Improving Neural Machine Translation of Indigenous Languages with Multilingual Transfer Learning

Wei-Rui Chen[1] **Muhammad Abdul-Mageed**[1,2]
[1]Deep Learning & Natural Language Processing Group, The University of British Columbia
[2]Department of Natural Language Processing & Department of Machine Learning, MBZUAI
{weirui.chen,muhammad.mageed}@ubc.ca

Abstract

Machine translation (MT) involving Indigenous languages, including endangered ones, is challenging primarily due to lack of sufficient parallel data. We describe an approach exploiting bilingual and multilingual pretrained MT models in a transfer learning setting to translate from Spanish into ten South American Indigenous languages. Our models set new SOTA on five out of the ten language pairs we consider, even doubling performance on one of these five pairs. Unlike previous SOTA that perform data augmentation to enlarge the train sets, we retain the low-resource setting to test the effectiveness of our models under such a constraint. In spite of the rarity of linguistic information available about the Indigenous languages, we offer a number of quantitative and qualitative analyses (e.g., as to morphology, tokenization, and orthography) to contextualize our results.

1 Introduction

Artificial intelligence (AI) is being widely integrated into many natural language processing (NLP) applications in our daily lives. However, these language technologies have focused almost exclusively on widely-spoken languages (Choudhury and Deshpande, 2021). Under-represented languages such as endangered languages are thus left out. For example, the Google machine translation (MT) system does not support any of the languages included in our current study.[1] Our objective in this work is hence to build machine translation (MT) models for Indigenous languages, which are by definition low-resource and possibly endangered. More specifically, we focus on South American Indigenous languages. In a MT scenario, a language pair is considered 'low-resource' if the parallel corpora consists of less than 0.5 million of parallel sentences and 'extremely low-resource' if less than 0.1 million of parallel sentences (Ranathunga et al., 2021). In this work, nine out of ten languages pairs we consider have under 0.1 million pairs of sentences (with only one language pair having roughly 0.1 million pairs of sentences). Developing MT systems for endangered languages can help preserve these languages.

Neural Machine Translation (NMT) is a branch of MT that leverages neural networks to build translation systems. Despite that NMT is able to produce powerful MT systems, it is data-hungry. That is, it requires large amounts of data to train a quality NMT model (Koehn and Knowles, 2017). Contemporary machine translation systems are oftentimes trained on over a million of parallel sentences (Fan et al., 2021; Tang et al., 2020) for high-resource language pairs. In contrast, the size of the dataset we have is limited. Transfer learning has been shown to help mitigate this issue by porting knowledge e.g. from a parent model to a child model (Zoph et al., 2016a). We leverage two types of pretrained MT models: *bilingual* models and a *multilingual* model. The overall training approach is illustrated in Figure 1. Our datasets are provided by AmericasNLP2021 (Mager et al., 2021) shared task. We compare our performance to the winner of the shared task (Vázquez et al., 2021).

The rest of this study is organized as follows: Section 2 is a literature review on Indigenous MT, transfer learning, the application of transfer learning to NMT, and the challenge of cross-lingual transfer. In Section 3, we describe our experimental settings. We present our results in Section 4, and provide discussions in Section 5. We conclude in Section 6.

2 Background

2.1 MT on Indigenous Languages

Languages are diverse. For example, in South America, there are 108 language families, 55 of which are in a language family with one single

[1]https://translate.google.com/about/languages/

Proceedings of the The Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023), pages 73–85

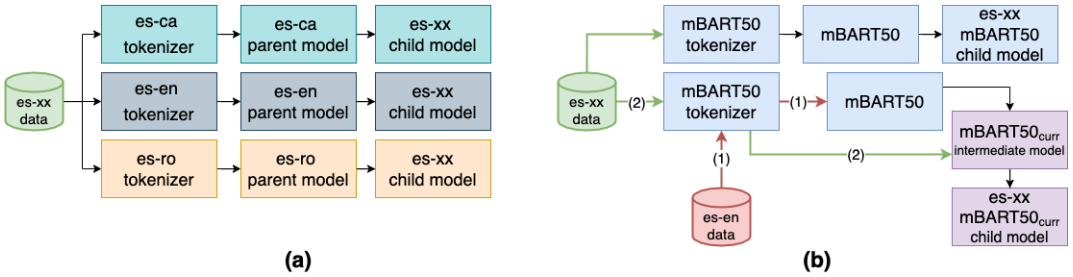

Figure 1: Model Training in **(a)** bilingual setting and **(b)** multilingual setting for one `es-xx` language pair. For both **(a)** and **(b)**, child models are those being used for prediction. xx represents arbitrary one of the ten South American Indigenous languages. For **(b)**, the blue es-xx mBART50 child model represents the model directly fine-tuned with `es-xx` data. The purple es-xx $\text{mBART50}_{\text{curr}}$ child model represents the model that is first being fine-tuned with `es-en` data to produce an intermediate model, indicated as (1). Afterwards, it is fine-tuned with `es-xx` data, indicated as (2).

member (i.e., language isolates) (Campbell et al., 2012). Due to this linguistic diversity, to the best of our knowledge, there is no single MT method that fits all Indigenous languages. However, since many Indigenous languages suffer the low-resource issue (Mager et al., 2018a), many researchers borrow ideas from low-resource MT to tackle the task of MT of Indigenous languages. We survey some approaches here.

Nagoudi et al. (2021) create models based on the T5 architecture (Raffel et al., 2019) and train it with monolingual Indigenous data before fine-tuning on parallel data, thus attempting to acquire knowledge of the Indigenous languages to benefit MT. Ngoc Le and Sadat (2020) focus on data preprocessing, and build a morphological segmenter for the source language Inuktitut to achieve better performance in Inuktitut-English translation. These aforementioned works all adopt methods invented to tackle the task of MT on low-resource languages.

2.2 Transfer Learning and NMT

It can sometimes be very expensive to collect data for MT. This is true especially for endangered languages when the number of speakers is decreasing. Therefore, many endangered languages suffer from the the low-resource issue. This motivates methods that can help port knowledge from existing resources to a down-stream task of interest with low-resources employing transfer learning methods. An additional motivation for studying and applying transfer learning is that human beings are able to apply knowledge/skills they acquired earlier from some jobs to better perform new related jobs with less efforts. An analogy is this: a person who has learned a music instrument may be able to pick up another instrument easier and quicker (Zhuang et al., 2020). When applying transfer learning in the context of NMT, a scenario can be as follows: a model previously trained on parent language pair(s) (called *parent model*) is further fine-tuned on child language pair(s) to form a *child model*. Under such a scenario, a parent language pair is one of the language pairs whose bilingual data is used to train a model from scratch and produce a parent model. A child language pair is one of the language pairs whose bilingual data is used to fine-tune a parent model and produce a child model. Again, the intuition here is that an experienced translator (pretrained MT model) on one language pair may be able to translate into another language pair with shorter time and less effort compared to a unexperienced person (new randomly-initialized model). The core idea is to retain the parameters of parent model as the starting point for the child model, instead of training from scratch where the parameters are randomly initialized (Zoph et al., 2016a; Kocmi and Bojar, 2018; Nguyen and Chiang, 2017).

2.3 Cross-lingual Transfer

One of the challenges of transfer learning in MT is the mismatch in parent and child vocabularies. Only when the parent language pair and child language pair are identical can there be no such issue. Otherwise, when at least one of the languages in child language pair is distinct from parent languages, such an issue would arise. This is the case since vocabulary is language-specific and discrete (Kim et al., 2019). For example, if a parent model has its vocabulary built upon Spanish-English text, the vocabulary will contain only Spanish and English tokens. It can be unpredictable

Figure 2: A map of the ten South American Indigenous languages in our data. The color for each country and each language is arbitrarily assigned.

when tokenizing French text with such a vocabulary.

Zoph et al. (2016b) tackle this challenge by retaining the token embeddings for their target language since the parent target language and child target language are the same in their work. For parent and child source languages, they randomly map tokens of parent source language to tokens of child source language. Kocmi and Bojar (2018) take another approach of vocabulary building: the vocabulary is built upon 50% of parallel sentences of the parent language pair and 50% of those of the child language pair, so the vocabulary will contain tokens of both parent and child language pairs. Kocmi and Bojar (2020) introduce yet another simpler idea named 'Direct Transfer' where the parent vocabulary is used to train a child model. Although the parent vocabulary is not optimized for child language pair and can oversegment words in child language pair to smaller pieces than necessary, such a method still shows significant improvement in many language pairs. Kocmi and Bojar (2020) suspect that this could be due to good generalization of the transformer architecture to short subwords.

3 Experiments

3.1 Dataset

Our dataset is from AmericasNLP 2021 Shared Task on Open Machine Translation, which was co-located with the 2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2021) (Mager et al., 2021). The dataset contains

Language	ISO	Major location	Speakers
Aymara	aym	Bolivia	1,677,100
Bribri	bzd	Costa Rica	7,000
Asháninka	cni	Peru	35,200
Guarani	gn	Paraguay	6,652,790
Wixarika	hch	Mexico	52,500
Nahuatl	nah	Mexico	410,000
Hñähñu	oto	Mexico	88,500
Quechua	quy	Peru	7,384,920
Shipibo-Konibo	shp	Peru	22,500
Rarámuri	tar	Mexico	9,230

Table 1: Overview of the ten Indigenous languages (Eberhard et al., 2021).

Language Pair	Train	Dev	Test
es-aym	6,531	996	1,003
es-bzd	7,506	996	1,003
es-cni	3,883	883	1,003
es-gn	26,032	995	1,003
es-hch	8,966	994	1,003
es-nah	16,145	672	996
es-oto	4,889	599	1,001
es-quy	125,008	996	1,003
es-shp	14,592	996	1,003
es-tar	14,720	995	1,003

Table 2: Number of parallel sentences

parallel data of 10 language pairs: from Spanish to Aymara, Asháninka, Bribri, Guarani, Hñähñu, Nahuatl, Quechua, Rarámuri, Shipibo-Konibo, and Wixarika. An overview of these 10 Indigenous languages is shown in Table 1. The geographical distribution of the languages is depicted in Figure 2. We offer information about the dataset splits as distributed by the shared task organizers in Table 2. The shared task has two tracks: **Track One**, where the training split (Train) involves an arbitrary portion of development set, and **Track Two**, where Train involves *no* development data. In this work, we take *Track One* as our main focus and concatenate 90% of Dev split to Train to acquire a bigger training set. We also conduct experiments for *Track Two*, and we put the results in Appendix.

3.2 Baselines

We compare our results with the winner of the shared task Vázquez et al. (2021) who achieve highest performance in evaluation metrics for all language pairs in Track One (and winning 9 out of 10 language pairs in Track Two). They augment the training data by (1) gathering external parallel data, e.g. Bibles and Constitutions (2) collecting monolingual data of Indigenous languages and adopt back-translation method to generate syn-

Pair	Source	Target
es-aym	Los artistas de IRT ayudan a los niños en las escuelas.	IRT artistanakax jisk'a yatiqañ utankir wawanakaruw yanapapxi.
	Los artistas de I RT ayudan a los niños en las escuelas .	I RT artist ana ka x ji sk ' a y ati qa ñ u tank ir wa wan aka ru w ya nap ap xi .
es-bzd	Fui a un seminario que se hizo vía satélite.	Ye' dë'rö seminario ã wéx yö' satélite kĩ.
	Fui a un seminario que se hizo vía satélite .	Ye ' d ë ' r ö seminar io ã w é x y ö ' sat éli te k ĩ .
es-cni	Pensé que habías ido al campamento.	Nokenkeshireashitaka pijaiti imabeyetinta.
	Pensé que había s ido al campamento.	No ken ke shire ashi t aka p ija iti im ab eye tin ta .
es-gn	Veía a su hermana todos los días.	Ko'êko'êre ohecha heindýpe.
	Ve ía a su hermana todos los días .	Ko ' ê ko ' ê re oh e cha he in d ý pe .
es-hch	Era una selva tropical.	pe h+k+t+kai metsi+ra+ ye tsie nieka ti+x+kat+.
	Era una selva tropical .	pe h + k + t + ka i met si + ra + ye t s ie nie ka ti + x + ka t + .
es-nah	Santo trabajó para Disney y operó las tazas de té.	zanto quitequitilih Disney huan quinpexontih in cafen caxitl
	Santo trabajó para Disney y o per ó las taza s de té .	zan to quite qui til ih Disney h uan quin pex on t ih in cafe n ca xi t l
es-oto	Otros continúan reconociendo nuestro éxito.	ymana ditantho anumahditho goma npâgu
	Otros continúan reconociendo nuestro éxito .	y man a di tant ho an um ah di th o go ma n p â gu
es-quy	De vez en cuando me gusta comer ensalada.	Yananpiqa ensaladatam mikuytam munani
	De vez en cuando me gusta comer ensalada .	Yan an pi qa en s ala data m m iku y tam mun ani
es-shp	El Museo se ve afectado por las inversiones.	Ja Museora en oinai inversionesbaon afectana.
	El Museo se ve afectado por las inversiones .	Ja Museo ra en o ina i in version es ba on a fect ana .
es-tar	Es un hombre griego.	Bilé rejói Griego ju
	Es un hombre griego .	Bil é re j ó i Gri ego ju

Table 3: Example sentences tokenized by `es-en` tokenizer. **Light blue** : Original sentences (source or target). **Light green** : tokenized sentenses with tokens separated by whitespace.

thetic parallel data. They build a 6-layered transformer (Vaswani et al., 2017) with 8 heads by first pretrain it with `es-en` parallel data and then finetune it with both internal dataset provided by the organizer and external augmented datasets of all 10 language pairs to produce a multilingual MT model. In this work, we leverage solely the dataset provided by the shared task organizer to test if our method works with scarce data.

3.3 Data Preprocessing

As mentioned in section 2.3, the cross-lingual challenge exists when one or both sides of child language pair is distinct from the parent languages which is the case to all of the our 10 language pairs. To tackle this, we opt for 'direct transfer' method, due to its simplicity, to exploit parent vocabulary for child model. As Kocmi and Bojar (2020) find that the words of child language are oversegmented with direct transfer, similar to their finding, we observe that the words of Indigenous language words can be oversegmented. As shown in Table 3, it can be seen that the source sentences are tokenized reasonably well with mostly one token per word. By contrast, the words of child target language are generally oversegmented into short subwords. The statistics of the tokenization is shown in Table 8. An analysis of oversegmentation phenomenon is given in section 5.3.

3.4 Parent Models

We offer two types of parent models, bilingual models and multilingual models.

Bilingual Models. For bilingual models, we leverage publicly accessible pretrained models from Huggingface (Wolf et al., 2020) as provided by Helsinki-NLP (Tiedemann and Thottingal, 2020). The pretrained MT models released by Helsinki-NLP are trained on OPUS, an open source parallel corpus (Tiedemann, 2012). Underlying these models is the Transformer architecture of Marian-NMT framework implementation (Junczys-Dowmunt et al., 2018). Each model has six self-attention layers in encoder and decoder parts, and each layer has eight attention heads. The three bilingual models we specifically use are each pretrained with OPUS Spanish-Catalan, Spanish-English, and Spanish-Romanian data.[2]

We choose these models because their source language is Spanish so they may have good Spanish subword embeddings. In this regard, as Adelaar (2012) point out, during the colonial period, Spanish grammatical concepts were introduced to some

[2]Tiedemann and Thottingal (2020) do not provide information about the size of OPUS data exploited in each of these models.

South American Indigenous languages. In addition, we pick Spanish-Catalan and Spanish-Romanian MT models because Catalan and Romanian are two languages in the same Romance language family as Spanish, and we suspect our ten Indigenous languages of South America may have some affinity to Spanish. We also choose Spanish-English as a contrastive model because English is in the Germanic language family rather than Romance and that the MT models built around English usually are well-performing due to its rich resource of parallel data.

Multilingual Models. For our multilingual models, we exploit mBART50 (Tang et al., 2020). mBART50 can be seen as an extension of mBART (Liu et al., 2020). mBART (or more specifically mBART25) is a multilingual sequence-to-sequence generative model pretrained on 25 monolingual datasets and fine-tuned on 24 bilingual datasets which cover all 25 languages used in pretraining. mBART50 takes mBART as a starting point and enlarges its embedding layers to accommodate tokens of 25 new languages to support 50 languages. mBART50 adopts multilingual fine-tuning under three scenarios: one-to-many, many-to-one, and many-to-many where 'one' represents English. We choose the one that is trained under many-to-many scenario to ensure (1) Spanish is fine-tuned as a source language so it may maintain a good representation for Spanish tokens (2) `es-en` language pair is covered so we can produce an intermediate model with `es-en` fine-tuning to test the effectiveness of curriculum learning.

Model	Target	Our BLEU	Our chrF	SOTA BLEU	SOTA chrF
es-ca		1.445	0.2344		
es-en		2.432	0.277		
es-ro	aym	2.009	0.2705	**2.8**	**0.31**
mBart50		2.017	0.2672		
mBART50$_{curr}$		2.23	0.2725		
es-ca		7.242	0.2378		
es-en		9.952	0.2753		
es-ro	bzd	10.278	0.2867	5.18	0.213
mBart50		**12.898**	**0.3082**		
mBART50$_{curr}$		12.495	0.3036		
es-ca		4.742	0.2984		
es-en		5.973	0.3367		
es-ro	cni	5.21	0.3229	6.09	0.332
mBart50		5.632	0.3183		
mBART50$_{curr}$		**6.255**	**0.3432**		
es-ca		4.395	0.2909		
es-en		5.918	0.3341		
es-ro	gn	5.853	0.3279	**8.92**	**0.376**
mBart50		6.329	0.3367		
mBART50$_{curr}$		6.449	0.3387		
es-ca		13.375	0.3061		
es-en		15.922	0.3461		
es-ro	hch	15.298	0.3444	15.67	**0.36**
mBart50		**16.731**	0.3397		
mBART50$_{curr}$		16.659	0.3391		
es-ca		1.95	0.2763		
es-en		2.045	0.2913		
es-ro	nah	1.734	0.2929	**3.25**	0.301
mBart50		2.422	0.2969		
mBART50$_{curr}$		2.947	**0.3015**		
es-ca		4.344	0.2268		
es-en	oto	6.414	0.2522		
es-ro		4.14	0.2315	5.59	0.228
mBart50		**7.504**	**0.265**		
mBART50$_{curr}$		7.489	0.2617		
es-ca		2.817	0.3449		
es-en		4.149	0.3788		
es-ro	quy	3.192	0.3718	**5.38**	**0.394**
mBart50		4.689	0.3928		
mBART50$_{curr}$		4.95	0.3881		
es-ca		5.184	0.2627		
es-en		7.664	0.3326		
es-ro	shp	6.663	0.32	**10.49**	**0.399**
mBart50		10.022	0.3556		
mBART50$_{curr}$		9.702	0.349		
es-ca		1.724	0.217		
es-en		2.432	0.248		
es-ro	tar	2.034	0.2358	**3.56**	**0.258**
mBart50		2.433	0.2396		
mBART50$_{curr}$		2.261	0.2362		

Table 4: Modeling results (of Track One). The bold-faced numeric values are the best performances. Source language is always Spanish so it is ignored. SOTA values represent the state-of-the-art performance which are all from Vázquez et al. (2021)

3.5 Training Approach

Bilingual Model Training. We fine-tune each of our three bilingual models for $60,000$ steps with Spanish-Indigenous data, acquiring performance on Dev every $1,000$ steps. The final model is the checkpoint that has the lowest validation/Dev loss, and it is what we use for predicting on Test. Our beam size (for beam search) (Reddy et al., 1977; Graves, 2012) is 6. We use a batch size [3] of 15 for our bilingual models. It takes $\sim$ 6 hours to train on four Nvidia V100-SXM2-16GB GPUs for each model per language pair.

Multilingual Model Training. For our multilingual setting, we train a model for each of the Spanish→Indigenous language pairs and it takes $\sim$ 12 hours to train on four NVIDIA Tesla V100 32GB NVLink GPUs for each model per language pair. We have two scenarios: mBART50 and mBART50$_{curr}$. Both of them have batch size [3] to be 5, and the beam size to be 6.

mBART50. For our first multilingual scenario, we fine-tune mBART50 on Spanish-Indigenous data immediately after tokenization. Similar to our bilingual models, we fine-tune the mBART50 model for $60,000$ steps, measuring performance on Dev every $1,000$ steps, and taking the checkpoint with the best validation loss as our final model used for prediction on Test.

mBART50$_{curr}$. For the second scenario, mBART50$_{curr}$ is first fine-tuned on `es-en` data for 300 steps. The validation is done every 20

[3] The batch sizes are small so the data can be loaded in the GPU memory.

Pair	Sentence
es-aym	nanakan utaxax khaysa Concord uksanx kimsatunka waranqa acres ukhamarac walja uywanakarakiw utjaraki.
	Concord markan nanakan utanx 30000 acre ukhamarak walja uywanaka utji.
es-bzd	Sa' ù Concord wã 30000 acres tã' nã tãîx íyiwak.
	Sa' ù ã Concord e' kĩ káx dör 20.000 acres tãîx íyiwak tãîx.
es-cni	Abanko Concordki otimi 30000 acres jeri osheki birantsipee.
	Ashi pankotsi Concordi timatsi 30000 acres aisati osheki piratsipee.
es-gn	Ore róga Concord-pe otroko 30000 acres ha hetaiterei orerymba.
	Ñane óga Concord-pe oreko 30000 acre ha hetaiterei mymba.
es-hch	ta kí wana Concord pe xeiya 30000 acres tsiere y+ wa+kawa yeuta meteu uwa.
	ta ki wana Concord pexeiya xeiya xeitewiyari acre meta wa+kawa te+teri.
es-nah	tochan Concord quipiya miyac tlalli nohiya miyac tlapiyalli.
	Tehuancalco Concord quipiah macualli tlatqui ihuan miyac yolcameh.
es-oto	mangû game ane Concord phodi 30000 yñi xi nã hmudi on yzuî
	Goma na madoongû ane Concord phodi 30000 yqhẽya xi na ngû on ybaoni
es-quy	Corcord nisqapi wasiykum kimsa chunka waranqa acres nisqan kan hinataq achkallaña uywa.
	Concordpi wasiykuqa 30000 acres hinaspa achka uywakunam
es-shp	Non xobo Concordainra 30000 acresya iki itan kikin icha yoinabo.
	Concordainra non xoboa riki 30000 acres itan kikin icha yoinabo.
es-tar	Tamó e'perélachi Concord anelíachi besá makói acres nirú a'lí weká namúti jákami shi.
	Concord anelíachi benéalachi, bilé mili akí weká nirú, wekabé namuti nirú.

Table 5: Example of ground truth and prediction of the Spanish sentence "Nuestra casa en Concord tiene 30000 acres y un montón de animales." (**Eng.** *Our home in Concord has 30,000 acres and lots of animals.*) by mBART50. The 'z' in 'yzuî' of es-oto is actually a Unicode character of code point U+0225 which is a 'z' with hook. **Light blue**: Ground Truth . **Light green**: Prediction .

steps where the checkpoint with lowest loss will be fine-tuned on Spanish-Indigenous language pair for $60,000$ steps, validated every $1,000$ steps to pick the best checkpoint with lowest validation loss. Our mBART50$_{\text{curr}}$ is inspired by the concept of *curriculum learning* (Soviany et al., 2021) where a model can possibly be improved when first trained on an easier task and followed by training on a harder task. In our case here, translating Spanish to English is considered an easier task because mBART50 is pretrained with `es-en` language pair; whereas Spanish to South American Indigenous languages is considered a more difficult job since mBART50 has not seen any of the 10 Indigenous languages before.

4 Results

We evaluate the translation performance with two automatic MT evaluation metrics: BLEU (Papineni et al., 2002) and chrF (Popović, 2015). chrF is an automatic evaluation metric for MT task which can be seen as a F-score for text and has value between 0 and 1. BLEU and chrF are the two metrics adopted by AmericasNLP 2021 Shared Task. We surpassed the winner of AmericasNLP2021 (Vázquez et al., 2021), in either or both metrics, for 5 language pairs with the following languages as target: Bribri (bzd), Asháninka (cni), Wixarika (hch), Nahuatl (nah), and Hñähñu (oto). Notably, we double the performance in BLEU score for `es-bzd`, increasing by about 7.7 BLEU scores and 0.1 chrF. We increase ~ 2 BLEU points in `es-oto` and ~ 1 BLEU points in es-hch. For both `es-cni` and `es-nah`, we slightly surpass their performance in both metrics. The performance of experiments are shown in Table 4. We also offer example predictions in Table 5.

All surpassing results are achieved by mBART50 or mBART50$_{\text{curr}}$. Surprisingly, mBART50$_{\text{curr}}$ does not consistently improve the performance if compared to mBART50; some of the best performances are achieved by mBART50 (`es-bzd`, `es-hch`, `es-oto`). Nevertheless, mBART50$_{\text{curr}}$ performs slightly better than mBART50 on average by 0.076 BLEU and 0.0034 chrF. Averagely, mBART50$_{\text{curr}}$ achieves 7.143 BLEU score and 0.3134 chrF while

mBART50 achieves 7.068 BLEU score and 0.31 chrF. Generally, multilingual models perform better than bilingual model despite that in some language pairs, es-en model performs nearly as good as multilingual models and outperform multilingual models in es-aym and es-tar. For 3 bilingual models, es-en model generally outperforms the other two es-ca and es-ro models.

5 Discussion

5.1 Comparisons to SOTA

We are able to surpass previous SOTA in five language pairs and mBART50$_{curr}$ achieves 7.143 BLEU and 0.3134 chrF on average, comparing to previous SOTA having 6.693 BLEU and 0.3171 chrF on average. It can be hypothesized that the reason why we are able to improve average BLEU score by 0.45, accomplish comparable average chrF, and surpass in five language pairs is because we use an MT model pretrained on 50 languages, while Vázquez et al. (2021) pretrain their model only on es-en. We suspect that there could be some languages, other than Spanish and English, which contribute to positive transfer to Indigenous languages. Unlike Vázquez et al. (2021), we do not leverage external data to build a larger train set. Nor do we build a single multilingual model for all 10 language pairs, but we rather train one model for each language pair (where every single language pair is independent from the other pairs). The approach of Vázquez et al. (2021) may be able to afford some positive transfer between different Indigenous languages, and hence can be one of our future directions.

5.2 Fusional to Polysynthetic Translation

There is literature showing that when translating between a polynthetic[4] and a fusional language, some morphological information of the polysynthetic language is 'lost'. This is especially relevant to our work since Spanish is a fusional language and many Indigenous languages in our work are polysynthetic (Mager et al., 2021). Mager et al. (2018b) carry out a morpheme-to-morpheme alignment between Spanish and polynthetic Indigenous languages, including Nahuatl (nah) and Wixarika (hch) which are both in our data and show that the meanings carried by some polysynthetic morphemes have no Spanish counterpart. This makes it difficult to translate from polysynthetic languages to fusional Spanish without losing some morphological information. This is also a challenge to translate from fusional Spanish to polysynthetic languages, as there may be no contexts provided to infer the missing parts. This is particularly the case for sentence-level (vs. document level) translation.

We hypothesize that if there is loss in morphological information when translating from a fusional to polysynthetic languages, either or both the sentence length and word length of prediction will be shorter than the gold standard because some parts in the prediction are left out while the ground truth may contain them. We therefore compare average sentence length and average word length between our gold standard and prediction as shown in Table 6. However, we find that this hypothesis does not hold for most language pairs as most of them are having similar average sentence and word lengths in gold standard and predictions. We suspect that this is because the test sets are translated from Spanish to Indigenous languages by human translators in a sentence-level fashion, the translators may leave out the missing morphological information when translating Spanish into Indigenous languages due to inability to infer the missing information. As Mager et al. (2018b) state:

> The important Wixarika independent asserters "p+" and "p" are the most frequent morphemes in this language. However, as they have no direct equivalent in Spanish, their translation is mostly ignored. ... This is particularly problematic for the translation in the other direction, i.e., from Spanish into Wixarika, as a translator has no information about how the target language should realize such constructions. Human translators can, in some cases, infer the missing information. However, without context it is generally complicated to get the right translation.

As this is a sentence-level translation task where contexts can be hard to infer, the gold standard may not contain these parts at the first place. However, a further qualitative linguistic investigation is required to spot the cause of this phenomenon.

[4]Polysynthetic languages generally have a more complex morphological system, possibly each word consisting of several morphemes (Haspelmath and Sims, 2013; Campbell et al., 2012).

Target	Sent (Gold)	Sent (Pred)	Word (Gold)	Word (Pred)
aym	6.71	7.97	7.83	5.88
bzd	11.66	10.83	3.79	3.86
cni	6.41	6.1	8.57	8.17
gn	6.46	6.66	6.5	6.46
hch	9.97	8.55	5.35	5.61
nah	6.7	6.9	7.11	7.16
oto	10.38	9.69	4.47	4.01
quy	6.73	6.04	7.71	8.19
shp	8.82	7.77	5.95	5.98
tar	9.36	8.75	5.15	4.86

Table 6: The averages of sentence and word length of test set. The predictions are produced by mBART50. Sent (Gold) and Sent (Pred) are the average sentence length of gold standard and prediction, respectively. Word (Gold) and Word (Pred) are the average word length of gold standard and prediction, respectively. Sentence length is calculated as number of words in each sentence (by splitting sentence with whitespace). Word length is calculated as number of characters in each word.

5.3 Tokenization with Parent Vocabulary

As discussed in Section 3.3, we re-use the tokenizer of parent models without building new ones for child language pairs. We observe that the tokens in target sentences tend to be very short. That is, tokens in these target sentences often consist of one or two characters as can be seen in Table 3. Hence, target sentences do seem to be encountering oversegmentation. This could be causing loss of meaning as these smaller segments differ from what would be suited for a given Indigenous language.

We further offer statistics related to tokenization with the calculation details provided in Appendix A.2, and results shown in Table 8 (in Appendix). The difference between the average length of tokens in source and target languages is quite large. For example, for the language pair `es-bzd`, when tokenized with the `es-en` tokenizer, average token length for the source language is 3.43 while that for the target language is 1.21. This indicates that tokens in source data consist averagely of $\sim$ 3.5 characters while tokens in target data consist averagely of $\sim$ 1.2 characters. For this particular `es-bzd` language pair whose words in target sentences are on average oversegmented into nearly one character per token, the performance is surprisingly better than the previous SOTA. For the other nine language pairs whose words in target sentences are segmented into tokens consisting of $\sim$ 1 to $\sim$ 2 characters, the models are still capable of reasonably carrying out the translation task. As Kocmi and Bojar (2020) conjecture, this may be a case in point where a model is able to simply generalize well to short subwords.

5.4 Non-Standard Orthography

Based on a pilot investigation, we find the lack of orthographic standardization to be potentially problematic. We place relevant sample predictions in Table 5. For example, for the prediction of `es-aym` pair, we find that a word is predicted nearly correctly with just a difference in one character: ground truth 'ukhamarac' is predicted to be 'ukhamarak'. As Coler (2014) point out, this may be an issue of non-standard orthography since some Aymara speakers do not consistently differentiate between 'c' and 'k'. It can be hypothesized that the model generalizes to the 'ukhamarak' as a translation of a phrase/word because of potentially relatively higher number of occurrences of 'ukhamarak' than 'ukhamarac' in training data. In fact, 'ukhamarak' (including its variants with characters following such as in 'ukhamaraki' and 'ukhamarakiw') appears 489 times in the training set while 'ukhamarac' appears zero time (it only exist in test set). Although 'ukhamarac' and 'ukhamarak' can be viewed as the same word, these are still not counted as a match by some automatic evaluation metrics (including metrics based on BLEU, which we adopt in this work). Interestingly, cases such as the current one illustrates a challenge for automatic MT metrics when evaluating on languages without standard orthography.

6 Conclusion

In this paper, we describe how we apply transfer learning to MT from Spanish to ten low-resource South American Indigenous languages. We fine-tune pretrained bilingual and multilingual MT models on downstream Spanish to Indigenous language pairs and show the utility of these models. We are able to surpass SOTA in five language pairs using multilingual pretrained MT models without leveraging any external data. Empirically, our results show that this method performs robustly even with an oversegmentation issue on the target side. We also discuss multiple issues that interact with our task, including translating between languages of different morphological structures, effect of tokenization, and non-standard orthography.

Limitations

One challenge for working on a wide host of Indigenous languages is insufficient knowledge of these languages, which also applies to us: We report models on ten different Indigenous languages none of which is the native tongue of us. In spite of this limitation, we strive to acquire linguistic knowledge about the languages we work on so that our arguments are informed. Regardless, we believe that lack of native knowledge of the languages remains a limitation at our side.

In section 5.3, our claim of potential oversegmentation is based on an assumption that human languages tend to not have morphemes with just a single character. That is, we assume that these languages should have longer morphemes in general. However, again, a more definitive approach to the problem would perhaps require expert linguistic knowledge of the languages under study. In absence of (detailed) linguistic analyses of the Indigenous language we treat, this again remains a constraint.

Ethics Statement

We develop methods for low-resource machine translation. Because our models are trained on limited amounts of data, and hence make frequent errors, they may not be immediately useful for the general public. However, our hope is that our work will propel MT progress on the ten Indigenous languages we tackle.

There are also some biases in the models and the textual data we use to train them. The datasets we use to train our models (Mager et al., 2021) is a translations of XNLI (Conneau et al., 2018), which itself is derived from MultiNLI (Williams et al., 2018). Our bilingual model for each pair is trained on OPUS corpus that is derived from different sources. The multilingual model mBART50 is also trained on multiple datasets, including IWSLT, WMT, and TED. Due to the complexity of neural models, it is hard to explicitly state how these biases can contribute to the failure modes. However, we explicitly state the existence of sources of potential biases to raise the awareness of the readers.

References

Willem FH Adelaar. 2012. Chapter historical overview: Descriptive and comparative research on south american indian languages. In *The indigenous languages of South America: A comprehensive guide*. De Gruyter.

Lyle Campbell, Verónica Grondona, and HH Hock. 2012. *The indigenous languages of South America*. de Gruyter.

Monojit Choudhury and Amit Deshpande. 2021. How linguistically fair are multilingual pre-trained language models? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12710–12718.

Matt Coler. 2014. *A grammar of Muylaq'Aymara: Aymara as spoken in Southern Peru*. Brill.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2021. Ethnologue: Languages of the world. twenty-fourth edition. Dallas, Texas. SIL International.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *CoRR*, abs/1211.3711.

Martin Haspelmath and Andrea Sims. 2013. *Understanding morphology*. Routledge.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Yunsu Kim, Yingbo Gao, and Hermann Ney. 2019. Effective cross-lingual transfer of neural machine translation models without shared vocabularies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1246–1257, Florence, Italy. Association for Computational Linguistics.

Tom Kocmi and Ondrej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. *CoRR*, abs/1809.00357.

Tom Kocmi and Ondřej Bojar. 2020. Efficiently reusing old models across languages via transfer learning. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 19–28, Lisboa, Portugal. European Association for Machine Translation.

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.

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

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. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018a. Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Manuel Mager, Elisabeth Mager, Alfonso Medina Urrea, Iván Meza, and Katharina Kann. 2018b. Lost in translation: Analysis of information loss during machine translation between polysynthetic and fusional languages. *CoRR*, abs/1807.00286.

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.

Oscar Moreno. 2021. The REPU CS' Spanish–Quechua submission to the AmericasNLP 2021 shared task on open machine translation. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 241–247, Online. Association for Computational Linguistics.

El Moatez Billah Nagoudi, Wei-Rui Chen, Muhammad Abdul-Mageed, and Hasan Cavusoglu. 2021. Indt5: A text-to-text transformer for 10 indigenous languages. *CoRR*, abs/2104.07483.

Tan Ngoc Le and Fatiha Sadat. 2020. Revitalization of indigenous languages through pre-processing and neural machine translation: The case of Inuktitut. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4661–4666, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. *CoRR*, abs/1708.09803.

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.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2021. Neural machine translation for low-resource languages: A survey.

D Raj Reddy et al. 1977. Speech understanding systems: A summary of results of the five-year research effort. *Department of Computer Science. Camegie-Mell University, Pittsburgh, PA*, 17:138.

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. 2021. Curriculum learning: A survey. *arXiv preprint arXiv:2101.10382*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

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

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

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

Raúl Vázquez, Yves Scherrer, Sami Virpioja, and Jörg Tiedemann. 2021. The Helsinki submission to the AmericasNLP shared task. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 255–264, Online. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2020. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016a. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016b. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575.

A Appendix

A.1 Additional Experiments

We conduct additional experiments for Track Two as mentioned in Section 3.1. This additional experiment have identical settings as Track One except that the train set does not involve sentences in development set. We surpass the state-of-the-art performance in 4 out of 10 language pairs in either or both BLEU and chrF. Similar to the results in Track One, multilingual MT models perform better than bilingual ones while there are no consistent winner between mBART50 and $\text{mBART50}_{\text{curr}}$.

A.2 Tokenization Output

As mentioned in Section 5.3, we calculate statistics related to tokenization on training data as shown in Table 8. To calculate these statistics, padding tokens, end of sentence tokens and the underscore (or more precisely, U+2581) prepended due to sentencePiece technique (Kudo and Richardson, 2018) are removed from the tokenized sentences. Sentence length is calculated as number of tokens in a sentence. Token length is calculated as the number of characters in a token. Average sentence length is calculated by averaging the sentence lengths of all sentences. Average token length is calculated as

$$\frac{\sum_{i=1}^{N}\sum_{j=1}^{n_i}|s_{ij}|}{\sum_{i=1}^{N} n_i}$$

where n_i denotes the number of tokens in i^{th} sentence and N denotes the number of sentences in training data. $|s_{ij}|$ denotes the length (number of characters) of j^{th} token in i^{th} sentence.

Model	Target	Dev BLEU	Dev chrF	Test BLEU	Test chrF	SOTA BLEU	SOTA chrF
es-ca		2.415	0.227	1.0	0.197		
es-en		2.503	0.261	1.253	0.22		
es-ro	aym	2.642	0.2666	1.369	0.2273	**2.29**	**0.283**
mBART50		3.105	0.275	1.38	0.236		
mBART50$_{curr}$		3.034	0.2679	1.37	0.2291		
es-ca		2.033	0.15	2.217	0.153		
es-en		2.987	0.168	3.437	0.178		
es-ro	bzd	2.803	0.1709	3.308	0.1816	2.39	0.165
mBART50		4.205	0.188	4.272	**0.197**		
mBART50$_{curr}$		4.072	0.1871	**4.438**	0.1911		
es-ca		2.628	0.212	2.429	0.201		
es-en		1.671	0.212	1.623	0.208		
es-ro	cni	1.639	0.2225	1.829	0.209	3.05	**0.258**
mBART50		3.074	0.26	**3.539**	0.25		
mBART50$_{curr}$		3.404	0.2573	3.537	0.2491		
es-ca		3.637	0.245	3.523	0.254		
es-en		4.206	0.282	4.217	0.297		
es-ro	gn	3.784	0.2771	4.699	0.291	**6.13**	**0.336**
mBART50		4.911	0.287	4.801	0.304		
mBART50$_{curr}$		4.496	0.2795	4.702	0.2918		
es-ca		5.618	0.191	7.595	0.197		
es-en		6.578	0.234	8.995	0.245		
es-ro	hch	7.536	0.2594	10.123	0.2732	9.63	**0.304**
mBART50		8.617	0.254	11.526	0.272		
mBART50$_{curr}$		9.067	0.2582	**11.539**	0.2731		
es-ca		0.753	0.239	0.705	0.222		
es-en		0.73	0.25	0.772	0.22		
es-ro	nah	1.06	0.2619	0.6983	0.2363	**2.38**	**0.266**
mBART50		1.69	0.281	1.497	0.255		
mBART50$_{curr}$		1.704	0.2731	1.78	0.2412		
es-ca		0.536	0.122	0.86	0.12		
es-en		0.745	0.124	1.039	0.121		
es-ro	oto	0.5125	0.1198	0.8811	0.1226	**1.69**	**0.147**
mBART50		0.816	0.133	1.354	0.132		
mBART50$_{curr}$		0.8851	0.1348	1.338	0.1331		
es-ca		2.199	0.322	2.191	0.328		
es-en		2.217	0.337	2.892	0.347		
es-ro	quy	2.081	0.3416	2.094	0.3539	2.91	0.346
mBART50		2.242	0.356	**3.167**	**0.366**		
mBART50$_{curr}$		2.516	0.355	3.038	0.3659		
es-ca		1.511	0.178	1.234	0.168		
es-en		2.134	0.21	2.017	0.196		
es-ro	shp	1.964	0.2205	1.43	0.2048	**5.43**	**0.329**
mBART50		2.131	0.194	2.013	0.185		
mBART50$_{curr}$		2.067	0.1947	1.809	0.1856		
es-ca		0.256	0.095	0.047	0.084		
es-en		0.034	0.057	0.023	0.05		
es-ro	tar	0.1583	0.094,38	0.2985	0.089,32	**1.07**	**0.184**
mBART50		0.09	0.093	0.073	0.101		
mBART50$_{curr}$		0.1212	0.094,63	0.090,13	0.1007		

Table 7: Modeling results of Track Two. The boldfaced numeric values are the best performances. SOTA values represent the state-of-the-art performance which are all from Vázquez et al. (2021) except that the es-quy SOTA chrF value is from (Moreno, 2021). Source language is always Spanish so it is ignored.

model	Target Lang	source avg sentence length	target avg sentence length	source avg token length	target avg token length
es-ca	aym	26.37	49.1	3.61	1.81
es-en		24.55	45.07	3.88	1.99
es-ro		25.74	47.9	3.71	1.91
mBART50		27.4	37.85	3.66	2.53
es-ca	bzd	9.42	22.42	3.24	1.28
es-en		8.9	21.43	3.43	1.21
es-ro		9.13	21.52	3.34	1.23
mBART50		10.75	19.67	3.3	1.54
es-ca	cni	17.6	30.56	3.33	1.92
es-en		16.72	27.78	3.51	2.12
es-ro		17.31	29.17	3.44	2.04
mBART50		19.38	23.9	3.27	2.69
es-ca	gn	31.89	50.6	3.69	2.01
es-en		30.15	50.77	3.9	2.0
es-ro		31.92	52.45	3.73	1.97
mBART50		33.79	41.34	3.63	2.6
es-ca	hch	11.15	23.01	3.24	1.68
es-en		10.49	21.56	3.44	1.79
es-ro		10.76	22.27	3.35	1.73
mBART50		13.34	20.14	3.08	2.17
es-ca	nah	33.7	51.39	3.03	1.83
es-en		34.36	49.58	2.96	1.94
es-ro		34.44	51.52	2.95	1.83
mBART50		36.78	45.54	2.87	2.32
es-ca	oto	18.0	37.72	3.14	1.64
es-en		18.2	36.06	3.1	1.51
es-ro		18.49	37.58	3.07	1.7
mBART50		20.62	32.91	2.98	1.82
es-ca	quy	20.16	42.8	3.65	1.83
es-en		19.26	37.68	3.82	2.08
es-ro		20.16	41.45	3.73	1.92
mBART50		22.96	31.47	3.42	2.65
es-ca	shp	9.71	16.53	3.19	1.75
es-en		9.06	15.56	3.41	1.85
es-ro		9.42	15.84	3.28	1.82
mBART50		11.12	13.54	3.23	2.5
es-ca	tar	12.48	19.4	2.97	1.48
es-en		12.83	18.32	2.89	1.57
es-ro		13.08	19.33	2.84	1.5
mBART50		14.15	15.64	2.98	2.16

Table 8: Token statistics for our Train set. The way of calculating these figures is presented in Appendix A.2. Since mBART50 and mBART50$_{curr}$ are having exactly same statistics as they use same tokenizer, the statistics of mBART50$_{curr}$ are ignored.