fra    | owa    | 34.21           | 17.17         | 48.95   | 58.11   |
| fra    | xho    | 21.73           | 2.27          | 40.06   | 73.72   |
| fra    | yor    | 11.42           | 1.16          | 27.67   | 75.33   |

Table 4: Evaluation Scores of the Flores M2M MMT model and MMTAfrica on FLORES Test Set.

### 6 Conclusion and Future Work

In this paper, we introduced MMTAfrica, a multilingual machine translation model on 6 African Languages, which outperformed the M2M MMT model Fan et al. (2020). Our results and analyses, including a new reconstruction objective, give insights on MMT for African languages for future research. Moreover, we plan to launch the model on Masakhane MT and FFRTranslate in order to get human evaluation feedback from the actual speakers of the languages in the Masakhane community (Orife et al., 2020) and beyond.

In order to fully test the advantage of MMTAfrica, we plan to finish comparing it on direct and pivot translations with the Masakhane benchmark models (Vo et al., 2020). We also plan to perform human evaluation. All test sets, results, code and checkpoints will be released at [https://github.com/edaofficial/mmtafrica](https://github.com/edaofficial/mmtafrica).

11[https://en.wikipedia.org/wiki/Benin](https://en.wikipedia.org/wiki/Benin) (Last Accessed: 30.08.2021).
| Source | Target | Test size | spBLEU↑ | spCHRF↑ | spTER↓ |
|--------|--------|-----------|---------|---------|--------|
| ibo    | swa    | 60        | 34.89   | 47.38   | 68.28  |
| ibo    | xho    | 30        | 36.69   | 50.66   | 59.65  |
| ibo    | yor    | 30        | 11.77   | 29.54   | 129.84 |
| ibo    | kin    | 30        | 33.92   | 46.53   | 63.21  |
| ibo    | eng    | 90        | 37.28   | 60.42   | 62.05  |
| ibo    | fra    | 60        | 30.86   | 44.09   | 69.53  |
| swa    | ibo    | 60        | 33.71   | 43.02   | 60.01  |
| swa    | xho    | 30        | 37.28   | 52.53   | 55.86  |
| swa    | yor    | 30        | 14.09   | 27.50   | 113.63 |
| swa    | kin    | 30        | 23.86   | 42.59   | 94.67  |
| swa    | eng    | 90        | 23.29   | 33.52   | 65.11  |
| swa    | fra    | 60        | 30.11   | 48.33   | 63.38  |
| xho    | ibo    | 30        | 33.25   | 45.36   | 62.83  |
| xho    | swa    | 30        | 39.26   | 53.75   | 53.72  |
| xho    | yor    | 30        | 25.11   | 43.19   | 74.80  |
| xho    | kin    | 30        | 31.81   | 47.43   | 63.39  |
| xho    | eng    | 90        | 18.71   | 41.73   | 93.00  |
| xho    | fra    | 30        | 15.44   | 30.97   | 90.57  |
| yor    | ibo    | 30        | 17.62   | 34.71   | 85.18  |
| yor    | swa    | 30        | 29.31   | 43.13   | 66.82  |
| yor    | kin    | 30        | 25.16   | 40.82   | 72.67  |
| yor    | eng    | 90        | 30.25   | 55.12   | 62.11  |
| yor    | fra    | 30        | 24.95   | 45.72   | 61.03  |
| kin    | ibo    | 30        | 31.25   | 42.36   | 66.73  |
| kin    | swa    | 30        | 33.65   | 46.34   | 72.70  |
| kin    | xho    | 30        | 20.40   | 39.71   | 89.97  |
| kin    | yor    | 30        | 18.34   | 33.53   | 70.43  |
| kin    | kin    | 30        | 22.43   | 32.49   | 67.26  |
| kin    | eng    | 60        | 15.82   | 43.10   | 96.55  |
| kin    | fra    | 30        | 16.23   | 33.51   | 91.82  |
| fon    | ibo    | 30        | 32.36   | 46.44   | 61.82  |
| fon    | swa    | 30        | 29.84   | 42.96   | 72.28  |
| fon    | xho    | 30        | 28.82   | 47.44   | 66.98  |
| fon    | yor    | 30        | 30.45   | 42.63   | 60.72  |
| fon    | kin    | 30        | 23.88   | 39.59   | 78.06  |
| fon    | eng    | 30        | 16.63   | 41.63   | 69.03  |
| fon    | fra    | 60        | 24.79   | 43.39   | 82.15  |
| eng    | ibo    | 90        | 44.24   | 54.89   | 63.92  |
| eng    | swa    | 60        | 49.94   | 61.45   | 47.83  |
| eng    | xho    | 120       | 31.97   | 49.74   | 72.89  |
| eng    | yor    | 90        | 23.93   | 36.19   | 84.05  |
| eng    | kin    | 90        | 40.98   | 56.00   | 76.37  |
| eng    | fon    | 30        | 27.19   | 38.66   | 62.54  |
| fra    | ibo    | 60        | 36.47   | 46.93   | 59.91  |
| fra    | swa    | 60        | 36.53   | 51.42   | 55.94  |
| fra    | xho    | 30        | 34.35   | 49.39   | 60.30  |
| fra    | yor    | 30        | 7.26    | 25.54   | 124.53 |
| fra    | kin    | 30        | 31.07   | 42.26   | 81.06  |
| fra    | fon    | 40        | 31.07   | 38.72   | 75.74  |

Table 5: Benchmark Evaluation Scores on MMTAfrica Test Set

References

Solomon Tefera Abate, Michael Melese, Martha Yi-firu Tachbelie, Million Meshesha, Solomon Atinafu, Wondwossen Mulugeta, Yaregal Assabie, Hafte Abera, Binyam Ephrem, Tewodros Abebe, Wondimagnenhue Tsegaye, Amanuel Lemma, Tsegaye Andargie, and Seifedin Shifaw. 2018. Parallel corpora for bi-lingual English-Ethiopian languages statistical machine translation. In Proceedings of the 27th International Conference on Computational Linguistics, pages 3102–3111, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jade Abbott and Laura Martinus. 2019. Benchmarking neural machine translation for Southern African languages. In Proceedings of the 2019 Workshop on Widening NLP, pages 98–101, Florence, Italy. Association for Computational Linguistics.

Jade Z. Abbott and Laura Martinus. 2018. Towards neural machine translation for african languages.

David I. Adelani, Dana Ruiter, Jesujoba O. Alabi, Damilola Adebonojo, Adesina Ayeni, Mofo Adeyemi, Ayodele Awokoya, and Cristina España-Bonet. 2021a. The effect of domain and diacritics in yorùbá-english neural machine translation.

David Ifeoluwa Adelani, Jade Z. Abbott, Graham Neubig, Daniel D’ouza, Julia Kreutzier, Constantine Lignos, Chester Palen-Michel, Happy Buzababa, Shruti Rijhwani, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwa, Catherine Gitau, Dergue Mbaye, Jesujoba O. Alabi, Seid Muhie Yimam, Tajudeen Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathon Mukibi, Verrah Otiende, Iroro Orife, Davis David, Sambha Ngom, Tosin P. Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwunekwe, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobias Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadou Mboup, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Oreaoghene Ahia, Bonaventure P. Dossou, Kelechi Ogweji, Thierno Ibrahima Diop, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021b. Masakhaner: Named entity recognition for african languages. CoRR, abs/2103.11811.

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation.
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 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Roei Aharoni, Melvin Johnson, and Wolfgang Macherey. 2019a. *The missing ingredient in zero-shot neural machine translation*. *CoRR*, abs/1903.07091.

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

Mikko Aulamo, Umut Sulubacak, Sami Virpioja, and Jörg Tiedemann. 2020. *OpusTools and parallel corpus diagnostics*. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3782–3789. European Language Resources Association.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. *Neural machine translation by jointly learning to align and translate*. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

S.T. Barnes. 1997. *Africa’s Ogun: Old World and New*. African systems of thought. Indiana University Press.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. *Findings of the 2020 conference on machine translation (WMT20)*. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55. Online. Association for Computational Linguistics.

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.

Lucas Biewald. 2020. *Experiment tracking with weights and biases*. Software available from wandb.com.

Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. *Semi-supervised learning for neural machine translation*. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1965–1974. Berlin, Germany. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merrienboer, Çağlar Güçlü, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. *Learning phrase representations using RNN encoder-decoder for statistical machine translation*. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734. ACL.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. *Unsupervised cross-lingual representation learning at scale*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. *BERT: pre-training of deep bidirectional transformers for language understanding*. *CoRR*, abs/1810.04805.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. *Multi-task learning for multiple language translation*. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732. Beijing, China. Association for Computational Linguistics.

Bonaventure F. P. Dossou and Chris C. Emezue. 2020. *FFR V1.0: fon-french neural machine translation*. *CoRR*, abs/2003.12111.

Bonaventure F. P. Dossou and Chris C. Emezue. 2021a. *Crowdsourced phrase-based tokenization for low-resourced neural machine translation: The case of fon language*. *CoRR*, abs/2103.08052.

Bonaventure F. P. Dossou and Chris C. Emezue. 2021b. Okwugbé: *End-to-end speech recognition for fon and igbo*. *CoRR*, abs/2103.07762.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig (eds.). 2020. *Ethnologue: Languages of the world. twenty-third edition*. 408
Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. CoRR, abs/1808.09381.

Ignatius Ezeani, Paul Rayson, Ikechukwu E. Onyenwe, Chinedu Uchechukwu, and Mark Hepple. 2020. Igbo-english machine translation: An evaluation benchmark. CoRR, abs/2004.00648.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. CoRR, abs/2010.11125.

Thibault Févry and Jason Phang. 2018. Unsupervised sentence compression using denoising auto-encoders. CoRR, abs/1809.02669.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016a. Multi-way, multilingual neural machine translation with a shared attention mechanism. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 866–875, San Diego, California. Association for Computational Linguistics.

Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016b. Zero-resource translation with multi-lingual neural machine translation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 268–277, Austin, Texas. Association for Computational Linguistics.

∀, Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasuabaa, Tajudeen Kolawole, Taiwo Fagbohunbge, Solomon Olouwle Akinola, Shamsudee Hassan Muhammad, Salomon Kabongo, Salomey Osei, et al. 2020. Participatory research for low-resourced machine translation: A case study in african languages. Findings of EMNLP.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc’Aurelio Ranzato, Francisco Guzman, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O.K. Li. 2019. Improved zero-shot neural machine translation via ignoring spurious correlations. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1258–1268, Florence, Italy. Association for Computational Linguistics.

Thanh-Le Ha, Jan Niehues, and Alexander H. Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. CoRR, abs/1611.04798.

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

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.

Pratik Joshi, Sebastian Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6282–6293, Online. Association for Computational Linguistics.

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.

Surafel Melaku Lakew, Mauro Cettolo, and Marcello Federico. 2018. A comparison of transformer and recurrent neural networks on multilingual neural machine translation. In Proceedings of the 27th International Conference on Computational Linguistics, pages 641–652, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Guillaume Lample, Ludovic Denoyer, and Marc’Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. CoRR, abs/1711.00043.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2017. Fully Character-Level Neural Machine Translation without Explicit Segmentation. Transactions of the Association for Computational Linguistics, 5:365–378.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. CoRR, abs/1711.05101.

Yichao Lu, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaoman Zhang, and Jason Sun. 2018. A neural interlingua for multilingual machine translation. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 84–92, Brussels, Belgium. Association for Computational Linguistics.

Laura Martinus and Jade Z. Abbott. 2019. A focus on neural machine translation for african languages. CoRR, abs/1906.05685.
Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, Xinyi Wang, and John Wieting. 2019. compare-mt: A tool for holistic comparison of language generation systems. CoRR, abs/1903.07926.

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

Rubungo Andre Niyongabo, Qu Hong, Julia Kreutzer, and Li Huang. 2020. KINNEWS and KIRNEWS: Benchmarking cross-lingual text classification for Kinyarwanda and Kirundi. In Proceedings of the 28th International Conference on Computational Linguistics, pages 5507–5521, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Linda Nordling. 2018. How decolonization could reshape south african science. Nature, 554(7691):159–162.

Evander Nyoni and Bruce A. Bassett. 2021. Low-resource neural machine translation for southern african languages. CoRR, abs/2104.00366.

Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamii Toure Ali, Jade Z. Abbott, Yukosi Maritime, Salomon Kabongo, Musie Meressa, Espoir Murhabazi, Orevagheine Ahia, Elan Van Biljon, Arshath Ramklowan, Adewale Akinfaderin, Alp Öktem, Wole Akin, Ghollah Kioko, Kevin Degila, Herman Kamper, Bonaventure Dossou, Chris Emezie, Kelechi Ogueji, and Abdallah Bashir. 2020. Masakhane - machine translation for africa. CoRR, abs/2003.11529.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. CoRR, abs/1910.10683.

Machel Reid, Junjie Hu, Graham Neubig, and Yutaka Matsuo. 2021. Afront: Pretraining strategies and reproducible benchmarks for translation of 8 african languages.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. CoRR, abs/1511.06709.

Shivachi Casper Shikali and Mokhosi Refuoe. 2019. Language modeling data for swahili.

Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Xu Chen, Sneha Reddy Kudugunta, Naveen Arivazhagan, and Yonghui Wu. 2020. Leveraging monolingual data with self-supervision for multilingual neural machine translation. CoRR, abs/2005.04816.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS’14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Xu Tan, Jiale Chen, Di He, Yingce Xia, Tao Qin, and Tie-Yan Liu. 2019. Multilingual neural machine translation with language clustering. 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 963–973, Hong Kong, China. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12), Istanbul, Turkey. European Language Resources Association (ELRA).

Daniel van Niekerk, Charl van Heerden, Marelle Davel, Neil Kleyhans, Oddur Kjartansson, Martin Jansche, and Linne Ha. 2017. Rapid development of TTS corpora for four South African languages. In Proc. Interspeech 2017, pages 2178–2182, Stockholm, Sweden.

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

Sarah Wild. 2021. African languages to get more bespoke scientific terms. Nature, 596(7873):469–470.

Ronald J. Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. Neural Computation, 1(2):270–280.

Linting Xue, Noah Constant, Adam Roberts, Mishir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 483–498, Online. Association for Computational Linguistics.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. CoRR, abs/2004.11867.
Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1535–1545, Austin, Texas. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.