Expanding the JHU Bible Corpus for Machine Translation of the Indigenous Languages of North America

Garrett Nicolai, Edith Coates, Ming Zhang and **Miikka Silfverberg**
Department of Linguistics
University of British Columbia
Vancouver, Canada
garrett.nicolai@ubc.ca, ecoates.bc@gmail.com,
mingz961018@gmail.com, miikka.silfverberg@ubc.ca

Abstract

We present an extension to the JHU Bible corpus, collecting and normalizing more than thirty Bible translations in thirty Indigenous languages of North America. These exhibit a wide variety of interesting syntactic and morphological phenomena that are understudied in the computational community. Neural translation experiments demonstrate significant gains obtained through cross-lingual, many-to-many translation, with improvements of up to 8.4 BLEU over monolingual models for extremely low-resource languages.

1 Introduction

In 2019, Johns Hopkins University collated a corpus of translations of the Christian Bible in more than 1500 languages - the largest such corpus ever collected (McCarthy et al., 2020). Its parallel structure allows for significant experimentation in cross-lingual and data augmentation methods, and provides data for many underserved languages of the world. However, even at its impressive size, the corpus only represents roughly 20% of the world's languages, and is relatively sparse in the Indigenous languages of North America. Despite Ethnologue listing 254 living languages on the continent, the corpus only contains translations for 6 of them.

In this paper, we describe an extension of the JHU Bible corpus - namely, the addition of translations in 24 Indigenous North American languages, and new translations in six more.[1] Our work continues a tradition of expanding Bible corpora to be more inclusive – Resnik et al. (1999)'s 13 parallel languages grew into Christodouloupoulos and Steedman (2015)'s 100. Mayer and Cysouw (2014) established a corpus that eventually grew to 1556 Bibles in 1169 languages (Asgari and Schütze, 2017), which was then subsumed by the 1611 language JHUBC (McCarthy et al., 2020).

Beyond contributing an important linguistic resource, our work also allows for development of computational tools for North American Indigenous languages - an important step in increasing the global presence of the language communities. We demonstrate the usefulness of our Indigenous parallel corpus by building multilingual neural machine translation systems for North American Indigenous languages. Multilingual training is shown to be beneficial especially for the most resource-poor languages in our corpus which lack complete Bible translations.

2 Corpus Construction

The Bible is perhaps unique as a parallel text. Partial translations exist in more languages than any other text (Mayer and Cysouw, 2014).[2] Furthermore, for nearly 500 years, the Bible has had a canonical hierarchical structure - the Bible is made up of 66 *books*, each of which contains a number of *chapters*, which are, in turn, broken down into *verses*. Each verse corresponds to a short segment – often no more than a sentence. Bible translations preserve this structure as much as possible, meaning that translations are much easier to parallelize than typical texts.

The first step in collecting Indigenous translations of the Bible is identifying existing translations. After first creating a list of Indigenous languages of North America, we searched existing Bible corpora online to obtain translations in as many languages as possible. For the majority of the collected Bibles, we obtained complete New Testament translations - consisting of 27 books of varying lengths. An additional 5 languages also contain complete Old Testament translations. The full list of languages is given in Table 1 and all the corpus data are available upon request. We emphasize that even incomplete translations - such as Siksika, which only has 2 translated books, are useful, particularly when they are in a parallel format with other related languages. Even a single book will typically contain a few hundred verses, which while small, can still be informative.

2.1 Sources

We collected Bibles from a variety of freely accessible online sources[3]: The Canadian Bible Society

[1] The corpus is available by request at https://github.com/GarrettNicolai/FirstNationsBibles

[2] Save, perhaps, the Universal Declaration of Human rights, which is much shorter.

[3] Most of the data we use are not in the public domain but our work falls under the fair use doctrine of North American copyright law.

Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages: Vol. 1 Papers, pages 1–5,
Online, March 2–3, 2021.

Family	Language	NT	OT	Books
Algic	Algonquin	Yes	No	27
Algic	Arapaho	No	No	1
Algic	Cree	Yes	No	27
Algic	Mikmaq	Yes	No	27
Algic	Moose Cree*	Yes	No	27
Algic	Naskapi	Yes	No	30
Algic	North-Eastern Ojibwa	Yes	No	41
Algic	Northern East Cree	No	No	8
Algic	Potawatomi	No	No	2
Algic	Siksika	No	No	2
Algic	Southern East Cree	Yes	No	27
Algic	Western Cree	Yes	Yes	67
Athabaskan	Carrier+	Yes	No	28
Athabaskan	Dane-Zaa	No	No	1
Athabaskan	Dogrib	Yes	No	28
Athabaskan	Gwich'in	Yes	No	27
Athabaskan	Navajo	Yes	Yes	67
Athabaskan	Western Apache	Yes	No	27
Athabaskan	Southern Carrier	Yes	No	27
Athabaskan	Tlicho	Yes	No	27
Athabaskan	Tsilqot'in	No	No	1
Haida	Haida	No	No	4
Inuit-Aleut	Central Alaskan Yupik	Yes	Yes	67
Inuit-Aleut	Central Siberian Yupik*	Yes	No	27
Inuit-Aleut	Inuinnaqtun	No	No	5
Inuit-Aleut	Inupiatum	Yes	No	27
Inuit-Aleut	Inuktitut+	Yes	Yes	67
Inuit-Aleut	Inuttitut	Yes	Yes	67
Iroquoian	Cherokee'	Yes	No	27
Tanoan	Tewa	No	No	17
Uto-Aztecan	Northern Paiute*	Yes	No	27
Zuni	Zuni	No	No	2

Table 1: Language Statistics of collected Bibles. * indicates new translations of languages that were in the JHUBC, while + indicates more complete translations.

(CBS) (biblesociety.ca) works to distribute the Bible to people in Canada and abroad and, therefore, also seeks to create and share translations of the Bible into Indigenous languages of Canada. Scripture Earth (SE) (scriptureearth.org) is a website sponsored by Wycliffe Canada (wycliffe.ca) with a mission statement to facilitate Bible translation among minority linguistic communities. Bible.com (bible.com) is an online Bible platform featuring Bibles for some 1200 languages, including several Indigenous languages of the Americas. The Digital Bible Society (dbs.org) and GospelGo (gospelgo.com) provide digital platforms for accessing Bibles in several languages.

2.2 Corpus Statistics

We extend the JHUBC by 24 languages in 8 language families (including 2 isolates), with new translations in an additional 6 languages. The breakdown of language families is illustrated in Table 1.

In Table 2, we demonstrate the type-to-token ratios for each language family in our corpus. We only include languages for which we have at least the New Testament, taking the largest translation that we have; we then average (weighted by number of verses) over each language family. A high TTR typically indicates a language with significant morphological productivity. As can be seen, the Indigenous languages in the corpus display high degrees of morphological productivity. Even the family having the lowest TTR, Uto-Aztecan still has four times as many types as English, and the Inuit-Aleut family, well-remarked for exhibiting productive synthetic morphology, will have 18 times the number of unique types as an English text of the same size.

Family	Average # of verses	Weighted TTR
Algic	13107	12.74
Athabaskan	12568	9.12
Inuit-Aleut	26458	36.92
Iroquoian	7957	22.04
Uto-Aztecan	7959	8.04
English	31088	2.22

Table 2: Weighted Type-to-Token ratios of collected language families.

The languages that we collect exhibit a wide range of interesting linguistic phenomena. Several of the languages are predominantly SVO languages (if all arguments occur in the sentence) (Schmirler et al., 2018) but we also include languages like Haida where SOV constructions are prevalent (Enrico, 2003). We also have examples of both nominative-accusative alignment and ergative-absolutive alignment exemplified by Inuktitut in the Inuit-Aleut family (Nowak, 2011). Additionally, the languages display a large variety of interesting morphological features. We find examples of predominantly suffixing morphology in the Algic languages and extensive use of prefixes encountered in Athabaskan languages. Furthermore, animacy is an important grammatical category which is morphologically marked in Plains Cree (Schmirler et al., 2018) and other Algic languages.

2.3 Verse Splitting

Although Bibles are readily parallelizable in general due to the canonical division into books, chapters and verses, translations sometimes combine several verses into one creating a discrepancy between the verse numbering in different Bible translations. JHUBC follows a convention presented by Mayer and Cysouw (2014): the combined verse is listed as the first verse in the sequence (ie, verse 16, if it spans 16-18), while the other verses are marked as “BLANK”. While reasonable, this convention can result in difficulties for cross-lingual training, as one verse on one side of data aligns with many verses on the other, and many verses must be discarded. We opt for a different approach and instead split combined verses apart. We identify separation points using a mixed Naive-Bayes classifier (Hsu et al., 2008) with two features: punctuation and token ratio. We assume that the relative length of the individual verses is likely to be similar across languages, and calculate the ratio of tokens between individual verses and the combined verse in our English Bible reference. An evaluation on artificially-combined verses demonstrates a macro-averaged F-score of 86% on identifying splitting points when two verses require splitting.

3 Experiments

We conduct a number of neural-MT experiments on the data. We investigate translation quality both for bilingual translation systems and for multilingual systems, while applying a number of variations to the training

```
Bible.Algonquin 2Bible.English apitc mois ka nodag ii , coda8innig ka ...
Bible.Cree 2Bible.English namawiya ēkosi ki ka itota@@ wāw kā tipēyihcikēt ki ...
Bible.Cree 2Bible.English ēkwa māka kiyām kanawāpa@@ mik ; cikēmā namawiya ki ka ...
Bible.English 2Bible.Algonquin when moses went into the tent of meeting to speak ...
```

Figure 1: Example of our training data format for many-to-many NMT experiments. The first symbol on each line (e.g. Bible.Algonquin) gives the language of the current sentence and the second one shows the language of the corresponding target or source sentence. This allows us to use each sentence both in the source and target set.

procedure of the NMT systems in order to improve translation quality. These are described in detail below.

Translation Scenarios We measure translation performance for three language families: the Algic, Athabaskan and Inuit-Aleut families. For each family, we evaluate performance on a few "high-resource" languages[4] which have complete Bible translations. Our high-resource languages are Plains Cree[5] for the Algic family, Navajo (NAV) for the Athabaskan family and Inuktitut (IKU) and Central Alaskan Yupik (ESU) for the Inuit-Aleut family. We also evaluate performance on a single lower-resource language from each family, which only has the NT available. Our lower-resource languages are Miḱmaq (Algic - MIC), Dogrib (Athabaskan - DGR), and Inupiatun (Inuit-Aleut - IKU). All of these translations, except for Inuktitut, are written in modified versions of the Latin script.

For each language family, we train (1) bilingual X-English NMT systems with a single source language X, (2) multilingual Family-English systems where we combine training examples from all the languages in the family into a joint training set, and (3) multilingual many-to-many NMT systems combining both Family-English and English-Family translation tasks for all the languages in the family.

Data Preprocessing We learn a joint Byte Pair Encoding (Sennrich et al., 2016) between source and target, experimenting with two vocabulary sizes: we try both 32,000 and 16,000 merge operations. In multilingual experiments we concatenate source and target language tags to our sentences in order to learn to translate into the appropriate language. Figure 1 shows a few multilingual training examples.

Model Details We use transformer systems for translation and train our models using the Fairseq toolkit (Ott et al., 2019), with 3 encoding and decoding layers, 4 attention heads, an embedding size of 512, and a maximum of 2000 tokens per batch[6]. Models are trained for 100 epochs. We set aside the book of Revelation as an evaluation set: the first 100 verses serve as a validation set, and the final 304 verses form a held-out test set.

Training Settings Preliminary experiments showed that multilingual systems trained on a single target corpus, i.e. the English Bible in our case, have a tendency to completely disregard the source sentence during test time and instead generate an unrelated English sentence as output. We dub this *target overfitting*. To counter this tendency, we employ four specialized training strategies: (1) *Single Source translation* (1Src) limits the number of training source texts to one even when we have multiple Bible translations in the same language [7]. (2) *Heterogeneous batching* (HB) (Aharoni et al., 2019) constructs minibatches by uniformly sampling sentences from the entire training data into each minibatch. In contrast, the common practice is to construct minibatches from training examples with similar length.[8] (3) We increase the amount of English target data available to the model by adding monolingual English training examples where the source and target sentence are identical (E2E).[9] (4) Finally, following Aharoni et al. (2019) we transform our many-to-English models into many-to-many models (M2M) by reversing the source and target language of our Bibles and combining the resulting data with our original training set.

4 Results and Discussion

Table 3 reports the tokenized, lower-case BLEU score for our experiments. Although Inuktitut is written in a different script than English, it translates relatively well – only transliterated Cree obtains a better BLEU score. When we extend our experiments to the entire Inuit-Aleut family, we see modest gains for both the Latin and Non-Latin languages. However, we also note that the translation quality collapses for the other language families. We suspect this may be due to a large BPE vocabulary - the Inuit-Aleut family, containing two scripts, is more likely to split words; the single script Athabaskan and Algic families, on the other hand, can simply memorize entire words, which may

[4] Relatively speaking. Of course all of our languages are low-resource but some still have more available resources than others.

[5] We use a version of the Plain Cree (CRK) Bible which has been transliterated into Latin script.

[6] These settings were established on a similar low-resource corpus

[7] Discussions of dialects and languages aside, we include the largest source which contains the language name - thus, we choose one source only from Western, Eastern, Plains, and Moose Cree, for example.

[8] According to our preliminary experiments, length-based batching can seriously harm the performance of MT models for X-English Bible translation

[9] To this end, we download the works of Martin Luther – which largely overlap in domain and size with the Bible (approximately 50,000 sentences) – from Project Gutenberg gutenberg.org.

Class	Setting	High-Resource					Low-Resource			
		CRE	NAV	ESU	IKU	Ave	MIC	DGR	IPK	Ave
Mono	32K Vocab	16.2	8.5	7.0	8.1	10.0	–	–	–	–
	16K Vocab	18.3	10.3	10.3	10.6	12.4	2.2	2.8	1.4	2.1
	+E2E	**18.5**	**11.0**	10.3	11.2	**12.8**	–	–	–	–
Multi	32K Vocab	2.8	2.4	8.7	9.4	5.8	1.3	1.6	6.5	3.1
	16K Vocab	1.8	2.3	7.8	8.7	5.2	1.5	1.4	6.9	3.3
	+1Src	12.6	5.9	9.4	9.8	9.4	**4.8**	4.7	6.9	5.5
	+HB	13.7	5.7	10.0	10.6	10.0	4.6	3.7	8.1	5.5
	+E2E	14.7	7.2	**11.5**	**11.8**	11.3	4.4	4.6	**9.8**	**6.3**
Many	+M2M	16.0	8.4	10.6	11.2	11.6	**4.8**	**5.6**	8.5	**6.3**

Table 3: Lowercase BLEU scores for NMT. The subsections correspond to monolingual, multilingual, and multilingual many-to-many translation. **Bolded** scores indicate the highest BLEU scores for the each language, as well as averages across high- and low-resource languages.

be less than beneficial for languages with high numbers of morphemes in each word.

When we reduce the BPE vocabulary, we see a large increase in translation quality for all monolingual experiments, as the system sees many more short sequences. Unfortunately, we fail to leverage the increase in data as we add more languages from the same family, with the Algic (Cree) and Athabaskan (Navajo) family models still collapsing, and the Inuit-Aleut slightly decreasing. This result is not entirely unforeseen, although we didn't expect it with such a small number of languages. Mueller et al. (2020) report that their models also completely devolved into translations that, while structurally fluent, were completely inadequate at representing the source translation. However, they did see small gains when the number of added languages was small.

We hypothesize that our results degrade because of a lack of complete Bible translations. Mueller et al. (2020) start with complete translations, and the numbers only start failing as incomplete translations are added. We see small gains for the Inuit family, for which we have multiple complete Bibles. We hypothesize that many copies of an identical target in the training data may be adversely affecting the multilingual models.

Reducing training data single source per language results in significant gains - multilingual training now clearly improves results for our four low-resource languages. The gains are encouraging, and the models are producing more adequate output. We thus maintain the single-source constraint for our other experiments – all following experiments are cumulative.

Heterogeneous batching also contributes modestly to the quality of translations, confirming our suspicion that certain batches were influencing the final results. Likewise, adding a purely English corpus increases BLEU notably.

Training a many-to-many model brings the scores on our high-resource languages nearly to the level of the monolingual models, but does not surpass them. We never expected much gain in the familial experiments in these languages – we already include the entire Bible as training, and the other languages are not introducing much new information. Where we expect to see gains is in the low-resource languages. And indeed we do. These three languages, containing only New Testament data, are not large enough to train monolingual NMT models. However, we see steady gains that, while not perfectly mirroring the results of the high-resource experiments, eventually results in translations that are 4.2 BLEU points, on average, better than the monolingual models. These languages are able to leverage the information of more complete Bibles in other related languages to improve substantially.

5 Conclusion

We have presented an extension to the JHU Bible corpus, expanding it by almost forty translations in thirty Indigenous languages. These languages represent only a fraction of the languages spoken in North America, but by presenting them in a parallel corpus, we hope to encourage computational research in these underrepresented languages. Based on our experiments, the benefits of cross-lingual training are clear. Our experiments have also uncovered a set of useful training strategies which counteract target overfitting in multilingual models which are trained using several source translations but only one target text.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Ehsaneddin Asgari and Hinrich Schütze. 2017. Past, present, future: A computational investigation of the typology of tense in 1000 languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 113–124, Copenhagen, Denmark. Association for Computational Linguistics.

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

John Enrico. 2003. *Haida syntax*, volume 1. U of Nebraska Press.

Chung-Chian Hsu, Yan-Ping Huang, and Keng-Wei Chang. 2008. Extended Naive Bayes classifier for mixed data. *Expert Systems with Applications*, 35(3):1080–1083.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3158–3163, Reykjavik, Iceland. European Language Resources Association (ELRA).

Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020. The Johns Hopkins university bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, Marseilles, France. European Language Resources Association (ELRA).

Aaron Mueller, Garrett Nicolai, Arya D. McCarthy, Dylan Lewis, Winston Wu, and David Yarowsky. 2020. An analysis of massively multilingual neural machine translation for low-resource languages. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, Marseilles, France. European Language Resources Association (ELRA).

Elke Nowak. 2011. *Transforming the images: Ergativity and transitivity in Inuktitut (Eskimo)*, volume 15. Walter de Gruyter.

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 NAACL-HLT 2019: Demonstrations*.

Philip Resnik, Mari Broman Olsen, and Mona Diab. 1999. The Bible as a parallel corpus: Annotating the 'book of 2000 tongues'. *Computers and the Humanities*, 33(1):129–153.

Katherine Schmirler, Antti Arppe, Trond Trosterud, and Lene Antonsen. 2018. Building a constraint grammar parser for Plains Cree verbs and arguments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

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.