Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3710–3718
Marseille, 11–16 May 2020

An Analysis of Massively Multilingual Neural Machine Translation for Low-Resource Languages

Aaron Mueller, Garrett Nicolai, Arya D. McCarthy, Dylan Lewis, Winston Wu,
and **David Yarowsky**
Center for Language and Speech Processing
Johns Hopkins University
(amueller, gnicola2, arya, dlewis77, wswu, yarowsky)@jhu.edu

Abstract

In this work, we explore massively multilingual low-resource neural machine translation. Using translations of the Bible (which have parallel structure across languages), we train models with up to 1,107 source languages. We create various multilingual corpora, varying the number and relatedness of source languages. Using these, we investigate the best ways to use this many-way aligned resource for multilingual machine translation. Our experiments employ a grammatically and phylogenetically diverse set of source languages during testing for more representative evaluations. We find that best practices in this domain are highly language-specific: adding more languages to a training set is often better, but too many harms performance—the best number depends on the source language. Furthermore, training on related languages can improve or degrade performance, depending on the language. As there is no one-size-fits-most answer, we find that it is critical to tailor one's approach to the source language and its typology.

Keywords: neural machine translation, low-resource, multilinguality, Bible

1. Introduction

Recently, machine translation (MT) has made significant progress in many language pairs by employing recurrent sequence-to-sequence neural networks (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014) in combination with attention (Bahdanau et al., 2015; Luong et al., 2015); even better performance has been achieved using self-attention (Vaswani et al., 2017). The neural approach has been effective because of the high fluency and adequacy of its output, as well as its language-agnostic methods. Indeed, given enough data, one need not know any linguistic features of one's source and target languages to effectively translate.[1] With the success of multilingual training (Johnson et al., 2017; Aharoni et al., 2019), the trend seems clear: more data generally results in better models for neural machine translation, regardless of the languages (or combinations thereof) used.

These results, however, have primarily been found for languages possessing large amounts of aligned data, and few such languages exist among the world's 7,000+ languages (Lewis et al., 2015). While high neural machine translation performance has been observed in some bilingual low-resource[2] contexts (Sennrich and Zhang, 2019), we have access to a multi-way aligned corpus of Bibles that allow us to perform a more in-depth analysis of the performance of multilingual low-resource NMT. This parallel corpus of Bible texts (McCarthy et al., 2020) contains a set of "mono-lingual" Bibles in 1,108 languages, all aligned by verse.

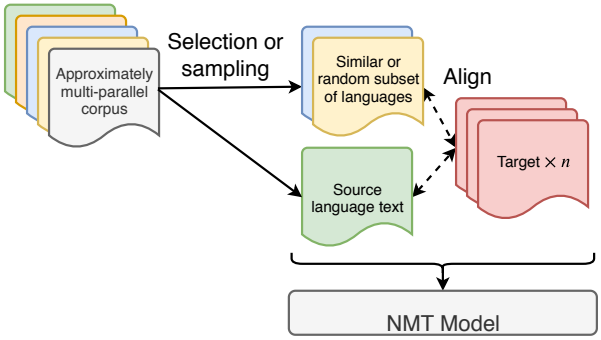

Figure 1: Our process for constructing training sets from a multi-parallel corpus. We select some $n-1$ languages that share either phylogeny or script or randomly sample $n-1$ languages to augment the source–target pair. Using a multi-parallel corpus controls for variations in data quality or domain; this lets us more cleanly assess our scientific questions, but it is not an engineering requirement for MT systems that use several helper languages. More details on sampling are given in Figure 2.

In this paper, we leverage the massively multilingual multi-way verse-aligned Bible corpus to investigate two questions:

1. Does the performance of multilingual neural machine translation in low-resource settings scale with the number of languages in the training corpus? Relatedly, when do we start to see diminishing returns or decreased performance, if ever?

2. Does the similarity of the languages in the multilingual training corpora matter, or does performance simply scale with the amount of data present?

We present the construction of 18 multilingual parallel corpora to evaluate various low-resource multilingual settings

[1] This is not to say that we no longer need linguistics; there are many cases in which knowing properties of one's evaluation language(s) can help one translate better (e.g., when working with morphologically complex languages). This statement simply implies that knowing properties of one's specific language pair is no longer necessary to achieve high performance, as it would be with rule-based and sometimes statistical methods.

[2] Sennrich and Zhang (2019) achieved higher scores with an NMT system than a phrase-based system on a training corpus containing only 100,000 tokens. This is comparable in size to the New Testament of the Bible, the most frequently translated section.

in neural machine translation. We translate from Arabic, German, Tagalog, Turkish, and Xhosa into English, using up to 1,106 helper languages. We also present the results of training translation models on these corpora using established low-resource hyperparameters, finding that the best approach depends to a large extent on the language pair. We hope that these results will help elucidate best practices for leveraging massively multilingual low-resource parallel datasets in future MT work.

2. Related Work

Neural machine translation (NMT) (Kalchbrenner and Blunsom, 2013) has become the state-of-the-art approach to MT in recent years, employing new innovations in machine learning to achieve high performance in many language pairs. This approach was first shown to be effective with respect to the previously state-of-the-art phrase-based statistical approach (Koehn et al., 2003) in Sutskever et al. (2014) and Bahdanau et al. (2015). Its success was due to the combination of the sequence-to-sequence neural network, the use of the LSTM (Hochreiter and Schmidhuber, 1997), and the introduction and refinement of the attention mechanism (Luong et al., 2015).

Although the main focus of investigation and improvement in NMT has been high-resource settings with millions of sentences, NMT has made great strides in low-resource settings as well. Initially considered ineffective without very large parallel corpora (Koehn and Knowles, 2017; Lample et al., 2018), NMT has achieved performance exceeding phrase-based MT in such settings by exploiting the same neural architectures as high-resource systems—primarily LSTMs (Hochreiter and Schmidhuber, 1997) and Transformers (Vaswani et al., 2017)—and by performing fine-grained hyperparameter tuning (Sennrich and Zhang, 2019). Another proposed strategy in for low-resource NMT is learning a latent variable NMT model with variational inference (McCarthy et al., 2019), though this has not been tried in a low-resource setting similar to ours. Unsupervised NMT (Artetxe et al., 2018) was suggested as a solution to the scarcity of parallel data for many languages, though this was found to be generally ineffective for low-resource and morphologically rich languages (Guzmán et al., 2019). Unsupervised phrase-based and neural models have also been proposed (Lample et al., 2018) and are reasonably effective when parallel data does not exist, though this is perhaps unrealistic given that the Bible exists in over 1,000 languages. Consequently, we aim to translate in a more supervised fashion.

One surprisingly effective approach in high-resource settings has been to train translation models on multiple language pairs at once (Johnson et al., 2017). This is done by concatenating the bitexts for multiple language pairs together into one large parallel corpus. Such approaches have enabled zero-shot translation into low- and no-resource languages, though the performance is understandably quite variable depending on the languages involved. Aharoni et al. (2019) have experimented with adding increasingly large numbers of languages to the multilingual parallel corpus, up 103 languages. These approaches have used exclusively high-resource or a mix of high-resource and low-resource languages. By comparison, we do not include auxiliary data to perform a controlled study in a limited setting; we just use the Bible, which is the same size and domain across all of our experiments. Of note, Arivazhagan et al. (2019) point to the difficulty of balancing data, which is less of an issue in our multi-parallel low-resource setting.

Recent studies have investigated transfer learning in the low-resource multilingual setting using multiple unrelated languages (Gu et al., 2018; Zoph et al., 2016), and some have done the similar work using multiple related languages (Nguyen and Chiang, 2017). These approaches employed a small number of helper languages during training, but no more than 5 to 10 at once. Prior work suggests that this could either result in transfer learning across languages, leading to increased performance (as in the above-cited works), or perhaps it could result in a bottleneck where there are too few parameters for too many languages (Sachan and Neubig, 2018; Wang et al., 2018). These latter works suggest that using more closely related languages will reduce the bottleneck effect and increase BLEU compared to using more unrelated languages. We aim to investigate this effect for our particular context in a more large-scale manner; we perform multilingual NMT by concatenating many low-resource corpora derived from the Bible, up to over 1,000 languages.

3. Data Preparation

The monolingual Bibles used in this study are from the Bible corpus of McCarthy et al. (2020), which contains over 4,000 translations (of varying lengths) in over 1,000 languages. This corpus is an aggregation of prior Bible corpora (Mayer and Cysouw, 2014; Asgari and Schütze, 2017; Black, 2019) and web-scraped Bible data, all post-processed to be in the same format. Namely, these monolingual corpora are all verse-aligned,[3] normalized to Unicode NFKC, modified such that archaic English forms are replaced with their contemporary equivalents (e.g., "thou" is changed to "you", and "-est" and "-eth" verb inflections are replaced with their modern "-es" or "-s" forms), tokenized,[4] and deduplicated. The Old Testament (OT) contains approximately 31,000 verses, and the New Testament (NT) contains approximately 8,000 verses. Some translations have the entire OT and NT, while others might have all of one or fragments of either.

3.1. Multilingual Many-Way Bible Corpora

We selectively concatenate subsets of the aforementioned set of 1,100+ monolingual verse-aligned corpora to form multilingual parallel corpora. These are aligned in such a way that one verse appears per line. We wish to evaluate whether adding more languages to the training/development sets results in monotonic performance increases, as well as whether using similar source languages

[3] A specific line number represents the same verse across all Bible translations for any given language. When a verse is absent from a particular translation, the line is blank.

[4] All Bibles from Mayer and Cysouw (2014) were already tokenized except Chinese, which we segment into individual characters. All other Bibles are tokenized using spaCy (`https://spacy.io/`).

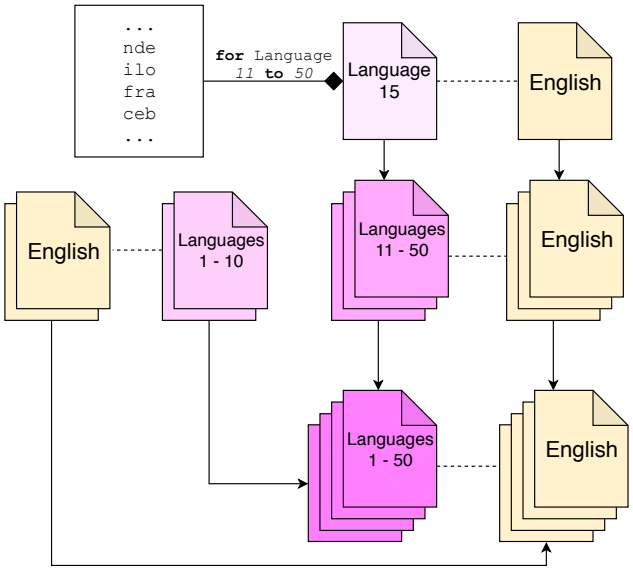

Figure 2: Visualization of the process of creating the 50-language multilingual corpus. Dotted lines connect parallel corpora. We use the list of languages (top-left) to determine which languages to add. For each specified language, we iterate over the verse-aligned `Language` and `English` corpora in parallel. If neither line is blank, we add to `Languages 11-50` and its respective `English` corpus. Once the loop is complete, we concatenate `Languages 1-10` to `Languages 11-50`; we also concatenate their respective `English` corpora. This makes the `Languages 1-10` corpus a subset of `Languages 1-50`.

is better than using unrelated source languages. Thus, we create a series of training datasets of varying sizes and linguistic compositions.

Our evaluation source languages of focus are Arabic (`arb`), German (`deu`), Xhosa (`xho`), Tagalog (`tgl`), and Turkish (`tur`). These were chosen to represent a variety of morphological complexities, syntaxes, and genealogical origins. Each of these languages has at least one Bible translation containing the Old and New Testaments; we select the translation with the most verses. We begin by creating bitexts for each of these languages, wherein the source language is one of the above-listed evaluation languages and the target language is always English. These are the datasets on which we will train our bilingual baselines. We then create a 5-language parallel corpus by concatenating each of these bitexts.

To create the 10-, 50-, 100-, 500-, and 1000+-language parallel corpora, we begin by randomly shuffling the remaining languages in the monolingual Bible datasets. We want each of the smaller corpora to be subsets of the larger corpora such that we are essentially comparing expanded versions of the same corpus. To ensure that the smaller corpora are a subset of the larger corpora, we sample ($\text{size}_n - \text{size}_{n-1}$) languages from the shuffled set of Bible languages, where size_n is the intended number of languages in the multilingual corpus and size_{n-1} is the number of languages included in the next-largest corpus. (This amounts to sampling languages uniformly without replacement.) Because we often have multiple Bible versions per language, we select the Bible version for which we have the most verses. These Bibles are then extracted in parallel with English using the same method as the bitexts and 5-language Bible sets. Finally, the size_{n-1} corpus and the new ($\text{size}_n - \text{size}_{n-1}$)-language corpus are concatenated to yield the final size_n corpus. See Figure 2 for a visual representation of this process. Table 1 gives the size (in verses and tokens) of all multilingual corpora.

No. Languages	Verses	Tokens
5	0.152 M	4.209 M
10	0.190 M	5.705 M
50	0.537 M	21.107 M
100	0.974 M	41.792 M
500	4.891 M	249.252 M
All	10.407 M	561.938 M

Table 1: The number of verses and tokens (in millions) for the source side of all multilingual corpora. For comparison, the number of verses in the German bilingual corpus is 0.031M, and the number of tokens in the source side is 0.831M. Note that the increase in size from 5 to 10 languages is not quite double because of the relatively small size of the monolingual corpora for languages 6–10; 4 of these languages have only the New Testament or Old Testament, but not both.

3.2. Similar-Language and Similar-Script Multilingual Corpora

To evaluate whether low-resource multilingual models benefit from training on similar languages (as opposed to unrelated languages), we create a separate parallel corpus for each of our evaluation languages. Each corpus contains 5 languages total: the evaluation language and 4 typologically and grammatically similar languages to the evaluation language. These are created for comparison with the 5-language corpus consisting of just the evaluation languages (i.e., the 5-language multilingual corpus from §3.1). For German, Xhosa, and Tagalog, we expect there to exist some subword vocabulary overlaps in their respective similar languages. See Table 2 for the list of languages included in their similar-language corpora.

Turkish and Arabic, however, are written in different scripts than most languages that are typologically similar to them, so we do not see much subword overlap between similar languages. Thus, we create **two** similar corpora each for Arabic and Turkish: one containing 4 other languages that are typologically and grammatically similar to the evaluation language (a *similar-language* corpus, as above), and one containing 4 other languages written in the same script as the evaluation language (a *similar-script* corpus). For the latter, we aim to include languages which are still typologically close to the evaluation language given the same-script constraint, though this is not always possible. See Table 2 for the list of languages included in Arabic's and Turkish's similar-language and similar-script corpora.

In total, we have 5 bilingual baseline corpora, 6 multilingual corpora of varying sizes featuring not-necessarily-

Language	Similar Languages	Similar Scripts
Arabic	Hebrew, Syriac, Amharic, Maltese	Azeri, Urdu, Farsi, Uyghur
German	Swedish, Danish, Dutch, Icelandic	
Tagalog	Javanese, Indonesian, Cebuano, Ilocano	
Turkish	Azeri, Gagauz, Uyghur, Kazakh	Zapotec, German, Hungarian, Spanish
Xhosa	Zulu, Swahili, North Ndebele, Kinyarwanda	

Table 2: The languages included in the similar-language and similar-script corpora for each evaluation language.

No. Languages	arb	deu	tgl	tur	xho
Bilingual	9.5	**14.6**	**15.4**	7.1	7.8
5	**10.6**	13.4	13.9	8.3	**10.4**
10	10.1	13.5	12.9	**8.8**	9.4
50	1.8	2.0	1.6	1.2	2.1
100	0.5	0.5	0.7	0.4	0.4
500	0.6	0.5	0.6	0.4	0.4
All	0.5	0.4	0.6	0.3	0.4

Table 3: BLEU scores for all n-language multilingual corpora, where n refers to the number of languages in the source side of a given corpus. "All" contains 1107 languages (everything except English). The BLEU scores for the bilingual and 5-language corpora are similar to those in (Sennrich and Zhang, 2019), but because we use a different dataset and training setup, these are not directly comparable.

related languages (from §3.1), 5 similar-language corpora, and 2 similar-script corpora. This yields 18 distinct corpora for training and thus 18 distinct neural translation models for comparison.

We wish to evaluate how the models generalize at test time to text from an unseen context in the same general domain as the training data; thus, rather than shuffle the corpus and randomly sample, we use a temporal train/development/test split. For each of our aforementioned corpora—bilingual *and* multilingual—we use all of the available Bible data except for the Book of Revelations for the training set. The first 100 lines of Revelations are the development set; for any given multilingual corpus, the development set includes all languages in its associated training set. The test sets are all bilingual; we simply create separate bitexts using the remaining lines from Revelations for each evaluation language, where the source language is the evaluation language and the target language is English.

4. Experiments

This study aims to discover the relationship between the number of languages in a multilingual dataset and the performance of NMT into English in a low-resource setting. We observe whether the performance increase is monotonic with respect to the number of languages, as well as at what point we begin to see diminishing returns. To do so, we compare the performance of neural models trained on our various multilingual corpora when translating from our evaluation languages to English. This is essentially the same as the many-to-one approach of Johnson et al. (2017).

All corpora have their tokens split into subwords using BPE (Sennrich et al., 2016). This is run jointly on all languages in the dataset, so each multilingual corpus will have different subword splits. We use 32,000 merge operations for all corpora despite their varying sizes; this results in more aggressive word splitting for larger corpora.

We use fairseq (Ott et al., 2019) to run both training and inference. Separate Transformer-based models are trained for each of our multilingual corpora using the low-resource hyperparameters from Sennrich and Zhang (2019).[5] Training is performed for a maximum of 100 epochs with early stopping after 10 epochs with no improvement in validation loss. All models were trained on one GTX 1080Ti each. The bilingual models converged in approximately one day, the 5-language and 10-language models in 1–2 days, the 50-language model in 4 days, and the 500-language and 1,107-language models required over a week. Inference is performed using a beam size of 5.

We also investigate whether the improved performance of NMT in multilingual settings is due to the similarity of the languages in the multilingual training set, or simply because the model observes more data during training.[6] We do so by comparing the performance of MT systems trained on the 5-language multilingual corpus and the similar-language and similar-script corpora for each evaluation language (all containing 5 source languages for comparability). These corpora and models are preprocessed and trained, respectively, according to the same procedure as in the previous experiment.

5. Results

Table 3 contains BLEU scores for the translation task for all models trained on varying numbers of source languages. All scores are for translations from the given evaluation language into English. We score on detokenized output translations using SacreBLEU (Post, 2018).

We first note that the BLEU increase/decrease with respect to the number of training languages is not uniform across evaluation languages. Indeed, for the 5-language model, we see notable gains over bilingual models for Arabic, Turkish,

[5] 1-layer encoder, 1-layer decoder, hidden size 1024, embedding size 512, tied decoder embeddings, hidden and embedding dropout 0.5, source and target word dropout 0.3, label smoothing 0.2, batch size 1000, initial learning rate 0.0005, Adam optimizer.

[6] More data in general may help for a variety of reasons that are not well understood, especially in low-resource settings. Potential causes include that more data may improve the target-language language model, reduce the extent of overparameterization in low-resouce settings, help the model understand language in general instead of one specific language, among others.

Similarity	arb	deu	tgl	tur	xho
Unrelated	**10.6**	13.4	13.9	**8.3**	10.4
Similar Language	7.4	**17.1**	**15.7**	6.0	**10.9**
Similar Script	9.3	-	-	8.0	-

Table 4: BLEU scores for all similar-language and similar-script corpora. "Unrelated" refers to the 5-language multilingual corpus of Table 3.

and Xhosa, but we also observe decreases for German and Tagalog. The gains for Arabic are surprising, as no other language in the 5-language corpus uses the same script—hence, we expect no shared vocabulary cross-lingually. Inversely, the decreased performance for German is surprising, given that three other languages use the same script and should thus have some shared subword vocabulary.

The trends are further muddled when observing the performance of the 10-language model, for Turkish sees improved performance while Arabic, Tagalog, and Xhosa see decreased performance. German has relatively unchanged performance compared to the 5-language model. So far, all we may say is that the performance when adding more languages is language-specific.

Performance in general drops quickly once we reach 50 languages, however, and all multilingual models trained on larger and larger multilingual corpora from this point achieve lower and lower scores. The potential reasons for this are many: interference across languages, underparameterized models, the evaluation languages forming too small a fraction of the overall corpus, among others.[7] The similar-language multilingual models should allow us to partially investigate the first possibility.

Table 4 contains BLEU scores for models trained on the similar-language and similar-script corpora. As expected, for languages whose similar languages do not share a script (i.e., Arabic and Turkish), we see that training on multilingual corpora consisting of languages sharing the same script is more effective than training on corpora consisting of languages sharing similar linguistic properties. However, for both of these languages, we see even better performance by simply training on a set of unrelated languages.

For German and Tagalog, however, we see notable gains when training on similar languages than on unrelated languages. These models outperform the bilingual models from Table 3 as well (and recall that the bilingual models outperformed the 5-language models for German and Tagalog therein). The similar-language Xhosa model obtains more modest gains over the unrelated-language model, especially considering the extent of the BLEU gains on German and Tagalog. Once again, it seems that the effectiveness of any particular multilingual approach is highly language-dependent.

6. Qualitative Analysis

Metrics do not fully encapsulate translation performance. They may not capture critical phenomena, and may not align with human judgments (Wang et al., 2019). Further, they do not give a clear understanding of *patterns* of errors. Thus, this section focuses on the analysis of system outputs.[8] We provide examples of translations from the multilingual experiments and the similar-language and similar-script experiments in Table 5.

First, we note that fluency is strongly preferred over adequacy, especially for the models that do not degenerate. That is, taking the perspective of the decoder as a conditional language model, in some settings we have developed strong English language models which largely ignore their source text. This effect is clearer for smaller corpora than larger ones, where translations seem to be more semantically similar to their respective references. There seems to exist a negative correlation between the number of languages in a multilingual corpus and the translation fluency starting at 10 languages, and this effect becomes quite obvious starting at 50 languages. At 100 languages or more, we reach a maximal point of degeneration, where translations no longer change significantly as we add more languages. However, this effect only seems to apply in certain cases: the 100-language output of the second Tagalog example in Table 5 seems fluent, but is nonetheless semantically distant from the reference.

Using similar languages rather than random languages seems to help most source languages, but not Arabic or Turkish. Thus, we investigate how these translations differ for the latter language. In Table 5, we see that the similar-language model uses very simple and repetitive sentence structures, and that these sentences are not very fluent or adequate. The similar-script model exhibits more subtle repetition, and it seems slightly more adequate. Nonetheless, using unrelated languages leads to the best performance: while the last two sentences are repeated, we have the greatest n-gram overlap with the reference and the greatest fluency (despite that the second clause of the first sentence has no verb). These trends are representative of what was found when investigating the output translations for Turkish: using similar languages greatly harms performance, while using similar scripts worsens performance in more subtle ways. Note, however, that we see the opposite trend for German, Tagalog, and Xhosa, where using similar languages improves performance: those languages tend to see more repetition when using the unrelated-language models, and more fluency and adequacy when using the similar-language models.

A shared theme among these translations is *neural text degeneration*. As explained by Holtzman et al. (2020), text generated by neural models in many domains tends to repeat itself "at the token, phrase, and sentence levels." While these problems might be mitigated by using better decoding methods or the likelihood function of Holtzman et al. (2020), it is clear that the models as-is are learning something fundamentally incorrect in the sequence modeling task; however, it is unclear why this is more prominent for the more multilingual corpora. Note, for example, the

[7] We go into further detail on these possibilities and suggest improvements for future work in §7.

[8] The examples we provide here are not representative; with such low BLEU scores, it is natural that in many cases, translations are incorrect in unrelated ways. Our examples represent specific instances of the broader trends we recognize.

	Xhosa–English
Source	Akusayi kuba sabakho nasinye isiqalekiso ; yaye itrone kaThixo neyeMvana iya kubakho kuwo ; abakhonzi bakhe bamkhonze ,
Reference	There will be no curse . The throne of God and of the Lamb will be in the city . His servants will serve him .
Bilingual	» ‘ You will not make a curse on the throne of God . It will be a descendant of God . They will serve him .
5	» There will no longer be a curse . The throne of God and the throne of God and his servants will serve them .
10	» There will be no longer be no longer a sign of God . His servants will serve them .
50	» The Son of God is the Son of God , the God of God , the God of God .
100	» If any man does not know that he does not know that he does not know that he does not know that he does not know him .
Source	Waza lowo uhleli phezu kwetrone wathi , Uyabona , izinto zonke ndizenza ntsha . Athi kum , Bhala ; ngokuba lawo ngamazwi ayinyaniso , athembekileyo .
Reference	He who sat on the throne said : " Behold , I make all things new . Then he told me to write : for these words are faithful and true . "
Bilingual	He who sits on the throne and said : » I will answer you . « He said to me : » Be careful to me . «
5	Then he sat on the throne and said : » See , I will do everything I have done . « They said to me : » These are the truth . «
10	Then he sat on the throne . He said : » Behold , I have done all these things . « He said to me : » This is the truth . «
50	» He said to them : ‘ I am going to tell you , ‘ I am going to you . ’ ’ ’ ’ ’ ’ ’ ’ ’ ’ «
100	Jesus said : » I tell you , « he said : » If you do not know me . «
	Tagalog–English
Source	At humihip ang ikaanim na anghel , at narinig ko ang isang tinig mula sa mga sungay ng dambanang ginto na nasa harapan ng Dios ,
Reference	The sixth angel blew his trumpet . I heard a voice from the golden altar that is before God .
Bilingual	The sixth angel blew the sixth voice , and heard a voice from the horns of gold altar .
5	The sixth angel blew the sixth horn and heard a voice from the horns of the altar in front of God .
10	The sixth hour came from the sixth hour and heard a voice from the horns of the altar of God .
50	The next day the next day he said to them : » The blood of the blood of the blood of God . «
100	The next day of the day of God has been taken away from the altar .
Source	Sila’y may pinakahari na anghel ng kalaliman : ang kaniyang pangalan sa wikang Hebreo ay Abaddon , at sa Griego ay may pangalan siyang Apolyon .
Reference	The angel of the bottomless pit was their king . His name in the Hebrew language is Abaddon , and in Greek , Apollyon .
Bilingual	» They are aX king of the deep tongue . His name is a Hebrew Hebrew tongue , and Greeks .
5	They were the king of the deep sea . His name was the Hebrew : » They were a Jew , and Greeks .
10	They were the king of the depths of the sea . His name was the Hebrews . They were called Apollos ; and there were three name .
50	The king of Judah went to the king of Judah . He went to Jerusalem . He went to Jerusalem . He went to Jerusalem .
100	The king of Israel said : » This is the son of David . «
	Turkish–English
Source	‹ ‹ Kaftanlarını yıkayan , böylelikle yaşam ağacından yemeye hak kazanarak kapılardan geçip kente girenlere ne mutlu !
Reference	" ' Blessed are those who obey his commandments . They have the right to the tree of life , and may enter in through the gates of the city .
Sim. Scripts	» ‘ Those who touch their clothes are not able to eat their food . Those who go into the city will enter the city . ’
Sim. Langs	» Blessed are those who live in the city of life . Blessed is the gate of the city .
Unrelated	» Blessed are those who go into the city of acacia wood and walk through the city .
Source	Kuzu altıncı mührü açınca , büyük bir deprem olduğunu gördüm . Güneş keçi kılından yapılmış siyah bir çul gibi karardı . Ay baştan aşağı kan rengine döndü .
Reference	When he opened the sixth seal I saw a great earthquake . The sun became black as sackcloth of hair and the moon became like blood .
Sim. Scripts	When I saw that the sixth hour came to the sixth hour . I saw that the sun was covered with sackcloth . The sun was covered with sackcloth and ashes .
Sim. Langs	» I saw a great earthquake . I saw a great earthquake . I saw a great fire . I saw a great fire . The sun was covered with fire .
Unrelated	When I saw that there was a great earthquake , a great crowd of goats’ hair . The sun was covered with sackcloth . The sun was covered with sackcloth .

Table 5: Example translations from multilingual Xhosa–English and Tagalog–English models, as well as similar-language/-script Turkish–English models. Prediction colors are from VizSeq (Wang et al., 2019); blue represents matched tokens, teal represents tokens that few models matched correctly, and red represents unmatched tokens.

Xhosa–English translations: we do not observe degeneration for multilingual corpora under 50 languages, but this effect quickly begins to compound as the data grows to that point. None of the output translations hit the maximum sequence length (a hyperparameter) of 200 tokens for models using fewer than 50 languages, but at 50 languages, 11 translations hit the sequence length limit. At 100 languages, 15 of them do. A similar trend appears for Tagalog: 14 translations hit the maximum sequence length at 50 languages; 19 sentences do at 100 languages.

To quantify the growing divergence in translation lengths, we calculate the Wasserstein distances between the empir-

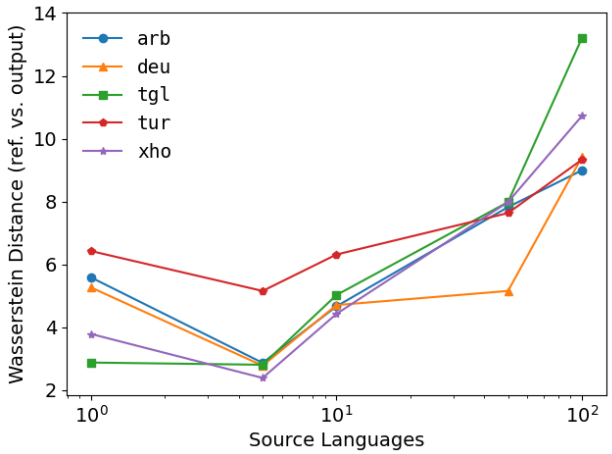

Figure 3: Wasserstein distance between the distributions of the reference translation lengths and predicted translation lengths, as a function of the number of source languages during training. All references and outputs are in English. Each line represents a different evaluation source language during testing. Note that the x-axis has a logarithmic scale.

ical length distributions of the reference translations and predicted translations. See Figure 3. There is a visible correlation between the number of source languages and the divergence between reference and output translation length. For every language–English pair, Pearson's $r > 0.88$ ($p < .05$); for Tagalog and Xhosa, $r > 0.95$ ($p < .01$), and Spearman's $\rho = 0.9$ ($p < .05$). Perhaps this indicates that we have increasing destructive interference as we increase the number of languages, or that the models are learning some unhelpful heuristics that apply to all languages. Perhaps these models would be able to tolerate more languages in more high-resource settings, as has been demonstrated in Aharoni et al. (2019). Further investigations will be necessary to determine the source of degeneration, though Holtzman et al. (2020) suggests that it is likely from a variety of sources.

One explanation for our too-powerful language model that ignores the source text is the amount of repetition in the training data: regardless of the number of training languages, only the English Bible is used as a target. Perhaps source language tags can alleviate this, because they condition the target language model on the source language's identity, at least partially. That might help explain why the "similar languages" group of Turkish maintains adequacy, but the "similar scripts" is not: with one alphabet on the source, it learns to produce the most-frequent LM prediction. BLEU is higher for similar scripts because common words are common, not because of any better-learned translation ability; BLEU does not convey the entire story.

7. Conclusions

We have constructed massively multilingual aligned datasets containing varying numbers of languages for training machine translation models. We have also performed some initial investigations into how to use these datasets for neural machine translation. These will hopefully give future researchers a starting point in performing much more large-scale experiments than were previously possible in the low-resource setting.

Our results suggest that low-resource multilingual NMT is highly variable in performance cross-lingually. In general, the best performance is achieved using either bilingual models, 5-language models, or 10-language models. Moreover, the relatedness of the multiple languages certainly has an effect on performance, though whether this effect is positive or negative is language-dependent. This implies that the idea of simply adding more data to a training corpus regardless of language will not always result in the best performance—though it sometimes does when done carefully. This highlights the need for multiple evaluation languages when working in the multilingual setting, for methods that work in one language may not generalize to others (Bender, 2009; Bender, 2011). If one is pursuing the best model for a specific low-resource language pair, it is imperative to try a variety of source language sets when training multilingually and to train a good bilingual baseline.

This variable cross-lingual performance could be due to the bottleneck effect mentioned in §2 or the destructive interference described in §6. Future work could treat the many-language corpora as high-resource datasets and use more typical high-resource hyperparameters instead of the low-resource hyperparameters used here. One could also investigate a larger variety of source language combinations in order to find some principled method to create multilingual corpora which improve NMT performance for a given language pair.

We encourage future researchers working with this massively multilingual aligned corpus to experiment with its many possible language subsets to discover the best settings for future NMT experiments. It is not infeasible that one could achieve an effective massively multilingual translation model using some hyperparameter tuning, a many-to-many setup instead of many-to-one, or some updated neural architecture that enables more effective parameter sharing cross-lingually.

Copyright restrictions limit our ability to publicly disseminate the data. The datasets used here are available by contacting the authors.

8. References

Aharoni, R., Johnson, M., and Firat, O. (2019). Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Arivazhagan, N., Bapna, A., Firat, O., Lepikhin, D., Johnson, M., Krikun, M., Chen, M. X., Cao, Y., Foster, G., Cherry, C., Macherey, W., Chen, Z., and Wu, Y. (2019). Massively multilingual neural machine translation in the wild: Findings and challenges. *CoRR*, abs/1907.05019.

Artetxe, M., Labaka, G., and Agirre, E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–

798, Melbourne, Australia, July. Association for Computational Linguistics.

Asgari, E. and Schütze, H. (2017). Past, present, future: A computational investigation of the typology of tense in 1000 languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 113–124, Copenhagen, Denmark, September. Association for Computational Linguistics.

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

Bender, E. M. (2009). Linguistically naïve != language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32, Athens, Greece, March. Association for Computational Linguistics.

Bender, E. M. (2011). On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6(3):1–26.

Black, A. W. (2019). CMU Wilderness Multilingual Speech Dataset. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Gu, J., Hassan, H., Devlin, J., and Li, V. O. (2018). Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354, New Orleans, Louisiana, June. Association for Computational Linguistics.

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 6097–6110, Hong Kong, China, November. Association for Computational Linguistics.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Holtzman, A., Buys, J., Forbes, M., and Choi, Y. (2020). The curious case of neural text degeneration. *International Conference on Learning Representations*.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., 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 Linguistics*, 5:339–351.

Kalchbrenner, N. and Blunsom, P. (2013). Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA, October. Association for Computational Linguistics.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, August. Association for Computational Linguistics.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

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

M Paul Lewis, et al., editors. (2015). *Ethnologue: languages of the world*. SIL International, Dallas, eighteenth edition.

Luong, T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, September. Association for Computational Linguistics.

Mayer, T. and Cysouw, M. (2014). Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3158–3163, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

McCarthy, A. D., Li, X., Gu, J., and Dong, N. (2019). Improved Variational Neural Machine Translation by Promoting Mutual Information. *arXiv e-prints*, page arXiv:1909.09237, September.

McCarthy, A. D., Wicks, R., Lewis, D., Mueller, A., Wu, W., Adams, O., Nicolai, G., Post, M., and Yarowsky, D. (2020). The Johns Hopkins University Bible Corpus: 1600+ tongues for typological exploration. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, Marseilles, France, May. European Language Resources Association (ELRA).

Nguyen, T. Q. and Chiang, D. (2017). Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Ma-*

chine Translation: Research Papers, pages 186–191, Belgium, Brussels, October. Association for Computational Linguistics.

Sachan, D. and Neubig, G. (2018). Parameter sharing methods for multilingual self-attentional translation models. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 261–271, Brussels, Belgium, October. Association for Computational Linguistics.

Sennrich, R. and Zhang, B. (2019). 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, July. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (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, Berlin, Germany, August. Association for Computational Linguistics.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wang, Y., Zhang, J., Zhai, F., Xu, J., and Zong, C. (2018). Three strategies to improve one-to-many multilingual translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2955–2960, Brussels, Belgium, October-November. Association for Computational Linguistics.

Wang, C., Jain, A., Chen, D., and Gu, J. (2019). VizSeq: a visual analysis toolkit for text generation tasks. 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): System Demonstrations*, pages 253–258, Hong Kong, China, November. Association for Computational Linguistics.

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, November. Association for Computational Linguistics.