TalaMT: Multilingual Machine Translation for Cabécar-Bribri-Spanish

Alex Jones
Dartmouth College
alexander.g.jones.23@dartmouth.edu

Rolando Coto-Solano
Dartmouth College
Department of Linguistics
rolando.a.coto.solano@dartmouth.edu

Guillermo González Campos
University of Costa Rica, Atlantic Branch, Turrialba
guillermo.gonzalezcampos@ucr.ac.cr

Abstract

In this paper, we experiment with building multilingual neural machine translation models to translate the extremely under-resourced Indigenous Costa Rican languages Cabécar and Bribri — members of the Viceitic branch of the Chibchan family — to and from Spanish. We explore a variety of techniques, including: (1) training trilingual models that can translate Bribri or Cabécar to and from Spanish; (2) performing self-supervised training, such as denoising autoencoding and masked sequence-to-sequence reconstruction; (3) adding data from a bilingual lexicon as additional parallel data; and (4) prepending indicator tokens to source sentences that tell the model which language it is translating to (<2tgt>) or from (<4src>). We observe some modest gains from self-supervised training and adding lexical data in this extremely under-resourced setting, and also find that trilingual models can outperform bilingual models, including models trained to translate in just one direction. We also see that prepending <2tgt> *and* <4src> tokens to source sentences yields modest gains. Our best model achieves around 26 ChrF averaged across the four directions (Spanish ↔ Cabécar, Bribri ↔ Spanish), despite being trained on only 8K parallel sentences for Bribri-Spanish and 4K for Cabécar-Spanish.

1 Introduction

This paper focuses on building neural machine translation (NMT) systems that translate two Indigenous Costa Rican languages to and from Spanish: Cabécar and Bribri. Cabécar and Bribri both fall under the Viceitic branch of the Chibchan language family. The Chibchan family is native to the Isthmo-Colombian Area, stretching from eastern Honduras to northern Colombia, including Costa Rica, Panama, and Nicaragua. There are hundreds of thousands of Chibchan speakers spread throughout this region. Along with Teribe, Cabécar and Bribri are the only living languages in the Viceitic branch. Cabécar and Bribri, like the other Chibchan languages, tend to have rich and complex morphology, compounding the challenge of building machine translation systems for them.

The Cabécar people live in the Chirripó and Talamanca regions in Eastern and Southern Costa Rica. As of 2011, the population numbered around 14,000 (INEC, 2011), and there are an estimated 11,100 native speakers of Cabécar presently. The Bribri people live in southern Costa Rica and northern Panama. Their population is around 17,000 (INEC, 2011), with approximately 7,000 speakers of the language. Both languages are classified as vulnerable (Moseley, 2010; Sánchez Avendaño, 2013).

There are a number of objectives we have in mind with this work, some of them purely technical and some of them related to language documentation and revitalization. On the technical side, we aim to see whether multilingual MT training and/or self-supervised training can improve translation performance for extremely under-resourced languages. Unlike other works that attempt these techniques at massive scale, involving hundreds of languages and billions of sentences, we wish to put multilingual training and self-supervision to the test using realistic under-resourced conditions: only three languages, four translation directions, and tens of thousands of parallel sentences. We hope that in training models with both Bribri *and* Cabécar the model will leverage linguistic similarity to improve performance in one or both languages.

On the documentation and revitalization side, we ultimately want to build systems that Indigenous people can use to engage with content in their community's language, e.g. by translating Spanish web text to Cabécar or Bribri. This capability becomes increasingly important as indigenous cultures adopt digital technologies and come into contact with content in other languages. If people cannot continue

Proceedings of the The 3rd Workshop on Multi-lingual Representation Learning (MRL), pages 106–117

using their culture's language in the digital age, the language may lose even more domains of usage and ultimately become dormant (Jany, 2018; Stern, 2018; Cruz and Waring, 2019; Zhang et al., 2022; Orynycz, 2022). On the flip side, translating in the other direction (e.g. {Bribri, Cabécar} → Spanish) can facilitate communication or help outsiders learn indigenous languages.
The contributions of this work are as follows:

1. We train and evaluate a multilingual NMT system that translates Cabécar and Bribri to and from Spanish. To our knowledge, we are the first to train and evaluate an MT system with Cabécar, and among the first to train multilingual NMT systems tailored to Indigenous languages of the Americas.

2. We compare a number of methods for enhancing multilingual NMT performance on extremely under-resourced languages, including self-supervised methods like denoising autoencoding and masked reconstruction, as well as other techniques like `<4src>` tagging or using bilingual lexicon entries as additional parallel data.

3. We provide comparisons between unidirectional bilingual models and bidirectional bilingual models, as well as between bilingual and trilingual models. Notably, we show that multilingual NMT models can beat bilingual models, even in an extremely resource-poor setting.

2 Related Work

2.1 MT and NLP for indigenous languages of the Americas

There are a number of previous efforts that have looked at machine translation and other NLP tasks for Indigenous languages of the Americas. For an extensive list of works in this area, we recommend the `Naki` GitHub page[1]. We will provide a brief overview of some recent work, with a focus on MT.

The closest work to ours, who our project is in part a follow-up to, is Feldman and Coto-Solano (2020), which experimented with training NMT models with back-translation for Bribri → Spanish and Spanish → Bribri. We use an extended version of Bribri-Spanish parallel dataset from their paper, but there are a number of differences: (1) we train on Cabécar-Spanish data as well; (2) we train multilingual, multidirectional models, rather than only unidirectional bilingual models; and (3) we experiment with self-supervised training on monolingual data.

There have been various other efforts at MT for other Amerindian languages. Some recent works include: Zhang et al. (2020), who work with Cherokee-English translation; Le and Sadat (2020), who work with Inuktitut-English translation; Montoya (2019), who work with Shipibo Konibo-Spanish translation; and Hois (2017), who work with Wixarika-Spanish translation. These works deploy a number of techniques for training low-resource MT models, such as incorporating language models and back-translation (Zhang et al., 2020), morphologically segmenting polysynthetic words before training (Le and Sadat, 2020), and leveraging related-language data from higher-resource languages to effect transfer learning (Montoya, 2019). Due to the extremely low level of resources for these languages, some of these works experiment with statistical machine translation, either in addition to NMT (e.g. Zhang et al. (2020)) or in place of it (e.g. Hois (2017)). In the AmericasNLP (Mager et al., 2021) shared task on MT for Indigenous languages of the Americas, various authors built and evaluated systems for a diverse set of languages, namely: Asháninka, Aymara, Bribri, Guarani, Nahuatl, Otomí, Quechua, Rarámuri, Shipibo-Konibo, and Wixarika.

Also of note is a recent collaborative effort between many NLP researchers who work on Indigenous languages of the Americas, called Americas-NLI (Ebrahimi et al., 2022). This paper examined the natural language understanding capabilities of pretrained multilingual models on Indigenous language data, investigating both zero-shot transfer and continued pretraining on these languages. They found that the pretrained multilingual models' performance was poor on the 10 Indigenous languages they examined, although continued pretraining offered substantial improvements. This is one of the few large-scale collaborative efforts for Indigenous NLP in the Americas, but there will hopefully be more projects of this sort that focus on other tasks such as MT.

2.2 Multilingual NMT

Multilingual NMT refers to training machine translation models on many languages, in many direc-

[1] https://github.com/pywirrarika/naki

tions, with a single set of parameters and a shared vocabulary. Currently, the largest industry labs with the most data and compute resources (e.g. Google, Meta, Microsoft) can train models capable of translating hundreds of directions, a procedure known as "massively multilingual machine translation" (Johnson et al., 2017; Aharoni et al., 2019; Fan et al., 2020; NLLB Team et al., 2022; Bapna et al., 2022). This is how state-of-the-art production MT systems are now trained.

Multilingual NMT has a number of appeals compared to training bilingual models. For one, the parameter efficiency is much greater. The number of possible language pairs scales quadratically with the number of languages, and if one wants the option of translating between all possible language pairs then the number of bilingual models required would scale quadratically as well. For instance, accommodating all possible language pairs for 30 languages would require 435 bilingual models. By contrast, a single model could be trained on all 30 languages, with parallel data for some language pairs, and then there is also the possibility of performing zero-shot translation for some of the language pairs not seen in training (Johnson et al., 2017). Multilingual models of course must be larger than bilingual models, but not so much larger that their use of parameters is less efficient.

Another appeal of multilingual MT systems is the potential for transfer learning. Specifically, it is possible for the model to improve on translating under-resourced languages by being trained on the rich data for higher-resource languages. Notably, however, this type of positive transfer is most likely to happen when the languages are closely related to each other genealogically (Ko et al., 2021; Khatri et al., 2021). In our case, we do not have a high-resource Chibchan language that we can use to bootstrap training for Cabécar and Bribri (and this is probably the case for most language families in the world). However, it is still theoretically possible to see gains on one or both languages due to their relatedness, even if they are both very under-resourced.

Although multilingual NMT has been spearheaded by large industry labs, there have been a number of recent efforts at training multilingual models specifically for low-resource languages. Among these are Yigezu et al. (2021), Emezue and Dossou (2022), and Vegi et al. (2022). All three of these papers build systems for African languages. Multilingual NMT hasn't been attempted for many Indigenous languages in other parts of the world, and certainly not for the Chibchan languages. It is promising, however, that industry labs are beginning to introduce Indigenous languages (of the Americas and elsewhere) into both research and production MT systems, e.g. Aymara and Guarani for Google Translate, and Yucatec Maya and Inuktitut for Microsoft Translator.

2.3 Self-supervised training

The other class of techniques we experiment with in this paper is self-supervised training. Self-supervised training refers to feeding the model some manipulated (e.g. noised or masked) form of monolingual sentences to the model and then tasking the model with reconstructing the original sentences. There are two types of self-supervised training methods we experiment with in this paper: denoising autoencoding and masked reconstruction[2].

The denoising autoencoding training we do is inspired by BART (Lewis et al., 2019) and mBART (Liu et al., 2020). In these works, sequence-to-sequence models are fed noisy (e.g. randomly shuffled) sentences and made to reconstruct the original sentences. By pretraining on this task in multiple languages, Liu et al. (2020) showed that the resulting model could be finetuned to perform well on MT.

The second self-supervised task we experiment with is MASS, or MAsked Sequence-to-Sequence pretraining (Song et al., 2019). In this method, the masked language modeling objective is generalized such that spans of arbitrary length are masked and the model has to predict either the masked tokens or reconstruct the entire original sentence. We opt for the latter approach (reconstructing the whole sentence), and try two different masking variants (see Section 4.2.2).

Self-supervised training has been shown to be successful in training massively multilingual NMT models, improving performance on low-resource and unsupervised languages in particular (Bapna et al., 2022; Siddhant et al., 2022; NLLB Team et al., 2022). A limited number of works have also looked at self-supervised training for MT in low-resource settings, and found it to be beneficial (Kuwanto et al., 2021; Dhar et al., 2022).

[2]Our masked sequence-to-sequence reconstruction task could be viewed as denoising autoencoding as well, but we keep it separate from our other denoising task for clarity.

3 Data

We have two parallel datasets at our disposal for this work: one for Bribri-Spanish, one for Cabécar-Spanish. The Bribri-Spanish dataset contains ≈ 8600 sentence pairs. These come from textbooks for Spanish speakers to learn Bribri (Constenla et al., 2004; Jara Murillo and García Segura, 2013), bilingual dictionaries (Margery, 2005), grammar books (Jara Murillo, 2018a), compilations of transcribed oral literature (Constenla, 2006, 1996; García Segura, 2016; Jara Murillo, 2018b), pedagogical textbooks (Sánchez Avendaño, 2020), and a digitized and transcribed oral corpus with traditional stories and songs (Flores Solórzano, 2017). Most of these sentences belong to general domains (e.g. *Ye' dör bua'ë* 'I am doing well'), but they also include technical passages from narrations about mythology and traditional practices. This corpus is available at the AmericasNLP 2021 repository[3].

The Cabécar-Spanish dataset contains ≈ 4200 sentence pairs. These come from the bilingual dictionary by González Campos and Obando Martínez (2020). This corpus is also composed of general sentences (e.g. *Yís sér dä él da* 'I live with my brother'). These were gathered from the authors' fieldwork and pedagogical books (González Campos et al., 2020; González Campos and Obando Martínez, 2018).

For both language pairs, we use a 90/5/5 train/validation/test split. Due to the lack of monolingual data for Bribri or Cabécar (besides Biblical data, which we deliberately do not use due to its linguistic and topical skew), we use the sentences from the parallel datasets as our monolingual data for the self-supervised (denoising/MASS) tasks as well. We also have a small bilingual lexicon available for Cabécar-Spanish, containing 1350 entries. We use this as additional parallel data in training a bidirectional Cabécar ↔ Spanish model (see Section 5.2).

4 Methods

4.1 Model

We use the OpenNMT (Klein et al., 2017) implementation of the Transformer (Vaswani et al., 2017) model for all our experiments. Each model has ≈ 50M parameters and we tokenize our data using the OpenNMT implementation of BPE (Sennrich et al., 2016) with `n_symbols` = 10000. Unless indicated otherwise, we train our models with Adam optimization (Kingma and Ba, 2015) for 4000 steps with a batch size of 4096, a learning rate of 2.0, 6 hidden layers, 8 attention heads, a hidden layer dimension of 512, a feedforward layer dimension of 2048, and a dropout probability of 0.1. We train on one NVIDIA A100 GPU provided by Google Colab, which took around 20-30 minutes per model. Full hyperparameters are given in Section B of the Appendix.

4.2 Training Techniques

We experiment with a variety of training techniques to arrive at the best method, or combination of methods. First, we train two types of *bilingual* models: unidirectional models, which only translate one language to another, and bidirectional models that translate two languages in both directions. Because we have Cabécar-Spanish bilingual lexicon data, we also experiment with adding that as additional parallel signal. Second, we experiment with training *trilingual* models, which translate Bribri ↔ Spanish and Cabécar ↔ Spanish.

Next, we experiment with several different self-supervised training schemes to improve the trilingual models. These methods are described below.

4.2.1 Multilingual Training

One of our main interests in this paper is training multilingual models that translate Bribri ↔ Spanish and Cabécar ↔ Spanish. The only modification we make to the training data for training the baseline trilingual model is prepending a `<2tgt>` token that tells the model which language to translate to, as in Bapna et al. (2022). For example, when translating Spanish to Cabécar we use the tag `<2cjp>`. The models are then trained in all four directions with a cross-entropy loss.

4.2.2 Self-supervised Training

We also experiment with self-supervised training using monolingual data (taken from the parallel datasets).

Denoising autoencoding One of the self-supervised tasks we try is denoising autoencoding, where the model is fed a noisy version of a sentence and has to reconstruct the original sentence. As our noising function, we randomly shuffle the order of words in a sentence, similar to Lewis et al. (2019); Liu et al. (2020). Once again following Bapna et al. (2022), we add a `<2task>` tag to all

[3] https://github.com/AmericasNLP/americasnlp2021

sentences in the dataset to help the model distinguish the denoising task from the MT task. In this case, that token is <2denoise> for the denoising task and <2translate> for the MT task.

MASS The second self-supervised training technique we experiment with is MASS (Song et al., 2019). This method involves masking tokens in the source sentence and having the model try to reconstruct the original sentence. Bapna et al. (2022); Siddhant et al. (2022) show this can be used to improve performance for many low-resource and unsupervised languages in massively multilingual MT systems. We employ two variants of MASS. In the first, text spans of arbitrary length in the source are replaced with a single [MASK] token (following Lewis et al. (2019)). In the second, each masked token is replaced with its own [MASK] token. In either case, we mask 50% of the words in each sentence and train on the task for all three languages. The <2task> token we use here is <2mass>.

4.2.3 Using bilingual lexicons

We also experiment with adding bilingual lexicon entries as extra parallel data. For this, we use a Cabécar-Spanish bilingual lexicon to help train a bidirectional Cabécar ↔ Spanish model. Once again, <2lang> tags are used so the model knows which language to translate to.

5 Experiments

All models use the hyperparameters described in Section 4.1 and Section B of the Appendix unless stated otherwise. We arrive at these hyperparameters through manual tuning of train_steps, learning_rate, warmup_steps, enc/dec_layers, heads, hidden_size, and transformer_ff. The remaining hyperparameters are left as the defaults selected by OpenNMT.

5.1 Unidirectional bilingual models

The simplest models we train are unidirectional bilingual models: models which just translate one language to one other language, e.g. Spanish → Bribri. These models act as baselines against which to compare our bidirectional bilingual models, described below. No modification to the training data is necessary for these models. The models here are referred to as **Cabécar → Spanish**, **Spanish → Cabécar**, **Bribri → Spanish**, and **Spanish → Bribri**.

5.2 Bidirectional bilingual models

The second type of models we train are bidirectional bilingual models, which translate two languages in both directions, e.g. Cabécar ↔ Spanish. For these models, we add a <2tgt> tag to the training data so the model knows which language to translate to. The models here are referred to as **Bribri+Spanish** and **Cabécar+Spanish**.

We also train a Cabécar ↔ Spanish model using bilingual lexicon entries as additional parallel data, which we will refer to as the **Cabécar+Spanish+bilingual lexicon data** model.

5.3 Trilingual models

We train multilingual models that translate Bribri ↔ Spanish and Cabécar ↔ Spanish as well.

Baseline In the baseline setup, we simply use the hyperparameters from 4.1 to train a three-language, four-directional model. This model is called **Trilingual baseline**. We also train two additional models, which are trained for 8000 steps and 12000 steps but otherwise use the same hyperparameters as the baseline. We do these as basic checks for approximately how long it takes the model to converge.

<4src> tagging Although all our trilingual models have <2tgt> tags to indicate which language to translate to, we also experiment with adding <4src> tags to tell the model which language it's translating *from* (e.g. <4cjp> when translating from Cabécar). The motivation here is that the model could potentially get confused between Cabécar and Bribri due to their similarity, and an explicit tag may mitigate some of this confusion. The source sentences for this model took the form <4src> <2tgt> word1 word2...wordN. This model is referred to as the **Baseline+<4src> tagging** model.

Joint denoising training We also experiment with jointly training the model on the denoising autoencoding task and the MT task. We try two variants of this: in the first, we simply train the model on both tasks simultaneously for 4000 steps. This model is called **Baseline+joint denoising training**. In the second variant, we do the same but then continue finetuning the model on the MT task, with the same data, for an extra 4000 steps. This variant is called **Baseline+joint denoising training, MT finetuning**.

Joint MASS training Additionally, we try jointly training the model on the MASS task and the MT task. We use two different variants of MASS: in the first, we replace spans of arbitrary length in the source with a single `[MASK]` token. This model is called **Baseline+joint MASS training (replace span)**. In the second, we replace *each* ablated token with a `[MASK]` token. This model is called **Baseline+joint MASS training (replace token)**.

6 Results

The results are summarized in Tables 1 and 2. Table 1 shows a comparison between the unidirectional and bidirectional bilingual models. Table 2 gives a comparison between the bilingual and trilingual models.

The first thing to note is that the bidirectional models outperform unidirectional models in all directions. Across all four directions, the average improvement (Δ CHRF) of the best-performing bidirectional model was $+4.9$. The model with bilingual lexicon data performs best on Spanish $\rightarrow$ Cabécar ($+5.2$ over unidirectional baseline), although it slightly underperforms the vanilla bilingual model on Cabécar $\rightarrow$ Spanish ($+0.1$ vs $+1.2$).

Next, there are a number of takeaways from the comparison between the bilingual and trilingual models. First, note that *at least one* trilingual model outperformed each bilingual baseline except in the Bribri $\rightarrow$ Spanish direction, where the next-best model got -5.7 CHRF relative to the bilingual Bribri+Spanish model. The reason for this deviation from the general trend is not clear to us. There were five trilingual models that improved over the bilingual baselines in at least one direction: **Trilingual baseline**, **Trilingual baseline+8000 steps**, **Trilingual baseline+12000 steps**, **Baseline+<4src> tagging**, and **Baseline+joint denoising training, MT finetuning**. The remaining models failed to improve over the bilingual baselines in any direction.

Looking at average CHRF across all four directions—denoted μ_4 in Table 2—we see a near three-way tie between **Baseline+joint denoising training, MT finetuning** (26.1 CHRF), **Baseline+8000 steps** (26.0 CHRF), and **Baseline+<4src> tagging** (25.9 CHRF). Just looking at the averages, it appears that these three techniques work pretty well in our training setting: (1) simply training the model a bit longer; (2) performing joint denoising training, followed by MT finetuning; and (3) adding `<4src>` tags to the beginning of source sentences.

Next, we examine each translation direction separately. For Cabécar-Spanish, the model with `<4src>` tagging wins in both directions, with gains of $+3.9$ CHRF in the Cabécar $\rightarrow$ Spanish direction and $+1.9$ in the Spanish $\rightarrow$ Cabécar direction. For Bribri-Spanish, the results are somewhat less clear-cut. For Bribri $\rightarrow$ Spanish, the bilingual baseline performs best, netting 30.8 CHRF. For Spanish $\rightarrow$ Bribri, the 8000 steps model does best, improving $+1.2$ CHRF over the bilingual baseline.

The models co-trained on the MASS task performed poorly, seeing huge losses across the board. There are a number of reasons why this might have happened. One is that we simply did not have enough data for the model to learn from the task effectively. The MASS task has been shown to work well for very high-resource settings on models with hundreds of millions or billions of parameters, and this result might simply not scale to the extremely low-resource, small model scenario. Another possibility is that there are different ways to implement MASS that would be more amenable to datasets of the size studied here. In personal correspondence with various authors on Bapna et al. (2022), we learned that the MASS task can be difficult to implement properly given the description in Song et al. (2019). Further experimentation with the MASS task in resource-poor settings is left for future work.

In regard to the denoising autoencoding task, it is interesting to note that while model performance decreased relative to the trilingual baseline using the **Baseline+joint denoising training** setup, we were able to see gains by adding in 4000 steps of MT finetuning following the joint dual-task training. It could be that this is a quirk of very low-resource training, as the extra finetuning step isn't necessary to see substantial improvements on large, high-resource, massively multilingual models (Bapna et al., 2022; Siddhant et al., 2022). In our setting, it seems that the model does indeed learn from the denoising task but that it needs more training passes on the MT data for it to really make use of those gains on unseen MT queries at inference time.

7 Discussion

There are a number of contributions that our experiments make from both a technical and a social angle. On the technical side, our experiments

	Cabécar → Spanish	Spanish → Cabécar	Bribri → Spanish	Spanish → Bribri
		Unidirectional		
Cabécar → Spanish	21.3	–	–	–
Spanish → Cabécar	–	23.8	–	–
Bribri → Spanish	–	–	24.9	–
Spanish → Bribri	–	–	–	21.2
		Bidirectional		
Cabécar+Spanish	**22.5**	26.4	–	–
+bilingual lexicon data	21.4	**29.0**	–	–
Bribri+Spanish	–	–	**30.8**	**28.6**

Table 1: A comparison between unidirectional and bidirectional bilingual models (CHRF). All models are trained for 4000 steps with identical hyperparameters. The "+bilingual lexicon data" model was trained with 1352 Cabécar-Spanish bilingual lexicon entries as additional parallel data.

	μ_4	cab → spa	spa → cab	bri → spa	spa → bri
			Bilingual		
Cabécar+Spanish (4000 steps)	–	22.5	26.4	–	–
+bilingual lexicon data	–	21.4	29.0	–	–
Bribri+Spanish (4000 steps)	–	–	–	**30.8**	28.6
			Trilingual		
Trilingual baseline (4000 steps)	24.2	21.8	28.8	18.9	27.3
Trilingual baseline with additional training (8000 steps)	26.0	24.2	29.3	20.5	**29.8**
Trilingual baseline with additional training (12000 steps)	25.1	24.2	28.3	19.6	28.2
Trilingual baseline+<4src> tagging	25.9	**26.4**	**30.9**	19.1	27.3
Trilingual baseline+joint denoising training	22.0	20.2	25.5	18.8	23.3
Trilingual baseline+joint denoising training, MT finetuning	**26.1**	22.1	29.5	25.1	27.7
Trilingual baseline+joint MASS training (replace span)	11.1	9.0	14.7	11.5	9.3
Trilingual baseline+joint MASS training (replace token)	8.6	6.7	9.5	9.8	8.5

Table 2: A comparison between the bilingual and trilingual models that translate Cabécar and Bribri to/from Spanish (performance is measured in CHRF). Green-colored indicate improvements over the baseline, with bright green cells being the best performers. Red-colored cells indicate losses relative to the bilingual baselines. μ_4 indicates the average performance across all 4 directions.

are noteworthy because they put to the test techniques that have been shown to work for giant-scale machine translation models trained with copious amounts of data, but haven't been rigorously examined in very under-resourced settings. Namely, the two classes of techniques we investigate here are (1) multilingual machine translation, and (2) self-supervised training, namely denoising autoencoding and masked reconstruction (MASS).

Our results show that we can get benefits from multilingual training even in this resource-scarce scenario, as well as from denoising autoencoding training. The first of these results suggests that there is some transfer learning happening between Bribri and Cabécar even with $<$ 10K sentences for each. Of course, these are closely related languages, and we would not expect such transfer to happen between distantly related languages with such little data. But this is a promising result for extremely low-resource MT nonetheless.

The fact that denoising autoencoding training did reasonably well, especially when followed by MT finetuning, is also interesting. The upshot here is that even a small amount of monolingual data for a low-resource language can potentially yield benefits on the MT task. By contrast, it is puzzling that our implementation of MASS yielded poor results. This could be an indication that the MASS task requires a certain amount of data to benefit MT training, and that we were well below that thresh-

old, but this hypothesis needs further investigation in future work. It is also possible that a different implementation of the MASS task could work better for extremely low-resource settings, e.g. one where only tokens at the beginning or end of source sentences are masked.

Lastly, although MT performance on under-resourced languages is far from where it needs to be to suit the demands of actual speakers, we see our work on these indigenous languages as a step in the right direction. Whenever an NLP method is shown to help high-resource, politically and economically dominant languages like English, Spanish, or Chinese, that same method should be tested on under-resourced languages, which constitute the vast majority of the world's languages (Joshi et al., 2020). If the method works, then that is a step toward making language technology better and more inclusive. If it doesn't, then that shows a fundamental limitation in state-of-the-art techniques, because it suggests they don't scale to down to the languages that much of the world speaks. What we have seen in this paper is a mixture of both these results. We hope that these findings are helpful for the research community and, ultimately, the indigenous speaker communities for whom this technology is made.

8 Conclusions

In this paper, we have experimented with training multilingual neural machine translation models that translate the indigenous Costa Rican languages Cabécar and Bribri to and from Spanish. First, we provide a comparison between unidirectional bilingual models and bidirectional bilingual models, showing that the latter can outdo the former in all directions. Next, we show that the trilingual models we train beat the bilingual baselines in all but one of the four translation directions (namely Bribri → Spanish). In training the trilingual models, we experiment with a number of variables: (1) training for more steps; (2) prepending a `<4src>` tag to source sentences to tell the model what language it's translating from, in addition to the `<2tgt>` tag we use for all multidirectional models; (3) adding in self-supervised training on monolingual data, either denoising autoencoding or masked reconstruction (MASS); and (4) finetuning models on the MT task following joint training on denoising autoencoding and MT. Out of these, the most promising findings are that `<4src>` tags appear useful (especially for Cabécar ↔ Spanish) and that joint denoising training followed by MT finetuning is an efficacious approach. We also show that adding bilingual lexicon entries as additional parallel data improves performance somewhat on Spanish → Cabécar.

Future work should look at combining these strategies with other techniques, such as back-translation. Additionally, with the increasing capabilities of Large Language Models as general NLP systems, much work must be done to see how their translation abilities on under-resourced languages can be evaluated and improved.

Limitations

One limitation of this work is the small number of languages explored. While it is important to examine the members of the Chibchan language family individually due to the extreme scarcity of attention they've been given in the NLP literature, it is true that the results in our paper are only directly applicable to Cabécar, Bribri, and Spanish. To mitigate this narrowness, future work should incorporate Chibchan languages into broader multilingual NLP efforts.

Another limitation of this work is the small amount of training data available. Of course, this is simply the state of affairs for extremely under resourced languages like Cabécar and Bribri, and it is part of the experimental design itself. However, future efforts should focus on data resource creation in addition to modeling in order to improve the state of technology for these languages.

Finally, a limitation of this work at present is the fact that some of the data we used is not yet open-source, due to intellectual property restrictions. However, it is our hope that all the data associated with this project will soon be released for public use.

Ethics Statement

Perhaps the greatest ethical concern in working on language technology for Indigenous languages is the European colonialist history that looms over these languages and their associated cultures. This history is one of violence, genocide, cultural theft and destruction, exploitation, and bigotry. Countless Indigenous languages across the world have been suppressed, stigmatized, diminished, or altogether wiped out in the wake of colonialism. These, of course, are only the linguistic consequences of a

history that has been violent in many distinct ways.

First and foremost, the purpose of building technology for Indigenous languages should be to benefit the speakers themselves. The features and potential applications of the technology should be guided by the speakers' needs and desires. It is our hope that our research will lead to technologies that the Cabécar, Bribri, and other peoples can use and benefit from, and that they can develop these tools themselves in the near future.

Building Indigenous language technologies ethically entails more than just constructing useful systems. It also entails respect for concerns such as data sovereignty and the ways in which the speakers want their language to be used (for instance, whether they would like outsiders to interact with their language). While some of these matters are not particular to Indigenous languages, they are especially pertinent to these languages because of the colonialist history described above.

Acknowledgments

We would like to thank Isaac Caswell, Ankur Bapna, and Xavier García on the Google Translate team for their correspondence regarding this project. We would also like to thank Franklin Morales, Freddy Obando, and Samantha Wray for their help.

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.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2022. Building machine translation systems for the next thousand languages.

Adolfo Constenla. 1996. *Poesía tradicional indígena costarricense*. Editorial Universidad de Costa Rica.

Adolfo Constenla. 2006. *Poesía bribri de lo cotidiano: 37 cantos de afecto, devoción, trabajo y entretenimiento*. Editorial Universidad de Costa Rica.

Adolfo Constenla, Feliciano Elizondo, and Francisco Pereira. 2004. *Curso Básico de Bribri*. Editorial de la Universidad de Costa Rica.

Hilaria Cruz and Joseph Waring. 2019. Deploying technology to save endangered languages.

Prajit Dhar, Arianna Bisazza, and Gertjan van Noord. 2022. Evaluating pre-training objectives for low-resource translation into morphologically rich languages. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4933–4943, Marseille, France. European Language Resources Association.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Chris C. Emezue and Bonaventure F. P. Dossou. 2022. Mmtafrica: Multilingual machine translation for african languages. *CoRR*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *CoRR*, abs/2010.11125.

Isaac Feldman and Rolando Coto-Solano. 2020. Neural machine translation models with back-translation for the extremely low-resource indigenous language Bribri. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3965–3976, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sofía Flores Solórzano. 2017. Corpus oral pandialectal de la lengua bribri.

Alí García Segura. 2016. *Ditsö Rukuö Identity of the Seeds: Learning from Nature*. IUCN.

Guillermo González Campos and Freddy Obando Martínez. 2018. *Fonología y ortografía del Cabécar*. Editorial de la Universidad Estatal a Distancia.

Guillermo González Campos and Freddy Obando Martínez. 2020. *Diccionario Escolar del Cabécar de Chirripó - Ditsá duchíwák kté chulíí i yuäklä*. Universidad de Costa Rica, Vicerrectoría de Acción Social, Sede del Atlántico.

Guillermo González Campos, Freddy Obando Martínez, and Arturo Peña Hurtado. 2020. *Itsó Pákë̈ - Historia de Itsó*. Universidad de Costa Rica, Vicerrectoría de Acción Social, Sede del Atlántico.

Jesús Manuel Mager Hois. 2017. *Traductor híbrido Wixarika - Español con escasos recursos bilingües*. Ph.D. thesis, Universidad Autónoma Metropolitana Azcapotzalco, Mexico City, Mexico.

INEC. 2011. Población total en territorios indígenas por autoidentificación a la etnia indígena y habla de alguna lengua indígena, según pueblo y territorio indígena. In Instituto Nacional de Estadística y Censos, editor, *Censo 2011*. INEC Costa Rica.

Carmen Jany. 2018. The role of new technology and social media in reversing language loss. *Speech, Language and Hearing*, 21(2):73–76.

Carla Victoria Jara Murillo. 2018a. *Gramática de la Lengua Bribri*. EDigital.

Carla Victoria Jara Murillo. 2018b. *I Ttè Historias Bribris*, second edition. Editorial de la Universidad de Costa Rica.

Carla Victoria Jara Murillo and Alí García Segura. 2013. *Se' ttö' bribri ie Hablemos en bribri*. EDigital.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Jyotsana Khatri, Nikhil Saini, and Pushpak Bhattacharyya. 2021. Language relatedness and lexical closeness can help improve multilingual NMT: IITBombay@MultiIndicNMT WAT2021. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 217–223, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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.

Wei-Jen Ko, Ahmed El-Kishky, Adithya Renduchintala, Vishrav Chaudhary, Naman Goyal, Francisco Guzmán, Pascale Fung, Philipp Koehn, and Mona Diab. 2021. Adapting high-resource NMT models to translate low-resource related languages without parallel data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 802–812, Online. Association for Computational Linguistics.

Garry Kuwanto, Afra Feyza Akyürek, Isidora Chara Tourni, Siyang Li, Alexander Jones, and Derry Wijaya. 2021. Low-resource machine translation for low-resource languages: Leveraging comparable data, code-switching and compute resources. *CoRR*, abs/2103.13272.

Tan Ngoc Le and Fatiha Sadat. 2020. Low-resource NMT: an empirical study on the effect of rich morphological word segmentation on Inuktitut. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 165–172, Virtual. Association for Machine Translation in the Americas.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *CoRR*, abs/2001.08210.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 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.

Enrique Margery. 2005. *Diccionario Fraseológico Bribri-Español Español-Bribri*, second edition. Editorial de la Universidad de Costa Rica.

Héctor Erasmo Gómez Montoya. 2019. *A crowd-powered conversational assistant for the improvement of a Neural Machine Translation system in native Peruvian language*. Ph.D. thesis, Pontifica Universidad Católica Del Perú, Lima, Peru.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*. Unesco.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Petro Orynycz. 2022. Say it right: AI neural machine translation empowers new speakers to revitalize Lemko. In *Artificial Intelligence in HCI: 3rd International Conference, AI-HCI 2022, Held as Part of the 24th HCI International Conference, HCII 2022, Virtual Event, June 26 – July 1, 2022, Proceedings*, page 567–580, Berlin, Heidelberg. Springer-Verlag.

Carlos Sánchez Avendaño. 2013. Lenguas en peligro en Costa Rica: vitalidad, documentación y descripción. *Revista Káñina*, 37(1):219–250.

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.

Aditya Siddhant, Ankur Bapna, Orhan Firat, Yuan Cao, Mia Xu Chen, Isaac Caswell, and Xavier Garcia. 2022. Towards the next 1000 languages in multilingual machine translation: Exploring the synergy between supervised and self-supervised learning. *CoRR*, abs/2201.03110.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: masked sequence to sequence pre-training for language generation. *CoRR*, abs/1905.02450.

Alissa J. Stern. 2018. Can the internet revitalize local languages? *Stanford Social Innovation Review*.

Carlos Sánchez Avendaño. 2020. *Se' Dalì Diccionario y Enciclopedia de la Agricultura Tradicional Bribri*. DIPALICORI.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Pavanpankaj Vegi, Sivabhavani J, Biswajit Paul, Prasanna K R, and Chitra Viswanathan. 2022. ANVITA-African: A multilingual neural machine translation system for African languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1090–1097, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Mesay Gemeda Yigezu, Michael Melese Woldeyohannis, and Atnafu Lambebo Tonja. 2021. Multilingual neural machine translation for low resourced languages: Ometo-english. In *2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 89–94.

Shiyue Zhang, Ben Frey, and Mohit Bansal. 2022. How can NLP help revitalize endangered languages? a case study and roadmap for the Cherokee language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1529–1541, Dublin, Ireland. Association for Computational Linguistics.

Shiyue Zhang, Benjamin Frey, and Mohit Bansal. 2020. ChrEn: Cherokee-English machine translation for endangered language revitalization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 577–595, Online. Association for Computational Linguistics.

A Appendix: Sample outputs

Table 3 shows some examples of outputs from each of our models in each direction.

B Hyperparameters

The full list of hyperparameters for all our models, except where stated otherwise, is as follows:

1. `train_steps = 4000`
2. `batch_size = 4096`
3. `valid_batch_size = 600`
4. `optimizer = adam`
5. `learning_rate = 2.0`
6. `warmup_steps = 8000`
7. `decay_method = noam`
8. `adam_beta2 = 0.998`
9. `label_smoothing = 0.1`
10. `position_encoding = true`
11. `enc_layers = 6`
12. `dec_layers = 6`
13. `heads = 8`
14. `hidden_size = 512`
15. `word_vec_size = 512`
16. `transformer_ff = 2048`
17. `dropout_steps = [0]`
18. `dropout = 0.1`
19. `attention_dropout = 0.1`
20. `share_vocab = true`
21. `share_embeddings = true`
22. `share_decoder_embembeddings = true`
23. `seed = 1234`

	Cabécar → Spanish
Source	¿Bikö matsílí ta Túrí rä?
Reference	¿A qué distancia queda Turrialba?
Unidirectional baseline	Vendí la carga para Turrialba.
Bidirectional bilingual baseline	¿Qué hora es?
Trilingual	¿Usted conoce la casa de Turrialba?
Trilingual + <4src> tagging	¿Cuánto es para Turrialba?
Trilingual + joint denoising training	¿La caña agria tiene hueba?
Previous model + MT finetuning	¿Juta tiene usted?
Trilingual + joint MASS training	rä?
Trilingual, 8K training steps	¿Cele con Turrialba.
Trilingual, 12K training steps	¿Qué tiene mucha saliva .
	Spanish → Cabécar
Source	Llegó un hombre con mucho tamaño.
Reference	Ékla jäyí dëju wákëi ta tái.
Unidirectional baseline	I jäyí bätsä káte.
Bidirectional bilingual baseline	Ékla jäyí dëju ju ska.
Trilingual	Jäyí dëkájuná tái.
Trilingual + <4src> tagging	Ékla jäyí dëju ju ska dí yäklä.
Trilingual + joint denoising training	jäyí júna kono wa.
Previous model + MT finetuning	Mulítä jénáká tái.
Trilingual + joint MASS training	I kjuátká ámijia.
Trilingual, 8K training steps	Jäyí butsaná tái.
Trilingual, 12K training steps	Jäyí butsaná tái.
	Bribri → Spanish
Source	E' kuéki e' mèkèattke se' ia, tö nai' rö se' kutà, kẽ rö katànok.
Reference	Por eso él nos dejó eso, que la danta es nuestra hermana, no es para comer.
Unidirectional baseline	eso ya iba a dejar eso establecido para nosotros, que la danta es nuestra hermana, no es para comer.
Bidirectional bilingual baseline	Por eso ya iba a dejar eso establecido para nosotros, que la danta es nuestra hermana, no es para comer.
Trilingual	Cuando el búho suena a los bejucos , para que se transformó en lengua ;
Trilingual + <4src> tagging	Al principio , por eso se debe decir que en la nariz , vea.
Trilingual + joint denoising training	A la hermana se les duelen las ví, las plantas.
Previous model + MT finetuning	por eso ahora , a partir de una persona , no eran para comer ,
Trilingual + joint MASS training	Por que majarse usa el cuerpo para bañar , y eso se usa la hermana ,
Trilingual, 8K training steps	¿Cuándo se apagan los bribris de monte?
Trilingual, 12K training steps	por eso las deidades siguen haciendo a la señora con un pedazo de piedra , porque era aprovechado
	Spanish → Bribri
Source	En la actualidad los jóvenes no conocen los taparrabos
Reference	Îñe ta se' duládulapa kẽ wa kipàdawo sùne ia.
Unidirectional baseline	iñ e alàrala i chèke. ema e' kuéki.
Bidirectional bilingual baseline	Skámoköl kë yö r ia dinamu súrule.
Trilingual	Nañéwe ta îñe kë ye' wa káse se se se lo que ".
Trilingual + <4src> tagging	Ká batá kë wa ya kë wa kapá taî táwa.
Trilingual + joint denoising training	Sä diëi yäklä ra, ká sá káwäta köchi chálí bu
Previous model + MT finetuning	Ká i' ki kë a' wa jóvenes ök..
Trilingual + joint MASS training	Chakì ye' chka' awá ta .
Trilingual, 8K training steps	Káwö wéle ta akëkëpa bák alambre yëuk.
Trilingual, 12K training steps	Skámoköl kë yör ktöm se' tabèla wa.

Table 3: Example model outputs. Green words are those that appear in the reference.

24. `valid_steps = 1000`
25. `accum_count = 3`
26. `accum_steps = 0`

These hyperparameters were passed to the `translate.py` function in OpenNMT-py[4].

[4] https://opennmt.net/OpenNMT-py/options/translate.html