MENYO-20k: A Multi-domain English–Yorùbá Corpus for Machine Translation and Domain Adaptation

David I. Adelani[13*], **Dana Ruiter**[1*], **Jesujoba O. Alabi**[23*], **Damilola Adebonojo**[3], **Adesina Ayeni**[4], **Mofe Adeyemi**[3], **Ayodele Awokoya**[35], **Cristina España-Bonet**[6]
[1]Spoken Language Systems Group (LSV), Saarland University, Germany
[2]Max Planck Institute for Informatics, Germany
[3]Masakhane NLP, [4]Yobamoodua Cultural Heritage (YMCH)
[5] University of Ibadan, Nigeria, [6] DFKI GmbH, Saarbrücken, Germany

Abstract

Massively multilingual machine translation (MT) has shown impressive capabilities, including zero and few-shot translation between low-resource language pairs. However, these models are often evaluated on high-resource languages with the assumption that they generalize to low-resource ones. The difficulty of evaluating MT models on low-resource pairs is often due the lack of standardized evaluation datasets. In this paper, we present MENYO-20k, the first multi-domain parallel corpus for the low-resource Yorùbá–English (*yo–en*) language pair with standardized train-test splits for benchmarking. We provide several neural MT (NMT) benchmarks on this dataset and compare to the performance of popular pre-trained (massively multilingual) MT models, showing that, in almost all cases, our simple benchmarks outperform the pre-trained MT models. A major gain of BLEU $+9.9$ and $+8.6$ (*en2yo*) is achieved in comparison to Facebook's M2M-100 and Google multilingual NMT respectively when we use MENYO-20k to fine-tune generic models.

1 Introduction

Machine translation (MT) is a Natural Language Processing (NLP) task which involves the automatic translation of sentences from a source language into a target language. In practice, training supervised MT models requires large amounts of parallel sentences but this is not always possible. Large and freely-available parallel corpora exist for a small number of high-resource pairs and domains. However, for low-resource languages such as Yorùbá (*yo*), one can only find few thousands of parallel sentences online[1]. In the best-case scenario, i.e. some amount of parallel data exists, one can use the Bible —the Bible is the most available resource for low-resourced languages (Resnik et al., 1999)— and JW300 (Agić and Vulić, 2019). Notice that both corpora belong to the religious domain and they do not generalize well to popular domains such as news and daily conversations (∀ et al., 2020).

In this paper, we address this problem for the Yorùbá–English (yo–en) language pair by creating a multi-domain parallel dataset, MENYO-20k, which we make publicly available[2] with CC BY-NC 4.0 licence. It is a heterogeneous dataset that comprises texts obtained from news articles, TED talks, movie and radio transcripts, science and technology texts, and other short articles curated from the web and translated by professional translators. Based on the resulting train-development-test split, we provide a benchmark for the yo–en task for future research on this language pair. This allows us to properly evaluate the generalization of MT models trained on JW300 and the Bible on new domains. We further explore transfer learning approaches that can make use of a few thousand sentence pairs for domain adaptation.

After giving a brief introduction to the Yorùbá language and its characteristics (section 2), we describe the collection and creation process of MENYO-20k together with an analysis of its different domains (section 3). In section 4 we present and analyze different MT benchmarks (neural MT with and without transfer learning) on this corpus. After discussing relevant related work (section 5), we conclude in section 6.

2 The Yorùbá Language

The Yorùbá language is the third most spoken language in Africa, and it is native to the south-western

* Equal contribution to the work
[1]http://opus.nlpl.eu
[2]https://github.com/dadelani/menyo-20k_MT

Nigeria and the Republic of Benin. It is one of the national languages in Nigeria, Benin and Togo, and it is also spoken in other countries like Ghana, Côte d'Ivoire, Sierra Leone, Cuba, Brazil and by a significant Yorùbá diaspora population in the US and United Kingdom mostly from the Nigerian ancestry. The language belongs to the Niger-Congo family, and it is spoken by over 40 million native speakers (Eberhard et al., 2019).

Yorùbá has several dialects but the written language has been standardized by the 1974 Joint Consultative Committee on Education (Asahiah et al., 2017). It has 25 letters without the Latin characters c, q, v, x and z, and with additional characters ẹ, gb, ṣ , ọ. There are 18 consonants (b, d, f, g, gb, j[dz], k, l, m, n, p[kp], r, s, ṣ, t, w y[j]), 7 oral vowels (a, e, ẹ, i, o, ọ, u), five nasal vowels, (an, ẹn, in, ọn, un) and syllabic nasals (m̀, ḿ, ǹ, ń). Yorùbá is a tonal language with three tones: low, middle and high. These tones are represented by the grave (e.g. "à "), optional macron (e.g. "ā") and acute (e.g. "á") accents respectively. These tones are applied on vowels and syllabic nasals, but the mid tone is usually ignored in writings. A few letters have underdots (i.e. "ẹ", "ọ", and "ṣ"), we refer to the tonal marks and underdots as diacritics. The tone information and underdots are important for the correct pronunciation of words.

As noted in Asahiah (2014), most of the Yorùbá texts found in websites or public domain repositories either use the correct Yorùbá orthography or replace diacriticized characters with un-diacriticized ones. Often, articles written online, including news articles such as BBC[3] ignore diacritics. Ignoring diacritics makes it difficult to identify or pronounce words except when they are in a context. For example, *èdè* (language), *edé* (crayfish), *ẹdẹ* (a town in Nigeria), *ẹ̀dẹ* (trap) and *ẹ̀dẹ̀* (balcony) will be mapped to *ede* without diacritics.

Machine translation might be able to learn to disambiguate the meaning of words and generate correct English even with un-diacriticized Yorùbá. However, one cannot generate correct Yorùbá if the training data is un-diacriticized. One of the purposes of our work is to build a corpus with correct and complete diacritization in several domains.

3 MENYO-20k

The dataset collection was inspired by the inability of machine translation models trained on JW300 to generalize to new domains (∀ et al., 2020). Although ∀ et al. (2020) evaluated this for Yorùbá with surprisingly high BLUE scores, the evaluation was done on very few examples from the COVID-19 and TED Talks domains with 39 and 80 sentences respectively. Inspired by the FLoRes dataset for Nepali and Sinhala (Guzmán et al., 2019), we create a high quality test set for Yorùbá with few thousands of sentences in different domains to check the quality of industry MT models, pre-trained MT models, and MT models based on popular corpora such as JW300 and the Bible.

3.1 Dataset Collection

Table 1 summarizes the texts collected, their source, the original language of the texts and the number of sentences from each source. We collected both parallel corpora freely available on the web and monolingual corpora we are interested in translating (e.g. the TED talks) to build the MENYO-20k corpus. Some few sentences were donated by professional translators such as "various texts" in Table 1. Our curation followed two steps: (1) Translation of monolingual texts crawled from the web by professional translators. (2) Verification of translation, orthography and diacritics for parallel texts obtained online and translated. Some texts obtained from the web that were judged by native speakers of high quality were verified once, the others were verified twice. The verification of translation and diacritics was done by professional translators and volunteers mostly from *Masakhane,*[4] a grassroots organisation whose mission "is to strengthen and spur NLP research in African languages, for Africans, and by Africans". We provide more specific description of the data sources below.

Jehovah Witness News We collected only parallel *"newsroom"* (or *"Ìròyìn"* in Yorùbá) articles from `JW.org` website to gather texts that are not in the religious domain. As shown in Table 1, we collected 3,508 sentences from their website, and we manually confirmed that the sentences are not in JW300. The content of the news mostly reports persecutions of Jehovah witness members around the world, and may sometimes contain Bible verses to encourage believers.

Voice of Nigerian News We extracted parallel texts from the VON website, a Nigerian Government news website that supports seven languages

[3] `https://www.bbc.com/yoruba`

[4] `https://www.masakhane.io`

Data name	Source	Language of source	No. Sentences
Jehovah Witness News	jw.org/yo/iroyin	en-yo	3,508
Voice of Nigeria News	von.gov.ng	en-yo	3,048
TED talks	ted.com/talks	en	2,945
Global Voices News	yo.globalvoices.org	en-yo	2,932
Proverbs	twitter.com/yoruba_proverbs	yo-en	2,700
Out of His Mind Book	Obtained from the author	en	2,014
Software localization	Obtained from Professional Translators	en	941
Movie Transcript ("Unsane")	youtu.be/hdWP0X5msZQ	yo-en	774
Various short texts	Obtained from Professional Translators	en	687
Radio Broadcast	Transcript from Bond FM 92.9 Radio	en	258
Creative Commons License	Obtained from Professional Translators	en	193
UDHR Translation	ohchr.org	en-yo	100
Total			20,100

Table 1: Dataset collection sources with source language(s) and the number of sentences contained.

with wide audience in the country (Arabic, English, Fulfulde, French, Hausa, Igbo, and Yorùbá). Despite the large availability of texts, the quality of Yorùbá texts is very poor, one can see several issues with orthography and diacritics. We asked translators and other native speakers to verify and correct each sentence.

Global Voices News We obtained parallel sentences from the Global Voices website[5] contributed by journalists, writers and volunteers. The website supports over 50 languages, with contents mostly translated from English, French, Portuguese or Spanish.

TED Talks Transcripts We selected 28 English TED talks transcripts mostly covering issues around Africa like health, gender equality, corruption, wildlife, and social media e.g "How young Africans found a voice on Twitter" (see the Appendix for the selected TED talk titles). The articles were translated by a professional translator and verified by another one.

Proverbs Yorùbá has many proverbs and culturally referred to words of wisdom that are often referenced by elderly people. We obtained 2,700 sentences of parallel yo–en texts from Twitter.[6]

Book With permission from the author (Bayo Adebowale) of the "Out of His Mind" book, originally published in English, we translated the entire book to Yorùbá and verified the diacritics.

Software Localization Texts (Digital) We obtained translations of some software documentations such as Kolibri[7] from past projects of professional translators. These texts include highly technical terms.

Movie Transcripts We obtained the translation of a Nigerian movie "Unsane" on YouTube from the past project of a professional translator. The language of the movie is Yorùbá and English, with transcription also provided in English.

Other Short Texts Other short texts like UDHR, Creative Commons License, radio transcripts, and texts were obtained from professional translators and online sources. Table 1 summarizes the number of sentences obtained from each source.

Table 2 (top) summarizes the figures for MENYO-20k dataset with 20,100 parallel sentences split into 10,070 training sentences, 3,397 development sentences, and 6,633 test sentences.

3.2 Baseline Corpora

For our experiments, we use two widely available parallel corpora from the religion domain, Bible and JW300 as shown in Table 2 (bottom). For the Bible, we align the verses from the New International Version (NIV) for English and the Bible Society of Nigeria version (BSN) for Yorùbá. Both versions are used due to the contemporary writing style used in both languages. After aligning the verses, we obtain 30,760 parallel sentences. Also, we download the JW300 parallel corpus which is available for a large variety of low resource language pairs. It has parallel corpora for English

[5] https://globalvoices.org

[6] Also available in https://github.com/Niger-Volta-LTI/yoruba-text

[7] https://learningequality.org/kolibri

Domain	Number of Sentences		
	Train. Set	Dev. Set	Test Set
MENYO-20k			
News	4,995	1,391	3,102
TED Talks	507	438	2,000
Book	-	1,006	1,008
Digital	356	312	273
Proverbs	2,200	250	250
Other Domains	2,012	250	250
Baselines			
Bible	30,760	–	–
JW300	459,871	–	–
TOTAL	500,701	3,397	6,633

Table 2: Number of sentences for baseline corpora (bottom) and MENYO-20k domains (top) in training, development and test splits.

to 343 languages containing religion-related texts. From the JW300 corpus, we get $459,871$ sentence pairs already tokenized with *Polyglot*[8] (Al-Rfou, 2015).

3.3 Dataset Domain Analysis

As shown in the previous section, MENYO-20k is highly heterogeneous. In this section we analyze the differences and how its (sub)domains depart from the characteristics of the commonly used Yorùbá–English corpora for MT.

Characterizing the domain of a dataset is a difficult task. Some previously used metrics need either large corpora or a characteristic vocabulary of the domain (Beyer et al., 2020; España-Bonet et al., 2020). Here, we do not have these resources and we report the overlapping vocabulary between training and test sets and the perplexity observed in the test sets when a language model (LM) is trained on the MT training corpora.

In order to estimate the perplexities, we train a language model of order 5 with KenLM (Heafield, 2011) on each of the 3 training data subsets (JW300, JW300+Bible, JW300+Bible+MENYO-20k). Following NMT standard processing pipelines (see subsection 4.2), we perform byte-pair encoding (BPE) (Sennrich et al., 2016) on the corpora to avoid a large number of out-of-vocabulary tokens which, for small corpora, could alter the LM probabilities. For each of the resulting language models, we evaluate their average **perplexity** on the different domains of the test set to evaluate *compositional* domain differences (Figure 1). As expected, the average perplexity drops when adding more training data. Due to the limited domain of both JW300 and Bible, a literary style close to the Books domain, the decrease in perplexity is small when adding additional Bible data to JW300, namely -8% (*en*) and -11% (*yo*). Interestingly, both JW300 and Bible also seem to be close to the TED domain (1st and 2nd lowest perplexities for *en* and *yo* respectively), which may be due to discourse/monologue content in both training corpora. Adding the domain-diverse MENYO-20k corpus largely decreases the perplexity across all domains with a major decrease of -66% on Digital (*yo*) and smallest decrease of -1% on Books (*en*). The perplexity scores correlate negatively with the resulting BLEU scores in Table 4, with a Pearson's r (r) of -0.367 (*en*) and -0.461 (*yo*), underlining that compositional domain differences between training and test subsets is the main factor of differences in translation quality.

Further, to evaluate *lexical* domain differences, we calculate the **vocabulary coverage** (tokenized, not byte-pair encoded[9]) of the different domains of the test set by each of the training subsets (Figure 2). The vocabulary coverage increases to a large extend when MENYO-20k is added. However, while vocabulary coverage and average perplexities have a strong (negative) correlation, $r = -0.756$ (*en*) and $r = -0.689$ (*yo*), a high perplexity does not necessarily mean low vocabulary coverage. E.g., the vocabulary coverage of the Digital domain by JW300 is high (0.91% for *en*) despite leading to high perplexities (765 for *en*). In general, vocabulary coverage of the test sets is less indicative of the resulting translation performance, showing only a weak correlation between vocabulary coverage and BLEU, with $r = 0.150$ and $r = 0.281$ for *en* and *yo* respectively.

4 Machine Translation

4.1 Machine Translation Architectures

Neural machine translation has been close to achieve human performance for high resourced language pairs such as English–Chinese (Hassan et al., 2018). Even if human performance is questionable (Läubli et al., 2018), it is clear that with huge amounts of data NMT achieves high quality results. NMT translates a text in a source language into a text in the target language using a single neural

[8] https://github.com/aboSamoor/polyglot

[9] We do not use byte-pair encoded data here, since, due to the nature of BPE, the vocabulary overlap would be close to 1 between all training and test sets.

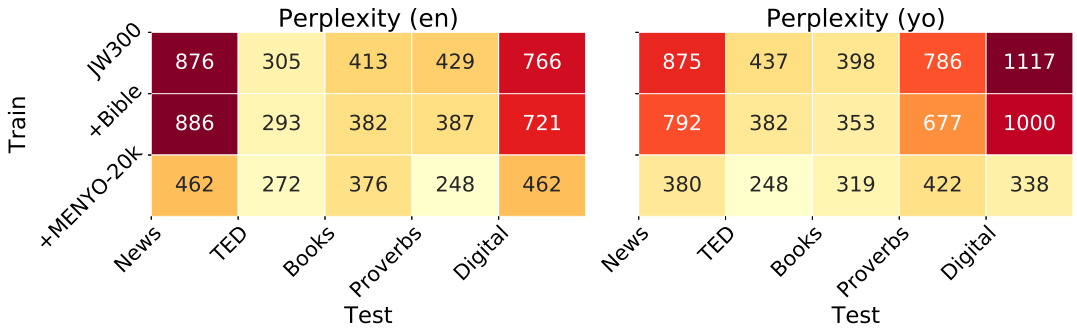

Figure 1: Perplexities of KenLM 5-gram language model learned on different training corpora and tested on subsets of MENYO-20k for English (left) and Yorùbá (right) respectively.

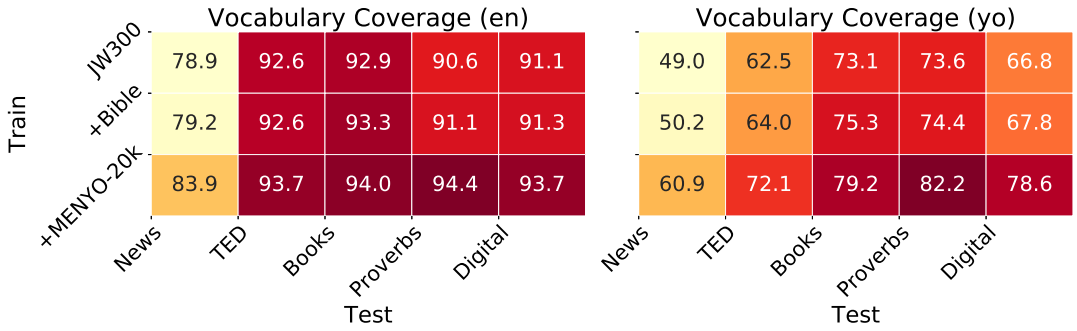

Figure 2: Vocabulary coverage (%) of different subsets of the MENYO-20k test set per training sets for English (left) and Yorùbá (right).

network which can be divided into two main components, an encoder and a decoder, jointly trained together. The encoder consumes the source sentence one token after the other and encodes it into a continuous semantic space which is fed into the decoder to generate the target sentence one token at a time. The decoder attends to the semantic representation of the source sentence (Bahdanau et al., 2015). Both encoder and decoder can be built with LSTM, GRU, or CNN networks (Cho et al., 2014; Sutskever et al., 2014; Gehring et al., 2017). More recently, the transformer architecture with only a combination of attention mechanisms (Vaswani et al., 2017) and without any of the recursion or convolution given by previous models has achieved state-of-the-art performance.

As any deep learning system, the performance of NMT heavily depends on the amount of data. Sennrich and Zhang (2019a) showed that with a correct hyperparameter selection, an RNN-based NMT system could outperform statistical machine translation systems even in low-resource settings.

In the following, we describe the setup of the NMT experiments (subsection 4.2), including the description of the data combinations used, their preprocessing and the NMT model specifications. Then, we compare the NMT model results (subsection 4.3) based on the data subsets used for training, the different domains included in the test set and the existence of Yorùbá diacritics during *yo2en* training. We further perform an external comparison with popular MT models that support *yo–en* translation.

4.2 Experimental Settings

Data and Preprocessing For the MT experiments, we use the training part of our MENYO-20k corpus and two other parallel corpora, Bible and JW300 (section 3). For tuning the hyperparameters, we use the dev split of the multi-domain data which has $3,397$ sentence pairs and for testing the test split with $6,633$ parallel sentences.

To ensure that all the parallel corpora are in the same format, we converted the Yorùbá texts in the JW300 dataset to Unicode Normalization Form Composition (NFC), the format of the Yorùbá texts in the Bible and multi-domain dataset. Our preprocessing pipeline includes punctuation normalization, tokenization, and truecasing. For punctua-

Model	*en2yo*	*yo2en*
Internal Comparison		
Bible	2.19	1.38 / 1.61
JW300	7.46	9.59 / 9.27
JW300+Bible	8.11	10.75 / 10.47
+MENYO-20k	10.93	13.97 / 13.95
+Transfer	**12.34**	13.19 / 13.91
External Comparison		
OPUS-MT	–	5.87 / –
Google MNMT	3.70	**22.39** / –
M2M-100	2.42	4.56 / –

Table 3: Tokenized BLEU scores over the full test for NMT models trained on different subsets of the training data (top), with top-scoring results underlined. For *yo2en* we report results for both diacritized / undiacritized Yorùbá training. Comparison to external models (bottom) included. Overall top BLEUs in bold.

tion normalization and truecasing we used *Moses* toolkit (Koehn et al., 2007) while for tokenization, we used *Polyglot*, since it is the tokenizer used in JW300. We apply joint BPE, with a vocabulary threshold of 20 and 40k merge operations.

In-house NMT Engines Specifications We train our NMT systems using the Transformer architecture proposed by Vaswani et al. (2017). We use the transformer-base model with 6 encoder and 6 decoder layers, hidden state of size 512, 8 attention heads, and feed-forward layer with 2048 units as implemented in Fairseq[10] toolkit (Ott et al., 2019). We set the drop-out at 0.3 and batch size at $10,240$ tokens. For optimization, we use *adam* (Kingma and Ba, 2015) with $\beta_1 = 0.9$ and $\beta_2 = 0.98$ and a learning rate of 0.0005. The learning rate has a warmup update of 4000, using label smoothed cross-entropy loss function with label-smoothing value of 0.1.

To evaluate our models, we pick only the best checkpoint on the development set and then evaluate the performance of our models on the test set. We used tokenized BLEU (Papineni et al., 2002) score implemented in *multi-bleu.perl*.

Publicly Available NMT Models We further evaluate the performance of three multilingual NMT systems on the test data: OPUS-MT (Tiedemann and Thottingal, 2020), Google multilingual NMT (MNMT) (Arivazhagan et al., 2019) and Facebook's M2M-100 (Fan et al., 2020). **Google MNMT** is a single NMT model that is capable of translating between 103 languages. The model is able to handle zero-shot translation as well as translation in low-resource settings due to parameter sharing. It has been trained using parallel sentences, originals and backtranslations, in 102 languages from and into English. We generated the translations of the test set using the *Google Translate*[11] interface. Very similar to Google MNMT, Facebook's **M2M-100** model is capable of translating between 100 languages. To train M2M-100, parallel sentences were extracted from large monolingual corpora in representative language pairs (inter- and intrafamilies) using LASER (Schwenk, 2018; Artetxe and Schwenk, 2019a). Backtranslation of monolingual corpora was also used to enlarge the training corpora. Translations of our test set are generated using the pre-trained models provided in *Fairseq*. **OPUS-MT** is a service that provides pre-trained MT models for over a thousand languages. The models comprising both bilingual MT and multilingual NMT are trained on parallel corpora from OPUS (Tiedemann, 2012). To evaluate OPUS-MT we translated the sentences in the test data using *Easy-NMT*.[12]

4.3 Results

Internal Comparison We train four basic NMT engines on different subsets of the training data: Bible, JW300, JW300+Bible and JW300+Bible+MENYO-20k. Further, we fine-tune the converged model trained on JW300+Bible on MENYO-20k (JW300+Bible+Transfer). This yields us five NMT models in total for *en2yo* and *yo2en* each. We evaluate their translation performance on the MENYO-20k test set (Table 3, top). The BLEU scores obtained after training on Bible only are low, with BLEU 2.19 and 1.39 for *en2yo* and *yo2en* respectively, which is due to its small amount of training data. Training on the larger JW300 corpus leads to higher scores of BLEU 7.46 (*en2yo*) and 9.59 (*yo2en*), while combining it with Bible only leads to a small increase of $+0.65$ and $+1.16$ for *en2yo* and *yo2en* respectively. When further adding MENYO-20k to the training data, the translation quality increases by $+2.82$ (*en2yo*) and $+3.22$ (*yo2en*). When, instead of adding MENYO-20k to the training pool, it is used to fine-tune the

[10] https://github.com/pytorch/fairseq

[11] https://translate.google.com/

[12] https://github.com/UKPLab/EasyNMT

converged JW300+Bible model, the increase in BLEU over JW300+Bible is even larger for *en*2*yo* (BLEU +4.23), which results in an overall top-scoring model with BLEU 12.34. For *yo*2*en* fine-tuning (BLEU 13.19) is slightly less effective than simply adding MENYO-20k to the training data (BLEU 13.97). As seen in subsection 3.3, perplexities and vocabulary coverage in English are not as distant among training/test sets as in Yorùbá, so the fine-tuning step resulted less efficient.

External Comparison We also evaluate the performance of popular multilingual engines introduced in the previous section on the full test set (Table 3, bottom). **OPUS-MT**, while having no model available for *en*2*yo*, achieves a BLEU of 5.87 for *yo*2*en*. Thus, despite being trained on JW300 and other available *yo*2*en* corpora on OPUS, it is largely outperformed by our NMT model trained on JW300 only (BLEU +3.72). This may be caused by some of the noisy corpora included in OPUS (CCaligned, Tatoeba etc.), which can depreciate the translation quality. Facebook's **M2M-100**, is also largely outperformed even by our simple JW300 baseline by 5 BLEU points. A manual examination of the *en–yo* LASER extractions used to train M2M-100 shows that these are very noisy, which explains the poor translation performance. **Google**, on the other hand, obtains impressive results for the *yo*2*en* direction, with BLEU 22.39. The opposite direction *en*2*yo*, however, shows a significantly lower performance (BLEU 3.70), being outperformed even by our simple JW300 baseline (BLEU +3.76). This difference in performance can be attributed to the highly multilingual but English-centric nature of the Google MNMT model. As already noticed by the authors (Arivazhagan et al., 2019), low-resourced language pairs benefit from multilinguality when translated into English, but improvements are minor when translating into the non-English language. Also notice that lots of diacritics are lost in Google translations, which damages the BLEU scores. It remains for future work the study of the effects of diacritics in OPUS-MT, M2M-100 and GMNMT translation quality.

Diacritization Alabi et al. (2020) found that diacritics are important for Yorùbá word embeddings. In order to verify whether the diacritics in Yorùbá MT, which are often ignored in popular multilingual models (e.g. multilingual BERT (Devlin et al., 2019)), can help disambiguate some translation choices, we additionally train a *yo*2*en* model trained on un-diacritized JW300, JW300+Bible and JW300+Bible+MENYO-20k (Table 3, top). Since one cannot generate correct Yorùbá text with training data without diacritics, *en*2*yo* systems are not trained.

Results for *yo*2*en* are not conclusive. Diacritization seems to be useful only with exclusively out-of-domain training data (JW300, JW300+Bible[13]). Since in this case the domain of the training data is very different to the domain of the test set, disambiguation is needed not to bias all the lexicon towards the religious domain. When we include in-domain data (JW300+Bible+MENYO-20k), both models perform equally well, with BLEU 13.97 diacritized vs. BLEU 13.95 undiacritized respectively. Diacritization is not needed when we fine-tune the model with data that shares the domain with the test set (JW300+Bible+Transfer). In this case, we obtain BLEU 13.19 diacritized vs. BLEU 13.91 undiacritized respectively.

In practice, this means that, when training data is far from the desired domain, investing work into having clean diacritized Yorùbá input can help improve the translation performance also for the *yo*2*en* direction. When more data is present, the diacritization becomes less important. This is only true when Yorùbá is not the target language, where diacritization is always needed.

Domain Differences In order to identify the domain-specific performance of the different NMT models, we evaluate each model on the different domain subsets of the test set (Table 4). The resulting BLEU scores indicate that the Proverb domain was especially difficult in both directions, as it shows the lowest translation performance across all domains, i.e. maximum BLEU of 9.04 (*en*2*yo*) and 8.74 (*yo*2*en*). This is due to the fact that proverbs often do not have literal counterparts in the target language, thus making them especially difficult to translate. The TED domain is the best performing test domain, with maximum BLEU of 16.12 (*en*2*yo*) and 16.81 (*yo*2*en*). This can be attributed to the decent base coverage of the TED domain by JW300 and Bible together (monologues) with the additional TED domain data included in the MENYO-20k training split (507 sentence pairs). Also, most BLEU results are on line with the LM perplexity results and conclusions drawn in subsec-

[13]We do not consider Bible alone. Due to its small data size, the BLEU scores are less indicative.

Data	*en2yo* Proverbs	News	TED	Book	Digital	*yo2en* Proverbs	News	TED	Book	Digital
Bible	0.82	1.66	3.05	3.44	1.51	1.11	0.89	2.06	2.38	0.91
JW300	2.22	6.37	9.79	9.78	4.83	2.64	8.40	13.12	9.55	6.96
JW300+Bible	3.49	6.70	10.70	11.26	4.90	4.76	9.51	14.35	10.86	7.83
+MENYO-20k	7.03	10.08	12.26	11.47	10.48	**8.74**	**13.54**	16.73	**11.64**	**12.39**
+Transfer	**9.04**	**10.24**	**16.12**	**15.00**	**11.77**	8.56	12.51	**16.81**	10.81	9.67

Table 4: Tokenized BLEU over different domains of the test set for NMT models trained on different subsets of the training data, with top-scoring results per domain in bold.

tion 3.3. Due to the closeness of Bible and JW300 to the book domain, we see only small improvements of BLEU on this domain, i.e. $+0.21$ (*en*2*yo*) and $+0.78$ (*yo*2*en*), when adding MENYO-20k to the JW300+Bible training data pool. On the other hand, the Digital domain benefits the most from the additional MENYO-20k data, with major gains of BLEU $+5.58$ (*en*2*yo*) and 4.56 (*yo*2*en*), owing to the introduction of Digital domain content in the MENYO-20k training data ($\sim 1k$ sentence pairs), which is completely lacking in JW300 and Bible.

5 Related Work

In order to make MT available for a broader range of linguistic communities, recent years have seen an effort in creating new **parallel corpora** for low-resource language pairs. Recently, Guzmán et al. (2019) provided novel supervised, semi-supervised and unsupervised benchmarks for Indo-Aryan languages {Sinhala,Nepali}–English on an evaluation set of professionally translated sentences sourced from the Sinhala, Nepali and English Wikipedias.

Novel parallel corpora focusing on **African languages** cover South African languages ({Afrikaans, isiZulu, Northern Sotho, Setswana, Xitsonga}–English) (Groenewald and Fourie, 2009) with MT benchmarks evaluated in Martinus and Abbott (2019), as well as multidomain (News, Wikipedia, Twitter, Conversational) Amharic–English (Hadgu et al., 2020) and multidomain (Government, Wikipedia, News etc.) Igbo–English (Ezeani et al., 2020). Further, the LORELEI project (Strassel and Tracey, 2016) has created parallel corpora for a variety of low-resource language pairs, including a number of Niger-Congo languages such as {isiZulu, Twi, Wolof, Yoruba}–English. However, these are not open-access. On the contrary, Masakhane (∀ et al., 2020) is an ongoing participatory project focusing on creating new freely-available parallel corpora and MT benchmark models for a large variety of African languages.

While creating parallel resources for low-resource language pairs is one approach to increase the number of linguistic communities covered by MT, this does not scale to the sheer amount of possible language combinations. Another research line focuses on **low-resource MT** from the modeling side, developing methods which allow a MT system to learn the translation task with smaller amounts of supervisory signals. This is done by exploiting the weaker supervisory signals in larger amounts of available monolingual data, e.g. by identifying additional parallel data in monolingual corpora (Artetxe and Schwenk, 2019b; Schwenk et al., 2019, 2020), or by including auto-encoding (Currey et al., 2017) or language modeling tasks (Gulcehre et al., 2015; Ramachandran et al., 2017) during training. Low-resource language pairs 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 (Lample and Conneau, 2019), or finding an optimal path of pivoting through related languages (Leng et al., 2019). By adapting the model hyperparameters to the low-resource scenario, Sennrich and Zhang (2019b) were able to achieve impressive improvements over a standard NMT system.

6 Conclusion

We present MENYO-20k, a novel *en–yo* multi-domain parallel corpus for machine translation and domain adaptation. By defining a standardized train-development-test split of this corpus, we provide several NMT benchmarks for future research on the *en–yo* MT task. Further, we analyze the domain differences on the MENYO-20k corpus and the translation performance of NMT models trained on different (popular) training *en–yo* training corpora, such as JW300 and Bible, across the

different domains. We show that, despite consisting of only $10k$ parallel sentences, adding the MENYO-20k corpus train split to JW300 and Bible largely improves the translation performance over all domains. Our benchmark NMT models trained on the concatenation of JW300, Bible and MENYO-20k outperform popular pre-trained MT models by far, including OPUS-MT and Facebook's M2M-100. It also outperforms Google MNMT in the low-resource *en*2*yo* direction.

Acknowledgements

We would like to thank Adebayo O. Adeojo, Babunde O. Popoola, Olumide Awokoya, Modupe Olaniyi, Princess Folasade, Akinade Idris, Tolulope Adelani and Oluyemisi Olaose for their support in translating English sentences to Yorùbá and the verification of Yorùbá diacritics. We thank Iroro Orife for providing the Bible corpus and Yorùbá Proverbs corpus. We are also thankful to Damyana Gateva for evaluations with open-source models. This project was funded by the AI4D language dataset fellowship.[14] CEB was funded by the German Federal Ministry of Education and Research under the funding code 01IW20010. The author is responsible for the content of this publication.

References

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

Rami Al-Rfou. 2015. *Polyglot: A massive multilingual natural language processing pipeline*. Ph.D. thesis, Stony Brook University.

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

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

[14] https://www.k4all.org/project/language-dataset-fellowship/

Mikel Artetxe and Holger Schwenk. 2019a. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203.

Mikel Artetxe and Holger Schwenk. 2019b. 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.

Franklin Asahiah, Odetunji Odejobi, and Emmanuel Adagunodo. 2017. Restoring tone-marks in standard yoruba electronic text: Improved model. *Computer Science*, 18(3):301–315.

Franklin O Asahiah. 2014. Development of a standard yoruba digital text automatic diacritic restoration system. *PhD. Thesis, Obafemi Awolowo University, Ile-Ife, Nigeria*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*, San Diego, CA.

Anne Beyer, Göran Kauermann, and Hinrich Schütze. 2020. Embedding space correlation as a measure of domain similarity. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2431–2439, Marseille, France. European Language Resources Association.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig (eds.). 2019. Ethnologue: Languages of the world. twenty-second edition.

Cristina España-Bonet, Alberto Barrón-Cedeño, and Lluís Màrquez. 2020. Tailoring and evaluating the wikipedia for in-Domain comparable corpora extraction. *arXiv e-prints 2005.01177*, pages 1–26.

Ignatius Ezeani, Paul Rayson, Ikechukwu E. Onyenwe, Uchechukwu Chinedu, and Mark Hepple. 2020. Igbo-english machine translation: An evaluation benchmark. In *Eighth International Conference on Learning Representations: ICLR 2020*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english–centric multilingual machine translation. *arXiv e-prints 2010.11125*.

∀, Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, ..., Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, .., and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in african languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252, International Convention Centre, Sydney, Australia. PMLR.

Hendrik J. Groenewald and Wildrich Fourie. 2009. Introducing the autshumato integrated translation environment. In *Proceedings of the 13th Annual conference of the European Association for Machine Translation*, pages 190–196, Barcelona, Spain. European Association for Machine Translation.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Asmelash Teka Hadgu, Adam Beaudoin, and Abel Aregawi. 2020. Evaluating Amharic Machine Translation. *arXiv e-prints 2003.14386*.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving Human Parity on Automatic Chinese to English News Translation. *arXiv e-prints 1803.05567*.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. 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 Linguist*, 5:339–351.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference for Learning Representations (ICLR)*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 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.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32, pages 7059–7069. Curran Associates, Inc.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Yichong Leng, Xu Tan, Tao Qin, Xiang-Yang Li, and Tie-Yan Liu. 2019. Unsupervised pivot translation for distant languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 175–183.

Laura Martinus and Jade Z. Abbott. 2019. A focus on neural machine translation for african languages. *arXiv e-prints 1906.05685*.

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 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Prajit Ramachandran, Peter Liu, and Quoc Le. 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.

Philip Resnik, Mari Broman Olsen, and Mona T. Diab. 1999. The Bible as a Parallel Corpus: Annotating the 'Book of 2000 Tongues'. *Computers and the Humanities*, 33:129–153.

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

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. *arXiv preprint arXiv:1907.05791*.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2020. Ccmatrix: Mining billions of high-quality parallel sentences on the web. *arXiv e-prints arXiv:1911.04944*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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.

Rico Sennrich and Biao Zhang. 2019a. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

Rico Sennrich and Biao Zhang. 2019b. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

Stephanie Strassel and Jennifer Tracey. 2016. LORELEI language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3273–3280, Portorož, Slovenia. European Language Resources Association (ELRA).

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2214–2218, Istanbul, Turkey. European Languages Resources Association (ELRA).

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

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

A Appendices

A.1 TED Talks titles

	Title	Topic
1	Reducing corruption takes a specific kind of investment	Politics
2	How young Africans found a voice on Twitter	Technology
3	Mothers helping mothers fight HIV	Health
4	How women are revolutionizing Rwanda	Gender-equality
5	How community-led conservation can save wildlife	Wildlife
6	How cancer cells communicate - and how we can slow them down	Health
7	You may be accidentally investing in cigarette companies	Health
8	How deepfakes undermine truth and threaten democracy	Politics
9	What tech companies know about your kids	Technology
10	Facebook's role in Brexit - and the threat to democracy	Politics
11	How we can make energy more affordable for low-income families	Energy
12	Can we stop climate change by removing CO2 from the air?	Climate
13	A comprehensive, neighborhood-based response to COVID-19	Health
14	Why civilians suffer more once a war is over	Human Rights
15	Lessons from the 1918 flu	Health
16	Refugees have the right to be protected	Human Rights
17	The beautiful future of solar power	Energy
18	How bees can keep the peace between elephants and humans	Wildlife
19	Will automation take away all our jobs?	Technology
20	A celebration of natural hair	Beauty
21	Your fingerprints reveal more than you think	Technology
22	Our immigration conversation is broken - here's how to have a better one	Politics
23	What I learned about freedom after escaping North Korea	Politics
24	Medical tech designed to meet Africa's needs	Health
25	What's missing from the American immigrant narrative	Education
26	A hospital tour in Nigeria	Health
27	How fake news does real harm	Politics
28	How we can stop Africa's scientific brain drain	Education

Table 5: TED talks titles.