Is Machine Translation Ready for Pandemic Response? Evaluation of current MT systems on COVID-19 data

Libbey Brown
University of Washington
egollhof@uw.edu

Vipasha Bansal
University of Washington
vipashab@uw.edu

1 Introduction

Are currently available machine translation (MT) systems ready for pandemic response? This paper aims to answer this question by evaluating the performance of two existing systems against the TICO-19 dataset, a corpus of content related to the COVID-19 pandemic translated into 38 languages. Two MT systems (Google and Microsoft) were used for translation, both from English to the target languages and from the target languages to English. Three different automatic evaluation metrics were used and analyzed (BLEU, BERTscore, and Comet), and a limited subset of translations were scored using human evaluation. In addition, we provide an analysis based on language status and region.

We found that MT systems provide good coverage translating from English for higher-resource (referred to as pivot) languages, and that MT systems perform better when translating into English. There are limitations inherent in all three automatic scoring systems, and ultimately BLEU is the most appropriate metric to use when evaluating translations that include lower-resource languages.

We also found that while there are clear patterns based on the political status of a language and the region it is spoken in, there is enough variation that a country-specific analysis is required to determine MT readiness. We found that our level of readiness varies widely, even amongst countries located in the same region.

2 Background

2.1 Crisis Informatics and Machine Translation

In a crisis scenario, access to information is vital, both for individuals affected by the crisis and for those attempting to provide aid. People use technology such as social media and SMS messages in crisis situations to request help, provide information, or seek news about the unfolding crisis (Lewis et al., 2011; Hagar, 2010). There are many different groups involved in any given crisis, including aid organizations, government and non-government agencies, citizens, and victims – all of which increase the amount of information produced (Hagar, 2014). It is important that this information is collected, organized, shared, and managed appropriately, and that relevant information can be disseminated to affected communities (Hagar, 2010). This is the focus of crisis informatics, which is defined as "the interconnectedness of people, organizations, information and technology during crises" (Hagar, 2010).

Unfortunately, those who are the most vulnerable to crises are often speakers of low-resource languages (Lewis et al., 2011). Aid organizations may not have providers fluent in the languages spoken by the affected individuals, and for many under-resourced languages, translators may be difficult to find (Anastasopoulos et al., 2020). Additionally, there is often too much information to be translated manually, particularly as human translation can be slow (Lewis et al., 2011). Although relevant news or public health information may be provided, it will not be accessible to individuals who do not speak the language used to create those resources. For example, during the COVID-19 pandemic, official communications from the World Health Organization were only available in majority languages (Anastasopoulos et al., 2020). An inability to communicate can hinder the delivery of aid.

Machine translation (MT) can be used to speed up the translation process and helps to bridge these gaps. It can allow more people to help with relief efforts and to effectively triage relevant information (Lewis et al., 2011). Therefore, MT is essential to facilitate communication where help is needed

most. Some examples where such technology was effectively deployed are the earthquakes in Haiti, Pakistan, and Ecuador (Lewis et al., 2011; Anastasopoulos et al., 2020).

2.2 The TICO-19 Dataset

COVID-19 was the worst pandemic in over a century, and had a global impact. In order to slow the spread of the virus and to reduce hospitalizations and death, it was vital to communicate information to vulnerable populations. The goal of TICO-19 was to translate COVID related content into a wider range of languages, so that technologies such as medical MT and other NLP tools can be developed (Anastasopoulos et al., 2020).

The dataset includes pandemic-related sentences translated from English into 38 different languages. This includes 9 pivot languages, which act as lingua francas around the world, 21 high-priority languages based on the translation requests received by Translators Without Borders (TWB), and 8 additional languages which are spoken by millions of people across Asia (Anastasopoulos et al., 2020). Many of the languages included in the dataset are extremely low-resource, and some have no currently existing MT systems.

The sentences to be translated were collected from various sources, including Wikipedia articles, PubMed, and the CMU English-Haitian Creole dataset (Anastasopoulos et al., 2020). This means that diverse domains, such as news, travel advisories, medical conversations, and academic content are covered. Both test and development datasets are available and identical sentences were used for each language. This allows for the development of MT between any language pair in the corpus.

2.3 Geographical and Political Background

The dominant regions represented by TICO-19 are South and Southeast Asia and Africa. Most (if not all) countries in these areas are extremely multilingual. For example, over 500 languages are spoken in Nigeria (Translators Without Borders, 2023) and over 1600 languages are spoken in India (Government Of India, 2021). Most of these countries were also victims of Western colonialism, and have retained the colonial languages in addition to indigenous ones. For example, in the late 1800s, Gabon was colonized by France, and gained independence in 1960 (Central Intelligence Agency, 2021). Today, French is an official language, spoken by at least 80% of the population (Chepkemoi, 2017). Similarly, Kenya was under British rule from the 1800s-1960s (Central Intelligence Agency, 2021), and English is now an official language spoken by over 4 million people (Eberhard et al., 2023). The Spanish Empire conquered much of Latin America, starting with Mexico in the 1500s, and Spanish is now the dominant language in a number of countries in the region (Magdoff et al., 2023).

Many countries have varying language policies designed to facilitate national communication within this multilingual setting. Some languages are official at a national level (usually the colonial language and one or more dominant indigenous languages), while other languages are official at a state or regional level (Laitin, 1989). Some languages are used as the medium of education, while others are taught in schools as a second language (Eberhard et al., 2023). Many languages, even if widely spoken, have no 'official' status at all, and are therefore not used in government or educational settings.

Languages differ in the political power or prestige based on their official status, which sometimes, but not always, corresponds to the number of speakers. While people often speak at least one of the official languages of a country, we cannot rely on this – in many countries there are large groups which do not speak any of the dominant languages (Anastasopoulos et al., 2020).

3 System Overview

Our system consisted of the following:

1. Machine Translation
2. Chinese Text Segmentation
3. Evaluation

Two MT systems were used to translate from English to the target language, and also to translate from the target language to English.

Following translation, text segmentation was used to segment the Chinese data.

Three automatic scoring methods were used to evaluate the performance of the MT systems: BLEU, BERTscore, and COMET. Additionaly, a smaller subset of translations were scored via human evaluation.

Further details are provided in the Methods section of this report.

4 Data

The TICO-19 test data consists of 2100 English sentences translated into 38 languages (Anastasopoulos et al., 2020). See Table 8 in Appendix for full list of languages. The data used in this study was obtained from the TICO-19 website[1].

Files for three languages were found to have significant errors in formatting (multiple lines were misaligned in the tables), so they were omitted from our study. Those languages were Kinyarwanda (rw), Kurdish Kurmanji (ku), and Kanuri (kr). While the TICO-19 paper mentions the inclusion of data for Congolese Swahili (Anastasopoulos et al., 2020), no Congolese Swahili files were in the data downloaded from the TICO-19 website, so it could not be included in our study.

5 Methods

5.1 Machine Translation

We chose to focus our investigation on two MT systems, Google Cloud's Translation AI[2] and Microsoft's Azure Cognitive Services Translator[3]. These MT systems will be referred to as Google and Microsoft, respectively, for the rest of the paper.

All sentences in the TICO-19 dataset were translated from English into the target language by the first author. Note that neither MT system differentiates between the two dialects of Tigrinya covered in the TICO-19 data, so the translations from English into the generic Tigrinya were used for both Ethiopian and Eritrean Tigrinya. The second author translated the reference sentences from the target language to English.

Not all languages in the TICO-19 dataset are covered by the MT systems used in this study. Please reference Table 8 in Appendix to see which languages were covered by which MT systems.

Some languages were translated one direction (English to target) but were not translated the other direction (into English) due to variations in the way the respective authors designed their code. The author translating from English iterated through all files in the TICO-19 dataset and sent sentences from each to the MT system, at which point the data was translated if the language was covered by that MT. The author translating to English checked the documentation for Microsoft and Google and only translated languages which were listed as being covered. As it appears that the documentation for both MT systems is out of date, this resulted in some languages lacking translation into English. Languages lacking English translations using Microsoft are Hausa and Lingala. Languages lacking English translations using Google are Kurdish Sorani, Luganda, and Tagalog.

5.2 Chinese Text Segmentation

Neither the TICO-19 Chinese reference sentences nor the Chinese outputs of the MT systems utilized character segmentation. The Python package Jieba[4] was used to segment the Chinese text, as segmentation is necessary for evaluation using BLEU.

5.3 MT Evaluation

5.3.1 BLEU

BLEU (Bilingual Evaluation Understudy) is a method of automatic machine translation that was introduced in 2002 (Papineni et al., 2002). It compares a reference (gold standard translation) sentence and a candidate translated sentence, and returns a score between 0 and 100 indicating how similar the two sentences are. The BLEU score is calculated based on two factors: n-gram overlap and a brevity penalty.

The n-gram overlap counts how many n-grams of length 1-4 in the candidate match those in the reference sentence; the n-gram counts are limited to the maximum count that occur in the reference. A high-scoring candidate will use the same words, in the same order, as the reference sentence. If a candidate sentence has words that do not appear in the reference, or has words that appear more frequently, it will have a lower score. The brevity penalty reduces the score for candidates that are too short relative to the reference, preventing a one-word candidate sentence from scoring well compared to a much longer reference (GoogleCloud, 2023).

BLEU scores were calculated for both translation directions (from English to the target language, and from the target language to English) for each supported MT system using the Python package sacrebleu[5] using the default parameters. Guidelines for interpretation of the BLEU scores were taken from Google's MT API documentation (GoogleCloud, 2023); see Table 1.

[1] https://tico-19.github.io/

[2] https://cloud.google.com/translate

[3] https://learn.microsoft.com/en-us/azure/cognitive-services/Translator/

[4] https://github.com/fxsjy/jieba

[5] https://github.com/mjpost/sacrebleu

BLEU Score	Interpretation
< 10	Almost useless
10-19	Hard to get the gist
20-29	Clear gist, grammatical errors
30-39	Understandable to good
40-49	High quality
50-59	Very high quality, fluent
> 60	Better than human

Table 1: Guidelines for Interpretation of BLEU scores

BLEU is both efficient and language-independent, making it a widely used method of MT evaluation (Papineni et al., 2002; GoogleCloud, 2023). However, there are some legitimate concerns regarding its effectiveness. The method of counting n-gram overlap fails to match paraphrases, and n-grams do not capture dependencies and so may not be sensitive to semantically-important changes in word order (Zhang* et al., 2020). Additionally, more modern neural MT translation approaches produce fluent outputs that are not simply lexical transfers of information between languages (Rei et al., 2020). These concerns have led to the development of newer scoring methods, such as BERTscore and COMET.

5.3.2 BERTscore

BERTscore is a method designed to evaluate language generation that utilizes information from pretrained BERT contextual embeddings (Zhang* et al., 2020). It calculates similarity scores for each token in both the candidate sentence and the reference sentence - each token in the reference is matched to a token in the candidate, and each token in the candidate is matched to a token in the reference. A greedy matching algorithm maximizes the matching similarity score, in which each token is matched to the most similar token in the other sentence (Zhang* et al., 2020).

The BERTscore method returns a score in the range of cosine similarity, between -1 and 1. The authors note that "in practice we observe scores in a more limited range potentially because of the learned geometry of contextual embeddings. While this characteristic does not impact BERTscore's capability to rank text generation systems, it makes the actual score less readable" (Zhang* et al., 2020).

To address this issue, the authors developed a method of normalizing scores so that they fall in the range of 0 to 1. One million candidate-reference pairs were made by grouping random sentences that have low lexical and semantic overlapping. A baseline was found by averaging the BERTscore on these (unrelated) sentence pairs, and the baseline was used to rescale the BERTscore to return a value between 0 (unrelated sentences) and 1 (identical sentences) (Zhang* et al., 2020).

The authors note that BERTscore addresses the pitfalls of the BLEU score by leveraging contextual embeddings, and found that their scores were better correlated with human evaluation. However, normalizing scores requires basline rescaling, and baselines are only available for 11 languages, preventing us from rescaling for the vast majority of the TICO-19 languages.

BERTscores were calculated for both translation directions (from English to the target language, and from the target language to English) for each supported MT system using the Python package BERTscore [6] with the default parameters. Scores were not rescaled when translating from English to the target language in order to more easily compare the scores across all languages. Scores for the translation into English were calculated both with and without baseline rescaling. The authors of BERTscore did not provide any guidelines mapping score number to translation quality, so we chose to use a score of 0.7 (70% similarity) as our cutoff for classifying a translation as sufficient for crisis response.

5.3.3 COMET

The COMET scoring method is similar to BERTscore, in that it leverages pretrained contextual embeddings to generate a similarity distance. COMET differs from BERTscore in that it utilizes multilingual modeling in order to leverage information from both the source sentence and the reference translation sentence to evaluate the quality of the candidate (Rei et al., 2020). As with BERTscore, the COMET model utilizes cosine similarity to calculate the distance between the candidate and the source and reference sentences. Unlike BERTscore, COMET automatically employs normalization for all languages to return scores within the range of 0 to 1.

One drawback of COMET is that it is only valid for languages which were used to train the pre-

[6] https://github.com/Tiiiger/bert_score

trained multilingual model (XLM-RoBERTa) (Rei et al., 2020). Because of this, translations for only 18 of the TICO-19 languages were able to have COMET scores evaluated.

COMET scores were calculated using the Python package COMET[7] with the default parameters. The authors of COMET did not provide any guidelines mapping score number to translation quality, so we chose to use the same criteria we set for BERTscore: 0.7 (70% similarity) indicates a translation as sufficient for crisis response.

5.3.4 Human Evaluation

A limited subset of translations were scored by the authors in order to incorporate human evaluation. The scoring methods for human evaluation were adopted following those laid out in (Callison-Burch et al., 2007); scores between 1 and 5 were given for adequacy (how much of the meaning from the reference sentence was included in the candidate sentence) and fluency (see Table 2).

Score	Adequacy	Fluency
1	None	Incomprehensible
2	Little	Disfluent
3	Much	Non-native
4	Most	Good
5	All	Flawless

Table 2: Guidelines for human evaluation scoring. Scores for adequacy indicate how much of the meaning expressed in the reference is expressed in the candidate sentence.

Each author scored 50 randomly selected English sentences translated by Microsoft and 50 randomly selected English sentences translated by Google from the languages French, Marathi, and Somali, so that a total of 100 translations were evaluated for each MT system for each of the selected languages. These languages were chosen based on a variety of criteria: they are each covered by both MT systems, they are each covered by all scoring systems, and they represent a variety of status positions and geographical locations.

6 Results

[7]https://github.com/Unbabel/COMET

6.1 BLEU scores

BLEU scores were highly variable across the TICO-19 languages. For all languages, however, they were better when translated into English rather than translated from English. See Table 9 in Appendix for all scores.

BLEU scores indicate adequate performance of both MT systems for all pivot languages translating both into the target language and into English, with scores of 30 and above. Scores for some non-pivot languages (particularly when translating from English), however, fall below 30, indicating the MT systems are too poor to be used for those particular TICO-19 languages. Table 3 categorizes the translation quality for each TICO-19 translation based on the best BLEU score for each translation direction.

6.2 BERTscores

BERTscores were calculated without baseline rescaling for all translations in both directions, as baseline rescaling is not available for most of the TICO-19 languages. BERTscores were calculated for the X to English translations a second time with baseline rescaling. See Table 10 in Appendix for all BERTscores. As with BLEU, scores were better when translated into English compared to translations from English.

As baseline rescaling was not available, sentences could not be evaluated in terms of adequacy for translation from English to the TICO-19 languages. However, translations from the TICO-19 languages to English were able to be evaluated in this way. All pivot languages demonstrate translations of sufficient quality, but not all non-pivot languages had scores falling in adequate ranges. Table 4 categorizes the translation quality for each TICO-19 language based on the best BERTscore for translation into English. We chose to define scores of .7 or higher as providing adequate translation for crisis response.

6.3 COMET scores

COMET scores were unable to be evaluated for all TICO-19 translations, as not all languages supported by the MT systems are supported by the COMET system (Rei et al., 2020). See Table 11 in Appendix for all COMET scores. While the scores for BLEU and BERTscores were universally better for translations into English, COMET scores were better for translation the other direction (i.e., from

Best BLEU Score	Target Language (English to Target translations)
< 10 (Almost useless)	Amharic, Myanmar, Oromo, Tigrinya (ET and ER)
10-19 (Hard to get the gist)	Marathi, Hausa, Nepali, Dari, Somali, Zulu, Luganda, Kurdish Sorani
20-29 (Clear gist, grammatical errors)	Bengali, Pashto, Urdu, Tamil, Farsi
30-39 (Understandable to good)	**Arabic**, Lingala, **Swahili**, **Russian**
40-49 (High quality)	**Chinese, Hindi, French**
50-59 (Very high quality, fluent)	Malay, **Portuguese (BR), Indonesian, Spanish (LA)**
> 60 (Better than human)	–
Best BLEU Score	Source Language (Source to English translations)
< 10 (Almost useless)	–
10-19 (Hard to get the gist)	Somali
20-29 (Clear gist, grammatical errors)	Tigrinya (ET and ER), Myanmar, Amharic, Khmer, Lingala
30-39 (Understandable to good)	Pashto, Dari, Tamil, Urdu, **Russian**, Kurdish Sorani, Hausa, Oromo
40-49 (High quality)	**Chinese**, Marathi, Farsi, Zulu, **Swahili**, Bengali, **Arabic, French**
50-59 (Very high quality, fluent)	Nepali, Malay, **Portuguese (BR), Indonesian, Spanish (LA), Hindi**
> 60 (Better than human)	–

Table 3: Best BLEU scores for TICO-19 languages. **Pivot** languages in bold

Best BERTScore	Source Language (Source to English translations)
0 - .39	Somali
.4 - .49	–
.5 - .59	Tigrinya (ET and ER)
.6 - .69	Amharic, Kurdish Sorani, Hausa, Khmer, Lingala, Myanmar, Oromo, Dari, Pashto, Tamil
	Scores greater than 0.7 are treated as adequate for crisis response
.7 - .79	**Arabic**, Bengali, Farsi, **French**, Marathi, **Russian, Swahili**, Urdi, **Chinese**, Zulu
.8 - .89	**Spanish (LA), Hindi, Indonesian**, Malay, Nepali, **Portuguese (BR)**
.9 - 1	–

Table 4: Best BERTscores for TICO-19 languages. **Pivot** languages in bold.

English to the target) for some languages.

COMET scores (for the languages covered by COMET) were indicative of higher performance than BLEU and BERT and showed excellent translations for all pivot languages for both translation directions. Table 5 categorizes the translation quality for each TICO-19 language based on the best COMET score.

6.4 Human Evaluation

Human evaluation was only done for three languages (French, Marathi, and Somali) and was evaluated for English translations only. The reported score is the average of the scores returned by the two evaluators. See Table 6.

Human evaluation scores showed excellent coverage for both English and Marathi. While the scores returned for Somali were lower, they did indicate moderate levels of both adequacy and fluency.

It should be noted that the evaluators found several sentences in which there was essentially no overlap in content or vocabulary between the reference (English source string) and candidate (English translation generated by MT from the provided target sentence in the non-English language). Upon reviewing those sentences in the original TICO-19 tables for the languages, it appeared that there may have been some lines in which the source string and the target string were not correctly aligned. Such sentences were found in both the Somali and the

Best COMET	Target Language (English to target translations)
0 - .59	–
.6 - .69	Amharic, Somali
	Scores greater than 0.7 are treated as adequate for crisis response
.7 - .79	Marathi, Pashto
.8 - .89	**Arabic**, Bengali, **Spanish (LA)**, Farsi, **French, Hindi**, Nepali, **Portuguese (BR)**, Urdu, **Russian**, **Swahili**, **Chinese**
.9 - 1	**Indonesian**, Malay
Best COMET	Source Language (Source to English translations)
0 - .59	–
.6 - .69	Somali
	Scores greater than 0.7 are treated as adequate for crisis response
.7 - .79	–
.8 - .89	Amharic, **Arabic**, **Spanish (LA)**, Farsi, **French**, Marathi, Pashto, **Russian**, **Swahili**, Urdu, **Chinese**
.9 - 1	Bengali, **Hindi**, **Indonesian**, Malay, Nepali, **Portuguese (BR)**
Not covered	*Tigrinya (ET and ER), Kurdish Sorani, Hausa, Khmer, Lingala Myanmar, Oromo, Dari, Tamil, Zulu*

Table 5: Best COMET scores for TICO-19 languages. **Pivot** languages in bold

Source	MT	Adequacy	Fluency
French	Microsoft	4.64	4.64
	Google	4.71	4.57
Marathi	Microsoft	4.21	4.22
	Google	4.38	4.46
Somali	Microsoft	3.27	3.53
	Google	3.68	4.21

Table 6: Human evaluation scores.

French subsets; none were found in the subset of Marathi sentences. For any sentence which the evaluator believed may have been generated from misaligned data, scores were only given for fluency, and that sentence was not considered for evaluation of adequacy.

7 Discussion

7.1 Comparison of scores across scoring systems

The three automatic scoring systems analyzed for this paper followed similar trends - languages which were the highest-scoring for one scoring system were high-scoring under the other systems as well. In general, however, BLEU scores demonstrated the poorest translation quality and COMET scores demonstrated the highest translation quality (for those languages which are covered by COMET), while BERTscores largely fell in the middle.

The developers of the BERTscore and COMET systems both noted concerns with BLEU's methodology, specifically that *n*-gram overlap will give poor scores to good translations that are paraphrases (Zhang* et al., 2020; Rei et al., 2020). This may explain why BLEU scores were lower than the other methods, in that BLEU is under-estimating translation quality.

COMET scores for covered languages were all higher than rescaled BERTscores. It is possible that COMET is scoring higher because it considers the source sentence in addition to the reference and the candidate. During our human evaluation, we found multiple sentences which seemed to be generated from misaligned datasets - meaning, the source sentence in language A did not match the content of the gold reference sentence in language B, so the candidate translation scored very poorly relative to the reference sentence. As COMET utilizes both the source and the gold reference in determining candidate similarity, the scores were almost certainly improved in the case where the gold reference sentence was not a valid translation

of the source sentence (but the candidate did align well with the meaning of the source).

Our limited human evaluation scores aligned better with COMET scores than with BLEU or BERTscores - our evaluators rated the French translations as excellent, the Marthi as very good and the Somali translations as acceptable. See Table 7. It is important to note that our human evaluators did not evaluate adequacy for the potentially misaligned sentences which were found in the French and Somali subsets, while those sentences were included in all automatic scoring methods.

	French	Marathi	Somali
Best BLEU	45.25	42.30	18.94
Best BERTscore	0.71	0.75	0.39
Best COMET	0.84	0.89	0.68
Best adequacy	4.71	4.38	3.68
Best fluency	4.64	4.46	4.21

Table 7: Comparison of scoring methods

While there are some legitimate concerns that BLEU scores are not the best way to measure machine translation, our experience working with other scoring algorithms has shown that BLEU continues to be an extremely useful metric. BLEU was the only automatic scoring system that was able to be applied in the same way for all translations both into and out of English; BERTscore can be applied for all translations, but the scores are challenging to interpret without rescaling, and COMET is not covered for many of the TICO-19 languages. Additionally, BLEU is far more efficient - for instance, obtaining BLEU scores for all translations from English took less than five minutes, while obtaining BERTscores or COMET scores took more than two hours. Ultimately, we determined that BLEU, despite limitations inherent in relying on *n*-gram overlap, remains the best way to evaluate MT across a wide variety of languages.

7.2 Comparison of BLEU Scores by Geography and Language Status

The results were also analyzed based on political and geographical factors. In this part of our analysis, we utilized only BLEU scores, as this was the only metric that covered every language under question. In order for translations to or from a language to be considered usable, they needed to have a BLEU score of at least 30.

In terms of status, languages were assigned one of four options: Colonial Official, Native Official, Local Official, or No Official Status. Colonial and Native official languages are official on a national scale, Local Official languages are those that are official on a regional level, or that are used as a medium of instruction in schools. The TICO-19 dataset did not contain any languages without any official status. For languages that are spoken in multiple countries, their highest official status was measured. For example, Pashto is the second most widely spoken language in Pakistan but has no official status there (Translators Without Borders, 2023). However, it is an official language in Afghanistan (Central Intelligence Agency, 2021; Eberhard et al., 2023). Therefore, it has been analyzed as a National Official language.

In this paper, Arabic has been analyzed as a Colonial Official language. While it is indigenous to many of the countries it is spoken in, there are also many countries where it spread due to trade and religion. Although this is not the same context as the colonialism undertaken by European countries, it is still an external language that was brought to or imposed on many countries.

Finally, for each language, the highest BLEU score available (from either Microsoft or Google) was used, since this is the maximum capability we have for translation between that language and English. The specific API which produced the highest score in each case was not deemed relevant for this section of our research, though it was noted that in general Google had slightly better performance and coverage.

7.2.1 Comparison of BLEU Scores by Language Status

As seen in Figure 1, Colonial Official languages have the best performance, with high quality translations both into the target language and into English. Spanish and Portuguese did particularly well with BLEU scores over 50 in each direction, indicating almost fluent translations. Translations into Arabic scored a little lower, but still met the criteria of scoring at least 30 in order to be usable. However, given that Arabic is a non-European language, and also utilizes a different script, this is unsurprising. Translation from Arabic to English was regardless high quality with a BLEU score over 40.

Figure 3 shows languages with Local Official

status. Translations from each of these languages into English were all usable (with the exception of Nigerian Fulfulde, which was not supported by either API). Bengali stood out as particularly high quality, with a BLEU score of 49.5. However, no language achieved a usable BLEU score for translation from English into the target language. It is also important to note that all Local Official languages in the TICO-19 dataset were from either India or Nigeria.

The Native Official languages (Figure 2) are where we see the most variation. All supported languages, except Eritrean Tigrinya, Somali, and Lingala, had usable translations into English, of varying quality. Eight languages also had usable translations from English into the target language, but it is clear that a more fine-grained, region specific analysis is needed here.

Overall, language status does appear to correlate with MT performance, given the high performance of Colonial Official languages as compared with Local Official languages. The variability of Native Official languages, however, indicates that other factors must also be taken into account.

7.2.2 Comparison of BLEU Scores by Continent/Region

The three major regions represented by the dataset were the Middle East/South Asia, Southeast Asia, and Africa. Translation for Southeast Asian languages (Figure 4) performed well. All translations from the target language into English were usable, with Malay receiving a particularly high score. While sentences were not translated from Tagalog into English due to errors in the documentation of the APIs, we can predict that the system would have comparable performance, based on the results when translating from English into Tagalog. With the exception of Khmer and Burmese, all languages also performed well translating from English into the target language, with BLEU scores above 40. The translations for Chinese scored somewhat lower than expected, but were still very much usable. This low score could be due to the fact that an additional step for text segmentation needed to be undertaken before running BLEU, increasing the possibility of errors or mismatches.

For the Middle Eastern and South Asian languages (Figure 5), translation from the target language into English generally did much better than the other direction. All translations into English were usable; Dari had the lowest score at 31.23. Hindi and Nepali performed particularly well, with BLEU scores over 50. Translation into the target language was only sufficient in Hindi and Arabic, as all other languages achieved BLEU scores under 30 (though Farsi and Urdu are close to usability, with scores of 29 and 28 respectively).

African languages (Figure 6) are where we see the most variability. As with the South Asian/Middle Eastern languages, translation into English generally did better than translation from English. However, the translations from Lingala, Somali, and Eritrean Tigrinya into English were not usable. When translating from English into the target language, only Swahili and Lingala had high enough BLEU scores to be usable. Amharic, Oromo, and both dialects of Tigrinya performed particularly poorly.

Lingala was an unusual result, in that translations from English into Lingala performed better than translations from Lingala into English, going against the overall trend. However, it is important to note that Lingala was not translated into English using the Microsoft MT system due to errors in their API documentation. Analyzing the BLEU scores for translation from English into Lingala, it is clear that Microsoft's MT system has significantly better performance for this language (the BLEU scores are 32.07 and 15.08 for Microsoft and Google, respectively). It is likely that the BLEU scores for Lingala to English translations would be much higher if Microsoft's MT had been used.

In general, the MT systems did not perform as well with African languages as they did with languages from other continents. The only languages where translations from the language into English were not usable are all African languages. Additionally, three of the languages in the TICO-19 dataset were not supported by either API – Nigerian Fulfulde, Nuer, and Dinka. Again, all three of these languages are spoken in Africa. These findings have significant implications for the readiness (or lack thereof) of MT systems for a future pandemic response in African countries.

Overall, Southeast Asian languages performed the best, while African languages had the weakest and most variable performance. In both African and South Asian/Middle Eastern countries we can generally expect MT to enable only one way communication, from the target language into English. Although this analysis does shed light some light on the performance of existing MT systems in various

regions, there is still enough variation in the data that a country specific analysis would be necessary in order to truly determine pandemic readiness.

7.2.3 Country Specific Analyses

Given the vast number of countries that speak some combination of the languages present in the TICO-19 dataset, it was not possible to analyze readiness in every relevant country. Therefore, just a few examples have been provided below, while an analysis of additional countries will be left for future research.

India

The official languages of India are English and Hindi, and in addition, every state within the country has its own official language (classified as Local Official in this analysis) (Laitin, 1989). Despite its official status however, Hindi is mainly spoken in the North, while the South favors Tamil and other Dravidian languages. India is an extremely multilingual country, and maintains a 3 language education policy. This means that everyone learns some amount of Hindi, English, and the language of the state, in addition to other languages that may be used in the home (Laitin, 1989). However, despite this, most people do not have enough proficiency in each of these languages to effectively communicate, and so we cannot assume that English and Hindi alone will be enough to reach the entire population.

There were four Indian languages included in the TICO-19 dataset - Hindi, Marathi, Bengali, and Tamil. Every language performed well when translated into English, and Hindi also performed well when translated from English. However, no other language has usable translations from English into the target language. This has aid providers well situated to understand incoming communication, from both the North and the South, but information can only be disseminated to those who speak Hindi or English. While this still accounts for millions of people, it is not enough to reach a majority of the population. Therefore, current MT systems are not yet fully prepared for pandemic response in India, though they are on their way to readiness. Improving performance of Tamil translations as well as including additional state languages would help reach a much larger segment of the population.

Ethiopia

Ethiopia is one of the few countries in the region to have escaped colonial occupation, and as such did not have to incorporate a colonial language into daily usage or official policy (Shaban, 2023). It has five official languages at the national level - Amharic, Oromo, Somali, Tigrinya, and Afar (Shaban, 2023). 4 of these languages are represented in the TICO-19 dataset (only Afar is missing). At the time TICO-19 was created, Amharic was the only official language. The other 4 were recently elevated from having regional status to national status (Shaban, 2023). Given this, it seems reasonable to assume that these languages are widely spoken in the country.

Aside from Somali, all represented languages had usable translations into English. However, there were no usable translations from English into the target language. This means that aid providers are equipped to understand incoming communications in 3 out of 5 official, and widely spoken, languages. Improving Somali performance should be a priority, and improving translation into the target languages in general would also be beneficial. However, our current MT systems have made a reasonable start at readiness in Ethiopia.

Democratic Republic of Congo (DRC)

There are four national languages in the DRC in addition to French (which was introduced as a result of Belgian colonization) (Translators Without Borders, 2023); Lingala, Swahili, Tschiluba, and Kituba. Of these, Lingala and Swahili are the most widely spoken (22 million and 14 million respectively), and cover the majority of the population in the northwest and east of the country (Translators Without Borders, 2023). They are both represented in the TICO-19 dataset alongside French, which is spoken by 74% of the population (Target Research and Consulting, 2021).

As previously mentioned, existing MT systems perform well when translating French and Swahili both into and out of English. We also have the ability to produce usable translations from English into Lingala, with a strong likelihood of achieving the same when translating from Lingala into English if the Microsoft API were used. Therefore, we seem reasonably well prepared for a future pandemic response in the DRC, even without support for Tschiluba and Kituba (though adding these would of course be beneficial).

Kenya and Tanzania

The official languages of both countries are English and Swahili (Central Intelligence Agency, 2021). Based on interviews with citizens of each country, the majority of the population is able to speak and understand Swahili. Given the high per-

formance of Swahili translation both to and from English, in these countries existing MT systems are reasonably prepared for a future pandemic response.

South Sudan

In South Sudan, the official languages are English and Arabic (Central Intelligence Agency, 2021), but most indigenous languages are now also used as national languages (Omondi, 2017). Of these, Dinka and Nuer are included in the TICO-19 dataset. These are also the two most widely spoken indigenous languages in the country (Omondi, 2017). However, neither language was supported by either MT system, and it is unclear how many people in the country can actually speak English or Arabic. Therefore, support for, as well as adequate translations in at least some of the indigenous languages of the country are essential to any future pandemic response in South Sudan. Existing MT systems are not yet ready.

Uganda

The official languages of Uganda are Luganda, English, and Swahili (Government Of Uganda, 2023), and these are also the three most commonly spoken languages in the country (Ager, 2023; Eberhard et al., 2023). Although the Luganda translation score was too low to be usable, current MT systems enable communication in two out of three of these languages. Therefore, we are on the way to being ready to respond to a future pandemic in Uganda. An improvement in Luganda performance would be extremely beneficial.

Gabon

As mentioned earlier, at least 80% of the population of Gabon speaks French (Chepkemoi, 2017). Given the good quality of translations both into and from French, our existing systems leave us well prepared for another pandemic response in Gabon.

Somalia

The official languages of Somalia are Somali and Arabic (Translators Without Borders, 2023). However, the majority of the population speaks Somali, while Arabic is mainly reserved for religious purposes (Translators Without Borders, 2023). While the translations for Arabic were usable in either direction, the translations for Somali were not. In fact, Somali was one of the few languages where even translation from the target language into English was not usable. Given this, we are not prepared to handle a pandemic response in Somalia based on current MT capabilities.

French and Arabic Speaking Countries

There are a large number of countries where both French and Arabic are official languages, likely due to a combination of colonialism and religion. Some examples include Chad and Djibouti (Eberhard et al., 2023). Given the good performance of MT in either direction in both these languages, it is reasonable to assume that we would be prepared for pandemic response in many of these countries. However, given that there are many regions where the population does not speak the dominant language, country specific analyses would still be needed to confirm this. However, having working MT systems in this language pair is a strong start.

8 Limitations and Future Directions

While the TICO-19 dataset provides a comprehensive corpus of COVID-19 related content, it is very likely that machine translation for future pandemics will need to cover additional vocabulary (for instance, if the symptoms, treatments, or infection pathways differ from those of COVID-19). Our study provides a comprehensive look at the TICO-19 corpus, but this performance may not generalize to a future pandemic. Additionally, the TICO-19 dataset covers languages that were specifically chosen based on the COVID-19 pandemic; a future pandemic may affect different regions and we cannot predict performance for languages that were not included in our dataset.

Another significant limitation is our human evaluation, as our human evaluators were unable to score translations for more than three languages, and were unable to score translations in languages other than English. We feel that a more thorough human evaluation process would allow us to better compare human evaluation to the different automatic scoring metrics and give us a more complete picture of those systems.

It is also important to note our concerns regarding the misaligned files in the TICO-19 dataset; if any of the files contained errors this certainly will have an impact on the validity of our results.

Additionally, errors in the documentation of both APIs led to some languages not being translated into English for this study. It would be important to have this data to fully test the MT systems. While this error seems small, it then leads to the question of what other information may be missing in the documentation, which may become relevant in the

future if we are to actually deploy these systems in a crisis or pandemic.

Finally, more research is needed on the linguistic situation in Eastern Europe and Latin America. These regions were represented in the TICO-19 dataset by Russian, Latin American Spanish, and Brazilian Portuguese. However, since the majority of the languages in the dataset were from Asia and Africa, these were the regions our analysis focused on.

9 Conclusion

We found that currently available MT systems are providing good coverage for pivot languages both into and out of English, indicating that we are ready for pandemic response for speakers of those languages. Many non-pivot languages have good coverage only when translating into English, meaning that English-speaking aid providers will be able to understand needs and provide assistance but it may be challenging to provide information to local individuals affected by the pandemic. There were several other languages in the TICO-19 corpus which did not have good MT performance in either direction, and in those cases MT is not capable of responding to a pandemic for any speakers.

Ultimately, MT readiness for pandemic response varies depending on the country in question. For some countries, including The Democratic Republic of Congo, Kenya, Tanzania, and Gabon, our systems are ready. In other countries, such as India, Ethiopia, Uganda, and the many countries which use both French and Arabic as official or national languages, we have reasonable performance and are well on the way to readiness, with just a few improvements needed. Finally, in some countries, including Somalia and South Sudan, our systems are not currently capable of handling a pandemic response, due to poor performance and lack of support for essential languages.

References

Simon Ager. 2023. Swahili (kiswahili). In *Omniglot - Writing Systems and Languages of the World*. Last accessed 02 June 2023.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the translation initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.

Central Intelligence Agency. 2021. The world factbook. Last accessed 02 June 2023.

Joyce Chepkemoi. 2017. What languages are spoken in gabon? In *World Atlas*. Last accessed 02 June 2023.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2023. *Ethnologue: Languages of the World*, twenty-sixth edition. SIL International, Dallas Texas. Last accessed 02 June 2023.

GoogleCloud. 2023. Evaluate models. Accessed on May 11, 2023.

Government Of India. 2021. Language education. Last accessed 02 June 2023.

Government Of Uganda. 2023. Facts and figures. Last accessed 02 June 2023.

Christine Hagar. 2010. Crisis informatics: Introduction. In *Bulletin of the American Society for Information Science and Technology*, pages 6–10.

Christine Hagar. 2014. Crisis informatics. *Journal of Geography and Natural Disasters*.

David D. Laitin. 1989. Language policy and political strategy in india. *Policy Sciences*, 22(3/4):415–436.

William Lewis, Robert Munro, and Stephan Vogel. 2011. Crisis mt: Developing a cookbook for mt in crisis situations. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 501–511.

Harry Magdoff, Charles E. Nowell, and Richard A. Webster. 2023. Western colonialism. In *Encyclopedia Britannica*. Last accessed 02 June 2023.

Sharon Omondi. 2017. What languages are spoken in south sudan? In *World Atlas*. Last accessed 02 June 2023.

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

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Abdur Rahman Alfa Shaban. 2023. One to five: Ethiopia gets four new federal working languages. *Africa News*. Last accessed 02 June 2023.

Target Research and Consulting. 2021. Target survey: French, the most spoken language in drc, far ahead of lingala. Last accessed 02 June 2023.

Translators Without Borders. 2023. Language data by country. Last accessed 02 June 2023.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Language	Code	Pivot	Location	Microsoft	Google
Amharic	am	No	Africa	Covered	Covered
Arabic	ar	Yes	Global	Covered	Covered
Bengali	bn	No	Asia	Covered	Covered
Chinese (simplified)	zh	Yes	Asia	Covered	Covered
Dari	prs	No	Asia	Covered	No coverage
Dinka	din	No	Africa	No coverage	No coverage
Farsi	fa	No	Asia	Covered	Covered
French	fr	Yes	Global	Covered	Covered
Hausa	ha	No	Africa	Covered	No coverage
Hindi	hi	Yes	Asia	Covered	Covered
Indonesian	id	Yes	Asia	Covered	Covered
Kanuri	kn	No	Africa	No coverage	No coverage
Khmer	km	No	Asia	Covered	Covered
Kinyarwanda	kr	No	Africa	No coverage	No coverage
Kurdish Kurmanji	ku	No	Asia	Covered	Covered
Kurdish Sorani	ckb	No	Asia	No coverage	Covered
Lingala	ln	No	Africa	Covered	Covered
Luganda	lg	No	Africa	Covered	No coverage
Malay	ms	No	Asia	Covered	Covered
Marathi	mr	No	Asia	Covered	Covered
Myanmar	my	No	Asia	Covered	Covered
Nepali	ne	No	Asia	Covered	Covered
Nigerian Fulfulde	fuv	No	Africa	No coverage	No coverage
Nuer	nus	No	Africa	No coverage	No coverage
Oromo	om	No	Africa	No coverage	Covered
Pashto	ps	No	Asia	Covered	Covered
Portuguese - Brazilian	pt-BR	Yes	Global	Covered	Covered
Russian	ru	Yes	Europe	Covered	Covered
Somali	so	No	Africa	Covered	Covered
Spanish - Latin American	es-LA	Yes	Global	Covered	Covered
Swahili	sw	Yes	Africa	Covered	Covered
Tagalog	tl	No	Asia	No coverage	Covered
Tigrinya - Eritrean	ti-ER	No	Africa	Covered*	Covered *
Tigrinya - Ethiopian	ti-ET	No	Africa	Covered*	Covered*
Urdu	ur	No	Asia	Covered	Covered
Zulu	zu	No	Africa	Covered	Covered

Table 8: TICO-19 Languages.
Covered*: Google and Microsoft cover a single generic dialect of Tigrinya

Language	Code	Microsoft to X	Google to X	Microsoft to E	Google to E
Amharic	am	8.67	11.89	26.49	36.51
Arabic	ar	30.26	29.15	44.88	46.90
Bengali	bn	19.29	22.95	44.38	49.46
Chinese (simplified)	zh	42.38	44.67	36.46	44.67
Dari	prs	17.32	–	31.21	–
Dinka	din	–	–	–	–
Farsi	fa	27.38	29.20	38.97	39.65
French	fr	45.61	34.14	45.25	44.25
Hausa	ha	15.98	25.87	–	36.17
Hindi	hi	43.24	46.00	53.17	56.24
Indonesian	id	54.39	55.07	51.73	51.15
Kanuri	kn	–	–	–	–
Khmer	km	2.18	18.14	27.41	36.31
Kinyarwanda	kr	–	–	–	–
Kurdish Kurmanji	ku	–	–	–	–
Kurdish Sorani	ckb	–	10.67	–	–
Lingala	ln	32.07	15.08	–	26.04
Luganda	lg	–	19.71	–	–
Malay	ms	51.06	53.33	50.88	57.63
Marathi	mr	15.52	16.75	37.21	42.30
Myanmar	my	1.87	8.15	23.82	32.39
Nepali	ne	16.71	25.44	48.38	53.76
Nigerian Fulfulde	fuv	–	–	–	–
Nuer	nus	–	–	–	–
Oromo	om	–	9.06	–	31.36
Pashto	ps	21.12	26.09	27.57	40.75
Portuguese - Brazilian	pt-BR	52.42	52.27	54.81	55.19
Russian	ru	36.88	36.71	36.06	37.14
Somali	so	18.95	9.49	17.68	18.94
Spanish - Latin American	es-LA	56.06	56.86	54.75	54.14
Swahili	sw	33.10	33.88	41.04	45.56
Tagalog	tl	–	48.46	–	–
Tigrinya - Eritrean	ti-ER	3.83	6.40	18.07	24.83
Tigrinya - Ethiopian	ti-ET	3.39	6.13	21.60	30.33
Urdu	ur	23.50	28.14	35.93	40.61
Zulu	zu	19.76	19.60	36.46	47.58

Table 9: BLEU scores
– indicates no translations were available for scoring

Language	Code	Microsoft to X	Google to X	Microsoft to English	Google to English	Microsoft to English rescaled	Google to English rescaled
Amharic	am	0.95	0.94	0.93	0.94	0.59	0.69
Arabic	ar	0.88	0.87	0.96	0.96	0.77	0.77
Bengali	bn	0.88	0.87	0.96	0.96	0.77	0.79
Chinese (simplified)	zh	0.89	0.90	0.96	0.95	0.74	0.73
Dari	prs	0.83	–	0.94	–	0.66	–
Dinka	din	–	–	–	–	–	–
Farsi	fa	0.888	0.88	0.95	0.95	0.75	0.74
French	fr	0.88	0.85	0.95	0.95	0.71	0.69
Hausa	ha	0.78	0.83	–	0.94	–	0.67
Hindi	hi	0.89	0.90	0.97	0.97	0.82	0.82
Indonesian	id	0.93	0.93	0.97	0.97	0.83	0.81
Kanuri	kn	–	–	–	–	–	–
Khmer	km	0.88	0.94	0.93	0.95	0.61	0.70
Kinyarwanda	kr	–	–	–	–	–	–
Kurdish Kurmanji	ku	–	–	–	–	–	–
Kurdish Sorani	ckb	–	0.86	–	–	–	–
Lingala	ln	0.85	0.79	–	0.93	–	0.61
Luganda	lg	–	0.81	–	–	–	–
Malay	ms	0.80	0.82	0.97	0.97	0.82	0.83
Marathi	mr	0.85	0.84	0.95	0.95	0.73	0.75
Myanmar	my	0.79	0.83	0.93	0.95	0.59	0.68
Nepali	ne	0.85	0.88	0.96	0.97	0.79	0.81
Nigerian Fulfulde	fuv	–	–	–	–	–	–
Nuer	nus	–	–	–	–	–	–
Oromo	om	–	0.79	–	0.94	–	0.65
Pashto	ps	0.85	0.85	0.93	0.95	0.59	0.72
Portuguese - BR	pt-BR	0.94	0.94	0.97	0.97	0.84	0.82
Russian	ru	0.89	0.89	0.96	0.95	0.74	0.73
Somali	so	0.78	0.75	0.90	0.90	0.38	0.39
Spanish - LA	es-LA	0.94	0.93	0.97	0.96	0.84	0.82
Swahili	sw	0.87	0.87	0.95	0.96	0.73	0.76
Tagalog	tl	–	0.90	–	–	–	–
Tigrinya - Eritrean	ti-ER	0.94	0.94	0.92	0.93	0.52	0.61
Tigrinya - Ethiopian	ti-ET	0.94	0.95	0.93	0.94	0.57	0.67
Urdu	ur	0.85	0.86	0.95	0.96	0.71	0.73
Zulu	zu	0.84	0.84	0.95	0.96	0.72	0.77

Table 10: BERTscores
– indicates no translations were available for scoring
Columns M to X and G to X provide BERTscores for translations using Microsoft MT and Google MT, respectively, without baseline rescaling. Columns M to E and G to E provide BERTscores for translations using Microsoft MT and Google MT, respectively, without baseline rescaling. The final two columns list the BERTscores with baseline rescaling.

Language	Code	Microsoft to X	Google to X	Microsoft to E	Google to E
Amharic	am	0.84	0.84	0.82	0.87
Arabic	ar	0.85	0.85	0.87	0.88
Bengali	bn	0.84	0.86	0.90	0.90
Chinese (simplified)	zh	0.88	0.89	0.87	0.88
Dari	prs	No coverage	–	No coverage	–
Dinka	din	–	–	–	–
Farsi	fa	0.86	0.87	0.88	0.88
French	fr	0.81	0.81	0.84	0.83
Hausa	ha	No coverage	No coverage	–	No coverage
Hindi	hi	0.79	0.80	0.90	0.91
Indonesian	id	0.92	0.80	0.90	0.90
Kanuri	kn	–	–	–	–
Khmer	km	No coverage	No coverage	No coverage	No coverage
Kinyarwanda	kr	–	–	–	–
Kurdish Kurmanji	ku	–	–	–	–
Kurdish Sorani	ckb	–	No coverage	–	–
Lingala	ln	No coverage	No coverage	–	No coverage
Luganda	lg	–	No coverage	–	–
Malay	ms	0.90	0.90	0.89	0.90
Marathi	mr	0.72	0.72	0.87	0.89
Myanmar	my	No coverage	No coverage	No coverage	No coverage
Nepali	ne	No coverage	No coverage	No coverage	No coverage
Nigerian Fulfulde	fuv	–	–	–	–
Nuer	nus	–	–	–	–
Oromo	om	–	No coverage	–	No coverage
Pashto	ps	0.78	0.79	0.81	0.86
Portuguese - Brazilian	pt-BR	0.90	0.90	0.90	0.90
Russian	ru	0.88	0.88	0.86	0.86
Somali	so	0.72	0.70	0.66	0.68
Spanish - Latin American	es-LA	0.89	0.89	0.90	0.90
Swahili	sw	0.81	0.82	0.84	0.86
Tagalog	tl	–	No coverage	–	–
Tigrinya - Eritrean	ti-ER	No coverage	No coverage	No coverage	No coverage
Tigrinya - Ethiopian	ti-ET	No coverage	No coverage	No coverage	No coverage
Urdu	ur	0.81	0.82	0.86	0.88
Zulu	zu	No coverage	No coverage	No coverage	No coverage

Table 11: COMET scores
– indicates no translations were available for scoring
No coverage indicates a translation was available but COMET does not support this language

Figure 1: BLEU Scores for Colonial Official Languages

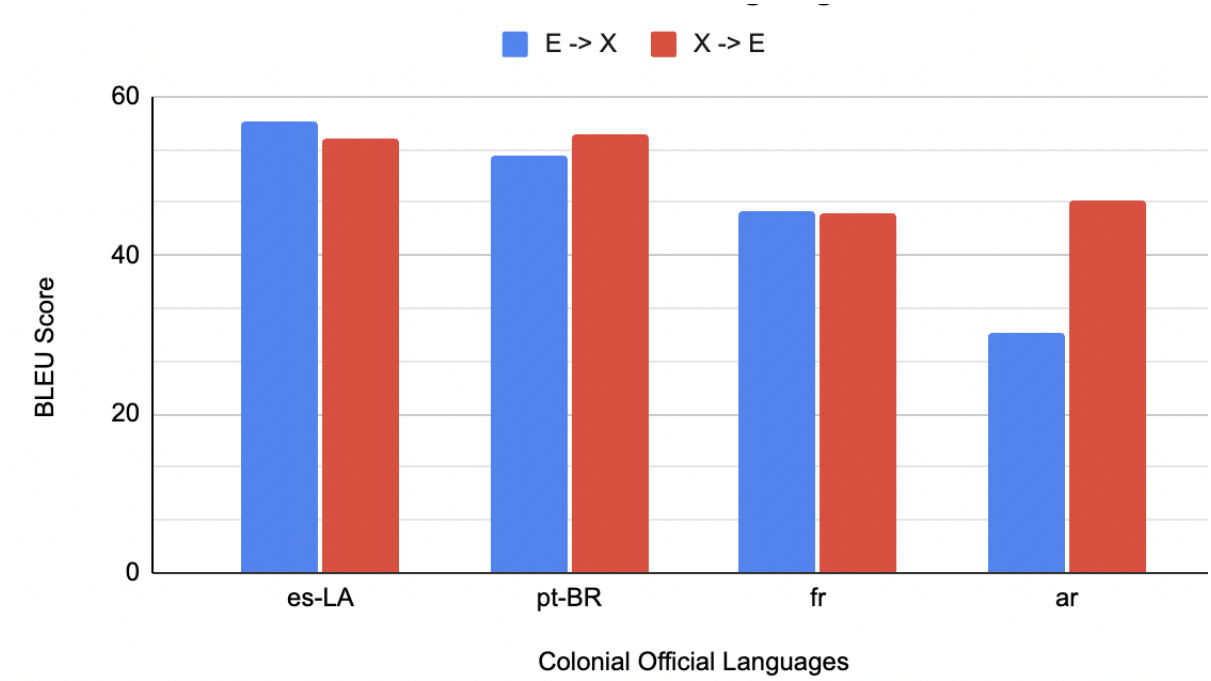

Figure 2: BLEU Scores for Native Official Languages

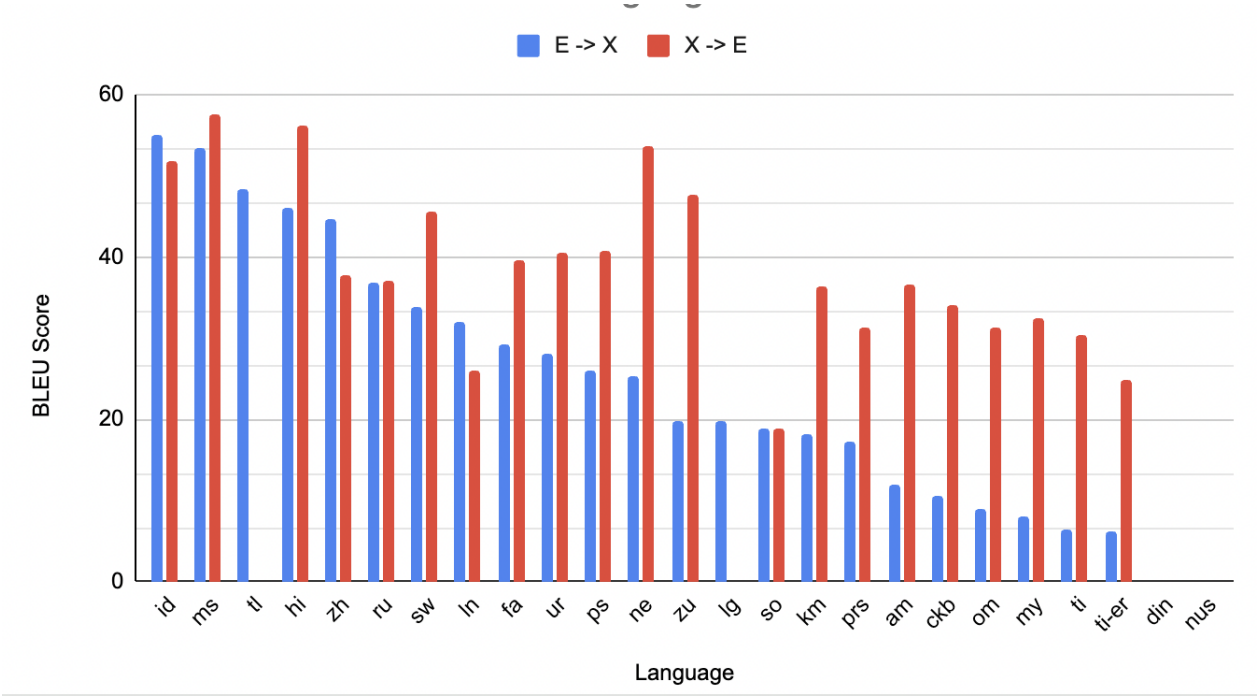

Figure 3: BLEU Scores for Local Official Languages

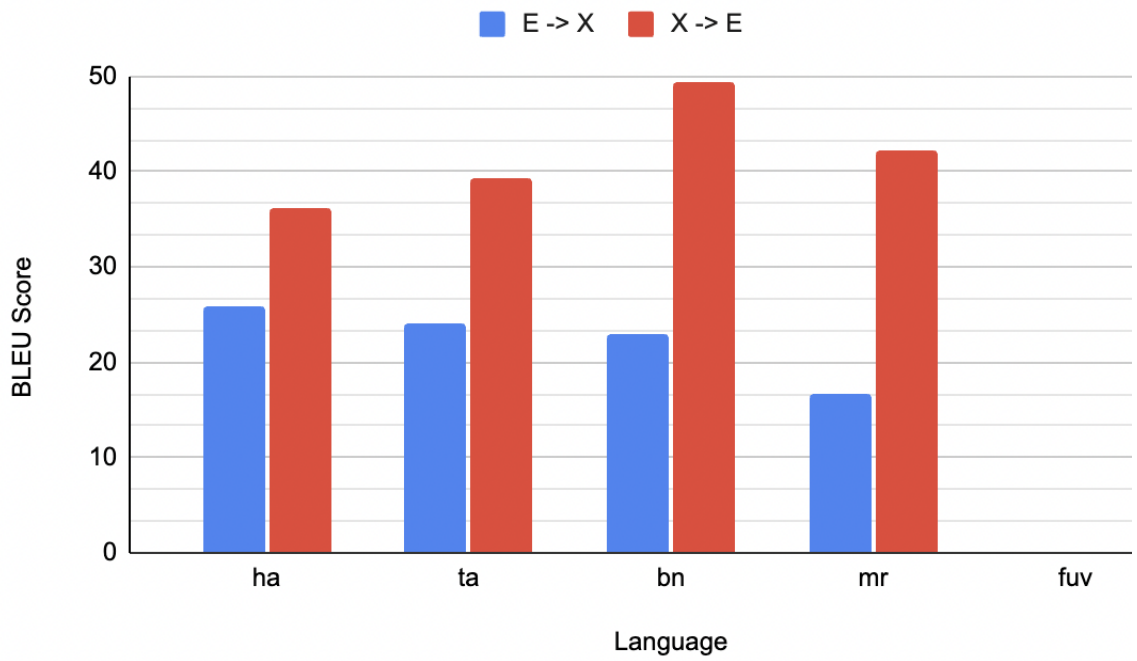

Figure 4: BLEU Scores for Southeast Asian Languages

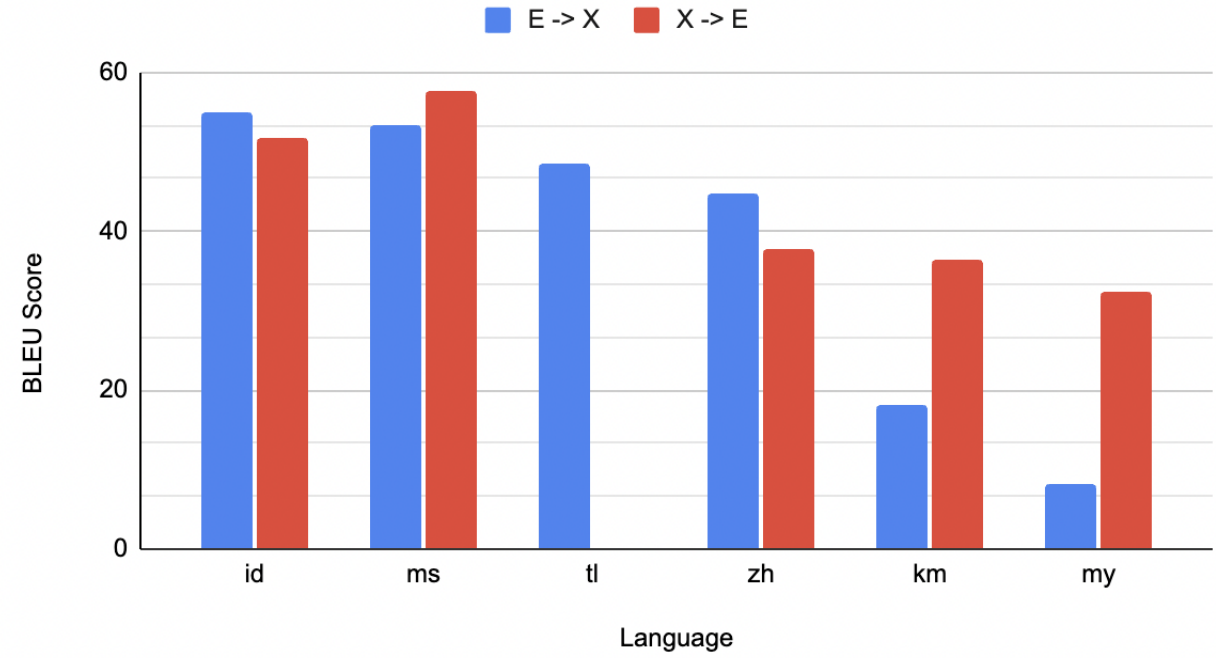

Figure 5: BLEU Scores for South Asian/Middle Eastern Languages

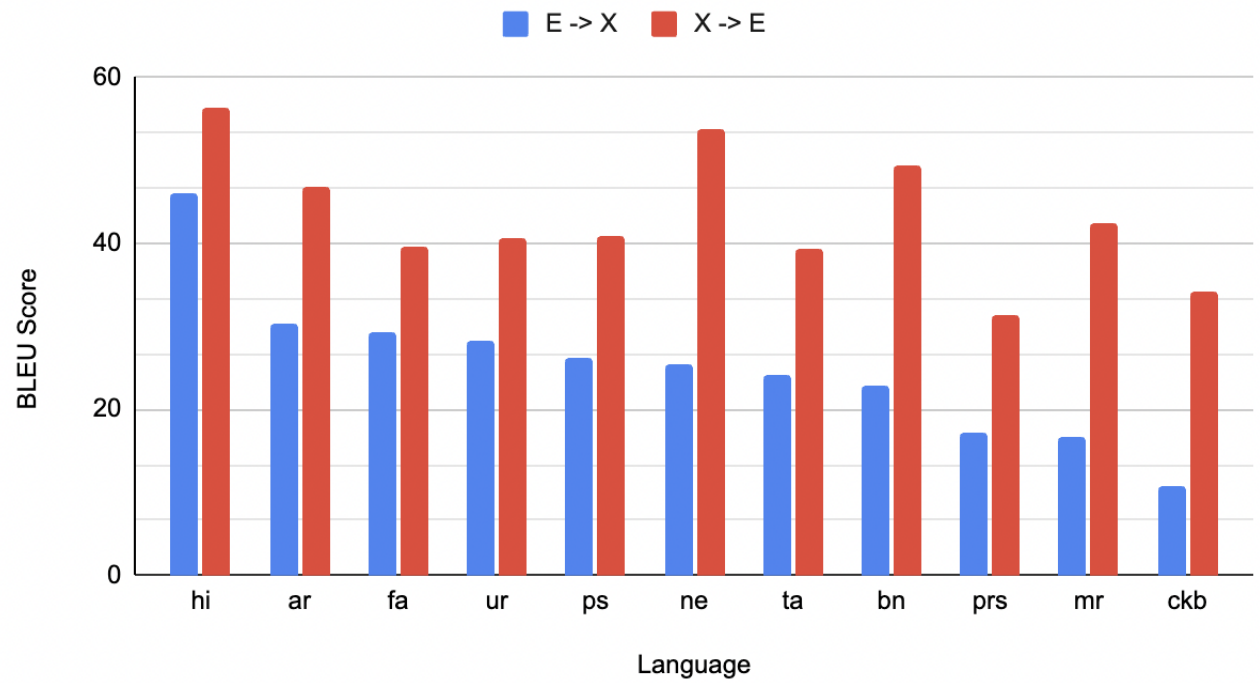

Figure 6: BLEU Scores for African Languages