Morphologically-Guided Segmentation For Translation of Agglutinative Low-Resource Languages

William Chen wchen6255@knights.ucf.edu
Brett Fazio brettfazio@knights.ucf.edu
University of Central Florida

Abstract

Neural Machine Translation (NMT) for Low Resource Languages (LRL) is often limited by the lack of available training data, making it necessary to explore additional techniques to improve translation quality. We propose the use of the Prefix-Root-Postfix-Encoding (PRPE) subword segmentation algorithm to improve translation quality for LRLs, using two agglutinative languages as case studies: Quechua and Indonesian. During the course of our experiments, we reintroduce a parallel corpus for Quechua-Spanish translation that was previously unavailable for NMT. Our experiments show the importance of appropriate subword segmentation, which can go as far as improving translation quality over systems trained on much larger quantities of data. We show this by achieving state-of-the-art results for both languages, obtaining higher BLEU scores than large pre-trained models with much smaller amounts of data.

1 Introduction

Subword segmentation is a common technique used to improve machine translation quality due to its ability to reduce the vocabulary size of input text. Unsupervised techniques such as Byte-Pair Encoding (BPE) (Sennrich et al., 2015) are prevalent in most NLP tasks. On the other side of the spectrum, state-of-the-art morphological segmentation is achieved using dedicated neural seq2seq models (Wang et al., 2016). Neither of these however, are well-suited for Low-Resource Languages (LRLs).

BPE was found to oversplit roots of infrequent words in both English and Japanese (Bostrom and Durrett, 2020). Lower BLEU scores in Quechua-Spanish models segmented by BPE (Ortega et al., 2021) suggest similar side-effects for Quechua. Neural morphological segmentation models require large amounts of morpheme-labeled training data, which often does not exist at all for LRLs. We propose the use of the Prefix-Root-Postfix-Encoding (PRPE) algorithm (Zuters et al., 2018) as an alternative for subword segmentation. PRPE is able to draw upon linguistic knowledge without needing large amounts of labeled training data, making it a middle-ground between BPE and neural seq2seq that is ideal for LRLs.

PRPE is a semi-supervised word segmentation algorithm that uses subword statistics to identify and learn the prefixes, roots, and postfixes of words in a corpus (Zuters et al., 2018), and can be guided using a language-specific heuristic. Using the generated lists of roots and affixes, the algorithm performs subword segmentation that only appears to be morphologically grounded; PRPE does not use any actual linguistic/morphological rules. This makes it well-suited for studying LRLs, as it only requires a surface level of understanding to tune the heuristic for a language instead of dedicated linguistic rules or large amounts of labeled training data.

We experiment with two distinct agglutinative LRLs, Indonesian and Quechua, as we hypothesized that PRPE would naturally work well in the morpheme-heavy environment of these languages. This is because words in agglutinative languages are constructed via a series of affixes, leading to large amounts of information expressed in a single word due to the presence of many morphemes. As such, machine translation for these languages is particularly challenging due to the increased vocabulary size and more frequent appearance of rare words (Koehn and Knowles, 2017). Quechua in particular is highly agglutinative; multiple suffixes are appended to modify a root to denote tense, mood, person, and number (Muysken, 1988).

To investigate the effectiveness of PRPE in improving machine translation quality, we conduct experiments using two distinct language pairs, Quechua to Spanish and Indonesian to English, across multiple domains of corpora. We accomplish this by training LSTM (Hochreiter and Schmidhuber, 1997) and Transformer (Vaswani et al., 2017) models on text segmented by PRPE and compare those with models trained on other segmentation methods. Our experiments show that PRPE subword segmentation can lead to significant improvements in machine translation performance, outperforming prior benchmarks with models pre-trained using masked language modeling (Guntara et al., 2020), transfer learning (Ortega et al., 2021), and models trained on much larger datasets (Guntara et al., 2020).

1.1 Contributions

Our contributions are outlined as followed: (1) we show the ease of extending the semi-supervised PRPE algorithm to new languages by applying it to Quechua and Indonesian; (2) we train several NMT models for those languages to demonstrate the effectiveness of PRPE in improving translation accuracy; (3) we re-introduce a general domain Quechua dataset for NMT by manually cleaning and re-aligning raw data used in early SMT experiments that was previously only available in parallel parse-tree format. Our code and dataset are available at https://github.com/wanchichen/morphological-nmt.

2 Related Work

The current segmentation standard for most NMT systems is the unsupervised method of Byte-Pair Encoding (Sennrich et al., 2015). BPE initially represents the corpus at a character level, after which pairs of the most frequently occurring symbols are iteratively merged together to form the vocabulary. However, recent works have shown that a unigram language model for segmentation (Kudo, 2018), another unsupervised method, appears to be the better alternative. Bostrom and Durrett (2020) found unigram models better preserved roots and split affixes compared to BPE in English and Japanese. Richburg et al. (2020) observed similar benefits for two LRLs: Swahili and Turkish.

Zuters et al. (2018) introduced the PRPE algorithm by experimenting with an English-Latvian pairing in both translation directions. PRPE utilizes a proposed 'Root alignment principle' - collecting statistics about prefixes and suffixes before aligning roots with the most frequent prefix and suffix. Aside from the differences in languages used, our work also differs from Zuters et al. (2018) in terms of how the algorithm is incorporated into the overall segmentation pipeline. They also did not consider running PRPE standalone, all text segmented with PRPE in their experiments were post-processed with BPE. In addition to this method, we also explore segmentation results obtained solely with PRPE, as well as those obtained from multiple iterations of PRPE.

There have been several studies on NLP for Quechua. Rios (2016) created a language toolkit for Quechua translation, which included a text normalizer, spell-checking, and morphological analyzer. Ortega et al. (2021) focused primarily on translation from Quechua to Spanish. They proposed BPE-Guided, a method to guide the BPE segmentation algorithm for Quechua by feeding BPE a dictionary of words to ignore during segmentation. They utilized transfer learning from Finnish (a high-resource agglutinative language) to obtain substantial improvements in BLEU. It was also in this study that Ortega et al. (2021) first suggested the use of PRPE for Quechua translation. Oncevay (2021) conducted research on multilingual translation for four Peruvian languages paired with Spanish: Aymara, Ashaninka, Quechua and Shipibo-Konibo. Pre-processing was done using the unigram model (Kudo (2018)) trained across a multilingual corpus. Quechua, the language with the most resources among the four, suffered in performance when trained on the multilingual task rather than solely Quechua and Spanish. Quechua was also recently featured as part of the AmericasNLP Shared Task (Mager et al., 2021), where participants were asked to translate Spanish text to Quechua among other indigenous American languages.

Compared to Quechua, Indonesian has enjoyed much more attention in NLP research. Extensive work has been done on computational approaches in Indonesian morphological analysis, such as MorphInd (Larasati et al., 2011) and later on MALINDO Morph (Nomoto et al., 2018), both of which created morphological dictionaries and supervised morphological analyzers for the language. Guntara et al. (2020) conducted extensive benchmarks of the current state of Indonesian-English NMT, evaluating performance across a multitude of domains such as news, religious text, and conversational text. Ariesandy et al. (2020) extended their work to improve translation for colloquial Indonesian to English by constructing a synthetic training corpus machine-translated from formal Indonesian.

3 PRPE

As suggested by its name, Prefix-Root-Postfix-Encoding (PRPE) separates a word into three main parts, a prefix, a root, and a postfix. Postfixes can be further broken down into a suffix and an ending. Instead of the character pairs of BPE, left and right substrings are used for segmentation. Left substrings are considered as potential prefixes, while right substrings are considered potential postfixes. Similar to BPE, PRPE is also split into a learning phase and application phase, the former of which Zuters et al. (2018) outlines as four main steps:

1. Collect the frequency of left and right substrings for each word.

2. Treat left substrings as potential roots and align them with the middle part of the word to extract potential prefixes.

3. Treat right substrings as potential roots and align them with the middle part of the word to extract potential postfixes.

4. Use obtained prefixes and postfixes to extract roots from left substrings by aligning them with the middle part of the word.

The learning phase of PRPE can also take in a heuristic to help it determine whether a subword unit is a good candidate for a certain affix type. In other words, a set of hyperparameters that determine the threshold for affix candidacy. For example, an English heuristic could help identify *statistically probable prefixes* by using a list of known *linguistic prefixes* (such as pre- or non-) and some maximum character length. A left substring would be considered a potential prefix if it is found in the list or is within the maximum character length. We apply the same principle when creating the Quechua and Indonesian heuristics by using a list of common prefixes and suffixes, obtained from Muysken (1988), Kinti-Moss and Perkins (2012), and Ortega et al. (2021) for Quechua and IndoDic [1] for Indonesian. The exact heuristic implementations

[1] http://indodic.com/affixeng.html

can be found online [2]. This approach is cross-dimensional, and thus makes the algorithm easily extendable to other languages due to the large amount of affix information publicly available.

This learning phase generates 6 files that are used during the application phase: ranked lists of prefixes, roots, postfixes, suffixes, endings, and words that the algorithm has learned. A word is segmented by generating all possible segmentations and choosing the highest ranked candidate. It is important to reiterate that PRPE is not attempting true morphological segmentation of a word into its linguistic morphemes, but rather into subword units that it deems statistically likely to be prefixes, roots, postfixes, suffixes, and endings. This constitutes the "morphologically guided" portion of PRPE, as it allows for sub-word tokenization that resembles morphologically motivated segmentation.

3.1 Additional Segmentation Methods

Different methods to implement PRPE into the larger pipeline were tested. One such method, denoted as PRPE+BPE, originated from Zuters et al. (2018). This technique is derived from the idea that we can obtain more accurate sub-word tokenization if the corpus is already segmented with a morphologically-driven heuristic. The implementation is simple, we first segment the corpus with PRPE. Then, we segment the PRPE-segmented corpus again using BPE.

We also devised Multi-PRPE - a segmentation method where PRPE is iteratively run n times, feeding the segmented text of each run as input to the next iteration. The intuition was that the rigid nature of PRPE (only one of each affix type) may not provide accurate segmentation for highly agglutinative languages. By running multiple times we continually break off affixes allowing new ones, should they exist, to be segmented off the root.

During development, we conducted a brief analysis of segmentation results of randomly sampled words from the training corpora for the sake of testing the different segmentation implementations. Comprehensive morphological analysis of the segmented text remained outside the scope of this paper. The segmentation methods used in the study were BPE (Sennrich et al., 2015) , SentencePiece Unigram (Kudo and Richardson, 2018), PRPE, PRPE+BPE, and Multi-PRPE (for $n = 2$, $n = 5$, and $n = 8$). The segmentation results were evaluated against morphological analyzers as gold standards due to the lack of labeled segmentation data, an analyzer by Rios Gonzales and Castro Mamani (2014) for Quechua and the MALINDO Morph analyzer (Nomoto et al., 2018) for Indonesian. Compared to the other methods, Multi-PRPE ($n = 5$) appeared to best match the gold standard across both languages. For example, in Table 1 PRPE segments the Quechua word *kausashanchej* as *kausashanchej* (no change), while Multi-PRPE segments it as *kausa - sha - nchej*: the exact output of the morphological analyzer.

Method	**Quechua Sample** 1	**Quechua Sample 2**	**Indonesian Sample 1**
Unsegmented	kausakusunman	kausashanchej	kebencian
BPE	kausa - kusunman	kausash - anchej	kebencian
Unigram	kausaku - sunman	kausasha - nchej	kebencian
PRPE	kausakusun - man	kausashanchej	ke - benci - an
PRPE+BPE	kausa - kusun - man	kausa - shanchej	ke - benci - an
Multi-PRPE	kausakusun - man	kausa - sha - nchej	ke - benc - i - an
Analyzer	kausa - ku - sun - man	kausa - sha - nchej	ke - benci - an

Table 1: Sample Segmentation Results

[2] https://github.com/wanchichen/morphological-nmt

4 Datasets

The main dataset used for analyzing performance was the JW300 texts (Agić and Vulić, 2019) from the Opus corpus (Tiedemann, 2012), comprised of Jehovah's Witness scripture across a variety of languages. Despite the dataset's domain specific content, it is also one of the largest parallel texts publicly available for Quechua. For the Quechua-Spanish pair, we used the version of the dataset made publicly available by Ortega et al. (2021), already split into 17,500 training sentences, 2,500 validation sentences, and 5,585 test sentences. For Indonesian-English we use the an altered version of JW300 provided by Guntara et al. (2020), which also includes Bible and Quran texts gathered from Bible-Uedin and Tanzil respectively (Christodouloupoulos and Steedman, 2015). This dataset is split into 579,544 training sentences, 5,000 validation sentences, 4,823 test sentences. We denote both of these as the Religious datasets.

To ascertain the effects of PRPE outside of the religious domain, we also conduct experiments using general language corpora. We include two different general language datasets for Quechua-Spanish. The first is comprised of financial news articles from DW News, originally created by Rios (2016) for statistical machine translation. However, it was only available as parallel parse trees in an XML format, rendering it unusable for NMT models trained on parallel plaintext. We manually align and clean the raw source text data, filtering out uncertain alignments. The entire cleaned 2,018 line corpus is denoted as the Financial dataset. The second corpus used was the 100 sentence Magazine dataset created by Ortega et al. (2021). Both general-language Quechua-Spanish datasets were used solely for testing due to their small size.

For the Indonesian-English pair, despite having access to much larger general language datasets, we chose to use the low-resource News dataset created by Guntara et al. (2020). This is because of the relatively large size of the Religious Indonesian-English dataset compared to many truly low-resource languages (although it is still substantially smaller than most high-resource language datasets); we wanted to examine the benefits of PRPE in a very low-resource setting for both language pairs. The News dataset is split into 38,469 training sentences, 1,953 validation sentences, and 1,954 test sentences.

5 Experimental Setup

We separate our experiments into two stages: development and testing. We use the development stage to experiment with different model architectures and parameter settings, the best performing of which were carried over to the testing stage. In the testing stage, we evaluate the models on both in-domain and out-of-domain corpora after they were trained on a specific dataset.

Our pipeline includes two pre-processing steps: tokenization and subword segmentation. Tokenization is done using Moses tokenizer (Koehn et al., 2007), demonstrated by Domingo et al. (2018) to be effective in translation tasks. Segmentation is done solely on the source language (either Indonesian or Quechua) text, using one of the methods described above: PRPE, Multi-PRPE, and PRPE+BPE. The target language is left unsegmented. To establish a baseline for comparison, we also conduct experiments with unsegmented text and text segmented with BPE and a unigram language model. We use the default SentencePiece vocabulary size of 8000 for all segmentation methods.

We use the sacreBLEU (Post, 2018) implementation of BLEU (Papineni et al., 2002) as our primary metric for evaluation of translation quality to allow for a comparison with other studies. Ortega et al. (2021) used BLEU to evaluate their NMT systems and found it to be correlated with human judgement for the Quechua-Spanish translation direction. Guntara et al. (2020) used BLEU scores from sacreBLEU to benchmark Indonesian-English translation.

5.1 Development Stage

Development was done on the Religious and News datasets for Indonesian-English, and solely on the Religious dataset for Quechua-Spanish. All datasets are used in their original data splits to allow for comparisons with Guntara et al. (2020) and Ortega et al. (2021). Our models were trained and evaluated in OpenNMT (Klein et al., 2017). It provides a variety of encoder and decoder types, however we focus on LSTM (Hochreiter and Schmidhuber, 1997) and Transformer (Vaswani et al., 2017) for better comparisons with previous works.

Most settings from OpenNMT were kept default to allow for comparisons with the results of Ortega et al. (2021), who used default parameters. The Transformer configuration was 6 encoder-decoder layers and 8 attention heads with size 512 word embeddings, a feed-forward network size of 2048, a learning rate of 2, a dropout of 0.1, and the ADAM optimizer (Kingma and Ba, 2014). These were obtained from the recommended OpenNMT Transformer settings also used by Ortega et al. (2021). The default configuration for LSTM is 2 layers with 500 hidden units, a learning rate of 1, a dropout of 0.3, and stochastic gradient descent as the optimizer.

The only parameters changed throughout development were batch size and training step count. The recommended 4096 batch size for the Transformer model resulted in poor performance (regardless of segmentation method used) and was adjusted to the default value of 64. Due to memory constraints, sentences longer than 50 tokens were filtered out when training on the Religious Indonesian-English corpus. We also found the default training step count of 100,000 to be unsuitable for the smaller training sets: the Quechua-Spanish Religious set and the Indonesian-English News set. We instead used values of 20,000 for Transformer and 60,000 for LSTM. Continued training beyond these values led to over-fitting: steady increases in validation perplexity and no increase in validation accuracy.

5.2 Testing Stage

The best performing models during development for each segmentation method on each validation set were carried over for evaluation on the testing set and out-of-domain corpora. During the development stage, translation models trained and tested on text trained with PRPE+BPE or Multi-PRPE performed consistently worse than models developed on text segmented solely with PRPE, although they still produced incremental to moderate gains over the baseline comparisons. Studying the segmented text led us to suspect that this was due to over-segmentation with regards to the task of translating to an unsegmented target language. As such, these segmentation methods were not included in experiments during the testing phase.

Models were tested on both in-domain and out-of-domain text. Quechua-Spanish models were evaluated using the Financial and Magazine corpora, as well as the testing split of the Religious text. Indonesian-English models were tested on both Religious and News text.

6 Results

BLEU scores from the development stage (Table 2) were encouraging, with the standalone PRPE algorithm outperforming the other segmentation algorithms in most instances, gaining as much as 1.9 BLEU compared to the next best method. Notably, the inclusion of PRPE significantly outperformed previous benchmarks on the Indonesian-English Religious validation set (26 BLEU vs 22.5 BLEU), which was established by Guntara et al. (2020) with a model pretrained with Masked Language Modeling (Devlin et al., 2019) and Translation Language Modeling (Lample and Conneau, 2019) using a training corpus of over 12.9 million non-segmented parallel sentences. A model with language pre-training from Guntara et al. (2020) trained on the exact same Religious training set as ours obtained a BLEU score of 20.2.

Perhaps most interestingly, the models trained on the smaller two training sets (the Quechua-Spanish Religious set and the Indonesian-English News set) consistently yielded better performance when trained on the LSTM architecture compared to the Transformer no matter the segmentation method used, with differences as high as 2.6 BLEU. Ortega et al. (2021) also observed the same pattern in their experiments with Quechua-Spanish translation. These results may suggest that the best performing architectures and techniques in high-resource settings may not be transferable to low-resource translation. We leave further evaluation to future studies. As such, the LSTM models trained on the Quechua-Spanish Religious set and the Indonesian-English News set, instead of their Transformer equivalents, were carried over for testing.

QZ-ES (Religious Validation Set)						
Architecture	Segmentation Method					
	None	BPE	Unigram	PRPE	PRPE+BPE	Multi-PRPE
LSTM	21.7	21.5	22.2	**23.7**	22.9	22.2
Transformer	20.24	19.74	21.1	21.8	20.27	21.03

ID-EN (Religious Validation Set)						
Architecture	Segmentation Method					
	None	BPE	Unigram	PRPE	PRPE+BPE	Multi-PRPE
LSTM	12.2	9.9	22.1	23.8	10.5	22.25
Transformer	19.8	18.7	23.4	**26**	22.4	24.1

ID-EN (News Validation Set)						
Architecture	Segmentation Method					
	None	BPE	Unigram	PRPE	PRPE+BPE	Multi-PRPE
LSTM	9.4	10.1	10.4	**12.2**	10.4	11.9
Transformer	9.8	9.3	9.9	10.1	9.4	9.8

Table 2: BLEU scores in the development stage. Multi-PRPE was run using *n* = *5* iterations. Models generally performed better when text is segmented with some implementation of PRPE.

QZ-ES						
Test Set	Segmentation Method					
	None	BPE	Unigram	PRPE	BPE-Guided*	BPE-Guided (qz-fi)-es*
Religious	20.14	16.5	20.48	**23.4**	17	22.5
Financial	**1.72**	1.06	0.76	1.4	N/A	N/A
Magazine	0.5	0.2	0.56	0.6	0.5	**0.7**

Table 3: In-domain and out-of-domain BLEU scores in the test stage for Quechua-Spanish. All models were LSTMs trained on the Religious training set. *Additional results were included from Ortega et al. (2021) for comparison.

In-domain testing results for Quechua-Spanish were consistent with the development stage with PRPE outperforming other segmentation methods (Table 3). Especially exciting was the PRPE model out-performing models enhanced with transfer learning from Finnish. Quechua-Finnish-Spanish models from Ortega et al. (2021) obtained BLEU scores of 22.9 and 22.5 with BPE and BPE-Guided respectively, whereas the PRPE model obtained a score of 23.4. However, out-of-domain performance for Quechua-Spanish models remained poor, similar to results obtained by Ortega et al. (2021). Segmentation with a unigram language model, the best performing baseline during development, performed especially poorly in out-of-domain evaluation. PRPE had better results than the other segmentation methods in Table 3, but was unable to outperform unsegmented data on the Financial set and transfer learning on the Magazine set.

ID-EN (Religious)				
Test Set	Segmentation Method			
	None	BPE	Unigram	PRPE
Religious	18.5	19.1	20.11	**24.6**
News	10.48	9.8	10.77	**11.47**

Table 4: In-domain and out-of-domain BLEU scores in the test stage for Indonesian-English for Transformer models trained on the Religious dataset.

ID-EN (News)				
Test Set	Segmentation Method			
	None	BPE	Unigram	PRPE
Religious	6.7	6.11	6.44	**7**
News	9.2	9.1	9.5	**10.8**

Table 5: In-domain and out-of-domain BLEU scores in the test stage for Indonesian-English for LSTM models trained on the News dataset.

PRPE performed well during both in-domain and out-of-domain testing for the Indonesian-English pair . In-domain results for the Religious dataset (Table 4) were especially strong, again out-performing a model with language pre-training and a much larger training corpus (24.6 BLEU for PRPE vs 22.1 BLEU from Guntara et al. (2020)). Results for in-domain News (Table 5) and out-of-domain evaluation (Tables 4 and 5) showed much more moderate improvements. A notable result was the poor performance of BPE: it performed worse than no segmentation by producing the lowest scores for 6/7 test sets. This result was surprising given its frequent use in NMT, although still somewhat expected given similar results obtained by Ortega et al. (2021).

Encouraged by the results on the Indonesian-English Religious set, we set up additional experiments using PRPE in an effort to match the Google Translate benchmarks obtained by Guntara et al. (2020) on the validation set (test set scores were not available), which obtained a BLEU score of 29.1 (Table 6). We added the high-resource 1.8 million sentence General dataset from Guntara et al. (2020) as additional training data. Fine-tuning parameters on the validation set, such as increasing word embedding size to 800 from 512 and increasing training steps to 200,000, led to our maximum score of 27.2 BLEU on the validation set and 25.6 BLEU on the testing set (Table 6). With these scores, our PRPE system outperformed the best results of 22.5 validation BLEU and 22.1 test BLEU obtained by Guntara et al. (2020), which was a masked language modelling pre-trained Transformer trained on much more data (10.1 million more sentences in addition to the General and Religious datasets), while performing almost comparably with Google Translate.

ID-EN			
Religious Dataset	Model		
	Our System	Guntara et al. (2020)*	Google Translate*
Validation	27.2	22.5	29.1
Test	25.6	22.1	N/A

Table 6: BLEU scores on the validation and test splits of the Religious set, after adding in additional training data to our system. Our system was a Transformer trained on PRPE-segmented-text from the Religious and General datasets. *Additional results were included from Guntara et al. (2020) for comparison.

Across all datasets, subword segmentation via PRPE consistently improved translation quality over other systems, whether it be unsegmented text or other segmentation methods. Our results demonstrated the importance of selecting suitable subword segmentation methods for low-resource translation. While out-of-domain evaluation remains a challenge for NMT systems, our experiments show that appropriate segmentation techniques can still lead to moderate gains in terms of BLEU. Results are more exciting for in-domain translation, as we show with PRPE that the same segmentation techniques can significantly improve translation quality in place of additional training data, making them especially useful in these low-resource settings.

7 Limitations

While we show that PRPE was able to bring substantial gains in translation quality, there are still constraints that limit its applicability. The most immediate is the pre-requisite for some amount of linguistic knowledge (a list of affixes) during the construction of its heuristics due to its semi-supervised nature. An extension of this limitation is that heuristics are thus language specific, making it less applicable in cross-lingual scenarios (although a multilingual heuristic could be developed to alleviate this problem). Finally, the effectiveness of the algorithm on non-agglutinative languages is unclear. While Zuters et al. (2018) showed that PRPE brought incremental improvements in BLEU score for the non-agglutinative English-Latvian pairing, their experiments also had text further segmented with BPE (which was shown in our results to decrease the benefits of PRPE for agglutinative languages).

8 Conclusion

We introduced the use of the PRPE algorithm for morphologically-guided subword segmentation and evaluate it on two distinct low-resource, agglutinative languages: Quechua and Indonesian. During the development of our experiments, we reintroduced datasets previously unavailable in parallel plaintext for NMT by manually re-aligning raw source data. We found that subword segmentation can have an especially large impact on low-resource translation; unsuitable segmentation methods can actually lower BLEU score when compared to unsegmented text, while effective segmentation can produce moderate to large gains. Our results show that segmentation using PRPE can lead to significant improvements in translation quality when evaluated via BLEU score, out-performing pre-trained, higher-resource models, making the algorithm ideal for low-resource languages that lack the large amounts of training data often necessary for neural machine translation.

References

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

Ariesandy, A. S., Amien, M., Aji, A. F., and Prasojo, R. E. (2020). Synthetic source language augmentation for colloquial neural machine translation.

Bostrom, K. and Durrett, G. (2020). Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624, Online. Association for Computational Linguistics.

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

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Domingo, M., Garcıa-Martınez, M., Helle, A., Casacuberta, F., and Herranz, M. (2018). How much does tokenization affect neural machine translation? *arXiv:1812.08621. Version 4.*

Guntara, T. W., Aji, A. F., and Prasojo, R. E. (2020). Benchmarking multidomain English-Indonesian machine translation. In *Proceedings of the 13th Workshop on Building and Using Comparable Corpora*, pages 35–43, Marseille, France. European Language Resources Association.

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

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *International Conference on Learning Representations, arXiv:1412.6980. Version 4.*

Kinti-Moss, N. and Perkins, J. (2012). *Imanhalla An Introduction to Quechua*. University of Kansas.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. (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.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation.

Kudo, T. (2018). Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Kudo, T. and Richardson, J. (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.

Lample, G. and Conneau, A. (2019). Cross-lingual language model pretraining.

Larasati, S. D., Kuboň, V., and Zeman, D. (2011). Indonesian morphology tool (morphind): Towards an indonesian corpus. In *International Workshop on Systems and Frameworks for Computational Morphology*, pages 119–129. Springer.

Mager, M., Oncevay, A., Ebrahimi, A., Ortega, J., Rios, A., Fan, A., Gutierrez-Vasques, X., Chiruzzo, L., Giménez-Lugo, G., Ramos, R., Meza Ruiz, I. V., Coto-Solano, R., Palmer, A., Mager-Hois, E., Chaudhary, V., Neubig, G., Vu, N. T., and Kann, K. (2021). Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217, Online. Association for Computational Linguistics.

Muysken, P. (1988). Affix order and interpretation: Quechua.

Nomoto, H., Choi, H., Moeljadi, D., and Bond, F. (2018). Malindo morph: Morphological dictionary and analyser for malay/indonesian. In *Proceedings of the LREC 2018 Workshop "The 13th Workshop on Asian Language Resources*, pages 36–43.

Oncevay, A. (2021). Peru is multilingual, its machine translation should be too? In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 194–201, Online. Association for Computational Linguistics.

Ortega, J., Castro Mamani, R., and Cho, K. (2021). Neural machine translation with a polysynthetic low resource language. *Machine Translation*.

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

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

Richburg, A., Eskander, R., Muresan, S., and Carpuat, M. (2020). An evaluation of subword segmentation strategies for neural machine translation of morphologically rich languages. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, pages 151–155, Seattle, USA. Association for Computational Linguistics.

Rios, A. (2016). A basic language technology toolkit for quechua. *Procesamiento del Lenguaje Natural*, 56:91–94.

Rios Gonzales, A. and Castro Mamani, R. A. (2014). Morphological disambiguation and text normalization for Southern Quechua varieties. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 39–47, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Sennrich, R., Haddow, B., and Birch, A. (2015). Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909.

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *arXiv:1706.03762. Version 5.*

Wang, L., Cao, Z., Xia, Y., and De Melo, G. (2016). Morphological segmentation with window lstm neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Zuters, J., Strazds, G., and Immers, K. (2018). Semi-automatic quasi-morphological word segmentation for neural machine translation. In Lupeikiene, A., Vasilecas, O., and Dzemyda, G., editors, *Databases and Information Systems*, pages 289–301, Cham. Springer International Publishing.