The Usefulness of Bibles in Low-Resource Machine Translation

Ling Liu and **Zach Ryan** and **Mans Hulden**
University of Colorado
{ling.liu,zachary.j.ryan,mans.hulden}@colorado.edu

Abstract

Bibles are available in a wide range of languages, which provides valuable parallel text between languages since verses can be aligned accurately between all the different translations. How well can such data be utilized to train good neural machine translation (NMT) models? We are particularly interested in low-resource languages of high morphological complexity, and attempt to answer this question in the current work by training and evaluating Basque-English and Navajo-English MT models with the Transformer architecture. Different tokenization methods are applied, among which syllabification turns out to be most effective for Navajo and it is also good for Basque. Another additional data resource which can be potentially available for endangered languages is a dictionary of either word or phrase translations, thanks to linguists' work on language documentation. Could this data be leveraged to augment Bible data for better performance? We experiment with different ways to utilize dictionary data, and find that word-to-word mapping translation with a word-pair dictionary is more effective than low-resource techniques such as backtranslation or adding dictionary data directly into the training set, though neither backtranslation nor word-to-word mapping translation produce improvements over using Bible data alone in our experiments.

1 Introduction

The Bible has been translated into a wide range of languages, including many low-resource ones, and a significant amount of Bible data is publicly available. For example, www.bible.com has a collection of 2,172 Bible versions in 1,482 languages. This provides valuable parallel text between languages for training machine translation models as well as for conducting other cross-lingual studies.

Neural machine translation (NMT) models have been pushing forward the state of the art in recent years (Bahdanau et al., 2014; Cho et al., 2014; Vaswani et al., 2017; Barrault et al., 2019). However, millions of parallel tokens are usually required in order to train a NMT model with high quality (Koehn and Knowles, 2017; Sennrich and Zhang, 2019), an amount much larger than the whole Bible. For example, the *Holy Bible, New Living Translation* on www.bible.com contains around 880,000 tokens. If we use Bible data to train a NMT model between English and a morphologically complex language like Basque or Navajo, what quality could we achieve and what can we do to improve the performance?

In the current work, we attempt to evaluate the Transformer architecture trained on Bible data for translating between Basque and English, and between Navajo and English. First, we experiment with different tokenization approaches for preprocessing the morphologically complex languages, including only separating the punctuation from the words, dividing words into syllables, and applying the popular byte-pair encoding (BPE) algorithm (Gage, 1994).

A novel result is that we found that data preprocessing with syllabification is promising for morphologically complex languages: preprocessing Navajo with the syllabification method always produces the best performance whether it is translating Navajo into English or the other way around, and it is the method which produces the second highest BLEU score when translating Basque from and into English.

Thanks to linguists' work on language documentation, dictionary data are also available for some endangered languages. Could this data be utilized to augment the Bible data in order to achieve better machine translation quality? To answer this question, we experiment with three dif-

Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages: Vol. 1 Papers, pages 44–50,
Online, March 2–3, 2021.

ferent ways of utilizing dictionary data to train the models. First, we add individual example pairs of sentences or phrases from the dictionary directly to the Bible training dataset. Second, in order to alleviate the cross-domain problem, we also experiment with only adding word pairs to the Bible data for training. Third, inspired by Nag et al. (2020), we experiment with the data augmentation approach of mapping additional English texts into Basque with the word-to-word dictionary data, which are then combined with the Bible data to train the Basque to English NMT model. This word-to-word translation data augmentation approach is compared with the commonly used backtranslation data augmentation method (Sennrich et al., 2016). However, neither the additional dictionary data nor any of the data augmentation method produces improvements in the BLEU score over using only the Bible data in our experiments, though the word-to-word mapping data augmentation method is the best of all the additional data approaches we apply.

2 Experiments

2.1 Data and data preprocessing

The site `www.bible.com` provides Bible texts in different languages, where the Bible text in each language is organized with book-title:chapter-id:verse-id[1] and the identifiers agree with each other between different languages. Therefore, we can create parallel text at the verse level with the identifiers. We experiment with Basque-English and Navajo-English machine translation. For English (`eng`), we use the *Holy Bible, New Living Translation (NLT)*, for Navajo (`nvj`), we use the *Navajo Bible (NVJOB)*, and for Basque (`eus`), we use the *Elizen Arteko Biblia (Biblia en Euskara, Traducción Interconfesional) (EAB)*. This gives us around 31.7k parallel Basque-English verses and around 31.8k parallel Navajo-English verses. More detailed statistics can be found in Table 1. We split the parallel verses into training, development and test sets with a ratio of 7:1:2.

The English data is preprocessed by segmenting punctuation from words. Considering that Basque and Navajo have very rich inflection patterns—for verbs in particular—we experiment with different tokenization methods for these two languages. The same tokenization method as English is used as a baseline, which is referred to as `tok`. In addition to that, we experiment with segmenting Basque or Navajo words by syllables, for which we manually develop finite-state transducers, using *foma* (Hulden, 2009), to break up words into syllables for each language. We refer to this method as `syl`. We also experiment with the byte-pair encoding (BPE) algorithm with a vocabulary size of 8,000 and 16,000 respectively, which are referred to as `bpe-8k` and `bpe-16k`.

		verse	token
bible	eus - eng	31.7k	eus 609.7k eng 836.8k
	nvj - eng	31.8k	nvj 716.1k eng 844.3k
dict	eus - eng	14.9k	eus 55.1k eng 76.7k
	nvj - eng	14.8k	nvj 43.0k eng 87.0k
w2w dict	eus - eng	80.6k word pairs	

Table 1: Data information

2.2 Syllabifier details

As in Agirrezabal et al. (2012) we treat Basque syllabification as following the maximum onset principle and a standard sonority hierarchy. Such syllabifiers can be built with a finite-state transducer (FST) constructed with a single rewrite rule which inserts syllable boundaries after legal syllables following a leftmost-shortest strategy (Hulden, 2005). Leftmost-shortest rewrite rule compilation is implemented in many standard FST toolkits such as *foma* (Hulden, 2009), the Xerox tools (Beesley and Karttunen, 2003), or *Kleene* (Beesley, 2012). Navajo has a relatively simple syllable structure, disallowing onsetless syllables (McDonough, 1990), and can thus be divided through a rule that always places the syllable boundaries before any CV sequence. Both languages have digraphs, such as Basque **tx** (=[t͡ʃ]), while Navajo also features trigraphs for ejectives such as **ch'** (=[t͡ʃ']), all of which need to be modeled as single consonants in the FST. Figure 1 shows example outputs of our syllabifiers in both languages, and Figure 2 shows the essential parts of the FST construction in *foma* code.

2.3 Model specifics and evaluation

For the neural machine translation model, we employ the self-attention Transformer architec-

[1] Each verse may or may not be one sentence, depending on the translation.

	Orthography	Syllabified
Basque	batzuetan	ba@ @tzu@ @e@ @tan
	oraindik	o@ @rain@ @dik
	Donostiarako	Do@ @nos@ @ti@ @a@ @ra@ @ko
	kotxearekin	ko@ @txe@ @a@ @re@ @kin
	ahizpa	a@ @hiz@ @pa
Navajo	łééchąą'í	łéé@ @chąą@ @'í
	shi'niiłhį	shi'@ @niił@ @hį
	náninichaadísh	ná@ @ni@ @ni@ @chaa@ @dísh
	nahwiilzhooh	na@ @hwiil@ @zhooh
	ch'éénísdzid	ch'éé@ @nís@ @dzid

Figure 1: Example outputs of FST syllabifiers for Basque and Navajo. The output (right) of the FST is in BPE-format with double @-signs for word-internal syllables, and single @-signs at the word edge.

ture (Vaswani et al., 2017) as implemented in the Fairseq toolkit (Ott et al., 2019).[2] The Transformer model we use has 4 encoding layers and 4 attention heads with an embedding dimension of 256 and hidden layer size of 1024. Its decoding layer, attention head and dimensions have the same setting. The model is trained with a batch size of 16 for 120k maximum number of updates. Beam search with a width of 5 is used for generation. More details on the hyperparameters and training heuristics are provided in Appendix A.1. The BLEU score (Papineni et al., 2002) is used as the metric to evaluate the NMT model generation output throughout. Note that in this paper the BLEU score is measured over word overlap, not word-piece overlap.

2.4 Using dictionary data

We experiment with incorporating dictionary data in three different ways. For Basque-English, we use the Elhuyar dictionary,[3] whose phrase examples give us around 14.9k parallel sentences or phrases, and around 80.6k word pairs. For Navajo-English, we extract around 14.8k parallel sentences or phrases from the Young & Morgan dictionary (Young and Morgan, 1987). We omit the word-pair experiments for Navajo.

The first way of incorporating the dictionary data is to add the sentence or phrase pairs from the dictionary to the Bible training set to train the NMT model. We refer to this method as `+dict`. This experiment is conducted for both Basque-English and Navajo-English.

Considering that the nature of the text used in the dictionary example data may be different from the Bible data, which can cause a cross-domain problem, we experiment with adding only word pairs from the dictionary for Basque-English translation, referred to as `+w2w`.

However, it's common in dictionaries that a word in one language is mapped to multiple words in another language or different word in one language is mapped to the same word in another language. For example, we get two English translations `polite` and `kind` for the Basque word `adeitsu` in the dictionary, and Basque words `adabatu`, `adaba` and `adabatzen` are all translated as `to patch` in English without keeping their inflection information. The dictionary lists these different stems to inform the reader what stem alternations occur in different verb inflections. To avoid such ambiguity, we experiment with randomly picking one word pair when multiple mappings are found in the dictionary data.[4] This is referred to as `+w2w-rnd` as opposed to `+w2w` which keeps the multiple mappings.

The third way of using the dictionary data is a combination of using dictionary data and monolingual data: We use the word pairs from the dictionary to translate extra monolingual English data (see section 2.5 for details) word by word into Basque, and augment the Bible training set with this translated data. When conducting the word-by-word translation from English to Basque, we randomly pick one translation if the dictionary provides multiple Basque translations for an English word. In cases where an English word does not appear in the dictionary, we copy the English word as Basque translation.

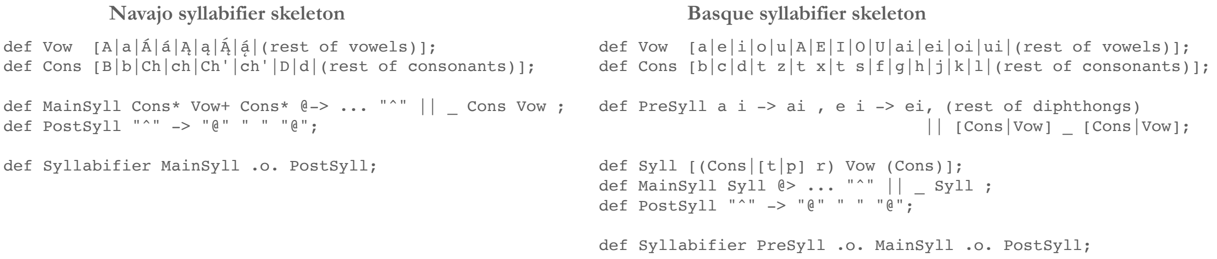
2.5 Using monolingual data

When the task is to translate a low-resource language to a high-resource language, the monolingual data for the high-resource language can be utilized to augment the training data (Przystupa and Abdul-Mageed, 2019; Chen et al., 2020; Edunov et al., 2020). One popular way to do it is through backtranslation (Sennrich et al., 2016). The way backtranslation works is first to train a model with the initial data for translation from the high-resource language to the low-resource language, then to utilize the model from the previous step to translate the monolingual data for the

[2] https://fairseq.readthedocs.io/en/latest/

[3] https://hiztegiak.elhuyar.eus/

[4] The way to choose words in such cases can be improved, for example, by conducting morphological analysis of the words since some of the multiple mappings are due to different inflected forms of the morphologically complex language.



Navajo syllabifier skeleton

```
def Vow  [A|a|Á|á|Ą|ą|Ą́|ą́|(rest of vowels)];
def Cons [B|b|Ch|ch|Ch'|ch'|D|d|(rest of consonants)];

def MainSyll Cons* Vow+ Cons* @-> ... "^" || _ Cons Vow ;
def PostSyll "^" -> "@" " " "@";

def Syllabifier MainSyll .o. PostSyll;
```

Basque syllabifier skeleton

```
def Vow  [a|e|i|o|u|A|E|I|O|U|ai|ei|oi|ui|(rest of vowels)];
def Cons [b|c|d|t z|t x|t s|f|g|h|j|k|l|(rest of consonants)];

def PreSyll a i -> ai , e i -> ei, (rest of diphthongs)
                                 || [Cons|Vow] _ [Cons|Vow];

def Syll [(Cons|[t|p] r) Vow (Cons)];
def MainSyll Syll @> ... "^" || _ Syll ;
def PostSyll "^" -> "@" " " "@";

def Syllabifier PreSyll .o. MainSyll .o. PostSyll;
```

Figure 2: Foma code example for the syllabifier. Some repetitive parts of vowel, consonant and diphthong specifications are omitted for compactness.

high-resource language to the low-resource language, and add the noisy translated data to the original training data to train the model for translation from the low-resource language to the high-resource language. We experiment with the back-translation data augmentation method for Basque to English and Navajo to English translation.

We also experiment with the word-to-word translation by utilizing word pairs from a dictionary (Nag et al., 2020) for Basque to English translation, as described in section 2.4.

The monolingual data we use is a collection of news data from the English Gigaword archive (5th edition) (Parker et al., 2011) (specifically nyt_eng_2010 and wpb_eng_2010), from which we randomly pick one (+1 ⋆) to seven (+7 ⋆) times the number of the Bible training set verses to compare leveraging different amounts of monolingual data.

		$* \rightarrow$ **eng**		
eus → **eng**	tok	syl	bpe-8k	bpe-16k
bible only	8.51	10.74	**11.36**	10.27
+dict	8.49	8.56	9.56	9.37
+w2w	1.34	2.06	2.00	2.07
+w2w-rnd	4.19	3.92	4.41	4.14
nvj → **eng**	tok	syl	bpe-8k	bpe-16k
bible only	8.94	**10.99**	10.91	9.56
+dict	7.27	8.63	8.54	9.02
		eng $\rightarrow *$		
eng → **eus**	tok	syl	bpe-8k	bpe-16k
bible only	**5.31**	5.28	2.65	2.42
+dict	4.36	4.20	1.92	1.89
+w2w	0.64	0.91	0.17	0.10
+w2w-rnd	2.42	1.79	1.02	1.00
eng → **nvj**	tok	syl	bpe-8k	bpe-16k
bible only	7.57	**8.69**	6.75	6.40
+dict	6.49	6.36	5.00	5.28

Table 2: BLEU scores on the test set for translating Basque or Navajo into or from English, with different tokenization methods for Basque and Navajo, and different training data. The English data is always tokenized by separating the word from the punctuation.

3 Results and discussion

Table 2 presents the BLEU scores on the test set for different tokenization methods and training data. When the translation is into English, BPE with a vocabulary size of 8,000 produces the best result for Basque, and syllabification tokenization produces the best performance for Navajo. When the translation is from English, only segmenting the punctuation from the words gives us the best performance for Basque, which is slightly higher than syllabification, and syllabification is the tokenization method which achieve the highest BLEU score for Navajo. As regards adding dictionary data to Bible training set, for the translation in both directions for both Basque-English and Navajo-English, adding additional dictionary data did not improve the BLEU score, though adding example sentence or phrase pairs turns out to be less harmful than adding only word pairs. Adding additional dictionary examples lowers the BLEU score, be it word pairs or sentence or phrase pairs, may be because the text used in the dictionary examples are different from Bible text. One reason why adding word pairs harms the performance so much may be related to the complex morphological inflection of Basque words: the additional dictionary forms of Basque words induce the model to tend to produce dictionary forms rather than particular inflected forms required by the context.

Figure 3(a) plots the performance of the NMT model for Basque to English translation when we augment the Bible training data by word-to-word translating different amounts of English news data to Basque. It shows that the performance is worse than using the Bible training data alone (Except that for `bpe-16k` tokenization, adding 1 time word-to-word translated data outperforms using

Bible training set alone, though the BLEU score is still lower than using Bible training set alone with bpe-8k tokenization), and the more translated data is added, the worse the performance becomes. One reason for the negative effect of leveraging monolingual data here may be that the news text and Bible text are quite different. Another reason may be that the dictionary data is too limited: around 43% of the English tokens can't be found in the word pairs we extract from the dictionary, which influences the translation quality. Other reasons may be that the word-to-word mapping translation does not inflect the Basque words, which is critical for a morphologically complex language, or that the randomly picked words do not match the context well. Therefore, though data augmentation by word-to-word mapping translation of monolingual data has been shown to be helpful in the literature (Nag et al., 2020), our experiments indicate that the helpfulness may depend on the size and quality of the word-to-word dictionary and the morphological complexity as well as the word ambiguity of the languages involved.

Our backtranslation results, shown in Figure3 (b) and (c), support the widely recognized fact about backtranslation that the quality of high-resource language to low-resource language translation model is positively related to backtranslation quality (Currey et al., 2017). When the high-resource to low-resource language translation model is too poor, adding backtranslated data actually creates more noise, and thus hurts performance. We see in our results that adding backtranslated data does not improve BLEU scores for Basque or Navajo over using Bible data alone, and we observe that the Basque-English results with backtranslation is even worse than adding the word-to-word translated data, indicating the poor quality of the backtranslated data.

4 Conclusion

We have performed a systematic evaluation on using Bible data for machine translation in two highly morphologically complex languages. The overarching result is that—while even unsupervised MT has been shown possible in some cases (Artetxe et al., 2018; Lample et al., 2017)—using only the Bible, together with possibly a phrase or word dictionary and standard tools of the trade such as backtranslation even with state-of-the-art sequence-to-sequence models is unlikely to produce very useful MT quality. The best BLEU score is only over 11 when the translation is into English and lower than 9 when the translation is out of English. Future work is needed to improve the machine translation quality in this scenario.

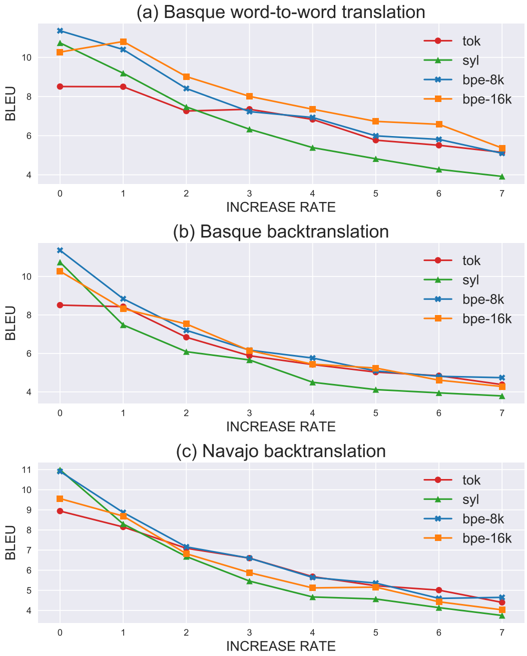


Figure 3: Comparison of $* \rightarrow English$ translation after adding different amounts of monolingual data. X-axis indicates the augmentation rate: 0 means using only Bible training set, 1 means adding 1 time translated data that of the Bible training set verses, etc.

An important finding in this paper is that syllabification as opposed to BPE or other sub-word tokenization methods may be helpful for morphologically complex languages like Basque and Navajo. We also experimented with leveraging dictionary data to increase the training data for translation in both directions or augmenting the training data for low- to high-resource language translation by utilizing monolingual data for the high-resource language, though we did not achieve improvements in the BLEU score over using Bible data alone. Data augmentation by word-to-word mapping translation of monolingual data with word pairs obtained from dictionaries outperforms backtranslation, but still did not achieve better performance than using only Bible training set. The word-to-word mapping augmentation approach can be improved by converting the mapped words to their correct inflected forms or selecting more context-appropriate candidates when multiple mappings are possible, which can be explored in future work.

References

Manex Agirrezabal, Iñaki Alegria, Bertol Arrieta, and Mans Hulden. 2012. Finite-state technology in a verse-making tool. In *Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing*, pages 35–39, Donostia–San Sebastián. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, Brussels, Belgium. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

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.

Kenneth R. Beesley. 2012. Kleene, a free and open-source language for finite-state programming. In *Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing*, pages 50–54, Donostia–San Sebastián. Association for Computational Linguistics.

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Publications, Stanford, CA.

Peng-Jen Chen, Ann Lee, Changhan Wang, Naman Goyal, Angela Fan, Mary Williamson, and Jiatao Gu. 2020. Facebook AI's WMT20 news translation task submission. In *Proceedings of the Fifth Conference on Machine Translation*, pages 113–125, Online. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156, Copenhagen, Denmark. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2020. On the evaluation of machine translation systems trained with back-translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2836–2846, Online. Association for Computational Linguistics.

Philip Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

Mans Hulden. 2005. Finite-state syllabification. In *International Workshop on Finite-State Methods and Natural Language Processing (FSMNLP)*, pages 86–96. Springer.

Mans Hulden. 2009. Foma: a finite-state compiler and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 29–32. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Joyce M. McDonough. 1990. *Topics in the phonology and morphology of Navajo verbs*. Ph.D. thesis, University of Massachusetts, Amherst.

Sreyashi Nag, Mihir Kale, Varun Lakshminarasimhan, and Swapnil Singhavi. 2020. Incorporating bilingual dictionaries for low resource semi-supervised neural machine translation. *arXiv preprint arXiv:2004.02071*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English gigaword fifth edition. *Linguistic Data Consortium*. Version 5.

Michael Przystupa and Muhammad Abdul-Mageed. 2019. Neural machine translation of low-resource and similar languages with backtranslation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 224–235, Florence, Italy. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Robert W. Young and William Morgan. 1987. *The Navajo language: A grammar and colloquial dictionary*. University of New Mexico Press, Albuquerque, NM.

A Supplemental Material

A.1 Hyperparameters

Here lists the hyperparameters we use for the Transformer model:

UNK threshold = 1,
encoder/decoder embedding dimension = 256,
encoder/decoder hidden layer size = 1024,
encoder/decoder number of layers = 4,
encoder/decoder number of attention heads = 4,
dropout = 0.3,
batch size = 16,
maximum updates = 120k,
warmup update = 4000,
learning rate = 0.001,
label smoothing = 0.1,
clip-norm = 1.0,
optimization function: adam,
adam-betas = (0.9, 0.98),
activation function: ReLU,
loss function: label smoothed cross entropy,
beam search for generation with width of 5.

A.2 Data augmentation results

Here are the results for adding monolingual data with word-to-word mapping translation by dictionary pairs and with backtranslation. These are the results used to created Figure 3.

eus → eng	tok	syl	bpe-8k	bpe-16k
+1*	8.50	9.19	10.40	10.81
+2*	7.26	7.47	8.41	9.02
+3*	7.35	6.33	7.23	8.01
+4*	6.83	5.39	6.93	7.35
+5*	5.77	4.82	5.99	6.73
+6*	5.51	4.28	5.81	6.58
+7*	5.15	3.92	5.10	5.37

Table 3: BLEU scores on the Basque test sets when models are trained on **Bible training set plus word-to-word mapping translated English news data**.

eus → eng	tok	syl	bpe-8k	bpe-16k
+1*	8.43	7.48	8.84	8.32
+2*	6.84	6.09	7.20	7.53
+3*	5.88	5.66	6.17	6.15
+4*	5.42	4.50	5.76	5.44
+5*	5.03	4.12	5.09	5.24
+6*	4.84	3.95	4.81	4.61
+7*	4.38	3.79	4.74	4.28
nvj → eng	tok	syl	bpe-8k	bpe-16k
+1*	8.15	8.30	8.87	8.69
+2*	7.09	6.68	7.16	6.82
+3*	6.60	5.46	6.60	5.88
+4*	5.68	4.67	5.64	5.12
+5*	5.23	4.57	5.36	5.16
+6*	5.01	4.14	4.60	4.43
+7*	4.40	3.75	4.65	4.03

Table 4: BLEU scores on the test sets for Basque and Navajo when models are trained on **Bible training set plus backtranslated English news data**.