MACHINE TRANSLATION FOR LOW-RESOURCE LANGUAGES:

USING RELATED LANGUAGE DATA FROM TAGALOG AND INDONESIAN TO IMPROVE ENGLISH TO CEBUANO TRANSLATION

By

Jade Athena Phoreman

Signature Work, submitted for
completion of the University Honors Program

15 June 2023

University Honors Program
University of California, Davis

APPROVED

06/19/2023

_____

Raul Aranovich, Ph.D.
Department of Linguistics, Letters and Sciences

_____

Kate Andrup Stephensen, M.Ed.
Undergraduate Education, Director of the University of California at Davis Honors Program

ABSTRACT

Machine translation models generally require a large amount of training data, but thousands of the world's languages are low-resource: they do not currently have a large amount of training data available. This project investigates whether additional parallel aligned training data can be generated using pivot languages for a high-resource and low-resource language pair. Specifically, we use Tagalog and Indonesian as pivot languages to create more parallel aligned data between English and Cebuano. The Bible is our starting point for a parallel corpus due to its availability in thousands of languages and because its verse structure is helpful for the purposes of alignment. We will be using the neural machine translation approach, specifically training a transformer model from OpenNMT. Finding ways to improve machine translation for low-resource languages has the potential to remove language barriers and provide billions of people with greater access to technologies and education.

INTRODUCTION

Machine translation is the process of using a computer to translate text or speech from one language to another. There have generally been three approaches to machine translation: rules-based, statistical, and neural. Rules-based machine translation (RBMT) relies on the existence of a complete bilingual dictionary and complete set of morphological, syntactic, and semantic rules mapping the source language to the target language. These dictionaries and rules are manually created, making RBMT very labor intensive [1]. Statistical machine translation (SMT) uses machine learning to train statistical models on bilingual text corpora. Informally, this means SMT automatically learns some of the language rules that would have been manually documented in RBMT. Challenges of SMT include finding enough bilingual corpora to train the model and dealing with language pairs that have significantly different word orders, such as Japanese and English [2]. Neural machine translation (NMT) specifically uses a neural net to learn the statistical model. NMT models are generally built from recurrent, convolutional, or self-attention-based architectures, all of which are considered encoder-decoder networks [3]. NMT has become industry standard because of improved translation quality, faster translation speed, and scalability.

However, a challenge of NMT is that a large amount of data is required to train the model. This is in large part due to the syntactic and semantic complexities of language and the variability of word choice or sentence structures within the same language across different contexts of use. The exact amount of data required to

train a neural machine translation model varies widely depending on the similarity of languages in the language pair, the domain of the translation task (e.g. healthcare, business, education), and the required quality of the translation. In general, however, a "large" amount of data refers to at least millions of words or sentences [4].

This poses a challenge because the vast majority of the world's 7000+ languages do not have millions of words or sentences collected into corpora; they are what we refer to as low-resource. Although there is not one universally accepted definition of a "low-resource" language, a common definition is a language that has limited dictionaries, grammars, annotated corpora, or other linguistic resources that are often necessary for training an accurate language model using machine learning. A language may be low-resource for a number of reasons, such as having a low number of speakers, having access to limited funding to develop linguistic resources or technologies, or lacking an orthography.

In response to this challenge, there has been a growing amount of research related to low-resource machine translation in recent years. This paper provides the background research and proposed design for two low-resource machine translation experiments. First, we summarize the findings of our literature survey and reflect on previous work in the Related Work section. The NMT Architectures section offers background knowledge on the types of neural networks that are typically used in machine translation, which provides a basis for our decision to use transformer-based architectures in our experiments. Next, the Experimental Design section outlines the details of our two experiments. Each is designed to test data

augmentation methods for improving low-resource machine translation, specifically between English and Cebuano. Experiment 1 deals with how Tagalog, a language closely related to Cebuano, can be used to augment the training data of the English to Cebuano translation model. Experiment 2 explores whether a more distantly related language, such as Indonesian, is more beneficial to the training data augmentation of the English to Cebuano translation model than Tagalog if there is more data from Indonesian used in the experiment than in the analogous Tagalog experiment. This section also includes information about the corpora we will be using and the relatedness of Cebuano, Tagalog, and Indonesian. We report the outcomes in Results and analyze factors that influenced the results in Discussion. Lastly, we offer recommendations for future work in Suggestions for Improvement and Future Research and discuss why the field of low-resource machine translation is important in the Impact section.

RELATED WORK AND TECHNICAL BACKGROUND

Similar to the methods in our experiments, many approaches to improving low-resource machine translation involve data augmentation. Data augmentation can include increasing the size of a dataset using a related language as well as adding more linguistic information to a dataset through part-of-speech (POS) tagging.

An example of a data augmentation strategy that increases the size of the training dataset can be seen in Irvine and Callison-Birch's experiment combining bilingual and comparable corpora [5]. Comparable corpora are comprised of texts from the source and target languages covering similar topics. However, the texts are not direct translations of one another. The authors found that training with comparable corpora in addition to bilingual corpora improves the accuracy and coverage of phrase-based SMT models.

Xia et al. [6] increased the size of their training datasets through a two-step pivoting process: back-translating from the target language to the low-resource source language and to a high-resource language related to the source, followed by translating a target-to-related language parallel corpus into a low resource-to-related language parallel corpus. Transformer models were used for these tasks. When these two steps were combined, the results showed improvement over simple unsupervised high-resource to low-resource translation by 2 to 10 BLEU points and over standard supervised back-translation by 1.5 to 8 BLEU points across all datasets that they used for testing.

Li et al. [7] increased the size of their training datasets by using restricted sampling in the decoder to generate pseudo parallel data on the source and target sides without using additional monolingual data. Transformer-based architectures were used to build their neural networks. Their approach improved upon baselines in low-resource translation tasks by 1.0 to 2.0 BLEU points.

Tars et al. [8] made use of multiple languages closely related to their extremely low-resource target language, as well as back-translation, to improve their neural machine translation model, built using a transformer-based encoder-decoder. Back-translation was used to produce synthetic data, and their experiments showed that combining synthetic data with the original parallel corpus as the training dataset generally produced better models.

Pourdamghani et al. [9] converted the parallel data between a high-resource language closely related to the low-resource source language and the target language, into parallel data between the low-resource source language and the target language. They used a statistical machine translation approach, and found that adding some converted parallel data almost always improves the SMT model's BLEU score, but adding too much converted parallel data sometimes decreases the model's BLEU score.

An example of a data augmentation strategy that adds linguistic information to a dataset can be seen in Hlaing et al.'s experiments [10]. The authors added POS tags to each word on either the source or target side, or both, in the proposed multi-source transformer and shared-multi-source transformer models. A multi-source

transformer is a transformer with multiple inputs and multiple encoders. In this case, one input would be the source text as a vector of words, and the other input would be a vector of the POS tags for each word in the text vector. A shared-multi-source transformer is a multi-source transformer where both input vectors are taken into the encoder and processed together, rather than each input being processed separately in its own encoder, allowing the model to learn the relationship between the inputs. The results of the experiments showed that adding linguistic features, such as POS-tagging, to the transformer-based models can improve the performance of low-resource NMT. Shared-multi-source transformer models with POS-tagging resulted in more significant increases in BLEU scores than the baseline transformer model.

Maimaiti et al. [11] used paraphrase embedding and POS-tagging to generate more monolingual data by replacing words based on parts of speech and semantic similarity. More specifically, nouns were replaced by similar nouns, verbs by similar verbs, adverbs by similar adverbs, and adjectives by similar adjectives. The results showed that this proposed data augmentation method, which is model transparent and was used with many different architectures in the experiments, can lead to NMT models that outperform previous state-of-the-art methods on low-resource language pairs in seven language pairs from four corpora by 1.16 to 2.39 BLEU points.

Although Adlaon and Marcos's [12] experiment does not relate to our data augmentation methods, the experiment used the same corpora and two of the same

languages that our experiments will use. The authors built a recurrent neural network, trained solely on the book of Genesis from our dataset, to implement a Cebuano to Tagalog translator. The distinctive characteristic of their experiments was their use of subword translation for verbs. They chose to use subword translation on verbs specifically since verbs are the most morphologically complex in Tagalog and Cebuano. Subword translation improved the translation results by almost 3 BLEU points. The authors recommended including more training data, which we do by using the entirety of the biblical texts available, rather than only the book of Genesis. However, our experiment differs in many other important ways from this previous work: we use a transformer model instead of an RNN, English is our source language, Cebuano is our target language, Tagalog and Indonesian are used as related supplementary languages; and we do not use subword translation. The details of our experimental setup are in the Experimental Design section.

NMT ARCHITECTURES

The most common and most basic NMT architecture is the encoder-decoder model, which is a sequence-to-sequence model most commonly implemented using a recurrent neural net (RNN) or a convolutional neural net (CNN) [13]. The encoder converts the input sequence from the source language into a vector that represents the meaning of the input. The decoder then converts the meaning vector into an output sequence in the target language.

Another type of RNN encoder-decoder-based model is the long short-term memory (LSTM) network, which is capable of remembering information over long

sequences [13]. This memory ability is helpful for learning dependencies and relationships that are far apart in sequences. Due to its memory cell, an LSTM-based model can store information that does not suffer from vanishing gradients.

Attention is another mechanism that can be paired with an RNN or CNN encoder-decoder architecture to improve the quality of machine translation [13]. Attention works by maintaining vectors for each word in the input sequence, rather than converting the input sequence into a single meaning vector as in a basic encoder-decoder architecture. With this attention mechanism, the decoder can then reference any of these word vectors at each step of the decoding process, and can also selectively decide which word vectors to focus on at each step using the attention weights that are learned in training.

However, attention can also be used alone, as in the transformer model [14]. The transformer model is based entirely on attention mechanisms and does not make use of RNN or CNN architectures. A huge benefit of the transformer model is its ability to produce high-quality translation models with relatively minimal computational resources and time, compared to the resources and time required to train RNNs or CNNs that produce similar translation quality. Transformer models are also easily parallelizable, which means that the computation can be divided across multiple processors to speed up training.

EXPERIMENTAL DESIGN

SUMMARY

We perform two different experiments, each designed to explore how data augmentation approaches using related languages can improve low-resource machine translation quality. First we train a baseline model without using any related language data, as shown in Figure 1. The baseline model is trained only on the English-Cebuano Bible corpus, described in more detail in Figure 1 Key and the Corpora Details section.

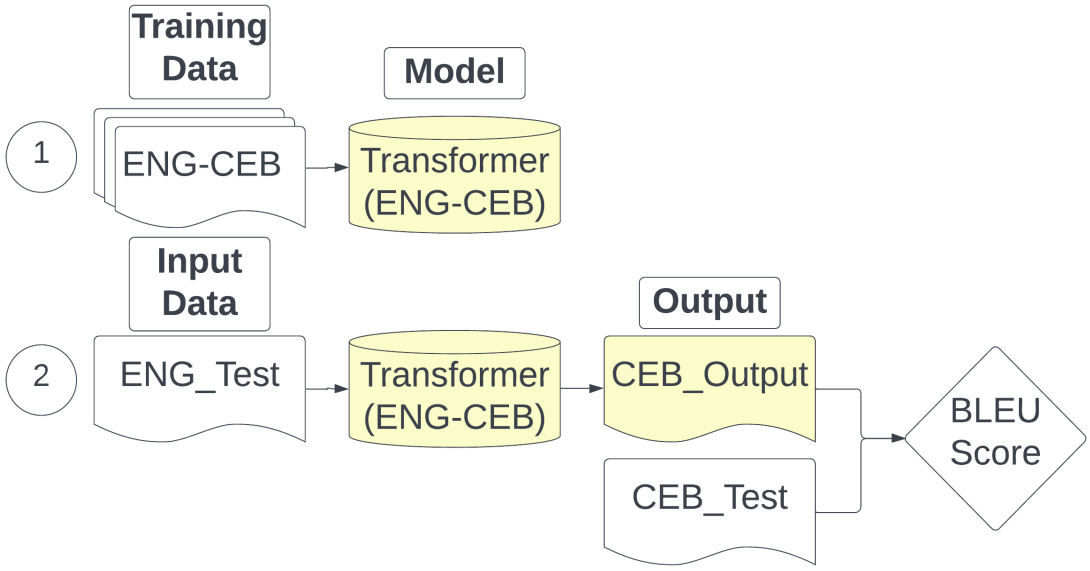

Figure 1. Baseline Low-Resource Translation Scenario from English to Cebuano. No related language data was used in the corpora.

Figure 1 Key

Label	Explanation	Total Sentence Pairs
ENG-CEB	English-Cebuano parallel Bible corpus [20]	31,034
ENG_Test	English sentences from English-Cebuano parallel corpus for testing; selected from English-Cebuano Tatoeba corpus [24]	524
CEB_Test	Cebuano sentences from English-Cebuano parallel corpus for testing; selected from English-Cebuano Tatoeba corpus [24]	524
CEB_Output	Cebuano sentences inferred by the baseline model from ENG_Test input. To be compared to CEB_Test to evaluate the model.	524

The BLEU score for the baseline model is obtained by comparing the gold standard translation, CEB_Test, to the model-inferred translation, CEB_Output.

EXPERIMENT 1

Does augmenting the resources of a low-resource language with a related language increase the quality of neural machine translation? Previous work [6], [15], [16] has shown affirmative results, but we would like to show similar results specifically for Cebuano and Tagalog, which were not studied in the aforementioned experiments. To answer this question, Experiment 1 will use English-Tagalog parallel

aligned data to produce additional English-Cebuano parallel aligned data to train our experimental model, as shown in Figure 2.

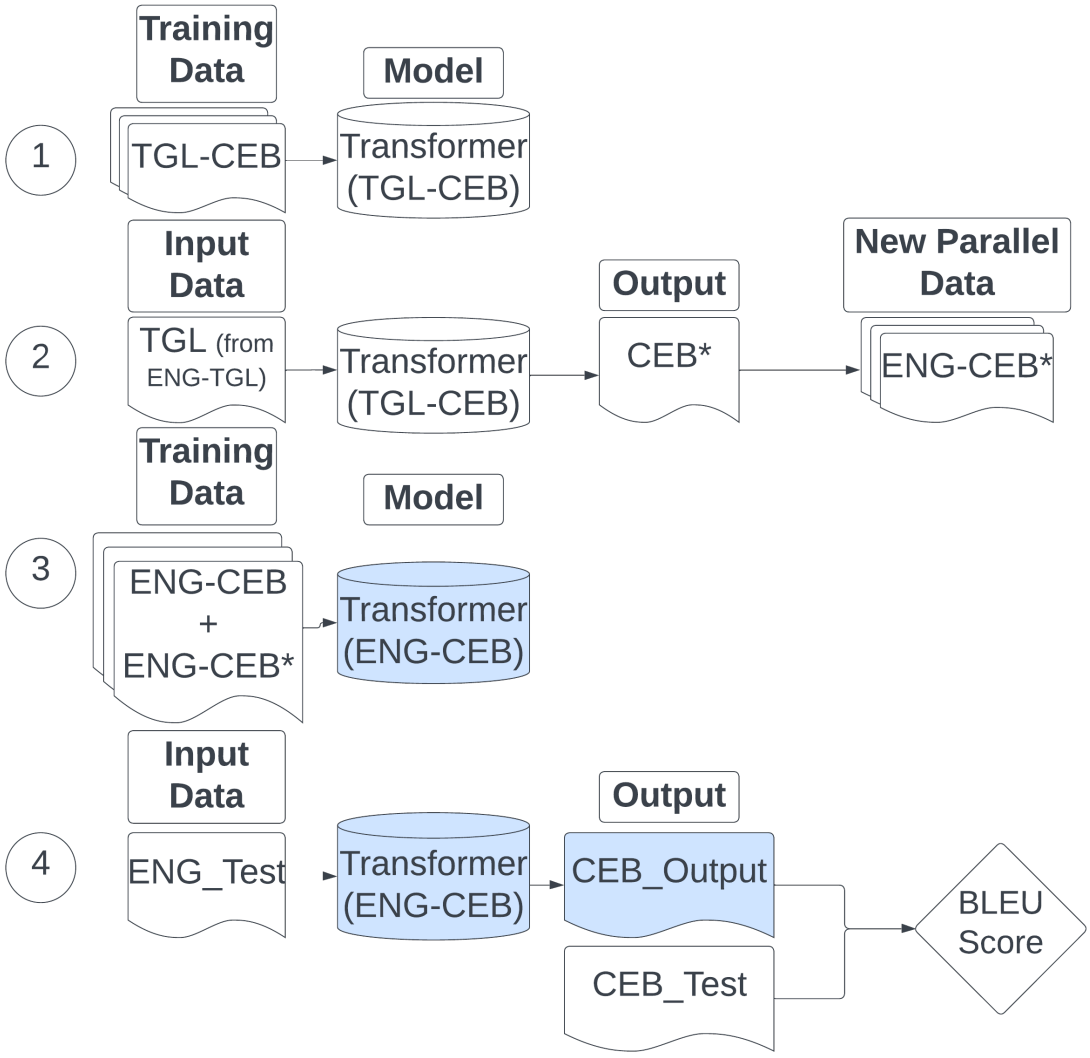

Figure 2. Experiment 1. We first train a model to translate from Tagalog to Cebuano, to produce additional English-Cebuano data from available English-Tagalog data (ENG-TGL, the source of TGL in step 2). Then, we train the English to Cebuano model. Compare with the BLEU score resulting from the baseline model in Figure 1.

Figure 2 Key

Label	Explanation	Total Sentence Pairs
TGL-CEB	Tagalog-Cebuano parallel Bible corpus [20]	31,035
ENG-TGL	English-Tagalog supplemental parallel corpus from QED, TED2020, and Tatoeba corpora [24], [25], [26], [27], [28]	34,077
ENG-CEB*	English-Cebuano supplemental parallel corpus, where CEB* was inferred by model from TGL in ENG-TGL	34,077
ENG-CEB	English-Cebuano parallel Bible corpus [20]	31,034
ENG_Test	English sentences from English-Cebuano parallel corpus for testing; selected from English-Cebuano Tatoeba corpus [24]	524
CEB_Test	Cebuano sentences from English-Cebuano parallel corpus for testing; selected from English-Cebuano Tatoeba corpus [24]	524
CEB_Output	Cebuano sentences inferred by the baseline model from ENG_Test input. To be compared to CEB_Test to evaluate the model.	524

The BLEU score for Experiment 1's model is obtained by comparing the gold standard translation, CEB_Test, to the model-inferred translation, CEB_Output. Then, we compare Experiment 1's BLEU score to the baseline's BLEU score to determine whether training with the supplemental data improved the final translation quality. More details about BLEU scoring can be found in the Results section.

EXPERIMENT 2

How similar to the low-resource target does a high-resource language need to be to have a noticeable effect on the resulting model's translation quality after being used to augment training data? Is there a tradeoff between the size of the training dataset and how closely related the target language and the supplementary language are? In other words, is it better to augment the training data using a more distant language that has more available data, or using a more closely related supplementary language that has less available data? To answer this question, Experiment 2 uses English-Indonesian parallel aligned data to produce additional English-Cebuano parallel aligned data to train our second experimental model, as shown in Figure 3 on the next page. It is almost identical to the setup of Experiment 1, only substituting the Tagalog corpora for Indonesian corpora that are larger.

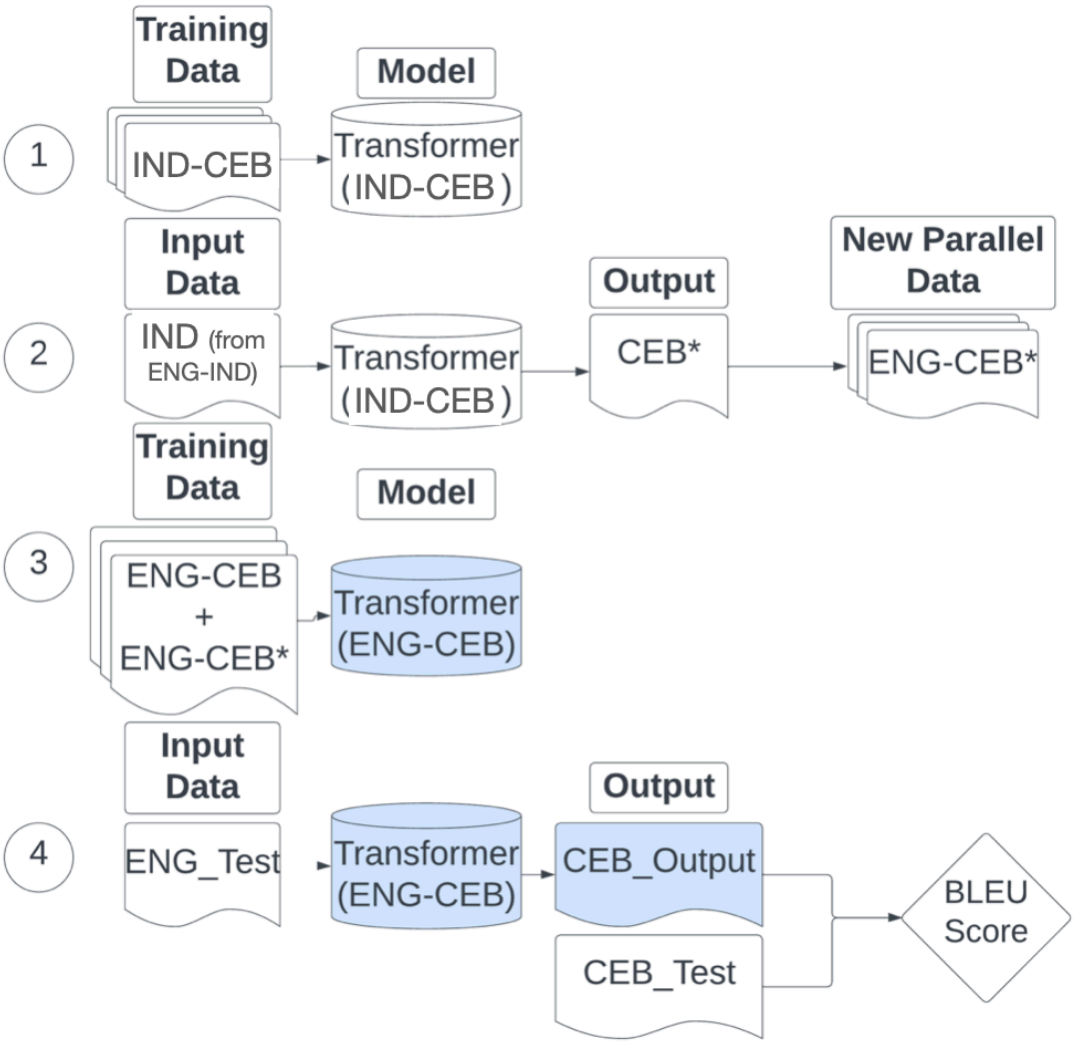

Figure 3. Experiment 2. We first train a model to translate from Indonesian to Cebuano, to produce additional English-Cebuano data from available English-Indonesian data (ENG-IND, the source of IND in step 2). Then, we train the English to Cebuano model. Compare with the BLEU score resulting from the baseline model in Figure 1.

Figure 3 Key

Label	Explanation	Total Sentence Pairs
IND-CEB	Indonesian-Cebuano parallel Bible corpus [20]	29,623
ENG-IND	English-Indonesian supplemental parallel corpus from QED and TED2020 corpora, [25], [26], [27], [28]	553,696
ENG-CEB*	English-Cebuano supplemental parallel corpus, where CEB* was inferred by model from IND in ENG-IND	553,696
ENG-CEB	English-Cebuano parallel Bible corpus [20]	31,034
ENG_Test	English sentences from English-Cebuano parallel corpus for testing; selected from English-Cebuano Tatoeba corpus [24]	524
CEB_Test	Cebuano sentences from English-Cebuano parallel corpus for testing; selected from English-Cebuano Tatoeba corpus [24]	524
CEB_Output	Cebuano sentences inferred by the baseline model from ENG_Test input. To be compared to CEB_Test to evaluate the model.	524

CORPORA

We selected the Bible corpus of Christos Christodouloupoulos and Mark Steedman as our baseline corpus [20]. Each language within the corpus has an XML file that organizes the Bible verses with tags specifying the book, chapter, and verse number. Utilizing the Bible as a corpus can often be advantageous for training machine translation models due to its rich linguistic diversity in genres and grammatical structures and its availability in thousands of languages, many of which are low-resource. Moreover, the consistent structure of the Bible facilitates alignment and comparison across different languages, enabling effective analysis and translation modeling.

Christodouloupoulos and Steedman acknowledge that one potential issue within their corpus and Bible corpora in general is that some languages may have been translated using the formal equivalence method and others may have been translated using the dynamic equivalence method [20]. With formal equivalence, an idiom in the source language may be translated literally into the target language, losing its intended meaning unless the reader has the resources to look into the cultural context of the source language. With dynamic equivalence, the idiom would be translated into its meaning, losing the direct relationship between the words in the source and the words in the target. For example, dynamic equivalence might translate “it’s raining cats and dogs” into “it was raining heavily.” Dynamic equivalence translations can be problematic for a machine translation model that is trying to learn mappings between words.

Other potential problems with the Bible corpus acknowledged by Christodouloupoulos and Steedman include the use of antiquated language, missing verses from some translations whose meanings may have been absorbed into the surrounding verses, and the fact that a verse does not always correspond to one sentence [20]. There may be partial sentences or multiple sentences within a verse. All of these problems make perfect alignment difficult.

For supplemental non-biblical data and testing data, we struggled to find high-quality corpora of any significant size. Ultimately, we chose TED2020, QED, and Tatoeba [24],[25],[26],[27], [28]. TED2020 can be found on OPUS and was compiled from a web crawl of about 4000 TED and TED-X transcripts from July 2020, and volunteers helped to translate the results of the crawl into many languages [24]. The TED2020 corpus contains several issues, including inconsistent timestamp XML tags across different languages, presence of empty files, and instances of time stamps with empty content specific to one language within a pair. Furthermore, additional problems include incomplete or incorrect translations, variations in subtitle/ transcript formatting, and occasional alignment discrepancies between the source and target languages. QED, the QCRI Educational Domain Corpus, can also be found on OPUS and was compiled from transcribed and translated subtitles for educational videos and lectures [28]. These subtitles are collaboratively generated and sourced from a web-based platform. The QED corpus contains all of the same issues as the TED2020 corpus. Lastly, Tatoeba, available on its own website, is a collection of individual sentences contributed by volunteers [24]. The Tatoeba corpus, organized in .tsv files, generally demonstrates good alignment due to the sentence pair format

in each row and language in each column. However, since it relies on volunteer contributions, the translations may not always be professional or completely accurate. The extent and significance of inaccuracies across our language pairs in the Tatoeba corpus remain uncertain.

ARCHITECTURE: OPENNMT TRANSFORMER

We decided to use transformer models due to their computational efficiency, parallelization capability, and availability on OpenNMT. Transformer models are ideal for low-resource scenarios because they can learn more quickly than other architectures can, and they can learn from smaller datasets. OpenNMT is an open-source NMT toolkit with prebuilt tools and models for building custom machine translation systems, which is a helpful starting place for our purposes.

The OpenNMT transformer requires word vocabularies for the source and target languages alongside the training data. The vocabularies should represent the most frequently used words in the training data. We can generate these vocabularies by using SentencePiece, an open-source unsupervised text tokenizer and detokenizer. SentencePiece works by splitting words into subwords and then constructing a vocabulary based on the frequency and diversity of those subwords in the training corpus. To ensure that both frequency and diversity are optimally balanced, SentencePiece uses algorithms like byte-pair encoding or unigram language model [29].

LANGUAGES

Language	Tagalog	Indonesian	Cebuano
HRL or LRL	High-resource	High-resource	Low-resource
Family	Austronesian - Central Philippine [1]	Austronesian - Malayo-Sumbawan [2]	Austronesian - Central Philippine [1]
Word order	verb-initial, but flexible [3]	subject-verb-object [2]	verb-subject-object [2]
Orthography	Latin script and diacritics	Latin script	Latin script and diacritics

We chose English as the source, Cebuano as the target, and Tagalog and Indonesian as pivot languages for several reasons. We needed a low-resource language that was closely related to a high-resource language, and distantly related to another high-resource language. We also wanted all languages to be in Latin script to reduce the complexity of the project.

The set of Cebuano, Tagalog, and Indonesian meet all of these criteria. Tagalog and Cebuano are much more closely related than Indonesian and Cebuano, as the Family row in the table above shows. These languages allow us to test our research questions and determine if there is a tradeoff between the amount of data available in a supplementary language and how closely related that language is to the target.

PROCEDURE

To train the baseline model, we followed these steps:

1. Convert the English and Cebuano Bible corpus .xml files to .txt files. If any verses were available in only one language, those verses were omitted from the resulting .txt files in both languages. The OpenNMT Transformer model requires that input data be formatted such that each line in a file corresponds to a sentence, and these lines must be aligned across source and target files. Because the corpus was aligned by verse and verses don't necessarily equate to sentences, we decided that each line in the source and target files would correspond to a verse to make alignment easier.

2. Randomize the order of the sentences in both the source and target .txt files in the same way such that the source and the target remain aligned. Randomization helps reduce the likelihood that the model will learn patterns of the training data file structure rather than the features that actually impact translation. This is especially important using a Bible corpus, since different themes, words, and styles are prominent in different books of the Bible.

3. Split the now randomized but aligned source and target .txt files into training and evaluation files using an 80-20 split. This results in 4 .txt files: source train, source evaluation, target train, and target evaluation.

4. Train a SentencePiece model and create a vocabulary based on the source train file using this command: **onmt-build-vocab --sentencepiece --size 5000 --save_vocab sp source_train.txt** [1]. Do the same for the target train file.

5. Create and edit the configuration file in .yml format. Below is our baseline model configuration file, called base.yml:

```
model_dir: base_model/

data:
    train_features_file: source_train.txt
    train_labels_file: target_train.txt
    eval_features_file: source_eval.txt
    eval_labels_file: target_eval.txt
    source_vocabulary: source.vocab
    target_vocabulary: target.vocab
    source_tokenization:
        type: SentencePieceTokenizer
        params:
            model: source.model
    target_tokenization:
        type: SentencePieceTokenizer
        params:
            model: target.model

train:
    save_checkpoints_steps: 100
    save_model: base_model/base

eval:
  scorers: bleu
  early_stopping:
    metric: bleu
    min_improvement: 0.2
    steps: 4
```

Note the use of early stopping. If there is not more than a 0.2 improvement in the loss for more than 4 steps, the model will stop training.

---

[1] https://opennmt.net/OpenNMT-tf/vocabulary.html

6. Start training using this command: **nohup onmt-main --model_type Transformer --config base.yml --auto_config train --with_eval --num_gpus 2 &** [2]. Note that we use 2 GPUs to speed up training time.

7. Once training stops, translate the English portion of the English-Cebuano Tatoeba corpus into Cebuano and save the result to a .txt file using this command: **onmt-main --config base.yml --auto_config infer --features_file eng_tatoeba.txt > ceb_tatoeba_inferred.txt** [3]**.** Note that eng_tatoeba.txt has 524 sentences.

8. Evaluate the translation with BLEU scoring using this command: **perl multi-bleu.perl ceb_tatoeba.txt < ceb_tatoeba_inferred.txt** [4]

To train the Tagalog-Cebuano and Indonesian-Cebuano models, later referred to as pivot models, that will produce the supplementary English-Cebuano corpus from the English-Tagalog or English-Indonesian corpora, we followed the steps 1-6 listed above for the baseline model, but substituting the appropriate source language Bible corpora. Once the models finished training, we used them to translate the Tagalog or Indonesian portions of the supplementary corpora into Cebuano, using the same command as step 7 above but substituting the appropriate file names.

To train the experimental models on the combination of the Bible and supplementary corpora, we followed the following steps:

---

[2] https://opennmt.net/OpenNMT-tf/training.html

[3] https://opennmt.net/OpenNMT-tf/inference.html

[4] https://forum.opennmt.net/t/how-to-compute-bleu-score/3457

1. Split the supplementary corpus into training and testing with an 80-20 split, similar to step 3 from the baseline.

2. Concatenate the supplementary source training file to the end of the Bible corpus source training file, and do the same for the other 3 pairs of corresponding files. Now source train, source evaluation, target train, and target evaluation contain both the Bible corpus data and the supplementary data that was generated by the pivot models.

3. Follow steps 4-8 from the baseline, substituting in the appropriate file names for your new source and target files. Note that for step 7, we use the same **eng_tatoeba.txt** as in the baseline so that we can achieve an accurate comparison.

RESULTS

BLEU stands for Bilingual Evaluation Understudy Score and measures how close a machine translation is to a human translation. The BLEU score is calculated by multiplying the brevity penalty by the geometric mean of the individual n-gram scores [30]. In simpler terms, the brevity penalty accounts for the length of the translation compared to the reference, and the geometric mean combines the scores of different n-gram lengths (such as unigrams, bigrams, trigrams, etc.). The resulting value represents the overall similarity between the translation and the reference, with higher scores indicating better alignment. Unigram precision refers to the percentage of single words that are the same between the predicted and reference translations. Similarly, bigram precision refers to the percentage of two contiguous words that are the same, and 3-gram and 4-gram precision refers to the percentage of three and four contiguous words respectively that are the same between the predicted and reference translations.

Based on the success of other experiments mentioned in the Related Work section that used related language data to improve translation quality, we expected that the experimental models would have higher BLEU scores than the baseline model. However, this was not the case, as shown in the table below. Note that the BLEU score is out of 100 and n-gram precisions are percentages. A brevity penalty of 1 means that the number of words in the predicted translation is greater than the number of words in the reference.

	BLEU	Unigram precision	Bigram precision	3-gram precision	4-gram precision	Brevity Penalty
Baseline	0.76	11.0	1.9	0.3	0.5	1.0
Experiment 1 (Tagalog)	0.00*	13.7	2.9	0.5	0.0	1.0
Experiment 2 (Indonesian)	0.00*	6.7	0.7	0.0	0.0	1.0

*these scores were accompanied by an error message saying "Use of uninitialized value in division (/) at multi-bleu.perl line 139". This error occurs simply because the BLEU score is too low. A 4-gram of 0.0 will result in a BLEU score of 0.0 because geometric mean requires all n-gram scores to be multiplied together.

The multi-bleu.perl script chose 4 as the maximum length, since it is a common convention in machine translation evaluation, but including 4-gram scores results in extremely low BLEU scores for our experiments. If we chose 3 instead, you can see from the table above that the supplementary training data produced from Tagalog actually did help improve the translation slightly. The supplementary training data produced from Indonesian made the translation quality worse.

Because our results were worse than expected, we decided to evaluate some of the translated sentences by hand to gain more insight into what went wrong. Since we did not have access to any native Cebuano speakers, we used ChatGPT, which supports English-Cebuano translation, to gain more insight into the meaning of the predicted translation. It is important to acknowledge that we do not know exactly how accurate ChatGPT's English-Cebuano translations are in all cases, but we do know that ChatGPT had access to much more training data than our models did.

To evaluate by hand, we randomly selected 11 sentence pairs. To get an idea of how accurate ChatGPT is, we asked it to translate the Cebuano reference translation back into English. Then we asked it to translate the Cebuano predicted translation back into English and compared the results to the original English sentence. Results of this process are listed in the tables below:

Sentence 1

English	Burj Khalifa is currently the tallest skyscraper in the world.
Cebuano Reference	Ang Burj Khalifa ang pinaka tag-as nga edipisyo sa kalibutan.
ChatGPT Translation of Reference	The Burj Khalifa is the tallest building in the world.
Baseline Prediction	Pinaagi sa atong sa Kawang,
ChatGPT Translation of Baseline Prediction	Through our Lord,
Experiment 1 (Tagalog Pivot) Prediction	Ang Burha ?? hata mao ang pagkatunga-gob sa itaas.
ChatGPT Translation of Experiment 1 Prediction	Ang Burha? That's the middle point above.
Experiment 2 (Indonesian Pivot) Prediction	Ang Jateo ni Carseph, kini maoy usa ka hataas nga torre sa mga langit.
ChatGPT Translation of Experiment 2 Prediction	The Jateo of Carseph is a tall tower in the sky.

Sentence 2

English	How?
Cebuano Reference	Unsaon?
ChatGPT Translation of Reference	How did it happen?
Baseline Prediction	Unsa? Unsaon man? Unsaon
ChatGPT Translation of Baseline Prediction	What? What do I do?
Experiment 1 (Tagalog Pivot) Prediction	Unsa? Unsaon man?
ChatGPT Translation of Experiment 1 Prediction	What? How can it be
Experiment 2 (Indonesian Pivot) Prediction	Unsaon man?"
ChatGPT Translation of Experiment 2 Prediction	How can it be done?

Sentence 3

English	She met a new friend in school today.
Cebuano Reference	Karong adlawa, nakakita siyag bag-ong higala sa tunghaan.
ChatGPT Translation of Reference	Today, he/she saw a new friend at school.
Baseline Prediction	Siya nakighibalag ug usa ka bag-ong higala sa lubagay sa kaudtohon.
ChatGPT Translation of Baseline Prediction	He made friends with a new person during the afternoon stroll.
Experiment 1 (Tagalog Pivot) Prediction	Adunay iyang hingkit-an nga mga bag-o diha sa pagmaymay.
ChatGPT Translation of Experiment 1 Prediction	He has new ideas in mind for the future
Experiment 2 (Indonesian Pivot) Prediction	Ug si Semaias migula ug usa ka bag-ong higala nga sandurot sa iyang mga higala.
ChatGPT Translation of Experiment 2 Prediction	And Semaias found a new friend who quickly became close with his friends.

Sentence 4

English	Do good to those who hate you.
Cebuano Reference	Minaayohi nang mga nasilag nimo.
ChatGPT Translation of Reference	Those who were born with you are blessed.
Baseline Prediction	Dawata ninyo ang mga nanagdumot kaninyo.
ChatGPT Translation of Baseline Prediction	Accept those who despise you.
Experiment 1 (Tagalog Pivot) Prediction	Buhata ninyo ang maayo sa mga nagadumot kaninyo.
ChatGPT Translation of Experiment 1 Prediction	Do good to those who hate you
Experiment 2 (Indonesian Pivot) Prediction	Buhata ninyo ang maayo kanila nga nanagdumot kanimo, ug nagdumot kanila;
ChatGPT Translation of Experiment 2 Prediction	Do good to those who hate you and despise you.

Sentence 5

English	To hear is to obey.
Cebuano Reference	Wala maminaw kon dili mosunod.
ChatGPT Translation of Reference	One won't be heard if they don't listen or follow.
Baseline Prediction	Ang pagpatalinghug kaniya mao ang pagsugot.
ChatGPT Translation of Baseline Prediction	Acknowledging him/her is the acceptance.
Experiment 1 (Tagalog Pivot) Prediction	Ang pagpatalinghug kanimo.
ChatGPT Translation of Experiment 1 Prediction	The respect for you.
Experiment 2 (Indonesian Pivot) Prediction	Ug ang mga igdulungog sa pagpatalinghug sa tingog sa mga pulong ni Jehova.
ChatGPT Translation of Experiment 2 Prediction	And the ears that listen to the uplifting of the voice of Jehovah's words.

Sentence 6

English	In the Animistic concept of panpsychism, everything and everybody have some aspect of "mind," at some level, even though sometimes scarcely.
Cebuano Reference	Sa animistikong konsepto sa pansikismo, ang tanang butang ug ang tanang tawo adunay pipila ka aspeto sa "hunahuna," sa usa ka lebel, bisan kung usahay dili kaayo.
ChatGPT Translation of Reference	In the animistic concept of pansychism, all things and all people possess certain aspects of "mind" on some level, even if not always very pronounced.
Baseline Prediction	Sa Anumim sa kakulbaan, may pagtinagdanon, ug ang tanang paagi sa kinabuhi, ang pipila sa mga nawong sa tawo nga ugmaon, apan ang uban sa pagkabingkil.
ChatGPT Translation of Baseline Prediction	In the face of uncertainty, there is resilience, and in all the ways of life, some people face the future, while others fear it.
Experiment 1 (Tagalog Pivot) Prediction	Diha sa unom ka bahin nga akong pagahumokan sa pasimo, tanang mga butang ug sa tanang tawo adunay usa ka bote nga hinabol, ug hunahuna, bisan pa kon pila man kini sa makausa ra.
ChatGPT Translation of Experiment 1 Prediction	In the six parts that I will conquer in silence, all things and all people have a coveted bottle, and mind you, even if it is only once.
Experiment 2 (Indonesian Pivot) Prediction	Ug ang Achor nagpakatawo sulod sa iyang kasingkasing, ug sa dakung kakugi ug sa katawohan, ug sa tanang mga butang nga nagakamang sa ibabaw sa yuta;
ChatGPT Translation of Experiment 2 Prediction	And Achor gave birth within his heart, with great passion and determination, to all things that thrive upon the earth.

Sentence 7

English	It kept raining for a week.
Cebuano Reference	Nigsige ug ulan sa tibuok semana.
ChatGPT Translation of Reference	It has been raining throughout the entire week.
Baseline Prediction	Kanang nagpaulan ug ulan tungod sa usa ka semana.
ChatGPT Translation of Baseline Prediction	That which caused continuous rain for a week.
Experiment 1 (Tagalog Pivot) Prediction	Ipaulan ang ulan sa ibabaw sa tanang dapit.
ChatGPT Translation of Experiment 1 Prediction	Let the rain fall upon all places
Experiment 2 (Indonesian Pivot) Prediction	Labut pa ang ulan sa nahaunang adlaw sa semana, ug ang ulan sa semana pagaisipon nga ulan.
ChatGPT Translation of Experiment 2 Prediction	Furthermore, the rain will continue on the first day of the week, and the rain throughout the week will be accumulated rainfall.

Sentence 8

English	I don't care what Tom eats.
Cebuano Reference	Wa koy labot kung unsay kan-on ni Tom. OR Wala koy labot kung unsa man ang gikaon ni Tom.
ChatGPT Translation of Reference	I have no idea what Tom will eat. OR I have no concern about what Tom is eating.
Baseline Prediction	Ug ang kabalaka akong ginabuhat unsaon.
ChatGPT Translation of Baseline Prediction	And the worry, I am dealing with it on my own.
Experiment 1 (Tagalog Pivot) Prediction	?? ala akoy igo, unsaon kong unsa ang pagkaon ni Tomis.
ChatGPT Translation of Experiment 1 Prediction	I don't have enough, how can I have Tomis' food?
Experiment 2 (Indonesian Pivot) Prediction	?? ala akoy nakita ko nga kalan-on sa Topheth."
ChatGPT Translation of Experiment 2 Prediction	I haven't seen any food in Topheth.

The English sentence was included multiple times in the corpus with multiple Cebuano reference translations.

Sentence 9

English	I love her.
Cebuano Reference	Gihigugma ko siya.
ChatGPT Translation of Reference	I love him/her.
Baseline Prediction	Ako nagahigugma kaniya.
ChatGPT Translation of Baseline Prediction	I am the one who loves him/her.
Experiment 1 (Tagalog Pivot) Prediction	Ug kay ako mahigugma man sa iyang kaugalingon.
ChatGPT Translation of Experiment 1 Prediction	And since I love myself.
Experiment 2 (Indonesian Pivot) Prediction	Ako nahigugma sa iyang kaugalingon; ako mahigugma.
ChatGPT Translation of Experiment 2 Prediction	I love myself; I have self-love.

Sentence 10

English	I want to apologize for the way I talked to you the last time we met.
Cebuano Reference	Gusto nako mangayog pasaylo sa di maayong batasan nga akong gipakita kadtong ngkita ta sa una.
ChatGPT Translation of Reference	I want to ask for forgiveness for the bad behavior I exhibited when we last met.
Baseline Prediction	Ako kawalad-on sa dalan sa pagpangahas sa pagpakigsulti kaninyo sa katapusang panahon;
ChatGPT Translation of Baseline Prediction	I apologize for my audacity in speaking to you during our last encounter.
Experiment 1 (Tagalog Pivot) Prediction	Gitug-an ko ang pagpasaylo sa dakung pagsalig sa akong pagsuok.
ChatGPT Translation of Experiment 1 Prediction	I humbly ask for forgiveness for the great trust I have betrayed
Experiment 2 (Indonesian Pivot) Prediction	Buot ko unta ang akong baba nga diyutay ra sa kaulahingan sa katapusan.
ChatGPT Translation of Experiment 2 Prediction	I want my smile to reach the corners of the last horizon

Sentence 11

English	May I know your name again?
Cebuano Reference	Kinsa to gani ka?
ChatGPT Translation of Reference	Who are you, really?
Baseline Prediction	Mamati ba ako sa inyong ngalan, unsaon ko man pag-usab ang inyong ngalan?
ChatGPT Translation of Baseline Prediction	Should I believe in your name? How can I change your name?
Experiment 1 (Tagalog Pivot) Prediction	Magakalipay ba ako sa mga ngalan?
ChatGPT Translation of Experiment 1 Prediction	Will I find joy in names?
Experiment 2 (Indonesian Pivot) Prediction	Ug ang imong ngalan nahibalo kanako nga ako dili na mahibalo sa imong ngalan.
ChatGPT Translation of Experiment 2 Prediction	And your name tells me that I no longer know your name.

We found that although the words used in the predicted translations are usually different from the words used in the reference translations, hence the low BLEU scores, sometimes the meaning is not completely incorrect. One limitation of BLEU scoring is that it cannot measure semantic similarity. Without a way to quantify semantic similarity, evaluating these results is subjective. However, it is clear that some of the meanings are semantically related to the reference translation, showing that our models did learn something even in our low-resource setting.

CONCLUSION

DISCUSSION

There are many possible reasons why our results were not as expected. Adlaon and Marcos's work formed part of the basis for our optimism, since their experiment used only the book of Genesis from the same Bible corpus that we used and yielded successful results in the context of their own research questions [12]. We thought that using the entire corpus rather than just the book of Genesis may be enough to yield successful results despite the difference in our methods and research questions. Their preprocessing methods differed from ours in the following ways: they converted all characters to lowercase, removed duplicate sentences, and aligned by sentences manually rather than by verses, while we did not take those steps. Although our corpus was larger, it was likely much noisier than theirs. They also used an RNN and subword unit translation, discussed in more detail in Related Work, which differed from our overall approach.

Due to time constraints, we only trained each model on one train-evaluation split, rather than using k-fold cross-validation and training each model on multiple train-evaluation splits. This greatly reduced the amount of learning the models could achieve. Another mistake we made was failing to randomize the supplementary corpora when appending it to the Bible corpora to train the experimental models. In hindsight, we should have appended and then randomized the whole dataset again to avoid any biases. Additionally, we struggled to find high-quality supplemental English-Tagalog and English-Indonesian corpora. We had originally planned to use

the CCMatrix corpus from OPUS, which contains an impressive 70.5 million sentence pairs in English-Indonesian, and 3.1 million sentences in English-Tagalog. However, upon sampling the corpus, we realized that it is very noisy, containing characters from other scripts and a large amount of blatantly obvious misalignments. The significantly smaller but cleaner corpora that we did use, QED and TED2020, still contained some level of noise such as occasional non-ASCII characters and varying capitalization that we did not clean. Another significant issue with the QED and TED2020 corpora was alignment. These corpora are aligned by time stamps. Across languages, some time stamps contained content in only one of the languages of the language pair, and oftentimes time stamps were not positioned consistently in relation to other xml tags across language files. We had to use a combination of ChatGPT-generated Python code and manual cleaning to resolve these issues.

Beyond time constraints and corpora issues, the nature of translating from English into Cebuano may have been more difficult than we initially expected. Cebuano is a highly inflectional and agglutinative language. This means that the translation model may have learned some word segments correctly, but if an affix is wrong, the word would not match the reference translation and thus BLEU scoring would not recognize that word as containing anything correct. Additionally, Cebuano is a verb-initial language, while English has a subject-verb-object word order. Although word order does not matter for BLEU scoring, this may have made it more complicated for the model to learn translation patterns, especially since we aligned the Bible corpus on the verse level and not the sentence level.

Low-resource translation tasks are always difficult, especially when the source and target languages are as distantly related as English and Cebuano, and when corpora are not available or cleaned properly.

SUGGESTIONS FOR IMPROVEMENT AND FUTURE RESEARCH

Given the mistakes and challenges we have discussed, we have several recommendations for what should be done differently if anyone were to repeat these experiments or build upon our work.

We do not recommend using corpora built from web-crawls in a low-resource setting since they tend to be highly noisy. We recommend investing the time it takes to align something like a Bible corpus at the sentence level rather than the verse level and converting all text to lowercase. When data is limited, quality is key.

K-fold cross-validation is another strategy that should be implemented to improve the translation models. This strategy involves splitting the dataset into k disjoint groups, and then training the model on all but one group, leaving that group as the evaluation data. Iterate so that each group acts as the evaluation data during one training session. K-fold cross-validation is a valuable technique for assessing the generalization ability and reliability of predictive models by simulating their performance on unseen data.

We had also hoped to experiment with POS-tagging, but time constraints prevented this. Based on previous work [10], [11], it is likely that POS-tagging will improve English to Cebuano translation quality. There are Cebuano POS taggers

available online [21].

Because evaluating translation quality with BLEU score alone has limitations, it would be beneficial to ask a native Cebuano speaker to evaluate at least a sample of the predicted translation to gain more insight into the quality of the translation. A native speaker would also be able to identify specific problems, such as a particular affix that is often incorrect. This feedback could help us improve the training process.

IMPACT

Why is it worthwhile to work towards improving machine translation for low-resource languages? As our world becomes increasingly more interconnected, translation is becoming increasingly more important to facilitate economic, political, and cultural exchanges and interactions. As these exchanges and interactions occur online more often, machine translation specifically is a vital tool for facilitating global communication.

Google Translate is a well-known example of a machine translation model serving to facilitate online communication. Speakers of the 133 languages supported by Google Translate [22] can have access to online resources from all around the world, even if those resources were not originally written in their languages since Google Chrome often automatically translates a webpage into the user's default language. Yet, there are over 7000 languages in the world. Speakers of the remaining languages have limited access to online materials which may include educational resources, public health information, news, literature, and social media. Although a tool like Google Translate is based on SMT, and our experiments focused on NMT, many of the underlying principles of data augmentation for low-resource machine translation remain relevant. It is also worth noting that Google Translate does not currently support Cebuano [23], so our experiments may contribute to finding ways to improve Cebuano translation models in general.

Finding ways to ensure that more and more of the world's languages can use various technologies, even if those languages have very low numbers of speakers,

also helps to preserve and celebrate language diversity. This has the unquantifiable impact of empowering communities to feel like their languages, cultures, and identities matter to the world.

ACKNOWLEDGMENTS

ChatGPT was used to aid in the author's understanding of many of the experiments discussed in Related Work and many of the basic NMT concepts discussed throughout the paper. A few short phrases throughout the paper have been crafted with the help of ChatGPT, but overall the writing is the author's original work. ChatGPT was also used to write the code for preprocessing the corpora (randomizing, train-evaluation split, stripping xml tags, etc.).

REFERENCES

[1] C. D. Manning and H. Schu¨tze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press, 1999.

[2] M. D. Okpor, “Machine translation approaches: Issues and challenges,” 2014.

[3] F. Stahlberg, “Neural machine translation: A review,” *Journal of Artificial Intelligence Research*, vol. 69, p. 343–418, 2020.

[4] O. Bojar, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, P. Koehn, and C. Monz, “Findings of the 2018 conference on machine translation (WMT18),” in *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Belgium, Brussels: Association for Computational Linguistics, Oct. 2018, pp. 272–303. [Online]. Available: https://aclanthology.org/W18-6401

[5] A. Irvine and C. Callison-Burch, “Combining bilingual and comparable corpora for low resource machine translation,” in *WMT@ACL*, 2013.

[6] M. Xia, X. Kong, A. Anastasopoulos, and G. Neubig, “Generalized data augmentation for low-resource translation,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 5786–5796. [Online]. Available: https://aclanthology.org/P19-1579

[7] Y. Li, X. Li, Y. Yang, and R. Dong, “A diverse data augmentation strategy for low-resource neural machine translation,” *Information*, vol. 11, no. 5, 2020. [Online]. Available: https://www.mdpi.com/2078-2489/11/5/255

[8] M. Tars, A. Ta¨ttar, and M. Fisˇel, “Extremely low- resource machine translation for closely related languages,” in *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*. Reykjavik, Iceland (Online): Linko¨ping University Electronic Press, Sweden, May 31–2 Jun. 2021, pp. 41–52. [Online]. Available: https://aclanthology.org/2021.nodalida-main.5

[9] N. Pourdamghani and K. Knight, “Neighbors helping the poor: Improving low-resource machine translation using related lan- guages,” *Machine Translation*, vol. 33, no. 3, p. 239–258, 2019.

[10] Z. Z. Hlaing, Y. K. Thu, T. Supnithi, and P. Netisopakul, “Improving neural machine translation with pos-tag features for low-resource language pairs,” *Heliyon*, vol. 8, no. 8, p. e10375, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405844022016632

[11] M. Maimaiti, Y. Liu, H. Luan, Z. Pan, and M. Sun, “Improving data augmentation for low-resource nmt guided by pos-tagging and paraphrase embedding,” *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 20, no. 6, aug 2021. [Online]. Available: https://doi.org/10.1145/3464427

[12] K. M. M. Adlaon and N. Marcos, “Neural machine translation for cebuano to tagalog with subword unit translation,” in *2018 International Conference on Asian Language Processing (IALP)*, 2018, pp. 328–333.

[13] G. Neubig, “Neural machine translation and sequence-to- sequence models: A tutorial,” 2017. [Online]. Available: https://arxiv.org/abs/1703.01619

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” *CoRR*, vol. abs/1706.03762, 2017. [Online]. Available: http://arxiv.org/abs/1706.03762

[15] M.Johnson,M.Schuster,Q.V.Le,M.Krikun,Y.Wu,Z.Chen, N. Thorat, F. Vie´gas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean, “Google’s multilingual neural machine translation system: Enabling zero-shot translation,” *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 339–351, 2017. [Online]. Available: https://aclanthology.org/Q17-1024

[16] G. Neubig and J. Hu, “Rapid adaptation of neural machine translation to new languages,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 875–880. [Online]. Available: https://aclanthology.org/D18-1103

[17] “Wals online - home.” [Online]. Available: https://wals.info/

[18] R.Blust, *TheAustronesianlanguages*. Asia-PacificLinguistics, School of Culture, History and Language, College of Asia and
the Pacific, The Australian National University, 2013.

[19] R. Garcia, J. E. Dery, J. Roeser, and B. Ho¨hle, "Word order preferences of tagalog-speaking adults and children," *First Language*, vol. 38, no. 6, pp. 617–640, 2018. [Online].
Available: https://doi.org/10.1177/0142723718790317

[20] C. Christodouloupoulos and M. Steedman, "A massively parallel corpus: The bible in 100 languages," *Language Resources
and Evaluation*, vol. 49, no. 2, p. 375–395, 2014.

[21] Rjrequina,"Rjrequina/cebuano-pos-tagger:Rule-basedcebuano pos tagger using constraint-based grammar." [Online].
Available: https://github.com/rjrequina/Cebuano-POS-Tagger

[22] N. S. Fatmi, "How many languages are now available on google translate?" Jun 2022. [Online]. Available: https://www.androidcentral.com/apps-software/how-many-languages-are-now-available-on-google-translate

[23] [Online].Available:https://translate.google.hu/?hl=enamp;tab=wT

[24] Tatoeba (Corpora download). Retrieved from https://tatoeba.org/en/downloads

[25] J. Tiedemann. 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. European Language Resources Association (ELRA).

[26] Opus - TED2020-v1. Retrieved from http://opus.nlpl.eu/TED2020-v1.php

[27] Opus - QED-v2.0a. Retrieved from http://opus.nlpl.eu/QED-v2.0a.php

[28] A. Abdelali, F. Guzman, H. Sajjad and S. Vogel, "The AMARA Corpus: Building parallel language resources for the educational domain", The Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14). Reykjavik, Iceland, 2014. Pp. 1856-1862. Isbn. 978-2-9517408-8-4.

[29] SentencePiece. Retrieved from https://github.com/google/sentencepiece

[30] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02), pp. 311-318, 2002. [Online]. Available: https://aclanthology.org/P02-1040.pdf