Integrating Unsupervised Data Generation into Self-Supervised Neural Machine Translation for Low-Resource Languages

Dana Ruiter druiter@lsv.uni-saarland.de
Dietrich Klakow dietrich.klakow@lsv.uni-saarland.de
Spoken Language Systems Group, Saarland University, Germany

Josef van Genabith josef.van_genabith@dfki.de
DFKI GmbH & Saarland University, Saarland Informatics Campus, Saarbrücken, Germany

Cristina España-Bonet cristinae@dfki.de
DFKI GmbH, Saarland Informatics Campus, Saarbrücken, Germany

Abstract

For most language combinations, parallel data is either scarce or simply unavailable. To address this, unsupervised machine translation (UMT) exploits large amounts of monolingual data by using synthetic data generation techniques such as back-translation and noising, while self-supervised NMT (SSNMT) identifies parallel sentences in smaller comparable data and trains on them. To date, the inclusion of UMT data generation techniques in SSNMT has not been investigated. We show that including UMT techniques into SSNMT significantly outperforms SSNMT and UMT on all tested language pairs, with improvements of up to $+4.3$ BLEU, $+50.8$ BLEU, $+51.5$ over SSNMT, statistical UMT and hybrid UMT, respectively, on Afrikaans to English. We further show that the combination of multilingual denoising autoencoding, SSNMT with backtranslation and bilingual finetuning enables us to learn machine translation even for distant language pairs for which only small amounts of monolingual data are available, e.g. yielding BLEU scores of 11.6 (English to Swahili).

1 Introduction

Neural machine translation (NMT) achieves high quality translations when large amounts of parallel data are available (Barrault et al., 2020). Unfortunately, for most language combinations, parallel data is non-existent, scarce or low-quality. To overcome this, unsupervised MT (UMT) (Lample et al., 2018b; Ren et al., 2019; Artetxe et al., 2019) focuses on exploiting large amounts of monolingual data, which are used to generate synthetic bitext training data via various techniques such as back-translation or denoising. Self-supervised NMT (SSNMT) (Ruiter et al., 2019) learns from smaller amounts of *comparable* data –i.e. topic-aligned data such as Wikipedia articles– by learning to discover and exploit similar sentence pairs. However, both UMT and SSNMT approaches often do not scale to low-resource languages, for which neither monolingual nor comparable data are available in sufficient quantity (Guzmán et al., 2019; España-Bonet et al., 2019; Marchisio et al., 2020). To date, UMT data augmentation techniques have not been explored in SSNMT. However, both approaches can benefit from each other, as i) SSNMT has strong internal quality checks on the data it admits for training, which can be

of use to filter low-quality synthetic data, and ii) UMT data augmentation makes monolingual data available for SSNMT.

In this paper we explore and test the effect of combining UMT data augmentation with SSNMT on different data sizes, ranging from very low-resource ($\sim 66k$ non-parallel sentences) to high-resource ($\sim 20M$ sentences). We do this using a common high-resource language pair (en–fr), which we downsample while keeping all other parameters identical. We then proceed to evaluate the augmentation techniques on different truly low-resource similar and distant language pairs, i.e. English (en)–{Afrikaans (af), Kannada (kn), Burmese (my), Nepali (ne), Swahili (sw), Yorùbá (yo)}, chosen based on their differences in typology (*analytic*, *fusional*, *agglutinative*), word order (*SVO*, *SOV*) and writing system (*Latin*, *Brahmic*). We also explore the effect of different initialization techniques for SSNMT in combination with finetuning.

2 Related Work

Substantial effort has been devoted to muster training data for **low-resource NMT**, e.g. by identifying parallel sentences in monolingual or noisy corpora in a pre-processing step (Artetxe and Schwenk, 2019a; Chaudhary et al., 2019; Schwenk et al., 2021) and also by leveraging monolingual data into supervised NMT e.g. by including autoencoding (Currey et al., 2017) or language modeling tasks (Gulcehre et al., 2015; Ramachandran et al., 2017). Low-resource NMT models can benefit from high-resource languages through transfer learning (Zoph et al., 2016), e.g. in a zero-shot setting (Johnson et al., 2017), by using pre-trained language models (Conneau and Lample, 2019; Kuwanto et al., 2021), or finding an optimal path for pivoting through related languages (Leng et al., 2019).

Back-translation often works well in high-resource settings (Bojar and Tamchyna, 2011; Sennrich et al., 2016a; Karakanta et al., 2018). NMT training and back-translation have been used in an incremental fashion in both unidirectional (Hoang et al., 2018) and bidirectional systems (Zhang et al., 2018; Niu et al., 2018).

Unsupervised NMT (Lample et al., 2018a; Artetxe et al., 2018; Yang et al., 2018) applies bi-directional back-translation in combination with denoising and multilingual shared encoders to learn MT on very large monolingual data. This can be done multilingually across several languages by using language-specific decoders (Sen et al., 2019), or by using additional parallel data for a related pivot language pair (Li et al., 2020). Further combining unsupervised neural MT with phrase tables from statistical MT leads to top results (Lample et al., 2018b; Ren et al., 2019; Artetxe et al., 2019). However, unsupervised systems fail to learn when trained on small amounts of monolingual data (Guzmán et al., 2019), when there is a domain mismatch between the two datasets (Kim et al., 2020) or when the languages in a pair are distant (Koneru et al., 2021). Unfortunately, all of this is the case for most truly low-resource language pairs.

Self-supervised NMT (Ruiter et al., 2019) jointly learns to extract data and translate from comparable data and works best on 100s of thousands of documents per language, well beyond what is available in true low-resource settings.

3 UMT-Enhanced SSNMT

SSNMT jointly learns MT and extracting similar sentences for training from comparable corpora in a loop on-line. Sentence pairs from documents in languages $L1$ and $L2$ are fed as input to a bidirectional NMT system $\{L1, L2\} \rightarrow \{L1, L2\}$, which filters out non-similar sentences after scoring them with a similarity measure calculated from the internal embeddings.

Sentence Pair Extraction (SPE): Input sentences $s_{L1} \in L1$, $s_{L2} \in L2$, are represented by the sum of their word embeddings and by the sum of the encoder outputs, and scored using the margin-based measure introduced by Artetxe and Schwenk (2019a). If a pair (s_{L1}, s_{L2}) is top

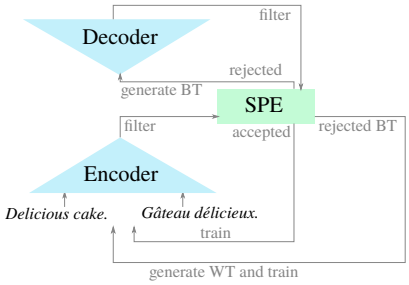

Figure 1: UMT-Enhanced SSNMT architecture (Section 3).

scoring for both language directions *and* for both sentence representations, it is accepted for training, otherwise it is filtered out. This is a strong quality check and equivalent to *system P* in Ruiter et al. (2019). A SSNMT model with SPE is our **baseline (B)** model.

Since most possible sentence pairs from comparable corpora are non-similar, they are simply discarded. In a low-resource setting, this potentially constitutes a major loss of usable monolingual information. To exploit sentences that have been rejected by the SSNMT filtering process, we integrate the following UMT synthetic data creation techniques *on-line* (Figure 1):

Back-translation (BT): Given a rejected sentence s_{L1}, we use the current state of the SSNMT system to back-translate it into s_{L2}^{BT}. The synthetic pair in the opposite direction $s_{L2}^{BT} \rightarrow s_{L1}$ is added to the batch for further training. We perform the same filtering process as for SPE so that only good quality back-translations are added. We apply the same to source sentences in $L2$.

Word-translation (WT): For synthetic sentence pairs rejected by BT filtering, we perform word-by-word translation. Given a rejected sentence s_{L1} with tokens $w_{L1} \in L1$, we replace each token with its nearest neighbor $w_{L2} \in L2$ in the bilingual word embedding layer of the model to obtain s_{L2}^{WT}. We then train on the synthetic pair in the opposite direction $s_{L2}^{WT} \rightarrow s_{L1}$. As with BT, this is applied to both language directions. To ensure sufficient volume of synthetic data (Figure 2, right), WT data is trained on without filtering.

Noise (N): To increase robustness and variance in the training data, we add noise, i.e. token deletion, substitution and permutation, to copies of source sentences (Edunov et al., 2018) in parallel pairs identified via SPE, back-translations and word-translated sentences and, as with WT, we use these without additional filtering.

Initialization: When languages are related and large amounts of training data is available, the initialization of SSNMT is not important. However, similarly to UMT, initialization becomes crucial in the low-resource setting (Edman et al., 2020). We explore four different initialization techniques: $i)$ no initialization (*none*), i.e. random initialization for all model parameters, $ii)$ initialization of tied source and target side word embedding layers only via pre-trained cross-lingual word-embeddings (WE) while randomly initializing all other layers and $iii)$ initialization of all layers via denoising autoencoding (DAE) in a bilingual and $iv)$ multilingual (MDAE) setting.

Finetuning (F): When using MDAE initialization only, the following SSNMT is multilingual, otherwise it is bilingual. Due to the multilingual nature of the SSNMT with MDAE initialization, the performance of the individual languages can be limited by the *curse of multilinguality* (Conneau et al., 2020), where multilingual training leads to improvements on low-resource languages up to a certain point after which it decays. To alleviate this, we finetune converged

	Comparable						Monolingual		
	# Art (k)	VO (%)	# Sent (k)		# Tok (k)		# Sent (k)	# Tok (k)	
en–L			en	L	en	L	en/L	en	L
en–af	73	7.1	4,589	780	189,990	27,640	1,034	34,759	31,858
en–kn	18	1.4	1,739	764	95,481	30,003	1,058	47,136	35,534
en–my	19	2.1	1,505	477	82,537	15,313	997	43,752	24,094
en–ne	20	0.6	1,526	207	83,524	7,518	296	13,149	9,229
en–sw	34	6.5	2,375	244	122,593	8,774	329	13,957	9,937
en–yo	19	5.7	1,314	34	82,674	1,536	547	17,953	19,370

Table 1: Number of sentences (Sent) and tokens (Tok) in the comparable and monolingual datasets. For comparable datasets, we report the number of articles (Art) and percentage of vocabulary overlap (VO) between the two languages in a pair. # Sent of monolingual data (en/L) is the same for en and its corresponding L due to downsampling of en to match L.

multilingual SSNMT models bilingually on a given language pair $L1$–$L2$.

4 Experimental Setting

4.1 Data

MT Training For training, we use Wikipedia (WP) as a comparable corpus and download the dumps[1] and extract comparable articles per language pair (*Comparable* in Table 1) using WikiExtractor[2]. For validation and testing, we use the test and development data from McKellar and Puttkammer (2020) (en–af), WAT2021[3] (en–kn), WAT2020 (en–my) (ShweSin et al., 2018), FLoRes (en–ne) (Guzmán et al., 2019), Lakew et al. (2021) (en–sw), and MENYO-20k (en–yo) (Adelani et al., 2021a). For en–fr we use *newstest2012* for development and *newstest2014* for testing. As the en–af data does not have a development split, we additionally sample 1 k sentences from CCAligned (El-Kishky et al., 2020) to use as en–af development data. The en–sw test set is divided into several sub-domains, and we only evaluate on the TED talks domain, since the other domains are noisy, e.g. localization or religious corpora.

MT Initialization We use the monolingual Wikipedias to initialize SSNMT. As the monolingual Wikipedia for Yorùbá is especially small (65 k sentences), we use the Yorùbá side of JW300 (Agić and Vulić, 2019) as additional monolingual initialization data. For each monolingual data pair en–$\{af,...,yo\}$, the large English monolingual corpus is downsampled to its low(er)-resource counterpart before using the data (*Monolingual* in Table 1).

For the word-embedding-based initialization, we learn CBOW word embeddings using `word2vec` (Mikolov et al., 2013), which are then projected into a common multilingual space via `vecmap` (Artetxe et al., 2017) to attain bilingual embeddings between en–$\{af,...,yo\}$. For the weak-supervision of the bilingual mapping process, we use a list of numbers (en–fr only) which is augmented with 200 Swadesh list[4] entries for the low-resource experiments.

For DAE initialization, we do not use external, highly-multilingual pre-trained language models, since in practical terms these may not cover the language combination of interest[5]. We therefore use the monolingual data to train a bilingual (en+$\{af,...yo\}$) DAE using BART-style

[1] Dumps were downloaded on February 2021 from `dumps.wikimedia.org/`

[2] `github.com/attardi/wikiextractor`

[3] `lotus.kuee.kyoto-u.ac.jp/WAT/indic-multilingual/index.html`

[4] `https://en.wiktionary.org/wiki/Appendix:Swadesh_lists`

[5] This is the case here: `MBart-50` (Tang et al., 2020) does not cover Kannada, Swahili and Yorùbá.

noise (Liu et al., 2020). We set aside $5\,k$ sentences for testing and development each. We use BART-style noise ($\lambda = 3.5$, $p = 0.35$) for word sequence masking. We add one random mask insertion per sequence and perform a sequence permutation. For the multilingual DAE (MDAE) setting, we train a single denoising autoencoder on the monolingual data of all languages, where en is downsampled to match the largest non-English monolingual dataset (kn).

In all cases SSNMT training is bidirectional between two languages en–{af,...,yo}, except for MDAE, where SSNMT is trained multilingually between all language combinations in {af,en,...,yo}.

4.2 Preprocessing

On the Wikipedia corpora, we perform sentence tokenization using NLTK (Bird, 2006). For languages using Latin scripts (af,en,sw,yo) we perform punctuation normalization and true-casing using standard Moses (Koehn et al., 2007) scripts on all datasets. For Yorùbá only, we follow Adelani et al. (2021b) and perform automatic diacritic restoration. Lastly, we perform language identification on all Wikipedia corpora using `polyglot`.[6] After exploring different byte-pair encoding (BPE) (Sennrich et al., 2016b) vocabulary sizes of $2\,k$, $4\,k$, $8\,k$, $16\,k$ and $32\,k$, we choose $2\,k$ (en–yo), $4\,k$ (en–{kn,my,ne,sw}) and $16\,k$ (en–af) merge operations using `sentence-piece`[7] (Kudo and Richardson, 2018). We prepend a source and a target language token to each sentence. For the en–fr experiments only, we use the data processing by Ruiter et al. (2020) in order to minimize experimental differences for later comparison.

4.3 Model Specifications and Evaluation

Systems are either not initialized, initialized via bilingual word embeddings, or via pre-training using (M)DAE. Our implementation of SSNMT is a transformer base with default parameters. We use a batch size of 50 sentences and a maximum sequence length of 100 tokens. For evaluation, we use BLEU (Papineni et al., 2002) calculated using `SacreBLEU`[8,9] (Post, 2018) and all confidence intervals ($p = 95\%$) are calculated using bootstrap resampling (Koehn, 2004) as implemented in `multeval`[10] (Clark et al., 2011).

5 Exploration of Corpus Sizes (en–fr)

To explore which technique works best with varying data sizes, and to compare with the high-resource SSNMT setting in Ruiter et al. (2020), we train SSNMT on en–fr, with different combinations of techniques (+BT, +WT, +N) over decreasingly small corpus sizes. The base (B) model is a simple SSNMT model with SPE.

Figure 2 (left) shows that translation quality as measured by BLEU is very low in the low-resource setting. For experiments with only $4\,k$ comparable articles (similar to the corpus size available for en–yo), BLEU is close to zero with base (B) and B+BT models. Only when WT is applied to rejected back-translated pairs does training become possible, and is further improved by adding noise, yielding BLEUs of 3.38[11] ($en2fr$) and 3.58 ($fr2en$). The maximum gain in performance obtained by WT is at $31\,k$ comparable articles, where it adds $\sim$ 9 BLEU over the B+BT performance. While the additional supervisory signal provided by WT is useful in the low and medium resource settings, up until $\sim$ $125\,k$ articles, its benefits are overcome by

[6] `https://github.com/aboSamoor/polyglot`

[7] `https://github.com/google/sentencepiece`

[8] `https://github.com/mjpost/sacrebleu`

[9] `BLEU+case.mixed+numrefs.4+smooth.exp+tok.intl+version.1.4.9`

[10] `https://github.com/jhclark/multeval`

[11] Note that such low BLEU scores should be taken with a grain of salt: While there is an automatically measurable improvement in translation quality, a human judge would not see a meaningful improvement between different systems with low BLEU scores.

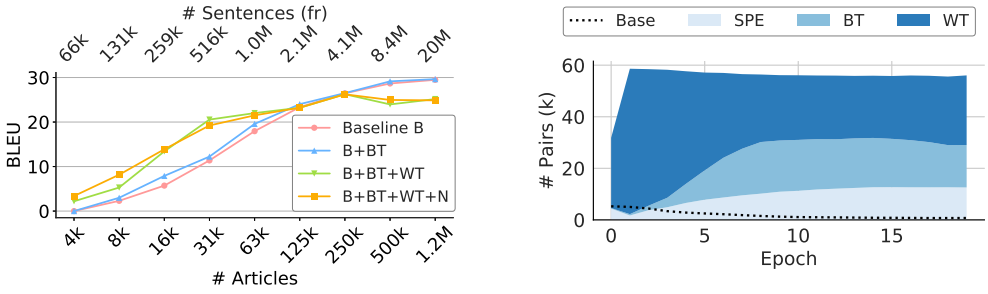

Figure 2: **Left**: BLEU scores ($en2fr$) of different techniques (+BT,+WT,+N) added to the base (B) SSNMT model when trained on increasingly large numbers en–fr WP articles (# Articles). **Right**: Number of extracted (SPE) or generated (BT,WT) sentence pairs (k) per technique of the B+BT+WT model trained on $4\,k$ comparable WP articles. Number of extracted sentence pairs by the base model is shown for comparison as a dotted line.

the noise it introduces in the high-resource scenario, leading to a drop in translation quality. Similarly, the utility of adding noise varies with corpus size. Only BT constantly adds a slight gain in performance of ∼1–2 over all base models, where training is possible. In the high resource case, the difference between B and B+BT is not significant, with BLEU 29.64 ($en2fr$) and 28.56 ($fr2en$) for B+BT, which also leads to a small, yet statistically insignificant gain over the en–fr SSNMT model in Ruiter et al. (2020), i.e. $+0.1$ ($en2fr$) and $+0.9$ ($fr2en$) BLEU.

At the beginning of training, the number of **extracted sentence pairs** (SPE) of the B+BT+WT+N model trained on the most extreme low-resource setting ($4\,k$ articles), is low (Figure 2, right), with $4\,k$ sentence pairs extracted in the first epoch. This number drops further to $2\,k$ extracted pairs in the second epoch, but then continually rises up to $13\,k$ extracted pairs in the final epoch. This is not the case for the base (B) model, which starts with a similar amount of extracted parallel data but then continually extracts less as training progresses. The difference between the two models is due to the added BT and WT techniques. At the beginning of training B+BT+WT is not able to generate backtranslations of decent quality, with only few (196) backtranslations accepted for training. Rejected backtranslations are passed into WT, which leads to large numbers of WT sentence pairs up to the second epoch ($56\,k$). These make all the difference: through WT, the system is able to gain noisy supervisory signals from the data, which leads to the internal representations to become more informative for SPE, thus leading to more and better extractions. Then, BT and SPE enhance each other, as SPE ensures original (clean) parallel sentences to be extracted, which improves translation accuracy, and hence more and better backtranslations (e.g. up to $20\,k$ around epoch 15) are accepted.

6 Exploration of Language Distance

BT, WT and N data augmentation techniques are especially useful for the low- and mid-resource settings of related language pairs such as English and French (both *Indo-European*). To apply the approach to truly low-resource language pairs, and to verify which language-specific characteristics impact the effectiveness of the different augmentation techniques, we train and test our model on a selected number of languages (Table 2) based on their typological and graphemic distance from English (*fusional*→*analytic*[12], SVO, Latin script). Focusing on similarities on

[12]English and Afrikaans are traditionally categorized as fusional languages. However, due to their small morpheme-word ratio, both English and Afrikaans are nowadays often categorized as analytic languages.

	English	Afrikaans	Nepali	Kannada	Yorùbá	Swahili	Burmese
Typology	fusional[9]	fusional[9]	fusional	agglutinative	analytic	agglutinative	analytic
Word Order	SVO	SOV,SVO	SOV	SOV	SOV,SVO	SVO	SOV
Script	Latin	Latin	Brahmic	Brahmic	Latin	Latin	Brahmic
sim(L–en)	1.000	0.822	0.605	0.602	0.599	0.456	0.419

Table 2: Classification (typology, word order, script) of the languages L together with their cosine similarity (sim) to English based on lexical and syntactic URIEL features.

the lexical and syntactic level,[13] we retrieve the URIEL (Littell et al., 2017) representations of the languages using `lang2vec`[14] and calculate their cosine similarity to English. Afrikaans is the most similar language to English, with a similarity of 0.822, and pre-BPE vocabulary (token) overlap of 7.1% (Table 1), which is due to its similar typology (*fusional→analytic*) and comparatively large vocabulary overlap (both languages belong to the West-Germanic language branch). The most distant language is Burmese (sim 0.419, vocabulary overlap 2.1%), which belongs to the Sino-Tibetan language family and uses its own (Brahmic) script.

We train SSNMT with combinations of BT, WT, N on the language combinations en–$\{af$,kn,my,ne,sw,$yo\}$ using the four different types of model initialization (none, WE, DAE, MDAE).

Intrinsic Parameter Analysis We focus on the intrinsic *initialization* and *data augmentation technique* parameters. The difference between no (*none*) and word-embedding (*WE*) **initialization** is barely significant across all language pairs and techniques (Figure 3). For all language pairs, except en–af, MDAE initialization tends to be the best choice, with major gains of $+4.2$ BLEU ($yo2en$, B+BT) and $+5.3$ BLEU ($kn2en$, B+BT) over their WE-initialized counterparts. This is natural, since pre-training on (M)DAE allows the SSNMT model to learn how to generate fluent sentences. By performing (M)DAE, the model also learns to denoise noisy inputs, resulting in a big improvement in translation performance (e.g. $+37.3$ BLEU, $af2en$ DAE) on the en–af and en–sw B+BT+WT models in comparison to their WE-initialized counterparts. Without (M)DAE pre-training, the noisy word-translations lead to very low BLEU scores. Adding an additional denoising task, either via (M)DAE initialization or via adding the +N data augmentation technique, lets the model also learn from noisy word-translations with improved results. For en–af only, the WE initialization generally performs best, with BLEU scores of 52.2 ($af2en$) and 51.2 ($en2af$). For language pairs using different scripts, i.e. Latin–Brahmic (en–$\{kn$,my,$ne\}$), the gain by performing bilingual DAE pre-training is negligible, as results are generally low. These languages also have a different word order (SOV) than English (SVO), which may further increase the difficulty of the translation task (Banerjee et al., 2019; Kim et al., 2020). However, once the pre-training and MT learning is multilingual (MDAE), the different language directions benefit from another and an internal mapping of the languages into a shared space is achieved. This leads to BLEU scores of 1.7 ($my2en$), 3.3 ($ne2en$) and 5.3 ($kn2en$) using the B+BT technique. The method is also beneficial when translating into the low-resource languages, with $en2kn$ reaching BLEU 3.3 (B).

B+BT+WT seems to be the best **data augmentation technique** when the amount of data is very small, as is the case for en–yo, with gains of $+2.4$ BLEU on $en2yo$ over the baseline B. This underlines the findings in Section 5, that WT serves as a crutch to start the extraction and training of SSNMT. Further adding noise (+N) tends to adversely impact on results on this

[13] This corresponds to `lang2vec` features `syntax_average` and `inventory_average`.

[14] `https://pypi.org/project/lang2vec/`

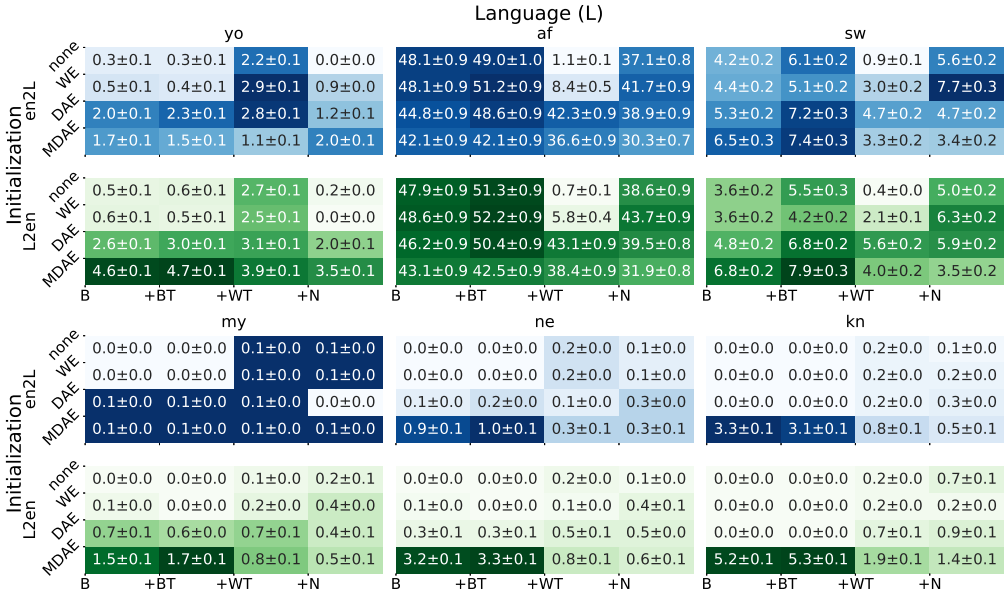

Figure 3: BLEU scores of SSNMT Base (B) with added techniques (+BT,+WT,+N) on low-resource language combinations $en2L$ and $L2en$, with $L = \{af, kn, my, ne, sw, yo\}$.

language pair. On languages with more data available (en–$\{af$,kn,my,ne,$sw\}$), +BT tends to be the best choice, with top BLEUs on en–sw of 7.4 ($en2sw$, MDAE) and 7.9 ($sw2en$, MDAE). This is due to these models being able to sufficiently learn on B (+BT) only (Figure 4), thus not needing +WT as a crutch to start the extraction and MT learning process. Adding +WT to the system only adds additional noise and thus makes results worse.

Extrinsic Parameter Analysis We focus on the extrinsic parameters *linguistic distance* and *data size*. Our model is able to learn MT also on **distant language pairs** such as en–sw (sim 0.456), with top BLEUs of 7.7 ($en2sw$, B+BT+W+N) and 7.9 ($sw2en$, B+BT). Despite being typologically closer, training SSNMT on en–ne (sim 0.605) only yields BLEUs above 1 in the multilingual setting (BLEU 3.3 $ne2en$). This is the case for all languages using a different script than English (kn,my,ne), underlining the fact that achieving a cross-lingual representation, i.e. via multilingual (pre-)training or a decent overlap in the (BPE) vocabulary (as in en–$\{af$,sw,$yo\}$) of the two languages, is vital for identifying good similar sentence pairs at the beginning of training and thus makes training possible. For en–my the MDAE approach was only beneficial in the $my2en$ direction, but had no effect on $en2my$, which may be due to the fact that my is the most distant language from en (sim 0.419) and, contrary to the other low-resource languages we explore, does not have any related language[15] in our experimental setup, which makes it difficult to leverage supervisory signals from a related language.

When the **amount of data** is small (en–yo), the model does not achieve BLEUs above 1 without the WT technique or without (M)DAE initialization, since the extraction recall of a simple SSNMT system is low at the beginning of training (Ruiter et al., 2020) and thus SPE fails to identify sufficient parallel sentences to improve the internal representations, which would then improve SPE recall. This is analogous to the observations on the en-fr base model B

[15]Both Nepali and Kannada share influences from Sanskrit. Swahili and Yorùbá are both Niger-Congo languages, while English and Afrikaans are both Indo-European.

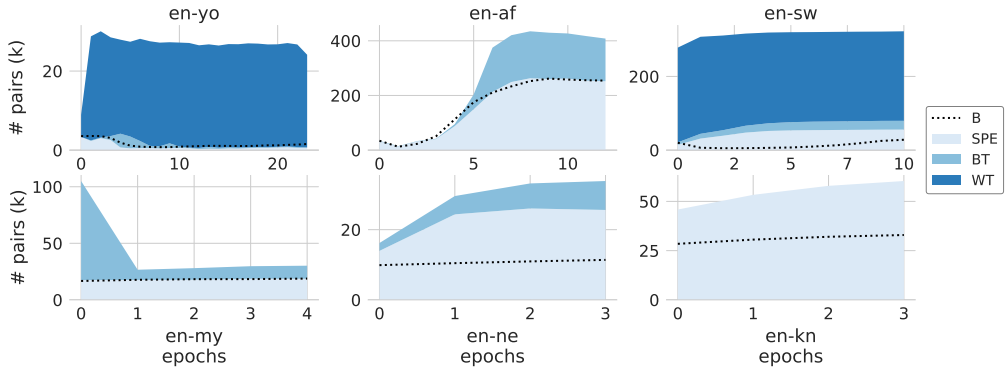

Figure 4: Number of extracted (SPE) or generated (BT,WT) sentence pairs (k) per technique of the best performing SSNMT model ($en2L$) per language L. Number of extracted sentence pairs by the base model (B) are shown for comparison as a dotted line.

trained on $4\,k$ WP articles (Figure 2). Interestingly, the differences between no/WE and DAE initialization are minimized when using WT as a data augmentation technique, showing that it is an effective method that makes pre-training unnecessary when only small amounts of data are available. For larger data sizes (en–$\{af,sw\}$), the opposite is the case: the models sufficiently learn SPE and MT without WT, and thus WT just adds additional noise.

Extraction and Generation The SPE extraction and BT/WT generation curves (Figure 4) for en–af (B+BT, WE) are similar to those on en–fr (Figure 2, right). At the beginning of training, not many pairs ($32\,k$) are extracted, but as training progresses, the model internal representations are improved and it starts extracting more and more parallel data, up to $252\,k$ in the last epoch. Simultaneously, translation quality improves and the number of backtranslations generated increases drastically from $2\,k$ up to $156\,k$ per epoch. However, as the amount of data for en–af is large, the base model B has a similar extraction curve. Nevertheless, translation quality is improved by the additional backtranslations ($+3.1$ BLEU). For en–sw (B+BT+WT+N, WE), the curves are similar to those of en–fr, where the added word-translations serve as a crutch to make SPE and BT possible, thus showing a gap between the number of extracted sentences (SPE) ($\sim\ 5.5\,k$) of the best model and those of the baseline (B) ($\sim$1–$2\,k$). For en–yo (B+BT+WT, WE), the amount of extracted data is very small ($\sim\ 0.5\,k$) for both the baseline and the best model. Here, WT fails to serve as a crutch as the number of extractions does not increase, but instead is overwhelmed by the number of word translations. For en–$\{kn,ne\}$ (MDAE), the extraction and BT curves also rise over time. For en–my, where all training setups show similar translation performance in the $en2my$ direction, we show the extraction and BT curves for B+BT with WE initialization. We observe that, as opposed to all other models, both lines are flat, underlining the fact that due to the lack of sufficiently cross-lingual model-internal representations, the model does not enter the self-supervisory cycle common to SSNMT.

Bilingual Finetuning The overall trend shows that MDAE pre-training with multilingual SSNMT training in combination with back-translation (B+BT) leads to top results for low-resource similar and distant language combinations. For en–afonly, which has more comparable data available for training and is a very similar language pair, the multilingual setup is less beneficial. The model attains enough supervisory signals when training bilingually on en–af, thus the additional languages in the multilingual setup are simply noise for the system. While the MDAE setup with multilingual MT training makes it possible to map distant languages into a

	en–af		*en–kn*		*en–my*		*en–ne*		*en–sw*		*en–yo*	
	→	←	→	←	→	←	→	←	→	←	→	←
Best*	**51.2**	**52.2**	0.3	0.9	0.1	0.7	0.3	0.5	7.7	6.8	**2.9**	3.1
MDAE	42.5	42.5	3.1	5.3	0.1	1.7	1.0	3.3	7.4	7.9	1.5	4.7
MDAE+F	46.3	50.2	**5.0**	**9.0**	**0.2**	**2.8**	**2.3**	**5.7**	**11.6**	**11.2**	**2.9**	**5.8**

Table 3: BLEU scores on the $en2L$ (→) and $L2en$ (←) directions of top performing SSNMT model without finetuning and without MDAE (Best*) and SSNMT using MDAE initialization and B+BT technique with (MDAE+F) and without finetuning (MDAE).

Pair	Init.	Config.	Best	Base	UMT	UMT+NMT	Laser	TSS	#P (k)
$en2af$	WE	B+BT	**51.2±.9**	48.1±.9	27.9±.8	44.2±.9	**52.1±1.0**	35.3	37
$af2en$	WE	B+BT	**52.2±.9**	47.9±.9	1.4±.1	0.7±.1	**52.9±.9**	–	–
$en2kn$	MDAE	B+BT+F	**5.0±.2**	0.0±.0	0.0±.0	0.0±.0	–	21.3	397
$kn2en$	MDAE	B+BT+F	**9.0±.2**	0.0±.0	0.0±.0	0.0±.0	–	40.3	397
$en2my$	MDAE	B+BT+F	**0.2±.0**	0.0±.0	0.1±.0	0.0±.0	0.0±.0	39.3	223
$my2en$	MDAE	B+BT+F	**2.8±.1**	0.0±.0	0.0±.0	0.0±.0	0.1±.0	38.6	223
$en2ne$	MDAE	B+BT+F	**2.3±.1**	0.0±.0	0.1±.0	0.0±.0	0.5±.1	8.8	–
$ne2en$	MDAE	B+BT+F	**5.7±.2**	0.0±.0	0.0±.0	0.0±.0	0.2±.0	21.5	–
$en2sw$	MDAE	B+BT+F	**11.6±.3**	4.2±.2	3.6±.2	0.2±.0	10.0±.3	14.8	995
$sw2en$	MDAE	B+BT+F	**11.2±.3**	3.6±.2	0.3±.0	0.0±.0	8.4±.3	19.7	995
$en2yo$	MDAE	B+BT+F	**2.9±.1**	0.3±.1	1.0±.1	0.3±.1	–	12.3	501
$yo2en$	MDAE	B+BT+F	**5.8±.1**	0.5±.1	0.6±.0	0.0±.0	–	22.4	501

Table 4: BLEU scores of the best SSNMT configuration (columns 2-4) compared with SSNMT base, USMT(+UNMT) and a supervised NMT system trained on Laser extractions (columns 5-8). Top scoring systems (TSS) per test set and the amount of parallel training sentences (#P) available for reference (columns 9-10).

shared space and learn MT, we suspect that the final MT performance on the individual language directions is ultimately being held back due to the multilingual noise of other language combinations. To verify this, we use the converged MDAE B+BT model and fine-tune it using the B+BT approach on the different en–$\{af$,...,$yo\}$ combinations individually (Table 3).

In all cases, the bilingual finetuning improves the multilingual model, with a major increase of $+4.2$ BLEU for en–sw resulting in a BLEU score of 11.6. The finetuned models almost always produce the best performing model, showing that the process of $i)$ multilingual pretraining (MDAE) to achieve a cross-lingual representation, $ii)$ SSNMT online data extraction (SPE) with online back-translation (B+BT) to obtain increasing quantities of supervisory signals from the data, followed by $iii)$ focused bilingual fine-tuning to remove multilingual noise is key to learning low-resource MT also on distant languages without the need of any parallel data.

7 Comparison to other NMT Architectures

We compare the best SSNMT model configuration per language pair with the SSNMT **baseline** system, and with Monoses (Artetxe et al., 2019), an **unsupervised** machine translation model in its statistical (USMT) and hybrid (USMT+UNMT) version (Table 4). Over all languages,

SSNMT with data augmentation outperforms both the SSNMT baseline and UMT models.

We also compare our results with a **supervised** NMT system trained on WP parallel sentences **extracted** by Laser[16] (Artetxe and Schwenk, 2019b) (en–$\{af,my\}$) in a preprocessing data extraction step with the recommended extraction threshold of 1.04. We use the pre-extracted and similarity-ranked WikiMatrix (Schwenk et al., 2021) corpus, which uses Laser to extract parallel sentences, for en–$\{ne,sw\}$. Laser is not trained on kn and yo, thus these languages are not included in the analysis. For en–af, our model and the supervised model trained on Laser extractions perform equally well. In all other cases, our model statistically significantly outperforms the supervised LASER model, which is surprising, given the fact that the underlying LASER model was trained on parallel data in a highly multilingual setup (93 languages), while our MDAE setup does not use any parallel data and was trained on the monolingual data of much fewer language directions (7 languages) only. This again underlines the effectiveness of joining SSNMT with BT, multilingual pre-training and bilingual finetuning.

For reference, we also report the **top-scoring system** (TSS) per language direction based on top results reported on the relevant test sets together with the amount of parallel training data available to TSS systems. In case of language pairs whose test set is part of ongoing shared tasks (en–$\{kn,my\}$), we report the most recent results reported on the shared task web-pages (Section 4). The amount of parallel data available for these TSS varies greatly across languages, from 37 k (en–af) to 995 k (often noisy) sentences. In general, TSS systems perform much better than any of the SSNMT configurations or unsupervised models. This is natural, as TSS systems are mostly supervised (Martinus and Abbott, 2019; Adelani et al., 2021a), semi-supervised (Lakew et al., 2021) or multilingual models with parallel pivot language pairs (Guzmán et al., 2019), none of which is used in the UMT and SSNMT models. For $en2af$ only, our best configuration and the supervised NMT model trained on Laser extractions outperform the current TSS, with a gain in BLEU of $+16.9$ (B+BT), which may be due to the small amount of parallel data the TSS was trained on (37 k parallel sentences).

8 Discussion and Conclusion

Across all tested low-resource language pairs, joining SSNMT-style online sentence pair extraction with UMT-style online back-translation significantly outperforms the SSNMT baseline and unsupervised MT models, indicating that the small amount of available supervisory signals in the data is exploited more efficiently. Our models also outperform supervised NMT systems trained on Laser extractions, which is remarkable given that our systems are trained on non-parallel data only, while Laser has been trained on massive amounts of parallel data.

While SSNMT with data augmentation and MDAE pre-training is able to learn MT even on a low-resource distant language pair such as en–kn, it can fail when a language does not have any relation to other languages included in the multilingual pre-training, which was the case for my in our setup. This can be overcome by being conscientious of the importance of language distance and including related languages during MDAE pre-training and SSNMT training. We make our code and data publicly available.[17]

Acknowledgements

We thank David Adelani and Jesujoba Alabi for their insights on Yorùbá. Part of this research was made possible through a research award from Facebook AI. Partially funded by the German Federal Ministry of Education and Research under the funding code 01IW20010 (Cora4NLP). The authors are responsible for the content of this publication.

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

[17] https://github.com/ruitedk6/comparableNMT

References

Adelani, D. I., Ruiter, D., Alabi, J. O., Adebonojo, D., Ayeni, A., Adeyemi, M., Awokoya, A., and España-Bonet, C. (2021a). MENYO-20k: A Multi-domain English-Yorùbá Corpus for Machine Translation and Domain Adaptation. *AfricaNLP Workshop, CoRR*, abs/2103.08647.

Adelani, D. I., Ruiter, D., Alabi, J. O., Adebonojo, D., Ayeni, A., Adeyemi, M., Awokoya, A., and España-Bonet, C. (2021b). The Effect of Domain and Diacritics in Yorùbá–English Neural Machine Translation. In *Proceedings of Machine Translation Summit (Research Track)*. European Association for Machine Translation.

Agić, Ž. and Vulić, I. (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.

Artetxe, M., Labaka, G., and Agirre, E. (2017). Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462.

Artetxe, M., Labaka, G., and Agirre, E. (2019). An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203, Florence, Italy. Association for Computational Linguistics.

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2018). Unsupervised neural machine translation. In *Proceedings of the Sixth International Conference on Learning Representations, ICLR*.

Artetxe, M. and Schwenk, H. (2019a). Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

Artetxe, M. and Schwenk, H. (2019b). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Banerjee, T., Murthy, V. R., and Bhattacharyya, P. (2019). Ordering matters: Word ordering aware unsupervised NMT. *CoRR*, abs/1911.01212.

Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Graham, Y., Guzman, P., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., and Negri, M., editors (2020). *Proceedings of the Fifth Conference on Machine Translation*. Association for Computational Linguistics, Online.

Bird, S. (2006). NLTK: The Natural Language Toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72, Sydney, Australia. Association for Computational Linguistics.

Bojar, O. and Tamchyna, A. (2011). Improving translation model by monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 330–336, Edinburgh, Scotland. Association for Computational Linguistics.

Chaudhary, V., Tang, Y., Guzmán, F., Schwenk, H., and Koehn, P. (2019). Low-resource corpus filtering using multilingual sentence embeddings. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 263–268, Florence, Italy. Association for Computational Linguistics.

Clark, J. H., Dyer, C., Lavie, A., and Smith, N. A. (2011). Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 176–181, Portland, Oregon, USA. Association for Computational Linguistics.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Conneau, A. and Lample, G. (2019). Cross-lingual language model pretraining. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Currey, A., Miceli Barone, A. V., and Heafield, K. (2017). Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156, Copenhagen, Denmark. Association for Computational Linguistics.

Edman, L., Toral, A., and van Noord, G. (2020). Low-resource unsupervised NMT: Diagnosing the problem and providing a linguistically motivated solution. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 81–90, Lisboa, Portugal. European Association for Machine Translation.

Edunov, S., Ott, M., Auli, M., and Grangier, D. (2018). Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

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

España-Bonet, C., Ruiter, D., and van Genabith, J. (2019). UdS-DFKI Participation at WMT 2019: Low-Resource (en–gu) and Coreference-Aware (en–de) Systems. In *Proceedings of the Fourth Conference on Machine Translation*, pages 382–389, Florence, Italy. Association for Computational Linguistics.

Gulcehre, C., Firat, O., Xu, K., Cho, K., Barrault, L., Lin, H.-C., Bougares, F., Schwenk, H., and Bengio, Y. (2015). On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*.

Guzmán, F., Chen, P.-J., Ott, M., Pino, J., Lample, G., Koehn, P., Chaudhary, V., and Ranzato, M. (2019). The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Hoang, V. C. D., Koehn, P., Haffari, G., and Cohn, T. (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.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F. B., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2017). Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguist*, 5:339–351.

Karakanta, A., Dehdari, J., and van Genabith, J. (2018). Neural machine translation for low-resource languages without parallel corpora. *Machine Translation*, 32(1):167–189.

Kim, Y., Graça, M., and Ney, H. (2020). When and why is unsupervised neural machine translation useless? In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 35–44, Lisboa, Portugal. European Association for Machine Translation.

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

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Koneru, S., Liu, D., and Niehues, J. (2021). Unsupervised machine translation on Dravidian languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 55–64, Kyiv. Association for Computational Linguistics.

Kudo, T. and Richardson, J. (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.

Kuwanto, G., Akyürek, A. F., Tourni, I. C., Li, S., and Wijaya, D. (2021). Low-resource machine translation for low-resource languages: Leveraging comparable data, code-switching and compute resources. *CoRR*, abs/2103.13272.

Lakew, S. M., Negri, M., and Turchi, M. (2021). Low Resource Neural Machine Translation: A Benchmark for Five African Languages. *AfricaNLP Workshop, CoRR*, abs/2003.14402.

Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. (2018a). Unsupervised machine translation using monolingual corpora only. In *Proceedings of the Sixth International Conference on Learning Representations, ICLR*.

Lample, G., Ott, M., Conneau, A., Denoyer, L., and Ranzato, M. (2018b). Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049. Association for Computational Linguistics.

Leng, Y., Tan, X., Qin, T., Li, X.-Y., and Liu, T.-Y. (2019). Unsupervised Pivot Translation for Distant Languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 175–183.

Li, Z., Zhao, H., Wang, R., Utiyama, M., and Sumita, E. (2020). Reference language based unsupervised neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4151–4162, Online. Association for Computational Linguistics.

Littell, P., Mortensen, D. R., Lin, K., Kairis, K., Turner, C., and Levin, L. (2017). URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Marchisio, K., Duh, K., and Koehn, P. (2020). When does unsupervised machine translation work? In *Proceedings of the Fifth Conference on Machine Translation*, pages 571–583, Online. Association for Computational Linguistics.

Martinus, L. and Abbott, J. Z. (2019). A Focus on Neural Machine Translation for African Languages. *CoRR*, abs/1906.05685.

McKellar, C. A. and Puttkammer, M. J. (2020). Dataset for comparable evaluation of machine translation between 11 South African languages. *Data in Brief*, 29:105146.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Niu, X., Denkowski, M., and Carpuat, M. (2018). Bi-directional neural machine translation with synthetic parallel data. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 84–91, Melbourne, Australia. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Ramachandran, P., Liu, P., and Le, Q. (2017). Unsupervised pretraining for sequence to sequence learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 383–391, Copenhagen, Denmark. Association for Computational Linguistics.

Ren, S., Zhang, Z., Liu, S., Zhou, M., and Ma, S. (2019). Unsupervised Neural Machine Translation with SMT as Posterior Regularization. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, Honolulu, Hawaii, USA*, pages 241–248. AAAI Press.

Ruiter, D., España-Bonet, C., and van Genabith, J. (2019). Self-supervised neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1828–1834, Florence, Italy. Association for Computational Linguistics.

Ruiter, D., van Genabith, J., and España-Bonet, C. (2020). Self-Induced Curriculum Learning in Self-Supervised Neural Machine Translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2560–2571, Online. Association for Computational Linguistics.

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., and Guzmán, F. (2021). WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. In Merlo, P., Tiedemann, J., and Tsarfaty, R., editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 1351–1361. Association for Computational Linguistics.

Sen, S., Gupta, K. K., Ekbal, A., and Bhattacharyya, P. (2019). Multilingual unsupervised NMT using shared encoder and language-specific decoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3083–3089, Florence, Italy. Association for Computational Linguistics.

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

Sennrich, R., Haddow, B., and Birch, A. (2016b). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

ShweSin, Y. M., Soe, K. M., and Htwe, K. Y. (2018). Large Scale Myanmar to English Neural Machine Translation System. In *2018 IEEE 7th Global Conference on Consumer Electronics (GCCE)*, pages 464–465.

Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., and Fan, A. (2020). Multilingual translation with extensible multilingual pretraining and finetuning. *CoRR*, abs/2008.00401.

Yang, Z., Chen, W., Wang, F., and Xu, B. (2018). Unsupervised neural machine translation with weight sharing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46–55. Association for Computational Linguistics.

Zhang, Z., Liu, S., Li, M., Zhou, M., and Chen, E. (2018). Joint training for neural machine translation models with monolingual data. In McIlraith, S. A. and Weinberger, K. Q., editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA*, pages 555–562. AAAI Press.

Zoph, B., Yuret, D., May, J., and Knight, K. (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.