Benchmarking Neural and Statistical Machine Translation on Low-Resource African Languages

Kevin Duh, Paul McNamee, Matt Post, Brian Thompson
Johns Hopkins University
Baltimore, MD, USA
{kevinduh, post}@cs.jhu.edu, {mcnamee,brian.thompson}@jhu.edu

Abstract

Research in machine translation (MT) is developing at a rapid pace. However, most work in the community has focused on languages where large amounts of digital resources are available. In this study, we benchmark state of the art statistical and neural machine translation systems on two African languages which do not have large amounts of resources: Somali and Swahili. These languages are of social importance and serve as test-beds for developing technologies that perform reasonably well despite the low-resource constraint. Our findings suggest that statistical machine translation (SMT) and neural machine translation (NMT) can perform similarly in low-resource scenarios, but neural systems require more careful tuning to match performance. We also investigate how to exploit additional data, such as bilingual text harvested from the web, or user dictionaries; we find that NMT can significantly improve in performance with the use of these additional data. Finally, we survey the landscape of machine translation resources for the languages of Africa and provide some suggestions for promising future research directions.

Keywords: machine translation, low-resource languages, evaluation

1. Introduction

Parallel text is an essential ingredient for building Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) systems. By definition, parallel text is a kind of corpus consisting of pairs of sentences, where one is written in the source language (e.g., Somali) and the other is its translation in the target language (e.g., English). This is an expensive resource to manually generate, requiring translators that are proficient in both languages.

Commercial SMT and NMT systems are often trained on millions to tens of millions of sentence pairs, if not more (Wu et al., 2016). It is unclear how these systems perform when the training data contains significantly fewer sentence pairs. For many languages in the world, and in particular for languages in the African continent, at present we cannot reasonably expect such a large amount of training data. While there is no established convention, we might consider systems that are trained on less than 100 thousand sentence pairs to be low-resource.

Previous work (Koehn and Knowles, 2017; Sennrich and Zhang, 2019) has established the idea that there is a cross-over point between NMT and SMT performance depending on the amount of training data. See Figure 1. The intuition is that NMT is data-hungry, so may perform worse than SMT in low-resource settings, but begins to excel when there is sufficient training data. With recent advances in NMT, the cross-over point has gradually decreased. Nevertheless, in general it is difficult to predict *a priori* whether we are on the left or right side of the cross-over point until we actually build the systems.

In this work, we perform a detailed evaluation of low-resource scenarios for SMT and NMT, focusing on Somali-to-English and Swahili-to-English translation with training data on the order of 24 thousand sentence pairs. We make two main observations:

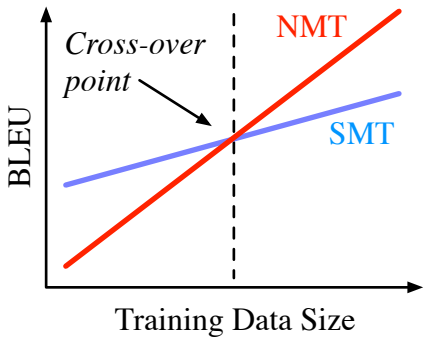

Figure 1: Illustration of the effect of training data size on Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) systems. Given the differences in terms of data requirements, there is a cross-over point that decides whether SMT or NMT has better translation quality (e.g., in terms of BLEU score). See (Koehn and Knowles, 2017; Sennrich and Zhang, 2019) for figures with example values.

1. We find that SMT and NMT perform similarly in these scenarios, but NMT importantly requires careful hyperparameter tuning to match SMT performance.

2. We find that both SMT and NMT can exploit additional data such as noisy parallel text harvested from the web, but NMT benefits significantly more from it.

Our goal is an empirical evaluation comparing standard models in SMT and NMT. In this respect, it is orthogonal to other work that propose novel methods to improve results under low-resource, for example by exploiting monolingual/synthetic corpora (Wang et al., 2019a; Fadaee et al., 2017), multilingual transfer (Zoph et al., 2016; Gu et al., 2018; Dabre et al., 2019; Kocmi and Bojar, 2018), or alternative modeling/training strategies (Zaremoodi and Haffari, 2019; Nguyen and Chiang, 2018; Neubig and Hu, 2018).

	Somali-English		Swahili-English	
	sentences	words	sentences	words
Train	24.8k	758k	24.9k	809k
Validation: Text	2.6k	68k	3.3k	87k
Test1: Text	4.0k	122k	6.7k	181k
Test2: Transcripts	7.1k	102k	5.8k	83k

Table 1: Data set sizes in sentences and words. Validation and Test1:Text sets consist of news, topical, and blog text. Test2:Transcripts consists of news broadcast, topical broadcast, and conversational telephony. The training set contains a mix of genres, but is most similar to Validation and Test1.

In the following, we first describe our low-resource condition for Swahili and Somali (Section 2). We then present our two main results: Section 3 compares SMT and NMT in this low-resource condition; Section 4 compares these systems in the case where additional data such as web-mined bitext can be exploited. Finally, we end with a discussion of the landscape of current resources for African language MT (Section 5), related work (Section 6), and potential future directions (Section 7).

2. Datasets and Low-Resource Condition

Our baseline training data consists of twenty-four thousand sentence pairs in both Swahili-English and Somali-English tasks. The data comes from the IARPA MATERIAL program and represents a diverse set of genres.[1]

Swahili is a Bantu language spoken widely in Eastern and Southeastern Africa. It exhibits agglutinative morphology and has a large number of noun classes (18), with which adjectives and verbs must agree. The dominant word order is SVO. Somali is an Afroasiatic language, and classified as part of the Cushitic branch. It is spoken in Somalia, Djibouti, and parts of Ethiopia and Kenya. It exhibits agglutinative morphology and SOV word order. Both Swahili and Somali are written in the Latin script.

The dataset sizes are given in Table 1. For robustness, we prepared two different testsets. The **Text testset** also comes from IARPA MATERIAL and represents a matched condition to our training and tuning data. The **Transcripts testset** are the reference speech transcripts, and evaluates how our systems tuned on text might perform with speech data. The validation set is used for MIRA tuning in SMT (for finding weights that tradeoff e.g., language model, translation model, and length penalty), and for early-stopping in NMT (for stopping the training run when perplexity fails to improve after a several consecutive checkpoint updates, which is effective against overfitting).

For both SMT and NMT, the data is uniformly preprocessed with the same Joshua tokenizer and then lower-cased. For NMT, we additionally segment words into subwords via Byte Pair Encoding (Sennrich et al., 2016). For a fair evaluation, all translation outputs are mapped backed to the raw untokenized forms, then evaluated via SacreBleu (Post, 2018)[2] to ensure that BLEU (Papineni et al., 2002) is computed using the same tokenization.

3. Comparing Standard SMT and NMT

Using the available training data, we built SMT systems using the Apache Joshua toolkit (Post et al., 2013) and NMT systems using the AWS Sockeye toolkit (Hieber et al., 2017).[3] Our Joshua system is a phrase-based model that represents the state of the art in SMT, with 4-gram KenLM language model and MIRA-based tuning. Our Sockeye system is a transformer model (Vaswani et al., 2017), which is among the strongest performers in the field of NMT. We vary the following hyperparameters:

- **Transformer Architecture**: number of layers (1, 2, 4, 6); embedding size (256, 512, 1024), number of hidden units in each layer (1024, 2048), number of heads in self-attention (8, 16).
- **Preprocessing**: number of Byte Pair Encoding (BPE) operations (1k, 2k, 4k, 8k, 16k, 32k)
- **Training configuration**: initial learning rate for the Adam optimizer ($3 \times 10^{-3}, 6 \times 10^{-3}, 10 \times 10^{-3}$)

This hyperparameter tuning leads to a large number of NMT models–approximately 600 per language pair. Our goal is to compare their performance in terms of BLEU scores with the SMT models.

Table 2 summarizes the results on one of the test sets. For our final models, we observe that SMT and NMT achieve similar BLEU scores: 15.1 vs 14.4 BLEU for Somali and 24.4 vs 24.8 for Swahili. There is a common expectation that low-resource settings pose challenges for NMT, because neural methods are data-hungry (Koehn and Knowles, 2017). The comparable results between SMT and NMT agree with more recent findings that NMT technology has advanced rapidly, and is increasingly capable of handling lower amounts of data (Sennrich and Zhang, 2019).

However, an important caveat worth emphasizing is that the positive NMT results do not come straight out of the box. For low-resource conditions, extensive hyperparameter tuning of the NMT models is necessary for good performance. NMT hyperparameters such as the number of neural layers, the type of neuron, and the learning rate for the training

[1] For the purpose of this benchmark, we use the Build Pack for training and the Analysis Packs for tuning and testing; we do not use other annotations such as domain or query relevance. For more details about the program and data, refer to (Rubino, 2018).

[2] https://github.com/mjpost/sacreBLEU

[3] Joshua: https://joshua.apache.org; Sockeye: https://github.com/awslabs/sockeye

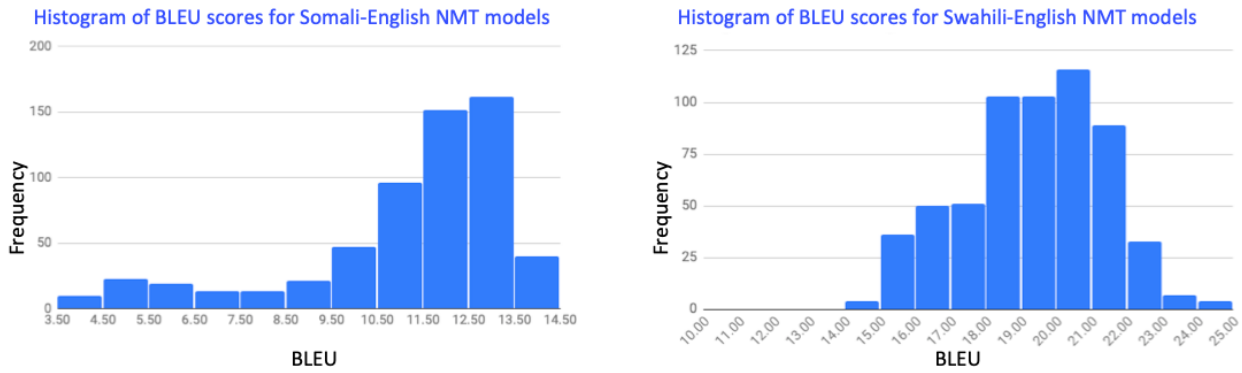

Figure 2: Histogram of testset BLEU scores for various NMT models with different hyperparameters. Left: Somali-English; Right: Swahili-English. Note the large variance, ranging from some state of the art NMT that is competitive to SMT, to a large number of underperforming NMT systems.

algorithm are all sensitive to the training data and require careful setting by the model developer. To illustrate this, we have plotted the distribution of BLEU scores for the 600 NMT models with different hyperparameter settings in Figure 1. Note that the majority of the models **underperform** SMT. The histogram is also summarized in the statistics in Table 2, which report their BLEU scores at the 75th, 50th, and 25th percentile. Note that an NMT model selected at random, performs much worse on average than SMT (e.g., 15.1 SMT vs 11.7 NMT for Somali, 24.4 SMT vs 18.7 NMT for Swahili). Additionally, the best hyperparameter setting for the Somali-English tasks is quite different from that of the Swahili-English task, so it is difficult to define standard defaults as best practice settings.

To summarize: state of the art SMT and NMT systems show comparable results for low-resource conditions (24k sentence training data), but NMT requires much more careful hyperparameter tuning by the model developer to achieve this result. We believe this observation is important for developing NMT for a new language pair: one must explore a large space of hyperparameters, since neural models are sensitive in low-resource conditions. While hyperparameter tuning can be expensive, it can be feasible when training data is scarce, which is exactly our low-resource scenario.

4. Exploiting Additional Data

There are two main research directions for solving low-resource problems: (a) develop new modeling techniques that require less data, and (b) devise ways to exploit additional opportunistic data sources. In this work we first focus on the latter.

We explore three types of resources:

1. *Dictionary*: Pre-existing dictionaries may be available from various sources (Ramesh and Sankaranarayanan, 2018; Thompson et al., 2019b). We define dictionaries as word-by-word or phrase-by-phrase translations, which are different in format from the sentence-by-sentence parallel data in that there may be less contextual information to learn from.[4]

2. *Found Bitext*: Pre-existing parallel sentences may be found via various sources (Tiedemann, 2012; Christodouloupoulos and Steedman, 2015), such as the Bible. These are relatively clean datasets that contain useful sentence-by-sentence translations, but may be in a different domain/genre from our baseline training set and testset.[5]

3. *Mined Bitext*: Parallel sentences can be mined by crawling the web, for example via Paracrawl[6]. We exploit the fact that various websites exist in multiple languages and devise methods to discover and extract these parallel sentences. Depending on the language-pair, large paracrawl corpora may be possible. The challenge with using this crawled data is that it can be more noisy (Koehn et al., 2019), i.e. automatically discovered parallel sentences may not always be true translations.[7]

For each of these resource types, there exist challenges in both the acquisition of the data itself and the integration thereof into existing MT training workflows. When successful, these additional resources may efficiently supplement the expensive baseline training data.

Table 3 shows the effect of Found Bitext, Paracrawl, and the Dictionary when added to our baseline training data. Similar to Table 2, the NMT results are obtained by careful tuning for each dataset condition (exploring approximately 60 models with different hyperparameters for each condition). We observe noticeable improvements for both

[4]We used Panlex as the dictionary for both languages.

[5]For Swahili, we employed found bitext from the DARPA LORELEI program, Global Voices, and the Tanzil corpus. For Somali, we employed parallel sentences from TED Talks, Tanzil, the Bible, and LORELEI.

[6]https://paracrawl.eu/

[7]Note one can define different versions of Paracrawl based on different ways to filter potential noise; we used a version with relatively aggressive cleaning.

	SMT	NMT				
		chosen	best	75%	50%	25%
Somali-English	15.1	14.4	14.4	12.7	11.7	9.9
Swahili-English	24.4	24.8	24.8	20.5	18.7	15.6

Table 2: **SMT vs. NMT: BLEU score on the Text Testset**. Models trained with 24k baseline dataset. For NMT, we trained approximately 600 systems with different hyperparameters. The "chosen" column shows the BLEU score on the test set based on a model chosen based on the validation set (which is a fair comparison to the SMT score), and the "best" column shows the best possible attainable score (in this case, chosen models happen to be the best models). We also show the 75, 50, 25 percentile of BLEU scores on the test set. **The wide range of scores for NMT indicates the sensitivity of NMT to design choices and the importance of careful tuning in low-resource scenarios.**

	Data	Test1: Text		Test2: Transcripts	
	Size	SMT	NMT	SMT	NMT
Somali-English baseline	24k	15.1	14.4	7.8	7.7
+ paracrawl	104k	15.7	20.2	8.8	10.5
+ dictionary	50k	15.4	14.3	8.3	7.9
+ dictionary + found-bitext	273k	16.8	24.4	9.4	13.3
+ dictionary + found-bitext + paracrawl	354k	**17.3**	**25.0**	**9.5**	**13.6**
Swahili-English baseline	24k	24.4	24.8	15.4	13.4
+ paracrawl	85k	24.2	26.6	14.5	15.1
+ dictionary	123k	24.6	25.3	15.5	13.1
+ dictionary + found-bitext	312k	25.5	33.3	**16.2**	18.7
+ dictionary + found-bitext + paracrawl	373k	**25.6**	**33.7**	15.9	**20.6**

Table 3: The effect of additional resource types for SMT and NMT. We show BLEU scores on the text and transcripts test sets. Data Size shows the number of segments used for training. The NMT BLEU scores correspond to those "chosen" on the validation set, and is a fair comparison with the SMT numbers. The best BLEU score in each column is boldfaced. The baselines are trained on the MATERIAL training data, taken from Table 2. Observe that adding paracrawl, dictionary and found-bitext to baseline tends to improve performance for both SMT and NMT, with NMT gaining significant benefits.

SMT and NMT. For example, on the Text Testset, SMT improved 2.2 BLEU points from 15.1 to 17.3 for Somali and 1.2 BLEU points from 24.4 to 25.6 for Swahili. For NMT, the improvement from additional data was much more significant: 10.6 BLEU points from 14.4 to 25.0 for Somali and 8.9 BLEU points from 24.8 to 33.7 for Swahili.

The trend is observed in the Transcripts test sets as well. In our experiments here, the models chosen on validation set, which are in the same domain as Text test, also worked well for Transcript test sets. In general, this is not always guaranteed and one may need to prepare a better-matching validation set, or employ domain adaptation techniques.

A factor to consider is whether to deploy a single model that serves any domain, or separate models that are optimized for each domain. Improvements in performance and robustness are possible depending on which scenario is chosen.

We conclude that exploiting additional data types is a fruitful research direction, especially for low-resource NMT.

5. Landscape of MT Resources for African Languages

We surveyed the resources available for various languages of Africa, to determine the feasibility of MT system development and additional data exploitation, as done for Somali and Swahili in previous sections.

The results are summarized in Table 4. Note that this table must be interpreted carefully for two reasons. First, the data conditions across languages are not directly comparable; for example, the apparently larger amount of Wikipedia articles in Yoruba than Somali does not imply that it is easier to build a Yoruba SMT or NMT system. Second, the statistics in the table are only meant as approximate numbers for reference: they are derived from complex calculations which are subject to change.

The table shows the top languages by the number of native speakers in Africa.[8] This is a diverse set of languages, including languages in the Afroasiatic, Niger-Congo, and Indo-European families. The columns CommonCrawl and Wikipedia indicate the amount of monolingual data on the web, which can be viewed as an indicator of the upper limit of how much web-crawled data we may be able to obtain. CommonCrawl[9] is a project that aims to archive all of the web, and the column in the table indicates our estimate of the number of webpages in its data-dump. More specifically, the number of webpages is estimated from the CC-MAIN-2019-35 datadump statistics[10]. The statistics report the percentage of webpages identified automatically by Compact Language Detector 2 (CLD2) into certain languages. We multiply this by the total datadump size (approximately 3 billion webpages) to obtain estimates for the

[8] Source: https://en.wikipedia.org/wiki/Languages_of_Africa

[9] https://commoncrawl.org/

[10] https://commoncrawl.github.io/cc-crawl-statistics/plots/languages

Language	Family	CommonCrawl (#documents)	Wikipedia (#documents)	OPUS (#sents)
Afrikaans (afr)	Indo-European	387k	84.0k	1.6m
Akan (aka)	Niger-Congo	3k	0.7k	0.2k
Amharic (amh)	Afroasiatic	66k	14.8k	1m
Arabic (ara)	Afroasiatic	17,772k	945.7k	70m
Berber (ber)	Afroasiatic	0	0	0.1m
Chewa (nya)	Niger-Congo	8k	0.5k	0.9m
Hausa (hau)	Afroasiatic	45k	3.7k	0.4m
Igbo (ibo)	Niger-Congo	8k	1.4k	0.5m
French (fra)	Indo-European	133,401k	2136.3k	180m
Fulani (ful)	Niger-Congo	0	0.2k	0.3k
Kinyarwanda (kin)	Niger-Congo	71k	1.8k	0.8m
Kirundi (run)	Niger-Congo	3k	0.6k	0
Malagasy (mlg)	Austronesian	126k	91.9k	0.9m
Mossi (mos)	Niger-Congo	0	0	0
Oromo (orm)	Afroasiatic	15k	0.8k	0.2m
Portuguese (por)	Indo-European	60,762k	1013.0k	72m
Shona (sna)	Niger-Congo	8k	4.8k	0.8m
Somali (som)	Afroasiatic	117k	5.4k	0.2m
Swahili (swa)	Niger-Congo	234k	53.7k	1.2m
Tigrinya (tir)	Afroasiatic	21k	0.2k	0.4m
Xhosa (xho)	Niger-Congo	12k	1.0k	1.5m
Yoruba (yor)	Niger-Congo	21k	31.9k	0.5m
Zulu (zul)	Niger-Congo	24k	1.3k	1.1m

Table 4: Potential digital resources for an abridged list of languages in Africa. We show the potential monolingual resources (Number of CommonCrawl and Wikipedia documents) and bilingual resources (Number of bilingual sentence pairs via OPUS). One can compare the low-resource condition of these languages, using Somali and Swahili as a reference point. Please refer to Section 5 for details, since these numbers need to be interpreted with care. Languages that are not on this list might have even fewer resources.

number of webpages per language. From these pages, we further identify candidates that are translations to create the Paracrawl resource.

The Wikipedia column lists the number of articles on Wikipedia, and is another way to estimate the extent of web presence for a language.[11] Note that some languages have reasonable web presence, e.g., 91.9 thousand (k) pages for Malagasy and 53.7k pages for Swahili, whereas others have literally none (e.g., Mossi, Berber).

Next, the table reports the potential amount of found bitext. The main statistic comes from OPUS, a project that aggregates datasets for MT research. In the OPUS column, we show the number of parallel sentence pairs between English and the African language in question, as available from OPUS (Tiedemann, 2012). For example, for Zulu we can obtain 1.1 million (m) sentences pairs of found bitext; compared to the datasizes (300k) in Table 3, so it may be feasible to explore SMT/NMT development for Zulu-English.[12] We note that found bitext is also available through some U.S. government research programs, either as training or test sets (e.g., LORELEI includes Arabic, Hausa, Yoruba, Amharic, Somali, Swahili, Akan, Zulu, Oromo, Kinyarwanda, and Tigrinya).

The table shows that the low-resource condition is quite complex for many of these African languages. Some languages have potentially exploitable monolingual resources, while others have existing found bitext. Further, some languages have apparently no resources whatsoever, so dataset creation by human translators will probably be a necessary first step.

6. Related Work

Low-Resource NMT While NMT models tend to be data-hungry, there is a growing body of research on improving NMT for low-resource conditions. One algorithmic method that has shown promise in moderate- or high-resource settings is *backtranslation*, using a baseline model and monolingual target language data to create "noisy" parallel data useful for training. For low-resource conditions, additional considerations are necessary to guarantee the quality of synthetic data (Wang et al., 2019a; Fadaee et al., 2017). Multilingual transfer is another approach to bootstrap MT in low-resource languages. One can combine multiple bitexts to train a single multilingual neural model (Arivazhagan et al., 2019; Johnson et al., 2017; Gu et

[11] Source: `https://meta.wikimedia.org/wiki/List_of_Wikipedias`, August 2019.

[12] All bitext reported by OPUS are available at `http://opus.nlpl.eu`. Note that for the majority of Niger-Congo languages in Table 4, the presence of bitext is due to a resource called JW300 (Agić and Vulić, 2019) released on August 28, 2019. This corpus contains magazine translations from `jw.org` for many low-resource languages; for future work, it will be promising to include this in our analysis of found bitext in Table 3.

al., 2018; Wang et al., 2019b) or design stage-wise transfer learning mechanisms (Zoph et al., 2016; Dabre et al., 2019; Kocmi and Bojar, 2018). MT performance may also be improved by using domain adaptation techniques (Thompson et al., 2019a) to deal with the problem of training on heterogeneous resource types and generalizing to new domains. Recent interest in unsupervised machine translation (c.f. (Artetxe et al., 2019)) promises to reduce the requirement for bitext, training only on monolingual data. Finally, general modeling improvements in NMT architectures can also help (Zaremoodi and Haffari, 2019; Nguyen and Chiang, 2018; Neubig and Hu, 2018). For example, results in other settings suggest that paraphrasing the English side of a bitext is a promising approach when translating into English in low-resource settings (Hu et al., 2019).

MT for African Languages We focused on Somali and Swahili in this paper. Other African languages have been explored in the context of both SMT and NMT. Hausa, a Chadic (Afroasiatic) language spoken mainly in Nigeria and Niger, was investigated in (Nguyen and Chiang, 2018; Zoph et al., 2016; Beloucif et al., 2016); (Murray et al., 2019) additionally perform experiments on Tigrinya, a Semitic (Afroasiatic) language spoken mainly in Eritrea and Ethiopia. The SMT work by (Tsvetkov and Dyer, 2015) focuses on out-of-vocabulary wods, with experiments in Swahili (but a different dataset from ours). Finally, there is a growing body of results in speech translation (Anastasopoulos and Chiang, 2018; Bansal et al., 2019; Inaguma et al., 2019), utilizing the Mboshi-French dataset of (Godard et al., 2018).

7. Conclusions and Future Directions

We performed an empirical comparison of SMT and NMT in two low-resource settings: Somali-to-English and Swahili-to-English. Our goal is to benchmark standard models and establish best practices. The two main findings are that (1) NMT can be made competitive with SMT in low-resource conditions, but only if sufficient hyperparameter tuning is performed; (2) NMT has the potential to benefit more than SMT from additional data such as mined bitext. Our NMT models and training recipes are publicly available.[13]

There are a number of promising directions for improving capabilities in African language translation:

Algorithm Improvement Synthetic data generation, multilingual transfer, and any other the methods described in Section 6. are prime candidates for improving low-resource MT in general. For African language translation, emphasis should be given to methods that are suitable for extreme low-resource setups. Also, the conditions are quite diverse as shown in Table 4, so we expect a multitude of methods being useful in practice.

Data Collection Languages that have little Web presence may prove particularly challenging, but our preliminary results suggest that even a few hundred thousand example translations can make a big difference with state of the art neural architectures. For this reason, we believe it is worth continuing to push the frontier of discovering and curating exploitable bitext for low-resource languages.

[13] https://github.com/kevinduh/sockeye-recipes/tree/master/egs/iarpa-material

Acknowledgments

We thank Dr. Carl Rubino and his team at the IARPA MATERIAL for providing this data. We also thank Philipp Koehn for help with the Paracrawl data.

Appendix: Example Translations

Example outputs of our SMT and NMT systems (under the baseline + dictionary + found-bitext + paracrawl condition) are shown here for the Text and Transcripts test sets. We also provide the input foreign sentence and English reference translation. We report the first 3 segments of each test set.

Somali-English Text Testset

1. **Input:** sannadihii 1914-kii ilaa 1918-kii waxaa soconaayay dagaalkii weynihii kowaad ee adduunka , waxayna xoogaggii ingiriiska iyo faransiisku dagaal ba ' an ku la jireen

   Reference: in the years 1914 to 1918 , the first world war was in progress , and the english and french forces were in fierce battle with

   NMT: in the year 1914 and 1918 , the first world war was going on , and the british forces and france were in a severe war .

   SMT: over the years until it is the first big 1914-kii 1918-kii which is going on in the world , and the united kingdom and france xoogaggii a severe with them

2. **Input:** khilaafaddii islaamka ee cusmaaniyiinta .

   Reference: the islamic caliphate of the ottomans .

   NMT: the islamic caliphate of the uzmaan .

   SMT: khilaafaddii of islamic ottoman empire .

3. **Input:** haddaba , xoogaggaas oo adeegsanaayay sarkaal ingiriisa oo la oran jiray lawrence of arabia , waxay bilaabeen dhagar ay isga horkeenayeen dawladdii islaamka iyo dadyowgii carbeed .

   Reference: so , the forces used by the english officer named lawrence of arabia began a plot to cause conflict between the islamic government and the arab people .

   NMT: therefore , the forces used a british official , called lawrence of arabia , have started a plan to confront the islamic government and the arabian people .

   SMT: and now , and a ingiriisa xoogaggaas adeegsanaayay and say that was lawrence of arabia , began , but they are horkeenayeen islam and the

arab peoples .

Somali-English Transcript Testset

1. **Input:** wafdi isku dhaaf ahoo kala socday xafiiska xoolaha beeraha iyo hormarinta reer miiga iyo maammulka magaalada dhagaxbuur ayaa kormeer indha indhayn ah ku soo maray musharicaha horumarineed ee ka socoda magaalada dhagaxbuur .

   Reference: a joint delegation from the office of livestock agriculture and rural development and the degehabur administration have supervised development projects that are going on in degehabur .

   NMT: a delegation from the bureau of livestock and meteorological development and the administration of dagabur has inspected the development candidate in dagahbur .

   SMT: a delegation from the office of the animals for the same drawing on agriculture and the development of the administration centre of the city , a blind miiga and dhagahbuur oversaw its through musharicaha development in that city ahmed mohamed hassan .

2. **Input:** faalo warkaas la xidhiidho waxaa i noo eegayaa wariye mohamed ciise .

   Reference: reporter mohamed isse will look at an analysis related to the news .

   NMT: the information related to that report is looked at us by reporter mohamed isse .

   SMT: comment that relates to me , it is for us at the journalist mohamed jesus .

3. **Input:** tababbarkan oo ay iska kaashadeen xafiiska waxbarashada heer deegaan iyo imminka aasaasidda lixaad ee xisbiga democratic-ga shacbiga soomaalida ethiopia .

   Reference: the training was jointly conducted by the district education office and the current sixth administration of the democratic ethiopian party of the somali people .

   NMT: the training , in collaboration with the office of education at a local level and now the sixth establishment of the democratic party of the somali people in ethiopia .

   SMT: tababbarkan to joint the office of the education system and now the establishment of the sixth level of the somali people democratic-ga ethiopia .

Swahili-English Text Testset

1. **Input:** jarida la wanawake : si vyema kufurahia wengine wanapopata changamoto

   Reference: a ladies magazine : it is not good to be happy when others get into challenges .

   NMT: women magazine : it 's not good to enjoy others when they get challenges

   SMT: the journal of women , " it is not good to others when they get the challenges

2. **Input:** imeandikwa na theopista nsanzugwanko

   Reference: it is written by theopista nsanzugwanko .

   NMT: written by theopista nsanzugwanko

   SMT: written by theopista nsanzugwanko

3. **Input:** imechapishwa : 25 septemba 2016

   Reference: published : 25 september 2016 .

   NMT: published : 25 september 2016

   SMT: published : 25 september 2016

Swahili-English Transcript Testset

1. **Input:** ahh sio hivyo hatujapotea kwa ubaya .

   Reference:ahh not that way we have n't been missing with badness .

   NMT: ahh is not so missing .

   SMT: ahh not so hatujapotea for evil .

2. **Input:** aai ni kwa ubaya .

   Reference: really it is with badness .

   NMT: indeed , he is evil .

   SMT: aai is for evil .

3. **Input:** < sta > kazi - kazi ndiyo imekuwa mingi < hes > sa labda tutafute nafasi tuje tuwatembelee .

   Reference: work - it is work that has become too much maybe if we get a chance to come visit you .

   NMT: <unk> sta > work - the work has been the many <unk> hes > sa maybe we will look for the opportunity to visit them .

   SMT: < sta > work - the work is has been many < hes > sa may have to find a position come tuwatembelee .

8. Bibliographical References

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

Anastasopoulos, A. and Chiang, D. (2018). Leveraging translations for speech transcription in low-resource settings. In *INTERSPEECH*.

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. (2019). An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203, Florence, Italy, July. Association for Computational Linguistics.

Bansal, S., Kamper, H., Livescu, K., Lopez, A., and Goldwater, S. (2019). Pre-training on high-resource speech recognition improves low-resource speech-to-text 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 58–68, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Beloucif, M., Saers, M., and Wu, D. (2016). Improving word alignment for low resource languages using English monolingual SRL. In *Proceedings of the Sixth Workshop on Hybrid Approaches to Translation (HyTra6)*, pages 51–60, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Christodouloupoulos, C. and Steedman, M. (2015). A massively parallel corpus: the bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395, Jun.

Dabre, R., Fujita, A., and Chu, C. (2019). Exploiting multilingualism through multistage fine-tuning for low-resource neural machine translation. 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 1410–1416, Hong Kong, China, November. Association for Computational Linguistics.

Fadaee, M., Bisazza, A., and Monz, C. (2017). Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada, July. Association for Computational Linguistics.

Godard, P., Adda, G., Adda-Decker, M., Benjumea, J., Besacier, L., Cooper-Leavitt, J., Kouarata, G.-N., Lamel, L., Maynard, H., Mueller, M., Rialland, A., Stueker, S., Yvon, F., and Zanon-Boito, M. (2018). A very low resource language speech corpus for computational language documentation experiments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

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.

Hieber, F., Domhan, T., Denkowski, M., Vilar, D., Sokolov, A., Clifton, A., and Post, M. (2017). Sockeye: A Toolkit for Neural Machine Translation. *arXiv preprint arXiv:1712.05690*, December.

Hu, J. E., Khayrallah, H., Culkin, R., Xia, P., Chen, T., Post, M., and Van Durme, B. (2019). Improved lexically constrained decoding for translation and monolingual rewriting. 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 839–850, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Inaguma, H., Duh, K., Kawahara, T., and Watanabe, S. (2019). Multilingual end-to-end speech translation. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

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.

Kocmi, T. and Bojar, O. (2018). Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252, Brussels, Belgium, 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., Guzmán, F., Chaudhary, V., and Pino, J. (2019). Findings of the WMT 2019 shared task on parallel corpus filtering for low-resource conditions. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 54–72, Florence, Italy, August. Association for Computational Linguistics.

Murray, K., Kinnison, J., Nguyen, T. Q., Scheirer, W., and Chiang, D. (2019). Auto-sizing the transformer network: Improving speed, efficiency, and performance for low-resource machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 231–240, Hong Kong, November. Association for Computational Linguistics.

Neubig, G. and Hu, J. (2018). Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels, Belgium, October-November. Association for Computational Linguistics.

Nguyen, T. and Chiang, D. (2018). Improving lexical choice in neural machine translation. 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 334–343, New Orleans, Louisiana, June. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Post, M., Ganitkevitch, J., Orland, L., Weese, J., Cao, Y., and Callison-Burch, C. (2013). Joshua 5.0: Sparser, better, faster, server. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 206–212, Sofia, Bulgaria, August. Association for Computational Linguistics.

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

Ramesh, S. H. and Sankaranarayanan, K. P. (2018). Neural machine translation for low resource languages using bilingual lexicon induced from comparable corpora. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 112–119, New Orleans, Louisiana, USA, June. Association for Computational Linguistics.

Rubino, C. (2018). Setting up a machine translation program for IARPA (Keynote speech at AMTA2018).

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*, Berlin, Germany, August. Association for Computational Linguistics.

Thompson, B., Gwinnup, J., Khayrallah, H., Duh, K., and Koehn, P. (2019a). Overcoming catastrophic forgetting during domain adaptation of 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 2062–2068, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Thompson, B., Knowles, R., Zhang, X., Khayrallah, H., Duh, K., and Koehn, P. (2019b). HABLex: Human annotated bilingual lexicons for experiments in machine translation. 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 1382–1387, Hong Kong, China, November. Association for Computational Linguistics.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Tsvetkov, Y. and Dyer, C. (2015). Lexicon stratification for translating out-of-vocabulary words. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 125–131, Beijing, China, July. Association for Computational Linguistics.

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 I. Guyon, et al., editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Wang, S., Liu, Y., Wang, C., Luan, H., and Sun, M. (2019a). Improving back-translation with uncertainty-based confidence estimation. 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 791–802, Hong Kong, China, November. Association for Computational Linguistics.

Wang, X., Pham, H., Arthur, P., and Neubig, G. (2019b). Multilingual neural machine translation with soft decoupled encoding. In *International Conference on Learning Representations (ICLR)*, New Orleans, LA, USA, May.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Łukasz Kaiser, Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation.

Zaremoodi, P. and Haffari, G. (2019). Adaptively scheduled multitask learning: The case of low-resource neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 177–186, Hong Kong, 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.