DenTra: Denoising and Translation Pre-training for Multilingual Machine Translation

Samta Kamboj, Sunil Kumar Sahu, Neha Sengupta
Inception Institute of Artificial Intelligence
Abu Dhabi, UAE
samta.kamboj@g42.ai, {sunil.sahu, neha.sengupta}@inceptioniai.org

Abstract

In this paper, we describe our submission to the *WMT-2022: Large-Scale Machine Translation Evaluation for African Languages* under the *Constrained Translation track*. We introduce DENTRA, a novel pre-training strategy for a multilingual sequence-to-sequence transformer model. DENTRA pre-training combines denoising and translation objectives to incorporate both monolingual and bitext corpora in 24 African, English, and French languages. To evaluate the quality of DENTRA, we fine-tuned it with two multilingual machine translation configurations, one-to-many and many-to-one. In both pre-training and fine-tuning, we employ only the datasets provided by the organisers. We compare DENTRA against a strong baseline, M2M-100, in different African multilingual machine translation scenarios and show gains in 3 out of 4 subtasks.

1 Introduction

Despite the compelling performance of machine translation (MT) in many European and Asian languages, their quality in African languages is relatively low. This is primarily because there are approximately 2000 known languages in the African continent, out of which very few languages have any significant presence on the Web (Eberhard et al., 2020; Emezue and Dossou, 2021; Adelani et al., 2022a). As a result, many African languages are not included in publicly available bitext resources, which are typically created by employing heuristics on large amounts of data crawled from the Web (Tiedemann, 2012; El-Kishky et al., 2020; Schwenk et al., 2021; Goyal et al., 2022).

To take a step towards addressing the under-representation of African languages in MT, WMT-2022 presented the Constrained Translation track under *Large-Scale Multilingual African Translation* (Adelani et al., 2022b), which releases bitext and monolingual corpora for 24 African languages, and participants are only allowed to use the provided data. Our submission is to the aforementioned track.

Roughly 34% of the provided data is monolingual, spread across 24 African languages pertaining to this task. Since the volume of bitext data provided is limited, our submission aims to leverage the monolingual data to improve the performance of a multilingual machine translation model. To leverage monolingual data in translation, pre-training the model is an obvious choice.

There are several existing multilingual pre-trained models such as mBART (Liu et al., 2020), mT5 (Xue et al., 2021), byT5 (Xue et al., 2022), mRASP (Pan et al., 2021), mRASP2 (Pan et al., 2021), and M2M-100[1] (Fan et al., 2021) that are trained on monolingual data, bitext data, or both, and have been demonstrated to improve translation performance for specific language pairs. But, these models do not include many of the African languages of interest. For example, the 50 languages covered by mBART50 include only two out of the 24 African languages in the shared task while M2M-100 includes 14. Moreover, all of these multilingual models rely on specially designated language id tokens to translate between each language pair. As a result, adding unseen languages requires pre-training again. Adelani et al. (2022a) investigated a way to leverage pre-trained models including M2M-100, mT5, byT5, and mBART for the translation of unseen languages. But the scarcity of African language texts in the pre-training corpora results in a marginal improvement in the translation quality of the fine-tuned model (Adelani et al., 2022a). Among the pre-trained models, the authors have noted that fine-tuning M2M-100 results in the best translation performance for African languages.

[1] Although M2M-100 is trained for many-to-many translation tasks, it has been used as a pre-trained model by Adelani et al. (2022a) for African MT. Therefore, we also consider it as a pre-trained model in this work.

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 1057 - 1067
December 7-8, 2022 ©2022 Association for Computational Linguistics

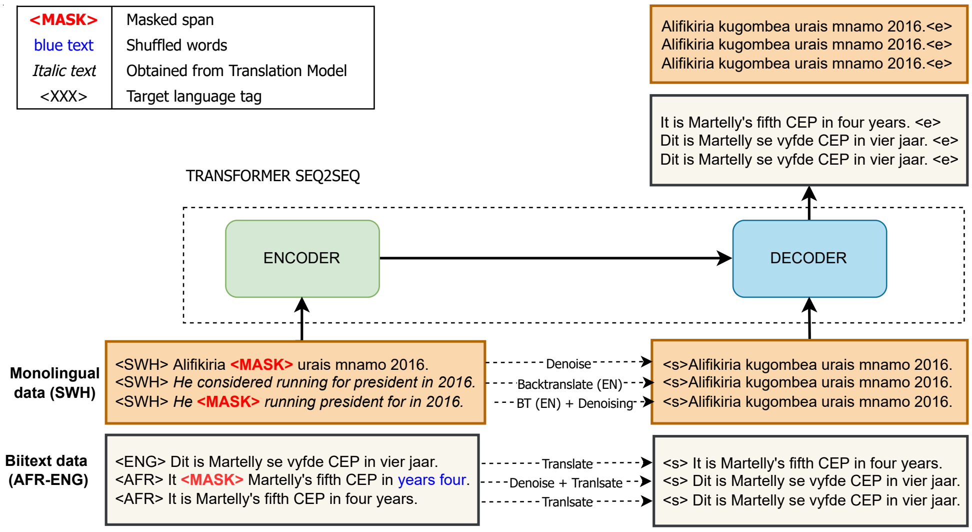

<MASK>
Masked span
blue text
Shuffled words
Italic text
Obtained from Translation Model
<XXX>
Target language tag
Alifikiria kugombea urais mnamo 2016.<e>
Alifikiria kugombea urais mnamo 2016.<e>
Alifikiria kugombea urais mnamo 2016.<e>
It is Martelly's fifth CEP in four years. <e>
Dit is Martelly se vyfde CEP in vier jaar. <e>
Dit is Martelly se vyfde CEP in vier jaar. <e>
TRANSFORMER SEQ2SEQ
ENCODER
DECODER
Monolingual data (SWH)
<SWH> Alifikiria <MASK> urais mnamo 2016.
<SWH> He considered running for president in 2016.
<SWH> He <MASK> running president for in 2016.
Denoise
Backtranslate (EN)
BT (EN) + Denoising
<s>Alifikiria kugombea urais mnamo 2016.
<s>Alifikiria kugombea urais mnamo 2016.
<s>Alifikiria kugombea urais mnamo 2016.
Biitext data (AFR-ENG)
<ENG> Dit is Martelly se vyfde CEP in vier jaar.
<AFR> It <MASK> Martelly's fifth CEP in years four.
<AFR> It is Martelly's fifth CEP in four years.
Translate
Denoise + Tranlsate
Tranlsate
<s> It is Martelly's fifth CEP in four years.
<s> Dit is Martelly se vyfde CEP in vier jaar.
<s> Dit is Martelly se vyfde CEP in vier jaar.

Figure 1: Pre-training Overview

In its multilingual pre-training strategy, mBART uses a denoising objective (Liu et al., 2020; Lewis et al., 2020) on combined monolingual corpora of several languages to train a Transformer sequence-to-sequence model (Vaswani et al., 2017). This strategy reduces the dependency on bitext by learning meaningful representations for multiple languages. The pre-trained model is fine-tuned for MT using bitext data. While this methodology does result in improved MT performance, the pre-training objective used does not induce the representation of similar sentences across languages to align, since it uses only monolingual data (Lin et al., 2020).

In contrast, mRASP2 combines the use of monolingual and bitext corpora. Their pre-training methodology is geared towards closing the representation gap across languages, bringing words and phrases with similar meanings across languages closer in the feature space. This results in better multilingual translation performance (Pan et al., 2021). To induce cross-language representations, mRASP2 uses word or phrase level dictionaries to augment both monolingual and bitext data, by replacing randomly chosen tokens in the source sentence by its corresponding words in another language. Since in this work we address primarily low-resource or under-represented languages, separate dictionaries are non-trivial to construct.

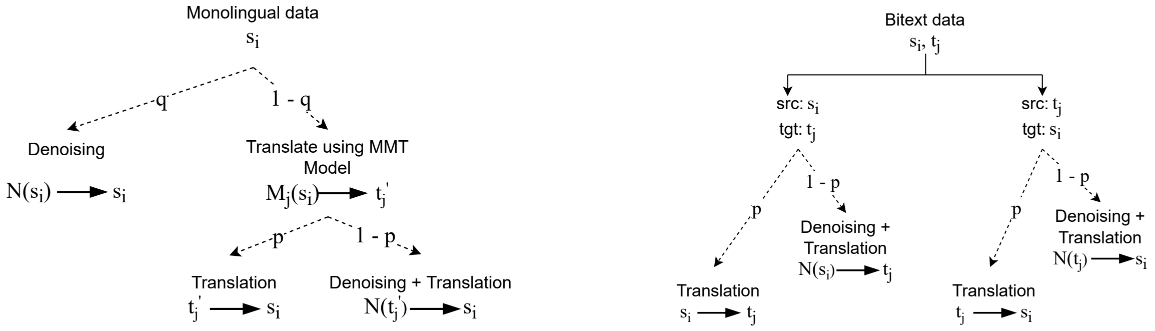
M2M-100 is a massively multilingual model trained on heuristically mined massive bitext corpora spanning 100 languages and 9,900 language pairs (Fan et al., 2021). Fine-tuning M2M-100 with bitext data from the shared task results in minimal improvement over a multilingual model trained from scratch (Section 6). We hypothesize that this is due to M2M-100's subword tokenizer. Since a majority of African languages are written in the Latin script, several subword units are common to many languages. For example, using the M2M-100 tokenizer in the WMT dataset, about 96% of the distinct subwords in African languages also appear in English. Since English corpora dominate the M2M-100 training dataset, the learnt representation of these common tokens are influenced majorly by English, limiting the contribution of African languages.

To address all of the above, we propose DENTRA, which uses a novel pre-training strategy and is trained exclusively on languages from the shared task using both monolingual and bitext data. Inspired by mRASP2 (Pan et al., 2021), our pre-training objective is also designed to explicitly reduce the representation gap between different languages. Figure 1 shows an overview of our pre-training technique.

To measure the effect of pre-training, we fine-tune the pre-trained model in one-to-many and many-to-one configurations of multilingual MT . In three out of four setups, average BLEU score of fine-tuned DENTRA exceeds that of fine-tuned M2M-100 by up to 1.56 points.

2 Definitions and Model Architecture

Task Description: The constrained translation track under WMT-2022 (Adelani et al., 2022b) consists of the following subtasks: English to 22 African languages (eng→{afs}), 22 African languages to English ({afs}→eng), French to 4



(a) Monolingual data

(b) Bitext data

Figure 2: Pre-training Data Preparation: Pre-training Objective for each example in the corpora are determined using the above trees. Solid lines indicate that both paths are executed. Dotted lines from a node indicate that one of its child nodes are selected at random, with the probability distribution along the edges.

African languages (fra→{afs}), 4 African languages to French ({afs}→fra) and 48 African to African languages within geographical and cultural clusters ({afs}→{afs}). In all tasks, training datasets are from the shared task while the validation and the test sets are from FLORES 200 (Goyal et al., 2022).

Multilingual Machine Translation (MMT): An MMT employs a sequence-to-sequence model to translate between arbitrarily many language pairs (Firat et al., 2016; Aharoni et al., 2019). We denote the set of languages in our corpora as $\mathcal{L} = \{l_1, l_2, \ldots l_n\}$ and the bitext data as $\mathcal{D} = \{\mathcal{D}(l_i, l_j), \quad l_i, l_j \in \mathcal{L}\}$ where $\mathcal{D}(l_i, l_j) = \{(s_i, t_j)\}$ is the parallel corpus for languages l_i and l_j. Monolingual data is denoted $\mathcal{M} = \{\mathcal{M}(l_i), l_i \in \mathcal{L}\}$, and $s_i \in \mathcal{M}(l_i)$ denotes an example in the monolingual corpus for language l_i. For training MMT on bitext data $\mathcal{D}$, an artificial token indicating the target language is prefixed to the source, so that $(s_i, t_j) \in \mathcal{D}(l_i, l_j)$ becomes (<J>s_i, t_j) (Johnson et al., 2017). MMT can be trained in three configurations: one-to-many (**1→M**), many-to-one (**M→1**) and many-to-many (**M→M**) (Tang et al., 2021).

Model Architecture: We use the Transformer *big* architecture described in (Vaswani et al., 2017), with 6 encoder and decoder layers, 16 attention heads, and 1024 model dimension. We train our models using FAIRSEQ (Ott et al., 2019) toolkit, and other hyperparameter values listed in Appendix A.1.

3 Methodology

Our overall methodology employs the pre-training followed by fine-tuning pipeline used in prior work in NMT. (Liu et al., 2020; Lin et al., 2020). We present the pre-training strategy used in this work in Section 3.1 and discuss the fine-tuning configurations used in our submission in Section 3.2.

3.1 Pre-training

In our pre-training, we combine monolingual and bitext data in the same corpus. The objective of pre-training is to either denoise, or translate, or both. For each individual example in the corpus, we randomly select which of these objectives to apply. By interleaving denoising and translation, our goal is to drive the model towards learning cross-lingual representations while at the same time learning robust semantic representations. The strong cross lingual representations enable better few-shot and zero-shot translation performance (Pan et al., 2021). Figure 2 illustrates our pre-training methods for both monolingual and bitext data. $N(\cdot)$ is the noising function which we describe in detail in Section 3.1.3, while $M_j(\cdot)$ denotes the translation function using an MMT model for translation to language l_j. In the remainder of this section, we describe each component of the pre-training individually.

3.1.1 Monolingual Data

Figure 2a shows the two ways in which we utilize monolingual data in pre-training. Independently, for each monolingual example $s_i \in \mathcal{M}(l_i)$ one of denoising or translation is selected with probability q and $1 - q$ respectively.

If denoising is selected, we apply a denoising objective similar to mBART (Liu et al., 2020) by masking or shuffling randomly chosen spans. If translation is selected, s_i is first translated to all languages l_j for which $\mathcal{D}(l_i, l_j) \in \mathcal{D}$. The translation is done by MMT models which are obtained by training Transformer models from scratch on $\mathcal{D}$, described in Section 5.1. Lets $M_j(s_i) = t'_j$ is the translation of s_i into language l_j. For the pair (s_i, t'_j), either translation only, or denoising + translation is selected with probabilities p and $1-p$ respectively. If translation is chosen, s_i must be reconstructed from t'_j, following traditional back-translation (Sennrich et al., 2016). If denoising + translation is selected, then s_i must be reconstructed from the noised version $N(t'_j)$. In this manner, the pre-training also incorporates back-translation for utilizing monolingual data.

3.1.2 Bitext Data

Figure 2b shows the bitext data usage in pre-training. Given a pair of sentences (s_i, t_j), the pre-training procedure treats s_i as source and t_j as target, and vice versa. Therefore, two pre-training examples are generated for each example in the bitext data. This is in contrast to pre-training with monolingual data described above, where only one pre-training example was generated per input example. Having designated either s_i or t_j as source, the pre-training procedure follows a path similar to that of monolingual pre-training with backtranslation.

3.1.3 Noising Function

The noising function $N(\cdot)$ largely follows the noising techniques used in Liu et al. (2020). Given an input sentence s, $N(s)$ randomly selects a noising type and applies it on s.

Mask only: With probability p_m, $N(s)$ applies span masking on s. It randomly samples spans of tokens from s, with length of the span drawn from a geometric distribution with parameter m_{sl}, and clipped at 3. The fraction of tokens thus masked is at most m_{sr}. Each masked span is either replaced by a single <MASK> token, or deleted, or replaced by randomly selected word in another language with equal probabilities.

Shuffle only: With probability p_s, $N(s)$ applies shuffling to s. An s_r fraction of tokens are selected at random from tokens in s, and permuted among each other, leaving the unsampled tokens intact.

Mask and Shuffle: With probability p_{ms}, $N(s)$ applies both masking and shuffling to s. First, masking is applied as described above. During the shuffling step, <MASK> tokens are excluded from the tokens to be sampled for shuffling.

None: With probability $1 - p_m - p_s - p_{ms}$, no noising is applied on the input.

3.1.4 Combining Datasets

We combine both $\mathcal{D}$ and $\mathcal{M}$ in pre-training. In order to balance the training dataset across language pairs, we apply temperature based sampling following (Fan et al., 2021) with one major change. Since we are operating in a data constrained setting, we do not reduce the size of any dataset.

Let $N_{(i,j)} = |\mathcal{D}(l_i, l_j)|$ the size of the bitext $\mathcal{D}(l_i, l_j)$, and $N_{\mathcal{D}} = \sum_{(i,j)} N_{(i,j)}$. Then the scaled proportion of language pair (l_i, l_j) is $\alpha'_{i,j} = \frac{\alpha_{(i,j)}}{\sum_{(i,j)} \alpha_{(i,j)}}$ where $\alpha_{(i,j)} = \left(\frac{N_{(i,j)}}{N_{\mathcal{D}}}\right)^{\alpha}$. The rescaled size of language pair (l_i, l_j) is then $R_{(i,j)} = max(N_{(i,j)}, \alpha'_{(i,j)} N_{(i,j)})$

We train the transformer network on the combined dataset until convergence, and then select the best checkpoints for further fine-tuning.

3.2 Fine-tuning

For fine-tuning our pre-trained models, we use bitext data $\mathcal{D}$. Unlike pre-training, we don't noise the source side at all. We apply fine-tuning in 1→M and M→1 settings. Following the pre-training setup, we continue to prefix the tag for the target language in the source side, and also rebalance the datasets as in Section 3.1.4. The checkpoint with best BLEU score on validation set is used for final translations. After fine-tuning, we have the following models. (i) eng→{afs}, (ii) {afs}→eng, (iii) fra→{afs}, and (iv) {afs}→fra. For the remaining pairs, i.e. between African languages, we use DENTRA directly.

4 Datasets and Pre-processing

For all experiments, we employ the datasets provided by the organizers, which mainly consist of datasets from Opus (Tiedemann, 2012), Mafand (Adelani et al., 2022a) and Web crawled aligned through LASER (Heffernan et al., 2022). It is worth mentioning that, in all pre-training and fine-tuning we only used a monolingual corpus of 26 languages and English and French-centric bitext. We have not used any African to African bitext in our experiments. Prior to using for pre-training or fine-tuning, datasets were filtered, cleaned, and preprocessed.

4.1 Data Filtering

Based on characteristics of the dataset and a few observed issues, we employed several heuristics to reject highly noisy examples from $\mathcal{D}$ and $\mathcal{M}$. Given an example $(s_i, t_j) \in \mathcal{D}(l_i, l_j)$, we reject it if (i) $|s_i| < 3$ or $|t_j| < 3$, (ii) $|s_i| > 1000$ or $|t_j| > 1000$, (iii) a character other than . appears at least 5 consecutive times in either s_i or t_j, (iv) a word other than . appears at least 3 consecutive times in either s_i or t_j, (v) s_i is identical to t_j, (vi) $|s_i|/|t_j| < 0.2$ or > 5, (vii) langid of s_i or t_j is not the expected langid with a confidence of at least 80% (where langid is computed using fasttext[2]), (viii) the fraction of characters not belonging in this language are more than 50% [3]. For monolinugal data $\mathcal{M}$, we apply rules corresponding to (i)-(iv) , and (vii)-(viii) above.

After filtering and pre-processing the size of the datasets (combined by English centric, French centric, or monolingual) obtained are shown in Table 1[4]. Full list is shown in Appendix A.2.

Dataset	# Datasets	Total Size	Min	Max	Δ
eng-{afs}	22	109.36	0.21	32.01	21.32 %
fra-{afs}	4	13.40	0.22	11.51	3.96 %
Mono	26	34.17	0.0	12.73	0.58 %

Table 1: Data set sizes specified in Million sentence pairs (or sentences). Δ refers to the percentage of sentence pairs (or sentences) rejected after filtering and pre-processing

Param	p	q	p_m, p_s, p_{ms}	m_{sl}	m_{sr}	s_r	α
Value	0.25	0.66	0.25	0.15	0.2	0.05	0.7

Table 2: Hyper-parameter values used in our data preparation

5 Experimental Setup

Table 2 specifies the hyperparameters we have used in pre-training (Section 3.1). We train the model for 6 epochs and select the best checkpoint based on pre-training task performance on a held out validation set.

The best pre-training checkpoint was subsequently fine-tuned for various tasks where we consider the concatenation of all FLORES 200 *dev sets* for the relevant translation directions as validation set. The FLORES *dev sets* have 997 examples for all language pairs.

For evaluation, we use two test sets. The FLORES *devtest* set (subsequently, we refer to it as FLORES *test set*), which has 1012 examples for all language pairs, and an in-domain test set, which is randomly sampled from the provided bitext data and has about 5000 examples from each language pair. Unless otherwise specified, we report performances on the FLORES *test set*.

We use tokenized BLEU from Moses[5] to measure the performance of all translations. Prior to computing BLEU we word tokenize all translations, also using Moses.

5.1 Models

We prepared following baselines for the comparison of our pre-trained and fine-tuned models:

MMT[6] is trained from scratch separately for four tasks with their corresponding bitext: eng→{afs}, {afs}→eng, fra→{afs} and {afs}→fra. We evaluate it in 1→M and M→1 setups.

M2M-100 is trained on many-to-many datasets of 100 languages. We use the trained version provided by the authors (Fan et al., 2021) and evaluate it in all setups, 1→M, M→1, and M→M. Note that M2M-100 does not support all language pairs in this task and thus, we report performance on only the common language pairs. In particular, M2M-100 includes 14/22 languages in eng↔{afs}, 3/4 in fra↔{afs} and 22/48 in {afs}→{afs}. In all M2M-100 experiments, we employ its 418M parameters checkpoint.

M2M_FT employs the pre-trained checkpoint of M2M-100 and fine-tunes it with bitext data. Similar to Adelani et al. (2022a), unseen African languages {*kam, kin, luo, nya, orm, sna, tso, umb*} are mapped to {*km , ht, lo, yi, fy, ba, kk, uz*} respectively for fine-tuning M2M-100. M2M_FT is used for evaluations in 1→M and M→1 setups only.

The following are the models trained in this work for demonstrating the importance of our pre-training and fine-tuning strategies.

DenTra is the pre-trained model described in Section 3, which we train using (i) monolingual data in 26 languages and (ii) bitext data for only English and French centric directions. We compare DenTra against the corresponding baselines in all setups, 1→M, M→1, and M→M.

DenTra_FT uses bitext data for English and French centric directions to fine-tune the DenTra

[2] https://fasttext.cc/docs/en/python-module.html

[3] Character sets for each language are built by manually curating distinct characters obtained from $\mathcal{D}$

[4] For the Kinyarwanda language, no monolingual data was provided. We reused the Kinyarwanda side from the Kinyarwanda-English bitext for this purpose. For English and French, we randomly sampled 1 million sentences from the combined English/French sides of the bitext datasets provided.

[5] https://github.com/moses-smt/mosesdecoder

[6] These models are also used for backtranslation described in Section 3.1.1

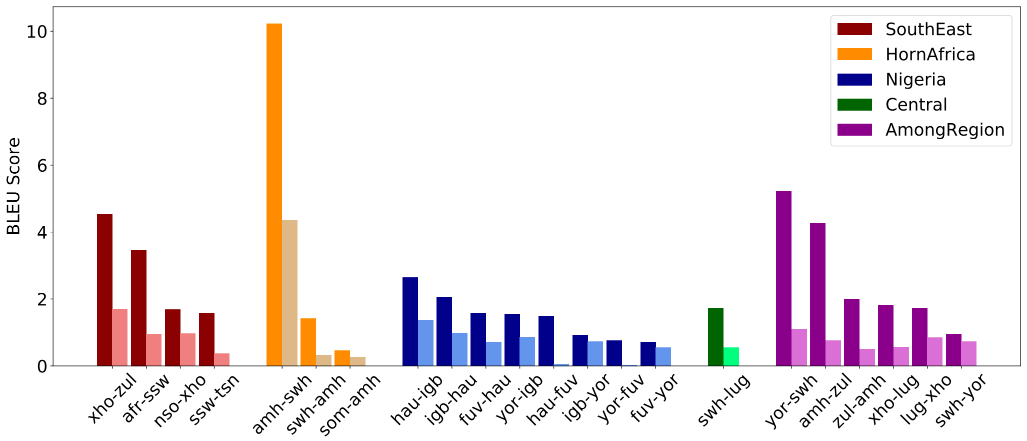


Figure 3: BLEU scores for M2M-100 and DenTra (zero-shot) among African languages on the Flores 200 test set. Dark colors represent DenTra and light colors represent M2M-100.

Tasks	M2M-100	DenTra
eng→{afs}	4.51	9.13
fra→{afs}	4.74	9.48
{afs}→eng	9.13	22.63
{afs}→fra	8.21	15.28

Table 3: Average performance for common languages of DenTra and M2M-100 before fine-tuning

model. DenTra_FT is trained and evaluated against the baselines in only the 1→M and M→1 setups.

6 Comparisons with Baselines

6.1 Without Fine-tuning

In this section, we will show the advantage of DenTra over M2M-100 for the 26 languages in the task without any fine-tuning on either models. As DenTra and M2M-100 both include bitext in their training, we can directly use them for translation.

Table 3 shows the average performance of M2M-100 and DenTra for the 14 English and 3 French centric tasks in M→1 and 1→M setups. In all four tasks, DenTra outperforms M2M-100 by significant margins.

Furthermore, in Figure 3 we display the performance of M2M-100 and DenTra on {afs}→{afs} tasks, i.e. translation between African languages. Similar to M→1 and 1→M setup, we include only the 22 common directions of DenTra and M2M-100 (Full performance list for DenTra is provided in Appendix A.3). Note this is the *zero-shot setting* (Johnson et al., 2017) for both[7]. However, in all translation directions, DenTra outperforms M2M-100 by large margins. These results shows the advantage of pre-training with combined monolingual and bitext data for only the desired set of languages, over pre-training with a large number of additional languages. This confirms our hypothesis discussed in Section 1.

6.2 With Fine-tuning

We evaluate DenTra after it is fine-tuned on the four M→1 and 1→M tasks. Tables 4, 5, 6, and 7 show the BLEU scores for DenTra_FT along with all baselines. Following conclusions may be drawn:

All model variants including DenTra_FT, M2M_FT and DenTra are significantly better than M2M-100 across all language pairs. The general trend of performance comparison in best to worst order is DenTra_FT, M2M_FT, MMT, DenTra, M2M-100, except {afs}→fra where M2M_FT outperforms DenTra_FT.

Improvement of M2M_FT and DenTra_FT over MMT shows the advantage of fine-tuning after pre-training in general. Moreover, since DenTra_FT typically outperforms M2M_FT also, this demonstrates the advantage of including the monolingual corpora and denoising objectives in the pre-training phase.

Generally, it has been shown that combining low resource and high resource languages in a single translation model benefits the low resource languages. We also observe similar behavior as shown by our MMT model. However, we find that extreme multilingual models like M2M-100 must necessar-

[7] Our assumption is that M2M-100 is zero shot in these settings.

Model	$xxx \rightarrow$ **eng**																							
	afr	amh	ful	hau	ibo	kam	kin	lug	luo	nso	nya	orm	sna	som	ssw	swh	tsn	tso	umb	xho	yor	zul	**AVG**	**MED**
MMT	**54.59**	22.47	5.51	25.85	19.67	8.38	22.9	15.77	16.58	30.78	21.21	10.61	21.7	19.6	23.36	37.35	21.21	23.51	5.27	30.38	**13.85**	31.12	21.89	21.46
M2M-100	43.48	6.64	1.98	6.4	5.62	-	-	2.63	-	4.05	-	-	-	2.94	4.95	25.62	0.78	-	-	10.35	1.93	10.4	9.13	5.28
M2M_FT	53.42	22.87	5.62	23.44	18.34	6.65	20.96	14.17	14.79	29.3	20.56	9.88	21.29	18.1	23.33	36.54	20.47	22.13	4.95	29.59	11.46	29.93	20.81	20.76
DenTra	51.65	18.85	5.36	23.24	16.73	8.59	22.24	15.12	14.81	27.85	19.57	9.34	19.77	16.77	20.95	34.16	18.5	21.72	4.9	27.86	11.65	28.08	19.9	19.21
DenTra_FT	54.46	**23.7**	**5.78**	**26.64**	**20.05**	**9.26**	**22.85**	**16.25**	**16.89**	**31.53**	**21.8**	**10.64**	**22.53**	**19.66**	**23.73**	**38.77**	**21.47**	**23.71**	**5.32**	**31.01**	13.72	**32.41**	**22.37**	**22.16**

Table 4: BLEU score on the Flores 200 test set, before and after fine-tuning for English centric MT. For each subtask, the **best model** is bold

Model	**eng** $\rightarrow xxx$																							
	afr	amh	ful	hau	ibo	kam	kin	lug	luo	nso	nya	orm	sna	som	ssw	swh	tsn	tso	umb	xho	yor	zul	**AVG**	**MED**
MMT	40.65	**9.12**	0.66	**22.64**	15.41	**3**	14.52	**6.04**	6.49	23.71	13.82	**2.1**	11.57	**9.83**	**7.48**	33.92	18.43	**16.3**	0.96	**15.76**	**2.52**	15.08	13.18	12.7
M2M-100	26.99	0.51	0.25	2.46	2.53	-	-	1.09	-	0.65	-	-	-	0.49	1.05	19.35	2.61	-	-	2.03	1.07	2.12	4.51	1.56
M2M_FT	38.62	6.85	0.64	18.58	11.72	2.93	14.22	5.93	6.94	**24.65**	12.38	1.91	10.33	9.15	7.21	29.43	17.17	16.24	1.15	14.23	2.2	12.96	12.07	11.03
DenTra	40.38	4.45	**0.76**	21.53	13.72	2.43	12.79	5.53	6.9	21.97	13.13	1.72	10.73	8.13	6.05	33.08	16.23	13.08	1	13.61	2.39	13.92	11.98	11.76
DenTra_FT	**40.95**	8.42	0.64	22.6	**15.69**	2.68	**14.54**	5.97	7.29	24.55	**14.14**	2.06	**11.68**	9.79	7.42	**34.69**	**18.45**	16.24	**1.16**	**15.76**	2.41	**15.32**	**13.29**	**12.91**

Table 5: BLEU score on the Flores 200 test set, before and after fine-tuning for English centric MT. For each subtask, the **best model** is bold

Model	$xxx \rightarrow$ **fra**					
	lin	kin	swh	wol	**AVG**	**MED**
MMT	15.46	17.61	27.29	9.69	17.51	16.54
M2M-100	2.78	-	19.79	2.07	8.21	2.78
M2M_FT	**17.02**	**18.4**	28.22	**11.44**	**18.77**	**17.71**
DenTra	14.38	16.13	22.14	9.33	15.5	15.25
DenTra_FT	16.27	18.28	**28.64**	11.28	18.62	17.27

Table 6: BLEU score on the Flores 200 test set, before and after fine-tuning for French centric MT. For each subtask, the **best model** is bold

Model	**fra** $\rightarrow xxx$					
	lin	kin	swh	wol	**AVG**	**MED**
MMT	13.4	10.08	21.45	4.56	12.37	11.74
M2M-100	0.93	-	12.88	0.42	4.74	0.93
M2M_FT	14.32	10.67	21.1	4.54	12.66	12.49
DenTra	4.97	9.76	21.02	2.46	9.55	7.365
DenTra_FT	**14.4**	**10.85**	**22.09**	**5.09**	**13.11**	**12.62**

Table 7: BLEU score on the Flores 200 test set, before and after fine-tuning for French centric MT. For each subtask, the **best model** is bold

ily have larger capacity to represent all languages in its corpora. In particular, if the languages of interest are restricted, it is better to also restrict pre-training to these languages only (Adelani et al., 2022a).

Finally, we note that the performance of translation models where African languages are the target language are generally lower than those where English or French are the target language. This is expected, since the volume of data where each individual African language appears on the target side is much lower than English or French.

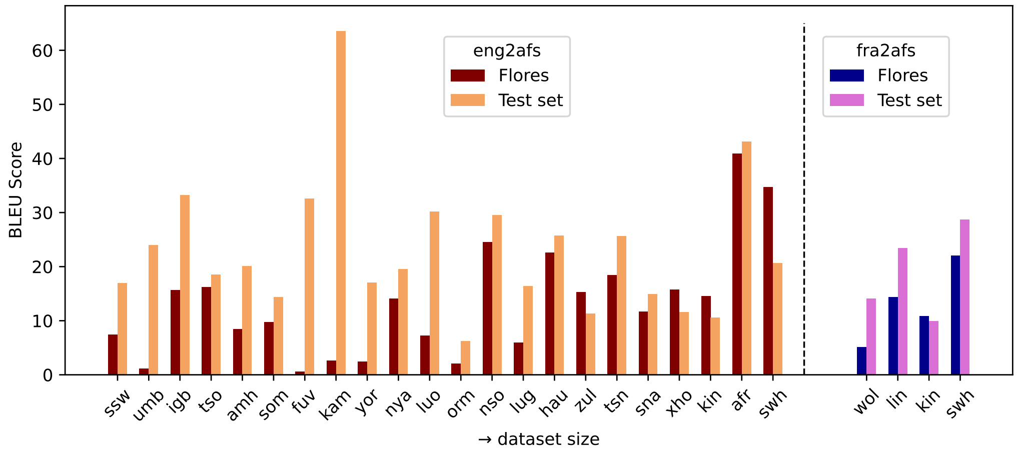


Figure 4: BLEU score comparison of the DENTRA_FT model for FLORES 200 and in-domain test set (isolated from bitext training data $\mathcal{D}$ for English/ French to African languages.

7 Analysis

In this section, we conduct a set of analytical experiments to better understand the datasets and what contributes to performance gains.

Figure 4 shows the BLEU scores of the DENTRA_FT model on {eng,fra}→{afs} translation directions, on the both FLORES 200 test set and the in-domain test set sampled from the bitext data prior to training. The language pairs on the horizontal axis are ordered by the dataset size (left to right in increasing order) independently for English and French centric directions. For the English centric translation (eng→{afs}), the BLEU scores on both test sets have little correlation with the dataset size, indicating noisy data. Some languages, such as *umb, fuv, kam,* and *yor* have stark difference between FLORES and in-domain test sets, indicating that these datasets may have predictable patterns that have no relevance to the translation of these languages. This is further exemplified by the comparison to *tso*, which has a smaller dataset yet exhibits better generalization.

Further investigations reveal two primary problems with the bitext data. First, some of these languages have several duplicates in the African side of the data. For example, for Kamba-English (*kam-eng*) dataset, the distinct number of Kamba sentences is less than 5% of the total dataset size. However, this is not consistent across all languages exhibiting overfitting on the training data, as the number of distinct Yoruba (*yor*) sentences in its bitext is about 95% of the total dataset.

Second, the African side of many datasets contain a large fraction of Indic languages from the Social Media domain. Strict heuristics designed based on manual inspection by the authors rejected about 637,000 examples as being clearly in the Hindi language. Clearly, neither langid nor the LASER encoder (Schwenk and Douze, 2017) are able to reliably detect and align data for these languages. We postulate that low resource languages form a vicious cycle for MT systems trained on bitext data created using multilingual encoders. This opens avenues for future work to explore bitext creation for low resource languages.

8 Conclusion

DENTRA has shown significant performance gains in Multilingual Machine Translation for African languages as demonstrated in this paper. DENTRA integrates denoising, backtranslation, and translation into the same pre-training setup, and has helped to improve MT performance after fine-tuning for both English and French centric translation. We have shown that massively multilingual models like M2M-100 may not be a good choice for fine-tuning when the languages to be translated from/to are restricted to a small set. Finally, we have studied the variation in performance and reported issues seen in heuristically created bitext data. While this is a known issue, we show this problem to be exacerbated for low-resource languages that share the alphabet with high-resource ones.

References

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022a. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

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

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2020. Ethnologue: Languages of the world. twenty-third edition. In *eds*.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Chris Chinenye Emezue and Bonaventure F. P. Dossou. 2021. MMTAfrica: Multilingual machine translation for African languages. In *Proceedings of the Conference on Machine Translation*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*

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

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. *arXiv preprint arXiv:2205.12654*.

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*.

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

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

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

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the Workshop on Representation Learning for NLP*.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the*

Annual Meeting of the Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems*.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*.

Linting Xue, Noah Constant, Adam Roberts, Mihir 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 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

A Appendix

A.1 Model Hyper-parameters

For pre-training and fine-tuning the DENTRA, we base our experiments on FAIRSEQ toolkit. Table 8 presents the hyper-parameter values used in all experiments. For fine-tuning, we employ the best checkpoint obtained from pre-training and continue to train them without resetting the *lr-scheduler*.

A.2 Languages in the Dataset

Table 9 shows the languages used in our experiments along with their bitext and monolingual data sizes.

A.3 Cluster wise Performance

Table 10 shows the performance of DENTRA and M2M-100 on the FLORES 200 test set for different African language pairs clustered geographically/culturally.

Params	Values
optimizer	adam
adam-betas	’(0.9, 0.98)’
clip-norm	0.0
lr	0.0005
lr-scheduler	inverse_sqrt
warmup-updates	4000
warmup-init-lr	1e-07
dropout	0.3
criterion	label_smoothed_cross_entropy
label-smoothing	0.1
max-tokens (batch size)	3584
num-updates (Pre-training)	1609115
num-updates (Fine-tuning) eng-{af}	548227
num-updates (Fine-tuning) fra-{af}	49600

Table 8: Model hyper-parameters and their values

Languages	ISO	Bitext Size	Monolingual Size	Rejected Bitext Size
Afrikaans	afr	13.9	12.732	0.18
Amharic	amh	1.02	0.006	0.11
Nigerian Fulfulde	fuv	1.3	0.255	0.15
Hausa	hau	3.6	3.513	5.5
Igbo	ibo	0.4	0.452	0.1
Kamba	kam	1.58	0.01	0.08
Kinyarwanda	kin	9.6 (eng)	-	0.36 (eng)
		1.2 (fra)	-	0.15 (fra)
Luganda	lug	3.39	0.11	0.11
Luo	luo	2.6	0.035	0.16
Northern Sotho	nso	2.9	0.018	0.18
Chichewa	nya	1.7	0.261	0.14
Oromo	orm	2.7	0.134	0.13
Swati	ssw	8.6	0.257	0.02
Shona	sna	1.25	0.007	0.3
Somali	som	0.2	-	0.15
Swahili	swh	31.7 (eng)	12.642	0.8 (eng)
		11.4 (fra)	-	0.3 (fra)
Tswana	tsn	5.6	0.04	0.43
Xitsonga	tso	0.6	0.037	0.05
Umbundu	umb	0.2	0.043	0.1
Xhosa	xho	9.3	0.308	19.78
Yoruba	yor	1.6	0.51	0.1
Zulu	zul	3.9	0.557	0.23
Lingala	lin	0.3	0.042	0.06
Wolof	wol	0.2	0.206	0.03
English	eng	-	1	0
French	fra	-	1	0

Table 9: Languages, their ISO codes used in the paper, and their corresponding data sizes (in Million sentences)
.

Cluster ID	Task	DenTra	M2M-100
A	xho→zul	4.54	1.7
	zul→sna	2.71	-
	sna→afr	10.65	-
	afr→ssw	3.47	0.95
	ssw→tsn	1.58	0.37
	tsn→tso	2.28	-
	tso→nso	3.8	-
	nso→xho	1.69	0.96
B	swh→am	1.41	0.33
	amh→swh	10.22	4.35
	luo→orm	0.63	-
	som→amh	0.46	0.26
	orm→som	0.85	-
	swh→luo	2.4	-
	amh→luo	3.51	-
	luo→som	2.37	-
C	hau→ibo	2.64	1.37
	ibo→yor	0.92	0.72
	yor→fuv	0.76	0.03
	fuv→hau	1.58	0.71
	ibo→hau	2.06	0.98
	yor→ibo	1.55	0.86
	fuv→yor	0.71	0.55
	hau→fuv	1.49	0.06
	wol→hau	2.11	-
	hau→wol	1.57	-
	fuv→wol	1.24	-
	wol→fuv	1.22	-
D	kin→swh	3.42	-
	lug→lin	1.87	-
	nya→kin	2.22	-
	swh→lug	1.73	0.55
	lin→nya	2.44	-
	lin→kin	2.45	-
	kin→lug	1.96	-
	nya→swh	3.08	-
E	amh→zul	4.27	0.75
	yor→swh	5.21	1.1
	swh→yor	0.95	0.72
	zul→amh	2	0.5
	kin→hau	2.53	-
	hau→kin	2.06	-
	nya→som	3.27	-
	smo→nya	4.28	-
	xho→lug	1.82	0.56
	lug→xho	1.73	0.85
	wol→swh	3.76	-
	swh→wol	1.71	-

Table 10: BLEU scores on the FLORES 200 test set for geographical/cultural clusters. A: South/South East Africa, B: Horn of Africa, C: Nigerian, D: Central African, E: Among the regions