Finding the Optimal Byte-Pair Encoding Merge Operations for Neural Machine Translation in a Low-Resource Setting

Kristine Mae M. Adlaon
University of the Immaculate Conception
F. Selga St. Davao City, Philippines
kadlaon@uic.edu.ph

Nelson Marcos
De La Salle University
Taft Avenue, Manila, Philippines
nelson.marcos@dlsu.edu.ph

Abstract

This paper investigates the impact of different Byte Pair Encoding (BPE) configurations, specifically, merge operations on neural machine translation (NMT) performance for the Filipino-Cebuano language pair across various text domains. Results demonstrate that smaller BPE configurations, notably 2k, 5k, and 8k consistently yield higher BLEU scores, indicating improved translation quality through finer tokenization granularity. Conversely, larger BPE configurations and the absence of BPE result in lower BLEU scores, suggesting a decline in translation quality due to coarser tokenization. Additionally, these findings help us understand how the size of the model and how finely we break down words affect the quality of translations. This knowledge will be useful for improving translation systems, especially for languages that don't have many parallel texts available for training.

1 Introduction

A high-quality, large-scale parallel corpus is undeniably crucial for enhancing neural machine translation. Unfortunately, parallel texts for many language pairs are scarce or nonexistent, limiting the effectiveness of many neural machine translation models. To improve translation for low-resource languages, several challenges must be addressed, including the scarcity of high-quality parallel data, the utilization of alternative sources such as monolingual resources and noisy comparable data, as well as parallel data in related languages, among other challenges.

An alternative method for enhancing translation quality in low-resource settings involves incorporating linguistic features, such as lemmas, part-of-speech (POS) tags, and dependency labels, into the NMT framework. This integration can aid in learning better token and sentence-level representations (Chakrabarty et al., 2020; Hoang et al., 2016; Li et al., 2018; Pan et al., 2020; Sennrich and Haddow, 2016). However, the availability of these linguistic features and resources is also scarce or not readily accessible for many languages.

Subword segmentation is a fundamental preprocessing step for NMT and has been shown to significantly impact the quality of the final output (Domingo et al., 2018; Gowda and May, 2020; Sennrich et al., 2016; Adlaon and Marcos, 2018; He et al., 2020). These methods enable NMT models to perform open-vocabulary translation by encoding rare and unknown words as sequences of subword units. This is particularly relevant for languages that form words through agglutination or compounding. Although subword segmentation does not typically adhere to morphological constraints, it mimics these processes by learning the optimal segmentation from training data, thereby creating vocabularies of subword tokens capable of generating new words not seen during training.

Subword segmentation has become the de facto standard in Neural Machine Translation (Bojar et al., 2018; Barrault et al., 2019). Known subword segmentation algorithms are SentencePiece Unigram LM (Kudo and Richardson, 2018), WordPiece algorithm (Song et al., 2021), and Byte Pair Encoding (BPE) (Sennrich et al., 2016) being the dominant approach (Provilkov et al., 2019). In this paper, we investigate the optimal configuration settings of BPE for translating from Filipino to Cebuano across various domains. We evaluated the results using intrinsic methods, such as the BLEU score, and extrinsic methods, including an examination of the morphological and syntactic divergence of the translations. To the best of our knowledge, there has been no comprehensive experimentation and analysis like this conducted for the Filipino-Cebuano language pair or for any of the major Austronesian languages.

Findings of the Association for Computational Linguistics: EACL 2024, pages 14673–14682

2 Related Work

This study was initiated after reviewing the research conducted by Ding et al. (2019). Their systematic exploration of various numbers of BPE merge operations shed light on how these operations interact with model architecture, vocabulary construction strategies, and language pairs. Their investigation aims to offer insights into selecting appropriate BPE configurations in future endeavors. Their findings demonstrate that LSTM-based architectures require experimentation with a wide range of BPE operations, as there is no single optimal configuration. Conversely, Transformer architectures tend to perform better with smaller BPE sizes. They emphasized the importance of careful consideration in selecting subword merge operations, as their experiments indicate that an ill-suited BPE configuration alone could lead to a decrease in system performance by 3–4 BLEU points.

Several studies have delved into the vocabulary generated by BPE merge operations. Cognetta et al. (2024)'s work has investigated threshold vocabulary trimming in BPE subword tokenization—a postprocessing step that replaces rare subwords with their component subwords. Their results suggest that while removing rare subwords is commonly recommended in machine translation implementations to reduce model size and enhance robustness, our experiments across a wide range of hyperparameter settings indicate that vocabulary trimming does not consistently improve performance. While the work of Saleva and Lignos (2023) presented two straightforward randomized adaptations of Byte Pair Encoding (BPE) and investigated their impact on a subsequent machine translation task. Their study concentrates on translating into languages with complex morphology, aiming to determine whether the method of selecting subwords significantly influences performance. Utilizing a Bayesian linear model for analysis, they find that one variant performs nearly identically to standard BPE, while the other exhibits less performance degradation than expected. This suggests that while standard BPE is prevalent, there are intriguing alternative variations worth exploring.

Gutierrez-Vasques et al. (2023) conducted detailed analyses across 47 typologically diverse languages and three parallel corpora, revealing that the types of recurrent patterns significantly influencing compression indicate morphological typology. For languages with richer inflectional morphology, early merges favor highly productive subwords, while languages with less inflectional morphology highlight more idiosyncratic subwords. Both pattern types enhance compression efficiency. They emphasized that contrary to the belief that BPE subwords lack linguistic relevance, they found cross-linguistic patterns resembling those in traditional typology.

3 Experimental Setup

Our experiments aim to understand the impact of adjusting the number of BPE merge operations across multiple domains in the Filipino to Cebuano Neural Machine Translation (NMT). We evaluated the outcomes using BLEU scores and by examining the actual translations.

3.1 Dataset

This work includes various text domains, such as encyclopedic content (Wikipedia), conversational text (Open domain), religious texts (Bible), and News Articles for the Filipino-Cebuano language pair.

Domain	Train	Dev	Test
Bible	25.0k	3.1k	3.1k
News Article	21.6k	2.7k	2.7k
Open Domain	17.6k	2.2k	2.2k
Wikipedia	20.8k	2.6k	2.6k
All	85.0k	10.6k	10.6k

Table 1: Number of sentence pairs across domains.

Both the Bible and Wikipedia are curated datasets taken from the work of Adlaon and Marcos (2019). The News article parallel corpus are scraped dataset from the web published by various news platforms in the Philippines such as Brigada News, Bandera, PhilStar, GMA News, and ABS-CBN News covering the months of November to December 2022. The sentences and translations for the Open Domain dataset which are mostly conversational in nature are from Tatoeba's [1] massive and awesome dataset, released under a CC-BY License. A copy of the curated parallel corpus used in the study is available here [2].

Table 1 datasets were used in the study of Baliber et al. (2020), specifically the DNMT model, which

[1] https://tatoeba.org/en/

[2] https://huggingface.co/datasets/jfernandez/cebuano-filipino-sentences

analyzes the performance of a multilingual model for Philippine languages and the work of Fernandez and Adlaon (2022) which attempted to develop a sentence aligner for automating the generation of parallel corpus given monolingual data.

Many of the most exciting studies in neural machine translation (NMT) concentrate on addressing open challenges related to long sentences rather than short ones. This focus stems from the intuition that, in terms of human learning and processing, short sequences are typically considered easier to handle (Wan et al.). With this, we would also like to see whether the average number of words in a sentence within the training data could be an important factor influencing the performance of NMT models, particularly considering the distribution of long sentences in the parallel corpora. As shown in Table 2, the Bible, News Article, and Wikipedia domains exhibit an average token length per training sample ranging from greater than 20 to less than 30 tokens. In contrast, the Open Domain has an average token length of 6. Notably, across all domains, less than 4% of sentences contain more than 50 tokens in a single training sample.

Domain		α	β
Bible	src	26	1104
	tgt	28	1507
News Article	src	20	213
	tgt	21	325
Open Domain	src	6	0
	tgt	6	0
Wikipedia	src	25	986
	tgt	25	1020

Table 2: α shows the average number of words(tokens) per sentence while β is the number of sentences with words(tokens) greater than 50.

3.2 BPE Segmentation

All our datasets were pre-tokenized prior to training. We applied Byte Pair Encoding (BPE) segmentation at the subword level using the subword-nmt tool (Sennrich et al., 2016). Separate subword vocabularies were learned for each language. BPE encoding was performed separately for each language, rather than jointly, to allow for a more in-depth examination of the segmentation specific to each language. Given that the number of merge operations is a hyperparameter, we experimented with values .5k, 2k, 5k, 8k, 10k, 15k, 20k, 32k adapted from the works of (Saleva and Lignos, 2023; Ding et al., 2019).

Table 3 presents an overview of source (src), target (tgt), and parameter (param) counts across different domains and BPE sizes, illustrating the differences and similarities in values where numbers in thousands are written in full and those in millions are abbreviated. In the Bible domain, the source values range from 605 (.5k BPE) to 42,328 (No BPE), while target values range from 622 to 45,925. The parameter count increases significantly from 45.1M to 112.8M as BPE size grows, indicating a higher complexity in model requirements. Similar trends are observed in the News Article and Wikipedia domains, where both source and target values increase progressively with larger BPE sizes, and parameter counts rise from around 45M to over 120M, reflecting the additional capacity needed for handling more extensive tokenization.

When considering all domains combined, source values range from 699 (.5k BPE) to 50,001 (No BPE), and target values follow a similar pattern, increasing from 697 to 50,001. The parameter count shows a consistent rise from 45.2M to 120.9M. These trends highlight the systematic relationship between BPE size and model complexity, emphasizing the need for more extensive models to manage larger vocabularies. Understanding these patterns is crucial for resource allocation in NLP model training, ensuring adequate computational power and memory.

3.3 NMT Architecture

We utilized the TransformerBase model implemented in OpenNMT-tf for our sequence-to-sequence tasks, particularly focusing on machine translation. The TransformerBase model follows the architecture proposed by Vaswani et al. (2023), which employs self-attention mechanisms to capture dependencies between input and output tokens more efficiently than traditional RNN-based models. This configuration in OpenNMT-tf, designed for optimal performance and computational efficiency, includes six layers each in the encoder and decoder. Each layer consists of multi-head attention and position-wise feed-forward networks, which are critical for learning complex linguistic patterns and structures. The adoption of this model allowed us to leverage its superior capability in handling parallel processing and long-range dependencies, thus significantly enhancing the accuracy and speed of our translation tasks (Klein et al., 2017,

Domain		.5k	2k	5k	8k	10k	15k	20k	32k	No BPE
Bible	src	605	2.2K	5K	8K	10K	14.8K	19.5K	29.9K	42.3K
	tgt	622	2.1K	5.1K	8K	10K	14.8K	19.6K	30.3K	45.9K
	param	45.1M	47.4M	51.9M	56.5M	59.5M	66.9M	74.2M	90.4M	112.8M
News Article	src	690	2.2K	5.2K	8.1K	10K	14.8K	19.5K	30K	50K
	tgt	687	2.2K	5.1K	8.1K	10K	14.8K	19.4K	29.8K	50K
	param	45.2M	47.5M	52.1M	56.6M	59.5M	66.9M	74.1M	90.1M	120.9M
Open Domain	src	644	2.1K	5K	7.8K	9.6K	13.3K	13.3K	13.3K	17.7K
	tgt	650	2.1K	5.1K	7.9K	9.7K	13.6K	13.9K	13.9	20.7K
	param	45.1M	47.4M	51.9M	56.2M	58.9M	64.9M	65.1M	65.1M	75.4M
Wikipedia	src	666	2.2K	5.1K	8.1K	10K	14.8K	19.4K	30K	50K
	tgt	667	2.2K	5.1K	8.1K	10K	14.8K	19.4K	29.8K	50K
	param	45.1M	47.5M	52.0M	56.6M	59.5M	66.9M	73.9M	89.9M	121M
All	src	699	2.2K	5.2K	8.2K	10.1K	15.1K	20K	31.7K	50K
	tgt	697	2.2K	5.2K	8.2K	10.2K	15.1K	20K	31.2K	50K
	param	45.2M	47.5M	52.1M	56.7M	59.7M	67.3M	74.9M	92.9M	120.9M

Table 3: Distribution of Source (src), Target (tgt), and Parameter (param) Counts Across Various Domains and BPE Sizes. Numbers in thousands are written in full, while those in millions are abbreviated. The table shows how increasing BPE sizes impact the complexity and size of datasets across Bible, News Article, Open Domain, Wikipedia, and combined domains, highlighting the corresponding rise in model parameters.

2020). Models were evaluated using BLEU score (Papineni et al., 2002).

3.4 Hardware

The models were trained using a Dell Precision 7770 workstation equipped with a 12th Gen Intel(R) Core(TM) i9-12950HX processor running at 2.30 GHz, 64.0 GB of RAM (63.7 GB usable), and an NVIDIA RTX A4500 Laptop GPU. On average, the model required approximately three hours to generate an inference every 5000 steps.

4 Results and Analysis

4.1 BLEU scores by Domain

Shown in Table 4, significant variations in BLEU scores were observed. In the Bible domain, BLEU scores ranged from 17.56 to 23.99, showcasing a clear trend of decreasing performance as BPE size increased, with the highest score recorded at 2k BPE and the lowest at 32k BPE. Similarly, in the News Article domain, BLEU scores spanned from 15.24 to 27.73, demonstrating a consistent pattern of decreasing scores with larger BPE configurations, with the highest performance seen at 2k BPE. Conversely, the Wikipedia domain displayed a more nuanced trend, with BLEU scores ranging from 31.34 to 36.24. Here, while smaller BPE configurations generally yielded superior results, some larger configurations performed comparably, highlighting domain-specific complexities influencing translation quality.

Generally, smaller BPE configurations, such as 2k and 5k, tend to yield higher BLEU scores across all domains, while larger configurations, such as 20k and 32k, result in lower scores. This suggests that finer tokenization granularity leads to better translation quality, likely due to improved handling of rare words and morphologically complex structures. Consistently, the top three models achieving the highest scores are the 2k, 5k, and 8k configurations.

4.2 Domain-Specific Variation

While smaller BPE configurations generally yield better results, the degree of improvement varies across domains. Higher δ values indicate greater variability in translation quality across BPE configurations, emphasizing the importance of selecting the appropriate BPE size for optimal performance. This variation could be attributed to differences in language complexity, domain-specific terminology, or sentence structures among the domains. Moreover, larger BPE configurations and the absence of BPE (No BPE) consistently result in lower BLEU scores. This indicates that coarser tokenization granularity leads to a loss of information during the translation process, resulting in reduced translation quality. Particularly noteworthy is the significantly lower performance observed in the "No

Domain	.5k	2k	5k	8k	10k	15k	20k	32k	No BPE	δ
Bible	22.33	**23.99**	21.38	19.61	18.78	17.86	17.72	17.56	17.84	6.43
News Article	22.91	**27.73**	27.07	24.05	22.53	20.65	18.70	16.71	15.24	12.49
Open Domain	**42.82**	37.63	33.93	31.49	30.70	27.56	27.57	27.39	33.52	15.43
Wikipedia	22.73	32.84	**36.24**	35.1	34.64	33.7	32.59	31.34	32.74	13.51
All	30.98	**39.56**	38.06	35.40	34.29	32.11	30.34	28.76	27.25	12.31

Table 4: The BLEU score is reported for Transformer architecture utilizing various BPE configurations. The δ symbol is the difference between the best and worst BLEU score of each row.

BPE" category across all domains, indicating that better translation (higher BLEU scores) is achieved with BPE segmentation.

Using the BLEU score interpretation by Lavie (2011), the Open Domain produces translations that range from understandable to high quality (27.39 - 42.82), whereas other domains produce translations that range from difficult to understand to clear but with significant grammatical errors (15.24 - 36.24). Overall, the translations generated across all domains are generally understandable (27.25 - 39.56). It is also noteworthy that none of the models across any of the domains generated translations with BLEU scores below 10, indicating that no translations were deemed almost useless.

4.3 Model Size

We have also recorded the sizes generated for each model. The size of the model significantly impacts scalability, making larger models more challenging to distribute across multiple machines or deploy in production environments, particularly when computational resources are limited. Furthermore, larger models typically lead to slower inference times, which can be critical in real-time applications where quick responses are essential.

Table 5 shows that model sizes grow with increasing BPE merges across all domains, implying a trade-off between model size and vocabulary granularity. Additionally, the difference in model size becomes more noticeable for models with more than 10k merge operations.This pattern is important for considerations in real-time applications and environments with computational constraints, where larger models might pose challenges for scalability and speed.

4.4 Morphological and Syntactic Divergence

The morphological divergence between Filipino and Cebuano manifests in their distinct approaches to affixation, reduplication, and verb focus/aspect marking (Cheng et al., 2017). In this section, we provide an examination on the quality of translation focusing on *affixation* morphological feature.

The analysis of different BPE configurations for translating "Mahiyain si Tom at hindi siya masyadong nagsasalita" (Tom is shy and doesn't talk much.) into Cebuano reveals varied performance in handling morphological structures. As shown in Table 6, without BPE segmentation the translation appears disjointed with redundant use of $siya$ (he) referring to Tom. The word $nagaistorya$ (speaking) correctly captures the progressive aspect (because of the prefix */naga-/* + root word *istorya*) of the verb but the translation lacks the nuance of $mahiyain$ (shy) implied in $nagsasalita$ (speaking).

The 0.5k BPE model accurately segmented and reconstructed $maulawon$ (shy) but slightly deviated with $niestorya$ which is in the past tense, instead of $nagaistorya$ which is progressive. The 2k BPE introduced $muraogulaw$ (somewhat shy), which, while correct, used the less natural $magstorya$ instead of $mosulti$. The 5k BPE model closely matched the reference with $maulawon$ and $masulti$, handling morphological rules effectively. The 8k BPE maintained accurate morphology with $moistorya$ fitting well contextually. The 10k BPE, while less nuanced with $maulaw$, still preserved grammatical correctness. The 15k BPE showcased the ability to handle complex morphemes with $kaistorya$ maintaining high fidelity to reference morphology. The 20k BPE accurately used $musulti$ consistently translating well at this segmentation level. Finally, the 32k BPE was not able to capture the correct tense of the verb $nagasulti$ or $nagaistorya$ in the progressive form but generated $nagsulti$ in the past tense form.

In the reference translation, *Manggiulawon si Tom ug dili siya kaayo motingog*, the word $manggiulawon$ reflects the trait of being shy or timid, capturing the essence of $mahiyain$. How-

Domain	.5k	2k	5k	8k	10k	15k	20k	32k	No BPE
Bible	175	184	201	218	230	259	286	348	434
News Article	175	184	202	218	230	258	286	347	465
Open Domain	175	184	201	218	228	238	252	252	287
Wikipedia	175	184	201	219	230	258	285	346	465
All	175	184	202	219	231	260	289	358	465

Table 5: Deployable model sizes in $megabyte(mb)$

Input	Mahiyain si Tom at hindi siya masyadong nagsasalita.	
Reference	Manggiulawon si Tom ug dili siya kaayo motingog.	
English	Tom is shy and doesn't talk much.	
No BPE	Si Tom siya ug dili kaayo siya nagaistorya.	
Models	**BPE Segmented**	**Reconstructed**
.5k	Ma@@ ula@@ w@@ on si T@@ om ug dili ka@@ ayo siya ni@@ es@@ tor@@ ya@@ .	Maulawon si Tom ug dili kaayo siya niestorya.
2k	M@@ u@@ ra og ula@@ w si Tom ug di kaayo siya ma@@ g s@@ tor@@ ya.	Mura og ulaw si Tom ug di kaayo siya mag storya.
5k	Ma@@ ula@@ won si Tom ug di kaayo siya ma@@ sulti@@ .	Maulawon si Tom ug di kaayo siya masulti.
8k	Ma@@ ula@@ won si Tom og dili kaayo siya mo@@ istor@@ ya.	Maulawon si Tom og dili kaayo siya moistorya.
10k	Ma@@ ulaw si Tom ug dili kaayo siya mu@@ istorya.	Maulaw si Tom ug dili kaayo siya muistorya.
15k	Ma@@ ula@@ won si Tom og dili kaayo siya ka@@ istorya.	Maulawon si Tom og dili kaayo siya kaistorya.
20k	Ma@@ ula@@ won si Tom ug dili kaayo siya mu@@ sulti.	Maulawon si Tom ug dili kaayo siya musulti.
32k	Ma@@ ula@@ won si Tom og dili kaayo siya nagsulti	Maulawon si Tom og dili kaayo siya nagsulti

Table 6: Short-text (< 15 words) translation results for different BPE configurations from Filipino to Cebuano.

ever, in the translated outputs, especially in the lower BPE configurations (e.g., 0.5k, 2k), alternative lexical choices like $maulawon$ and *mura og ulaw* are used, deviating from the reference. Additionally, while the reference employs *kaayo motingog* to convey *not very talkative*, the translations vary in the expression of this concept, with some models using *dili kaayo siya muistorya* (not very talkative) and others opting for *dili kaayo siya moistorya* (not very inclined to speak).

In a longer sentence shown in Table 7, the No BPE translation correctly uses $Gipangita$ and $gihatag$ accurately reflecting the verbs $searched$ and $gave$ with appropriate actor-focus and recipient-focus affixes. However, the use of $nila$ instead of $kanila$ indicates a minor error in pronoun usage. The noun $pinuy-anan$ accurately translates to $residence$, showing correct lexical handling. Syntactically, the structure remains close to the reference, but the pronoun error affects clarity. The translation in .5k model introduces unnecessary elements like $grupo$ (group) and incorrectly contextualizes $nagtuon$ (studied) for $students$. The verb $Naggitabonan$ (covered) is a morphological error, deviating from the intended meaning of $searched$. The translation of $modules$ to $modyul$ is correct. Improvement in morphological handling with correct use of $Gipangita$ and $gihatag$ can be seen in model 2k. However, the phrase *sa mga modyul* should be *kanila ang mga modyul* to correctly reflect the recipient-focus morphology. The segmentation properly handles af-

fixes but still misses finer morphological distinctions. Noticeably in the 8k model and 15k model, a significant morphological and contextual error was observed with the introduction of the words $Biyernes$ (Friday) and *ilang interes* (their interest) respectively from the intended $modules$.

Overall, models 5k and 8k BPE balanced segmentation granularity and morphological fidelity best.

5 Conclusions

In this work, we have explored the efficacy of various BPE configurations on neural machine translation performance across multiple text domains for the Filipino-Cebuano language pair. The analysis demonstrated that smaller BPE configurations, particularly 2k and 5k, consistently yielded higher BLEU scores across all domains, highlighting the importance of finer tokenization granularity in enhancing translation quality. Conversely, larger BPE configurations and the absence of BPE resulted in lower BLEU scores, indicating a loss of translation quality due to coarser tokenization.

The performance variations across different domains underscore the domain-specific challenges in translation. For instance, the Bible and News Article domains exhibited the highest BLEU scores at smaller BPE sizes, while the Wikipedia domain presented more nuanced trends, with some larger configurations performing comparably to smaller ones. This indicates the complexity of domain-specific terminologies and structures in influencing NMT performance.

The morphological and syntactic analysis revealed that the models' ability to handle affixation and other linguistic features varied with BPE size. Smaller BPE configurations better preserved morphological integrity and syntactic accuracy, whereas larger configurations occasionally introduced errors, particularly in handling complex linguistic patterns. This underscores the critical role of appropriate BPE segmentation in maintaining linguistic fidelity in translations.

Furthermore, the study highlighted the significant impact of model size on scalability and deployment. Larger models, associated with higher BPE configurations, pose challenges for distribution and real-time application due to increased computational and memory requirements. This necessitates a careful balance between tokenization granularity and model complexity to optimize performance while managing resource constraints.

Overall, the findings emphasize the importance of selecting optimal BPE configurations tailored to specific domains and linguistic features to enhance NMT performance. The study contributes valuable insights into the balance between tokenization granularity, model size, and translation quality, guiding future developments in both bilingual and multilingual NMT systems.

6 Limitations and Future Work

The study's exploration of BPE configurations for Filipino to Cebuano Neural Machine Translation (NMT) reveals several limitations and areas for future work. Firstly, the variability in BLEU scores across domains highlights the need for domain-specific optimizations, as the optimal BPE size can differ significantly between text types. Furthermore, while smaller BPE configurations generally resulted in higher BLEU scores, suggesting better handling of rare words and morphological complexity, larger BPE sizes and the absence of BPE consistently underperformed, indicating a trade-off between model complexity and token granularity that needs further refinement. Another limitation is the hardware dependency, as the study utilized high-end computational resources, which may not be readily available in all research settings, potentially limiting the replicability and scalability of the approach.

Future work could include exploring alternative evaluation metrics that better account for semantic fidelity and fluency. There is also a need to investigate the impact of different BPE configurations on other linguistic features, such as syntactic structures and morphological features especially that Filipino and Cebuano are morphologically-complex languages, to develop more robust translation models. In addition to evaluating translations from Filipino to Cebuano, it is imperative to conduct evaluations in the reverse direction, from Cebuano to Filipino, to ensure bidirectional translation quality, including other Philippine languages. Lastly, integrating more diverse datasets and leveraging advanced architectures, such as transformer variants or hybrid models, could further enhance translation performance and provide deeper insights into the complexities of translating between low-resource languages.

Furthermore, future work will ideally involve a comparative analysis of morphological and se-

Input	Hinanap ng mga guro ang tirahan ng mga mag-aaral at ibinigay sa mga ito ang modules.
Reference	Gipangita sa mga magtutudlo ang kwarter sa mga estudyante ug gihatag kanila ang mga modyul.
English	The teachers searched the students' residences and gave them the modules.
No BPE	Gipangita sa mga magtutudlo ang pinuy-anan sa mga estudyante ug gihatag nila sa mga modyul.

Models	**BPE Segmented**	**Reconstructed**
.5k	Nag@@ gi@@ tab@@ on@@ an og mga g@@ ru@@ p@@ o ang pin@@ u@@ y@@ -@@ anan sa mga nag@@ tu@@ on ug gi@@ hatag kini sa mga mo@@ d@@ y@@ ul@@ .	Naggitabonan og mga grupo ang pinuy-anan sa mga nagtuon ug gihatag kini sa mga modyul.
2k	Gi@@ pan@@ gita sa mga magtu@@ tudlo ang pu@@ y-anan sa mga es@@ tud@@ yante ug gihatag sa mga mod@@ y@@ ul@@ .	Gipangita sa mga magtutudlo ang puy-anan sa mga estudyante ug gihatag sa mga modyul.
5k	Gi@@ pangita sa mga magtutudlo ang puy-anan sa mga estudyante ug gihatag sila sa mga mod@@ y@@ ul@@ .	Gipangita sa mga magtutudlo ang puy-anan sa mga estudyante ug gihatag sila sa mga modyul.
8k	Gi@@ pangita sa mga magtutudlo ang puy-anan sa mga estudyante ug gihatag kanila ang Biyern@@ es.	Gipangita sa mga magtutudlo ang puy-anan sa mga estudyante ug gihatag kanila ang Biyernes.
10k	Nahitabo sa mga magtutudlo ang puy-anan sa mga estudyante ug gihatag sa mga mod@@ y@@ ul@@ .	Nahitabo sa mga magtutudlo ang puy-anan sa mga estudyante ug gihatag sa mga modyul.
15k	Gi@@ pangita sa mga magtutudlo ang puy-anan sa mga estudyante ug gihatag ang ilang inter@@ es.	Gipangita sa mga magtutudlo ang puy-anan sa mga estudyante ug gihatag ang ilang interes.
20k	Ang mga magtutudlo nakakita sa puy-anan sa mga estudyante ug gihatag sa ila ang mga mod@@ y@@ ul.	Ang mga magtutudlo nakakita sa puy-anan sa mga estudyante ug gihatag sa ila ang mga modyul.
32k	Gipangita sa mga magtutudlo ang puy-anan sa mga estudyante ug gihatag ngadto sa mga mody@@ ul.	Gipangita sa mga magtutudlo ang puy-anan sa mga estudyante ug gihatag ngadto sa mga modyul.

Table 7: Long-text (> 15 words) translation results for different BPE configurations from Filipino to Cebuano.

mantic characteristics in relation to frequency, embeddings, untrained tokens, and alignment across various languages. A more comprehensive investigation of Byte-Pair Encoding (BPE) merging operations is also planned, with a focus on the role of intermediate tokens and the influence of token merge order on morphological structures.

7 Acknowledgment

The researchers express their gratitude to the Department of Science and Technology - Philippine Council for Industry, Energy and Emerging Technology Research and Development (DOST-PCIEERD) for providing the funding to purchase the equipment used in training the models.

References

Kristine Mae M. Adlaon and Nelson Marcos. 2018. Neural machine translation for cebuano to tagalog with subword unit translation. In *2018 International Conference on Asian Language Processing (IALP)*, pages 328–333.

Kristine Mae M. Adlaon and Nelson Marcos. 2019. Building the language resource for a cebuano-filipino neural machine translation system. In *Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval*, NLPIR '19, page 127–132, New York, NY, USA. Association for Computing Machinery.

Renz Iver Baliber, Charibeth Cheng, Virgion Mamonong, and Kristine Mae Adlaon. 2020. Bridging Philippine languages with multilingual neural machine translation. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 14–22, Suzhou, China. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019.

Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Abhisek Chakrabarty, Raj Dabre, Chenchen Ding, Masao Utiyama, and Eiichiro Sumita. 2020. Improving low-resource nmt through relevance based linguistic features incorporation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4263–4274.

Charibeth Cheng, Kristine Mae Adlaon, Maristella Aquino, Ervin Fernandez, and Kevin Villanueva. 2017. Mag-tagalog: A rule-based tagalog morphological analyzer and generator. *n Proceedings of the 17th philippine computing*, pages 171–178.

Marco Cognetta, Tatsuya Hiraoka, Naoaki Okazaki, Rico Sennrich, and Yuval Pinter. 2024. An analysis of bpe vocabulary trimming in neural machine translation. *arXiv preprint arXiv:2404.00397*.

Shuoyang Ding, Adithya Renduchintala, and Kevin Duh. 2019. A call for prudent choice of subword merge operations in neural machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 204–213, Dublin, Ireland. European Association for Machine Translation.

Miguel Domingo, Mercedes García-Martínez, Alexandre Helle, Francisco Casacuberta, and Manuel Herranz. 2018. How much does tokenization affect neural machine translation? *CoRR*, abs/1812.08621.

Jenn Leana Fernandez and Kristine Mae M. Adlaon. 2022. Exploring word alignment towards an efficient sentence aligner for Filipino and Cebuano languages. In *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*, pages 99–106, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Thamme Gowda and Jonathan May. 2020. Finding the optimal vocabulary size for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3955–3964, Online. Association for Computational Linguistics.

Ximena Gutierrez-Vasques, Christian Bentz, and Tanja Samardžić. 2023. Languages Through the Looking Glass of BPE Compression. *Computational Linguistics*, 49(4):943–1001.

Xuanli He, Gholamreza Haffari, and Mohammad Norouzi. 2020. Dynamic Programming Encoding for Subword Segmentation in Neural Machine Translation. page 3042–3051.

Cong Duy Vu Hoang, Gholamreza Haffari, and Trevor Cohn. 2016. Improving neural translation models with linguistic factors. In *Proceedings of the Australasian Language Technology Association Workshop 2016*, pages 7–14.

Guillaume Klein, François Hernandez, Vincent Nguyen, and Jean Senellart. 2020. The OpenNMT neural machine translation toolkit: 2020 edition. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 102–109, Virtual. Association for Machine Translation in the Americas.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Alon Lavie. 2011. Evaluating the output of machine translation systems.

Qiang Li, Derek F Wong, Lidia S Chao, Muhua Zhu, Tong Xiao, Jingbo Zhu, and Min Zhang. 2018. Linguistic knowledge-aware neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(12):2341–2354.

Yirong Pan, Xiao Li, Yating Yang, and Rui Dong. 2020. Dual-source transformer model for neural machine translation with linguistic knowledge. *Preprints*, page 2020020273.

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.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2019. Bpe-dropout: Simple and effective subword regularization. *arXiv preprint arXiv:1910.13267*.

Jonne Saleva and Constantine Lignos. 2023. What changes when you randomly choose BPE merge operations? not much. In *Proceedings of the Fourth Workshop on Insights from Negative Results in NLP*, pages 59–66, Dubrovnik, Croatia. Association for Computational Linguistics.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. *arXiv preprint arXiv:1606.02892*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. 2021. Fast WordPiece tokenization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2089–2103, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Yu Wan, Baosong Yang, Derek Fai Wong, Lidia Sam Chao, Liang Yao, Haibo Zhang, and Boxing Chen. Challenges of neural machine translation for short texts. In *Computational Linguistics 2022*, page 321–342. Association for Machine Translation in the Americas.