Multilingual unsupervised sequence segmentation transfers to extremely low-resource languages

Anonymous ACL submission

Abstract

We show that unsupervised sequence-segmentation performance can be transferred to extremely low-resource languages by pre-training a Masked Segmental Language Model (Downey et al., 2021) multilingually. Further, we show that this transfer can be achieved by training over a collection of low-resource languages that are typologically similar (but phylogenetically unrelated) to the target language. In our experiments, we transfer from a collection of 10 Indigenous American languages (AmericasNLP, Mager et al., 2021) to K'iche', a Mayan language. We compare our multilingual model to a monolingual (from-scratch) baseline, as well as a model pre-trained on Quechua only. We show that the multilingual pre-trained approach yields consistent segmentation quality across target dataset sizes, exceeding the monolingual baseline in 6/10 experimental settings. Our model yields especially strong results at small target sizes, including a zero-shot performance of 20.6 F1. These results have promising implications for low-resource NLP pipelines involving human-like linguistic units, such as the *sparse transcription* framework proposed by Bird (2020).

1 Introduction

Unsupervised sequence segmentation (at the word, morpheme, and phone level) has long been an area of interest in languages without whitespace-delimited orthography (e.g. Chinese, Uchiumi et al., 2015; Sun and Deng, 2018), morphologically complex languages without rule-based morphological anlayzers (Creutz and Lagus, 2002), and automatically phone-transcribed speech data (Goldwater et al., 2009; Lane et al., 2021) respectively. It has been particularly important for lower-resource languages in which there is little or no gold-standard data on which to train supervised models (Joshi et al., 2020).

In modern neural end-to-end systems, unsupervised segmentation is usually performed via information-theoretic alogrithms such as BPE (Sennrich et al., 2016) and SentencePiece (Kudo and Richardson, 2018). However, the segmentations they produce are largely non-sensical to humans (Park et al., 2021). The motivating tasks listed above instead require unsupervised approaches that correlate more closely with human judgements of the boundaries of linguistic units. For example, in a human-in-the-loop framework such as the *sparse transcription* proposed by Bird (2020), lexical items are automatically proposed to native speakers for confirmation, and it is important that these candidates be (close to) sensical, recognizable pieces of language.

In this paper, we investigate the utility of recent models that have been developed to conduct unsupervised segmentation as a byproduct of a language modeling objective (e.g. Kawakami et al., 2019; Downey et al., 2021, see Section 2). The key idea is that recent breakthroughs in crosslingual language modeling and transfer learning (Conneau and Lample, 2019; Artetxe et al., 2020, *inter alia*) can be leveraged to facilitate transferring unsupervised segmentation performance to a new target language, using these types of language models.

Specifically, we investigate the effectiveness of multilingual pre-training in a Masked Segmental Language Model (Downey et al., 2021) when applied to a low-resource target. We pre-train our model on the ten Indigenous languages of the 2021 AmericasNLP shared task dataset (Mager et al., 2021), and apply it to another low-resource, Indigenous, and morphologically complex language of Central America: K'iche' (quc), which at least phylogenetically is unrelated to the pre-training languages (Campbell et al., 1986).

We hypothesize that multilingual pre-training on similar, possibly contact-related languages, will outperform both a monolingual baseline trained from scratch and a model pre-trained on a single language (Quechua) with the same amount of pre-training data. We also expect that the pre-trained models will perform increasingly better than the monolingual baseline the smaller the target corpus is.

Indeed, our experiments show that a pre-trained multilingual model provides stable performance across all dataset sizes and far exceeds the monolingual baseline at low-to-medium target sizes. We additionally show that the multilingual model achieves a zero-shot segmentation performance of 20.6 F1 on the K'iche' data, where the monolingual baseline yields a score of zero. These results suggest that transferring from a multilingual model can greatly assist unsupervised segmentation in very low-resource languages, even those that are morphologically rich. It also supports the idea that transfer from multilingual models works at a more moderate scale than is typical for recent crosslingual models.

In the following section, we overview work relating to unsupervised segmentation, crosslingual pre-training, and transfer-learning (Section 2). We then introduce the multilingual data used in our experiments, and the additional pre-processing we performed to prepare the data for pre-training (Section 3). Next we provide a brief overview of the type of Segmental Language Model used in our experiments, as well as our multilingual pre-training process (Section 4). After this, we describe our experimental process applying the pre-trained and from-scratch models to varying target data sizes (Section 5). Finally, we discuss the results of our experiments and their significance for low-resource pipelines, both within unsupervised segmentation and for other NLP tasks more generally (Sections 6 and 7).

2 Related Work

Work related to the present study largely falls either into the field of (unsupervised) word segmentation, or the field(s) of crosslingual language modeling and transfer learning. To our knowledge, we are the first to propose a multilingual model for unsupervised word/morpheme-segmentation.

Unsupervised Segmentation Current state-of-the-art unsupervised segmentation has largely been achieved with Bayesian models such as Hierarchical Dirichlet Processes (Teh et al., 2006; Goldwater et al., 2009) and Nested Pitman-Yor (Mochihashi et al., 2009; Uchiumi et al., 2015). Adaptor Grammars (Johnson and Goldwater, 2009) have been successful as well. Models such as *Morfessor* (Creutz and Lagus, 2002), which are based on Minimal Description Length (Rissanen, 1989) are also widely used for unsupervised morphology.

As Kawakami et al. (2019) note, most of these models have weak language modeling ability, being unable to take into account much other than the immediate local context of the sequence. Another line of techniques has focused on models that are both strong language models and good for sequence segmentation. Many are in some way based on Connectionist Temporal Classification (Graves et al., 2006), and include Sleep-WAke Networks (Wang et al., 2017), Segmental RNNs (Kong et al., 2016), and Segmental Language Models (Sun and Deng, 2018; Kawakami et al., 2019; Wang et al., 2021; Downey et al., 2021). In this work, we conduct experiments using the Masked Segmental Language Model of Downey et al. (2021), due to its good performance and scalability, the latter usually regarded as an obligatory feature of multilingual models (Conneau et al., 2020a; Xue et al., 2021, *inter alia*).

Crosslingual and Transfer Learning Crosslingual modeling and training has been an especially active area of research following the introduction of language-general encoder-decoders in Neural Machine Translation, offering the possibility of zero-shot translation (i.e. translation for language pairs not seen during training; Ha et al., 2016; Johnson et al., 2017).

The arrival of crosslingual language model pretraining (XLM, Conneau and Lample, 2019) further demonstrated that large models pre-trained on multiple languages yielded state-of-the-art performance across an abundance of multilingual tasks including zero-shot text classification (e.g. XNLI, Conneau et al., 2018), and that pre-trained transformer encoders provide great initializations for MT systems and language models in very low-resource languages.

Since XLM, numerous studies have attempted to single out which components of crosslingual training contribute to transferability from one language to another (e.g. Conneau et al., 2020b). Others have questioned the importance of multilingual training, and have instead proposed that even monolingual pre-training can provide effective transfer to new languages (Artetxe et al., 2020). Though some like

Lin et al. (2019) have tried to systematically study which aspects of pre-training languages/corpora enable effective transfer, in practice the choice is often driven by availability of data and other ad-hoc factors.

Currently, large crosslingual successors to XLM such as XLM-R (Conneau et al., 2020a), MASS (Song et al., 2019), mBART (Liu et al., 2020), and mT5 (Xue et al., 2021) have achieved major success, and are the starting point for a large portion of multilingual NLP systems. These models all rely on an enormous amount of parameters and pre-training data, the bulk of which comes from very high-resource languages. In contrast, in this paper we assess whether multilingual pre-training on a suite of very low-resource languages, which combine to yield a moderate amount of unlabeled data, can provide good transfer to similar languages which are also very low-resource.

3 Data and Pre-processing

We draw data from three main datasets. We use the AmericasNLP 2021 open task dataset (Mager et al., 2021) to pre-train our multilingual models. The multilingual dataset from Kann et al. (2018) serves as segmentation validation data for our pre-training process in these languages. Finally, data from Tyers and Henderson (2021) is used as the training set for our experiments transferring to K'iche', and Richardson and Tyers (2021) provides the validation and test data for these experiments.

AmericasNLP 2021 The AmericasNLP data consists of train and validation files for ten low-resource Indigenous languages of Central and South America: Asháninka (cni), Aymara (aym), Bribri (bzd), Guaraní (gug), Hñähñu (oto), Nahuatl (nah), Quechua (quy), Rarámuri (tar), Shipibo Konibo (shp), and Wixarika (hch). For each language, AmericasNLP also includes parallel Spanish sets, which we do not use. The data was originally curated for the AmericasNLP 2021 shared task on low-resource Machine Translation. (Mager et al., 2021).[1]

We augment the Asháninka and Shipibo-Konibo training sets with additional available monolingual data from Bustamante et al. (2020),[2] which is linked in the official AmericasNLP repository. We add both the training and validation data from this corpus to the *training* set of our splits.

To pre-process for a multilingual language modeling setting, we first remove lines that contain urls, copyright boilerplate, or that contain no alphabetic characters. We also split lines that are longer than 2000 characters into sentences/clauses where evident. Because we use the Nahuatl and Wixarika data from Kann et al. (2018) as validation data, we remove any overlapping lines from the Americas-NLP set. We create a combined train file as the concatenation of the training data from each of the ten languages, as well as a combined validation file likewise.

Because the original ratio of Quechua training data is so high compared to all other languages (Figure 1), we downsample it to 2^{15} examples, the closest order of magnitude to the next-largest training set. A plot of the balanced (final) composition of our AmericasNLP train and validation sets is seen in Figure 2.

To compare the effect of multilingual and monolingual pre-training, we also pre-train a model on Quechua alone, since it has by far the most data (Figure 1). However, the full Quechua training set has about 50k fewer lines than our balanced AmericasNLP set (Figure 2). To create a fair comparison between multilingual and monolingual pre-training, we additionally create a downsampled version of the AmericasNLP set of equal size to the Quechua data (120,145 lines). The detailed composition of our data is available in Appendix A.

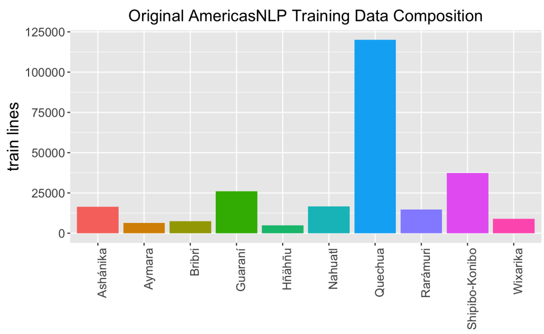

Figure 1: Original (imbalanced) language composition of the AmericasNLP training set

Kann et al (2018) The data from Kann et al. (2018), originally curated for a segmentation task on polysynthetic low-resource languages, contains morphologically segmented sentences for Nahuatl and Wixarika. We use these examples as valida-

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

[2] https://github.com/iapucp/multilingual-data-peru

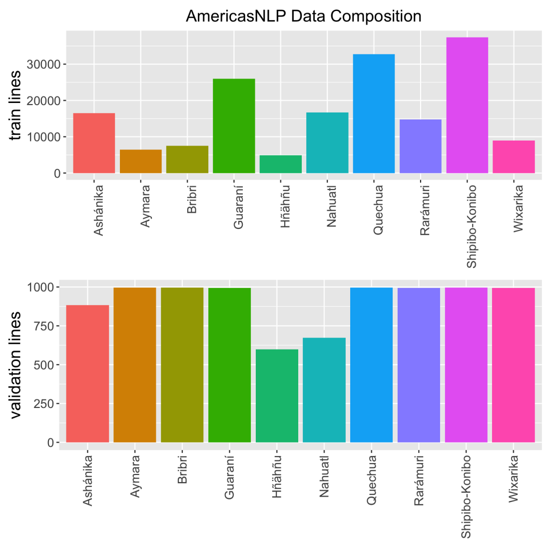

Figure 2: Final language composition of our AmericasNLP splits after downsampling Quechua

tion data for segmentation quality during the pre-training process. We clean this data in the same manner as the AmericasNLP sets.

K'iche' data The K'iche' data used in our study was curated for Tyers and Henderson (2021). The raw (non-gold-segmented) data, used as the training set in our transfer experiments, comes from a section of this data web-scraped by the Crúbadán project (Scannell, 2007). This data is relatively noisy, so we clean it by removing lines with urls or lines where more than half of the characters are non-alphabetic. We also remove duplicate lines. The final data consists of 47,729 examples and is used as our full-size training set for K'iche'. Our experiments involve testing transfer at different resource levels, so we also create smaller training sets by downsampling the original to lower orders of magnitude.

For evaluating segmentation performance on K'iche', we use the segmented sentences from Richardson and Tyers (2021),[3] which were created for a shared task on morphological segmentation. These segmentations were created by a hand-crafted FST, then manually disambiguated. Because gold-segmented sentences are so rare, we concatenate the original train/validation/test splits and then split them in half into final validation and test sets.

[3] https://github.com/ftyers/global-classroom

4 Model and Pre-training

This section gives an overview of the Masked Segmental Language Model (MSLM), introduced in Downey et al. (2021), along with a description of our pre-training procedure.

MSLMs An MSLM is a variant of a Segmental Language Model (SLM) (Sun and Deng, 2018; Kawakami et al., 2019; Wang et al., 2021), which takes as input a sequence of characters $\mathbf{x}$ and outputs a probability distribution for a sequence of segments $\mathbf{y}$ such that the concatenation of $\mathbf{y}$ is equivalent to $\mathbf{x}$: $\pi(\mathbf{y}) = \mathbf{x}$. An MSLM is composed of a Segmental Transformer Encoder and an LSTM-based Segment Decoder (Downey et al., 2021). See Figure 3.

The MSLM training objective is based on the prediction of masked-out spans. During a forward pass, the encoder generates an encoding for every position in $\mathbf{x}$, for a segment up to k symbols long; the encoding at position $i-1$ corresponds to every possible segment that starts at position i. Therefore, the encoding approximates

$$p(\mathbf{x}_{i:i+1}, \mathbf{x}_{i:i+2}, \ldots, \mathbf{x}_{i:i+k} | \mathbf{x}_{<i}, \mathbf{x}_{\geq i+k})$$

To ensure that the encodings are generated based only on the portions of $\mathbf{x}$ that are outside of the predicted span, the encoder uses a Segmental Attention Mask (Downey et al., 2021) to mask out tokens inside the segment. Figure 3 shows an example of such a mask with $k = 2$.

Finally, the Segment Decoder of an SLM determines the probability of the j^{th} character of the segment of $\mathbf{y}$ that begins at index i, $\mathbf{y}_j^i$, using the encoded context:

$$p(\mathbf{y}_j^i | \mathbf{y}_{0:j}^i, \mathbf{x}_{<i}, \mathbf{x}_{\geq i+k}) = Decoder(h_{j-i}^i, y_{j-1}^i)$$

The output of the decoder is not conditional on the determination of other segment boundaries. The probability of $\mathbf{y}$ is modeled as the marginal probability over all possible segmentations of $\mathbf{x}$. Because directly marginalizing is computationally intractable, the marginal is computed using dynamic programming over a forward-pass lattice. The maximum-probability segmentation is determined by Viterbi decoding. The training objective optimizes language-modeling performance, which is measured in Bits Per Character (bpc).

Pre-training Procedure In our experiments, we test the transferability of multilingual and monolingual pre-trained MSLMs. The multilingual models

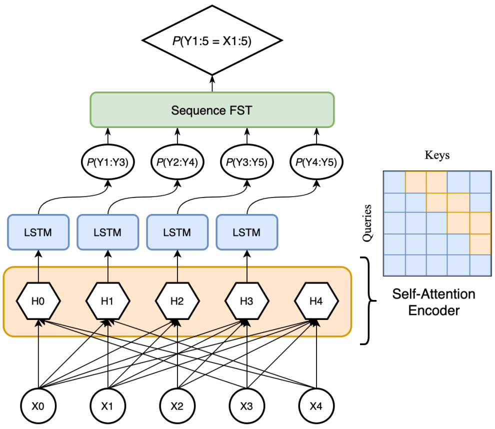

Figure 3: Masked Segmental Language model (left) and Segmental Attention Mask (right). (Figure 3 in Downey et al., 2021)

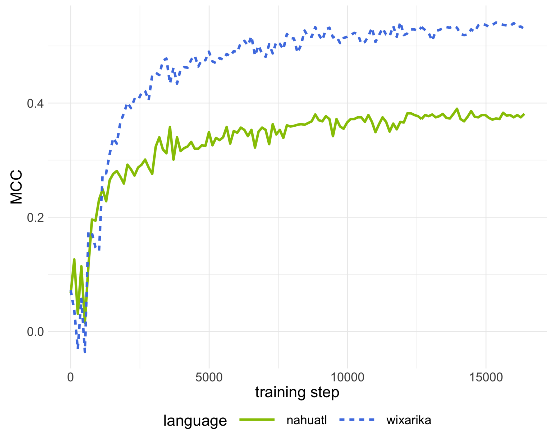

Figure 4: Plot of segmentation quality for Nahuatl and Wixarika during multilingual pre-training (measured by Matthews Correlation Coefficient with gold segmentation)

are trained on the AmericasNLP 2021 data (see Section 3). Since SLMs operate on plain text, we can train the model directly on the multilingual concatenation of this data, and evaluate it by its language modeling performance on the concatenated validation data. As mentioned in Section 3, we create two versions of the multilingual pre-trained model: one trained on the full AmericasNLP set (∼172k lines) and the other trained on the downsampled set, which is the same size as the Quechua training set (∼120k lines). We designate these models MULTI-PT$_{full}$ and MULTI-PT$_{down}$ respectively. Our pre-trained monolingual model is trained on the full Quechua set (QUECHUA-PT).

Each model is an MSLM with four encoder layers, hidden size 256, feedforward size 512, and four attention heads. Character embeddings are initialized using Word2Vec (Mikolov et al., 2013) over the training data. The maximum segment size is set to 10. The best model is chosen as the one that minimizes the Bits Per Character (bpc) loss on the validation set. For further pre-training details, see Appendix B.

To evaluate the effect of pre-training on the segmentation quality for languages within the pre-training set, we also log MCC between the model output and gold-segmented secondary validation sets available in Nahuatl and Wixarika (Kann et al., 2018, see Section 3). Figure 4 shows the unsupervised segmentation quality for Nahuatl and Wixarika almost monotonically increases during pre-training (MULTI-PT$_{full}$).

5 Experiments

We evaluate whether multilingual pre-training facilitates effective low-resource transfer learning for segmentation. To do this, we pre-train SLMs on one or all of the AmericasNLP 2021 languages (Mager et al., 2021) and transfer it to a new target language: K'iche' (Tyers and Henderson, 2021). As a baseline, we train a monolingual K'iche' model from scratch. We evaluate performance with respect to the size of the target training set, simulating varying degrees of low-resource setting. To do this, we downsample the K'iche' training set to 8 smaller sizes, for 9 total: {256, 512, ... 2^{15}, 47,729 (full)}. For each size, we both train a monolingual baseline and fine-tune the pre-trained models we describe in Section 4.[4]

Architecture and Modeling All models are Masked Segmental Language Models (MSLMs) with the architecture described in Section 4. The only difference is that the baseline model is initialized with a character vocabulary *only* covering the particular K'iche' training set (size-specific). The character vocabulary of the K'iche' data is a subset of the AmericasNLP vocabulary, so we are able to transfer the multilingual models without changing the embedding and output layers. The Quechua vocabulary is *not* a superset of the K'iche', so we add the missing characters to the Quechua model's embedding block *before* pre-training (these are randomly initialized). The character embeddings

[4] All of the data and software required to run these experiments can be found at (url redacted)

for the baseline are initialized using Word2Vec (Mikolov et al., 2013) on the training set (again, size-specific).

Evaluation Metrics SLMs can be trained in either a fully unsupervised or "lightly" supervised manner (Downey et al., 2021). In the former case, only the language modeling loss (Bits Per Character, bpc) is used to pick parameters and checkpoints. In the latter, the segmentation quality on gold-segmented validation data can be considered. Though our validation set is gold-segmented, we pick the best parameters and checkpoints based on bpc only, simulating the unsupervised case. However, to monitor the change in segmentation quality during training, we also use Matthews Correlation Coefficient (MCC). This measure frames segmentation as a character-wise binary classification task (i.e. boundary vs. no boundary), and measures correlation with the gold segmentation.

To make our results comparable with the wider word-segmentation literature, we use the scoring script from the SIGHAN Segmentation Bakeoff (Emerson, 2005) for our final segmentation F1. For each model and target size, we choose the best checkpoint (by bpc), apply the model to the combined validation and test set, and use the SIGHAN script to score the output.

For comparison to the Chinese Word-Segmentation and speech literature, any whitespace segmentation in the validation/test data is discarded before it is fed to the model. However, SLMs can also be trained to treat spaces like any other character, and thus could be able to take advantage of existing segmentation in the input. We leave this for future work.

Parameters and Trials For our training procedure (both training the baseline from scratch and fine-tuning the pre-trained models) we tune hyperparameters on three of the nine dataset sizes (256, 2048, and full) and choose the optimal parameters by bpc. For each of the other sizes, we directly apply the chosen parameters from the tuned dataset of the closest size (on a log scale). We tune over five learning rates and three encoder dropout values. As in pre-training, we set the maximum segment length to 10. For more details on our training procedure, see Appendix B.

6 Results

The results of our K'iche' transfer experiments at various target sizes can be found in Table 1. In general, the (full) pre-trained multilingual model (MULTI-PT$_{full}$) demonstrates good performance across dataset sizes, with the lowest segmentation performance (20.6 F1) being in the zero-shot case and the highest (40.7) achieved on 2^{14} examples. The monolingual baseline outperforms MULTI-PT$_{full}$ at the two largest target sizes, as well as at size 4096 (achieving the best overall F1 of 44.8), but performs very poorly under 2048 examples, and has no zero-shot ability.

Interestingly, other than in the zero-shot case, QUECHUA-PT and the comparable MULTI-PT$_{down}$ perform very similarly to each other. However, the zero-shot transferability of MULTI-PT$_{down}$ is almost twice that of the model trained on Quechua only. MULTI-PT$_{full}$ exceeds both MULTI-PT$_{down}$ and QUECHUA-PT by a wide margin in every setting. Finally, all models show increasing performance until about size 4096, after which more target examples don't provide a large increase in segmentation quality.

Interpretation These results show that MULTI-PT$_{full}$ provides consistent performance across target sizes as small as 512 examples. Even for size 256, there is only a 9% (relative) drop in quality from the next-largest size. Further, the pre-trained model's zero-shot performance is impressive given the baseline is effectively 0 F1.

On the other hand, the performance of the monolingual baseline at larger sizes seems to suggest that given enough target data, it is better to train a model devoted to the target language only. This is consistent with previous results (Wu and Dredze, 2020; Conneau et al., 2020a). However, it should also be noted that MULTI-PT$_{full}$ never trails the baseline by more than 5.2 F1.

One less-intuitive result is the dip in the baseline's performance at sizes 8192 and 2^{14}. We believe this discrepancy may be partly explainable by sensitivity to hyperparameters in the baseline. Though the best baseline trial at size 2048 exceeds MULTI-PT$_{full}$ by a small margin, the baseline shows large variation in performance across the top-four hyperparameter settings at this size, where MULTI-PT$_{full}$ actually performs better on average and much more consistently (Table 2). We thus believe the dip in performance for the baseline

Model	Target Language Segmentation F1									
	0	256*	512	1024	2048*	4096	8192	2^{14}	2^{15}	47,729 (full)*
MULTI-PT$_{full}$	**20.6**	**34.0**	**37.4**	**37.4**	38.2	40.5	**38.6**	**40.7**	38.9	38.2
MULTI-PT$_{down}$	15.0	25.1	25.7	29.3	32.5	33.2	33.3	31.5	33.6	31.9
QUECHUA-PT	7.6	29.9	31.0	30.4	30.7	31.0	29.9	33.6	31.8	33.3
MONOLINGUAL	0.002	4.0	3.3	10.3	**39.2***	**44.8**	29.4	39.5	**44.1**	**43.2**

Table 1: Segmentation quality on the combined validation and test set for each model, at each target training set size. Star indicates size at which hyperparameter tuning is conducted. For tuned sizes, showing only the performance of the model with the best bpc. *See Table 2: the best baseline trial achieved slightly better performance than MULTI-PT$_{full}$, but the former is far more sensitive to variation due to hyperparameters at this size

Model	Target Language Segmentation F1		
	256*	2048*	47,729 (full)*
MULTI-PT$_{full}$	**34.2** ± **0.6** (1.8%)	**38.1** ± **0.4** (1.0%)	39.4 ± 1.1 (2.8%)
MULTI-PT$_{down}$	25.7 ± 0.6 (2.3%)	30.5 ± 2.3 (7.5%)	31.7 ± 0.6 (1.9%)
QUECHUA-PT	30.1 ± 0.2 (0.7%)	31.4 ± 0.6 (1.9%)	32.7 ± 0.7 (2.1%)
MONOLINGUAL	4.2 ± 0.5 (11.9%)	36.5 ± 6.8 (18.6%)	**44.7** ± **2.0** (4.5%)

Table 2: Variation of segmentation quality across the best four hyperparameter combinations for a single size (by bpc; mean ± standard deviation (stdev ÷ mean); models ranked by mean minus stdev)

at sizes 8192 and 2^{14} may be due to an inability to extrapolate hyperparameters from other experimental settings.

7 Analysis and Discussion

Standing of Hypotheses Within the framework of unsupervised segmentation, these results provide strong evidence that relevant linguistic patterns can be learned over a collection of low-resource languages, and then transferred to a new language without much (or any) target training data. Further, it is shown that the target language need not be (phylogenetically) related to any of the pre-training languages, even though details of morphological structure are ultimately language-specific.

The hypothesis that multilingual pre-training yields increasing advantage over a from-scratch baseline at smaller target sizes is also strongly supported. This result is consistent with related work showing this to be a key advantage of the multilingual approach (Wu and Dredze, 2020).

The hypothesis that multilingual pre-training also yields better performance than monolingual pre-training given the same amount of data seems to receive mixed support from our experiments. On one hand, the comparable multilingual model has a clear advantage over the Quechua model in the zero-shot setting, and outperforms the latter in 5/10 settings more generally. However, because the Quechua data lacks several frequent K'iche' characters (and these embeddings remain randomly initialized), it is unclear how much of this advantage comes from the multilingual training *per-se*. Instead, the advantage may be due to the multilingual model's full coverage of the target vocabulary—an advantage which may disappear at larger target sizes. Further analysis of this hypothesis will require additional investigation.

Significance The above results, especially the strong zero-shot transferability of segmentation performance, suggest that the type of language model used here learns some abstract linguistic pattern(s) that are generalizable across languages, and even to new ones. It is possible that these generalizations could take the form of abstract stem/affix or word-order patterns, corresponding roughly to the lengths and order of morphosyntactic units. Because MSLMs operate on the character level (and in these languages orthographic characters mostly correspond to phones), it is also possible the model could recognize syllable structure in the data (the ordering of consonants and vowels in human languages is relatively constrained), and learn to segment on syllable boundaries.

It is also helpful to remember that we select the training suite and target language to have some characteristics in common that may help facilitate transfer. The AmericasNLP languages are almost all morphologically rich, with many considered polysynthetic (Mager et al., 2021), a feature that K'iche' shares (Suárez, 1983). Further, all of the languages, including K'iche', are spoken in countries where either Spanish or Portuguese are the official language, and have very likely had close contact with these Iberian languages and borrowed lexical items. Finally, the target language family (Mayan) has also been shown to have close historical contact with the families of several of the AmericasNLP set (Nahuatl, Rarámuri, Wixarika, Hñähñu), forming a Linguistic Area or *Sprachbund* (Campbell et al., 1986).

It is possible that one or several of these shared characteristics facilitates the strong transfer shown here, in both our multilingual and monolingual pre-trained models. However, our current study does not conclusively show this to be the case. Lin et al. (2019) show that factors like linguistic similarity and geographic contact are often not as important for transfer success as non-linguistic features such as the raw size of the source dataset. Indeed, the fact that our Quechua pre-trained model performs similarly to the comparable multilingual model (at least at larger target sizes) suggests that the benefit to using MULTI-PT$_{full}$ could be interpreted as a combined advantage of pre-training data size and target vocabulary coverage.

The nuanced question of whether multilingual pre-training *itself* enables better transfer than monolingual pre-training requires more study. However, taking a more pragmatic point of view, multilingual training can be seen as a methodology to 1) acquire more data than is available from any one language and 2) ensure broader vocabulary overlap with the target language. Our character-based model is of course different from more common word- or subword-based approaches, but with these too, attaining pre-trained embeddings that cover a novel target language is an important step in cross-lingual transfer (Garcia et al., 2021; Conneau et al., 2020a; Artetxe et al., 2020, *inter alia*)

Future Work We believe some future studies would shed light on the nuances of segmentation transfer-learning. First, pre-training either multilingually or monolingually on languages that are *not* linguistically similar to the target language could help isolate the advantage given by pre-training on *any* language data (vs. similar language data).

Second, we have noted that monolingual pre-training on a language that does not have near-full vocabulary coverage of the target language leaves some embeddings randomly initialized, yielding worse performance at small target sizes. Pre-training a model on a single language that happens to have near-complete vocabulary coverage of the target could give a better view of whether multilingual training intrinsically yields advantages, or whether monolingual training is disadvantaged mainly due to this lack of vocabulary coverage.

8 Conclusion

This study has shown that unsupervised sequence segmentation ability can be transferred via multilingual pre-training to a novel target language with little or no target data. The target language also need not be from the same family as a pre-training language for successful transfer. While training a monolingual model from scratch on large amounts of target data results in good segmentation quality, our experiments show that pre-trained models, especially multilingual ones, far exceed the baseline at small target sizes ($\leq$1024), and seem to be much more robust to hyperparameter variation at medium sizes (2048, 8192, 2^{14}).

One finding that may have broader implications is that pre-training can be conducted over a set of low-resource languages with some typological or geographic connection to the target, rather than over a crosslingual suite centered around high-resource languages like English and other European languages. Most modern crosslingual models have huge numbers of parameters (XLM has 570 million, mT5 has up to 13 billion, Xue et al., 2021), and are trained on enormous amounts of data, usually bolstered by hundreds of gigabytes in the highest-resource languages (Conneau et al., 2020a).

In contrast, our results suggest that effective transfer may be possible at smaller scales, by combining the data of low-resource languages and training moderately-sized, more targeted pre-trained multilingual models (our model has 3.15 million parameters). Of course, this study can only support this possibility within the unsupervised segmentation task, so future work will be needed to investigate whether transfer to and from low-resource languages can be extended to other tasks.

References

Željko Agić and Ivan Vulić. 2019. JW300: A Wide-Coverage Parallel Corpus for Low-Resource Languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the Cross-lingual Transferability of Monolingual Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Steven Bird. 2020. Sparse Transcription. *Computational Linguistics*, 46(4):713–744.

David Brambila. 1976. *Diccionario Rarámuri-castellano (Tarahumar)*. Obra Nacional de la Buena Prensa.

Gina Bustamante, Arturo Oncevay, and Roberto Zariquiey. 2020. No Data to Crawl? Monolingual Corpus Creation from PDF Files of Truly low-Resource Languages in Peru. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2914–2923, Marseille, France. European Language Resources Association.

Lyle Campbell, Terrence Kaufman, and Thomas C. Smith-Stark. 1986. Meso-America as a Linguistic Area. *Language*, 62(3):530–570. Publisher: Linguistic Society of America.

Luis Chiruzzo, Pedro Amarilla, Adolfo Ríos, and Gustavo Giménez Lugo. 2020. Development of a Guarani - Spanish Parallel Corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2629–2633, Marseille, France. European Language Resources Association.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual Language Model Pretraining. In *Advances in Neural Information Processing Systems*, volume 32, Vancouver, Canada. Curran Associates, Inc.

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.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging Cross-lingual Structure in Pretrained Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

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

Mathias Creutz and Krista Lagus. 2002. Unsupervised Discovery of Morphemes. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*, pages 21–30. Association for Computational Linguistics.

Rubén Cushimariano Romano and Richer C. Sebastián Q. 2008. Ñaantsipeta asháninkaki birakochaki. Diccionario Asháninka-Castellano. Versión preliminar.

C. M. Downey, Fei Xia, Gina-Anne Levow, and Shane Steinert-Threlkeld. 2021. A Masked Segmental Language Model for Unsupervised Natural Language Segmentation. *arXiv:2104.07829 [cs]*. ArXiv: 2104.07829.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir, Gustavo A. Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando A. Coto Solano, Ngoc Thang Vu, and Katharina Kann. 2021. AmericasNLI: Evaluating Zero-shot Natural Language Understanding of Pretrained Multilingual Models in Truly Low-resource Languages. *arXiv:2104.08726 [cs]*. ArXiv: 2104.08726.

Thomas Emerson. 2005. The Second International Chinese Word Segmentation Bakeoff. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*.

I. Feldman and R. 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.

Sofía Flores Solórzano. 2017. Corpus Oral Pandialectal de la Lengua Bribri.

Ana-Paula Galarreta, Andrés Melgar, and Arturo Oncevay. 2017. Corpus Creation and Initial SMT Experiments between Spanish and Shipibo-konibo. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 238–244, Varna, Bulgaria. INCOMA Ltd.

Xavier Garcia, Noah Constant, Ankur Parikh, and Orhan Firat. 2021. Towards Continual Learning for Multilingual Machine Translation via Vocabulary Substitution. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1184–1192, Online. Association for Computational Linguistics.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. 2009. A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54.

Alex Graves, Fernández Santiago, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA.

Ximena Gutierrez-Vasques, Gerardo Sierra, and Isaac Hernandez Pompa. 2016. Axolotl: a Web Accessible Parallel Corpus for Spanish-Nahuatl. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4210–4214, Portorož, Slovenia. European Language Resources Association (ELRA).

Thanh Le Ha, Jan Niehues, and Alexander Waibel. 2016. Toward Multilingual Neural Machine Translation with Universal Encoder and Decoder. In *Proeedings of the 13th International Conference on Spoken Language Translation*.

Diego Huarcaya Taquiri. 2020. *Traducción Automática Neuronal para Lengua Nativa Peruana*. Bachelor's Thesis, Universidad Peruana Unión.

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

Carla Victoria Jara Murillo. 2018b. *I Ttè Historias Bribris*, 2 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.

Mark Johnson and Sharon Goldwater. 2009. Improving nonparameteric Bayesian inference: experiments on unsupervised word segmentation with adaptor grammars. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 317–325, Boulder, Colorado. Association for Computational Linguistics.

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.

Katharina Kann, Jesus Manuel Mager Hois, Ivan Vladimir Meza-Ruiz, and Hinrich Schütze. 2018. Fortification of Neural Morphological Segmentation Models for Polysynthetic Minimal-Resource Languages. 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 47–57, New Orleans, Louisiana. Association for Computational Linguistics.

Kazuya Kawakami, Chris Dyer, and Phil Blunsom. 2019. Learning to Discover, Ground and Use Words with Segmental Neural Language Models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6429–6441, Florence, Italy. Association for Computational Linguistics.

Diederik Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings*, San Diego, CA, USA.

Lingpeng Kong, Chris Dyer, and Noah A Smith. 2016. Segmental Recurrent Neural Networks. In *4th International Conference on Learning Representations, ICLR 2016, Conference Track Proceedings*, San Juan, Puerto Rico.

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.

William Lane, Mat Bettinson, and Steven Bird. 2021. A Computational Model for Interactive Transcription. In *Proceedings of the Second Workshop on Data Science with Human in the Loop: Language Advances*, pages 105–111, Online. Association for Computational Linguistics.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing Transfer Languages for Cross-Lingual Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. 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.

James Loriot, Erwin Lauriault, and Dwight Day. 1993. *Diccionario Shipibo-Castellano*. Ministerio de Educación.

Manuel Mager, Diónico Carrillo, and Ivan Meza. 2018. Probabilistic Finite-State Morphological Segmenter for Wixarika (Huichol) Language. *Journal of Intelligent & Fuzzy Systems*, 34(5):3081–3087.

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*, 2 edition. Editorial de la Universidad de Costa Rica.

Elena Mihas. 2011. *Añaani katonkosatzi parenini, El idioma del alto Perené*. Milwaukee, WI: Clarks Graphics.

Tomas Mikolov, Kai Chen, G. Corrado, and J. Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations, ICLR 2013, Workshop Track Proceedings*, Scottsdale, AR, USA.

Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. 2009. Bayesian Unsupervised Word Segmentation with Nested Pitman-Yor Language Modeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 100–108, Suntec, Singapore. Association for Computational Linguistics.

Héctor Erasmo Gómez Montoya, Kervy Dante Rivas Rojas, and Arturo Oncevay. 2019. A Continuous Improvement Framework of Machine Translation for Shipibo-Konibo. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 17–23, Dublin, Ireland. European Association for Machine Translation.

John Ortega, Richard Alexander Castro-Mamani, and Jaime Rafael Montoya Samame. 2020. Overcoming Resistance: The Normalization of an Amazonian Tribal Language. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 1–13, Suzhou, China. Association for Computational Linguistics.

Hyunji Hayley Park, Katherine J. Zhang, Coleman Haley, Kenneth Steimel, Han Liu, and Lane Schwartz. 2021. Morphology Matters: A Multilingual Language Modeling Analysis. *Transactions of the Association for Computational Linguistics*, 9:261–276. _eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00365/1924158/tacl_a_00365.pdf.

Prokopis Prokopidis, Vassilis Papavassiliou, and Stelios Piperidis. 2016. Parallel Global Voices: a Collection of Multilingual Corpora with Citizen Media Stories. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 900–905, Portorož, Slovenia. European Language Resources Association (ELRA).

I. Richardson and F.M. Tyers. 2021. A morphological analyser for Kiche. *Procesamiento de Lenguaje Natural*, 66:99–109.

Jorma Rissanen. 1989. *Stochastic Complexity in Statistical Inquiry*, volume 15 of *Series in Computer Science*. World Scientific, Singapore.

Kevin Scannell. 2007. The Crúbadán Project: Corpus building for under-resourced languages. *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*.

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.

K. Song, X. Tan, Tao Qin, Jianfeng Lu, and T. Liu. 2019. MASS: Masked Sequence to Sequence Pretraining for Language Generation. In *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, CA.

Zhiqing Sun and Zhi-Hong Deng. 2018. Unsupervised Neural Word Segmentation for Chinese via Segmental Language Modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4915–4920, Brussels, Belgium. Association for Computational Linguistics.

George Suárez. 1983. *The Mesoamerican Indian Languages*. Cambridge Language Surveys. Cambridge University Press, Cambridge.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

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).

Francis Tyers and Robert Henderson. 2021. A corpus of K'iche' annotated for morphosyntactic structure. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 10–20, Online. Association for Computational Linguistics.

Kei Uchiumi, Hiroshi Tsukahara, and Daichi Mochihashi. 2015. Inducing Word and Part-of-Speech with Pitman-Yor Hidden Semi-Markov Models. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1774–1782, Beijing, China. Association for Computational Linguistics.

Chong Wang, Yining Wang, Po-Sen Huang, Abdelrahman Mohamed, Dengyong Zhou, and Li Deng. 2017. Sequence Modeling via Segmentations. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3674–3683, International Convention Centre, Sydney, Australia. PMLR.

Lihao Wang, Zongyi Li, and Xiaoqing Zheng. 2021. Unsupervised Word Segmentation with Bi-directional Neural Language Model. *arXiv:2103.01421 [cs]*. ArXiv: 2103.01421.

Shijie Wu and Mark Dredze. 2020. Are All Languages Created Equal in Multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

A AmericasNLP Datasets

Composition The detailed composition of our preparation of the AmericasNLP 2021 training and validation sets can be found in Tables 3 and 4 respectively. `train_1.mono.cni`, `train_2.mono.cni`, `train_1.mono.shp`, and `train_2.mono.shp` are the additional monolingual sources for Asháninka and Shipibo-Konibo obtained from Bustamante et al. (2020). `train_downsampled.quy` is the version of the Quechua training set downsampled to 2^{15} lines to be more balanced with the other languages. `train.anlp` is the concatenation of the training set of every language before Quechua downsampling, and `train_balanced.anlp` is the version after Quechua downsampling. `train_downsampled.anlp` is the version of our multilingual set downsampled to be the same size as `train.quy`. MULTI-PT$_{full}$ is pre-trained on `train_balanced.anlp`, MULTI-PT$_{down}$ is pre-trained on `train_downsampled.anlp`, and QUECHUA-PT is pre-trained on `train.quy`.

Citations A more detailed description of the sources and citations for the AmericasNLP set can be found in the original shared task paper (Mager et al., 2021). Here, we attempt to give a brief listing of the proper citations.

All of the validation data originates from AmericasNLI (Ebrahimi et al., 2021) which is a translation of the Spanish XNLI set (Conneau et al., 2018) into the 10 languages of the AmericasNLP 2021 open task.

The training data for each of the languages comes from a variety of different sources. The **Asháninka** training data is sourced from Ortega et al. (2020); Cushimariano Romano and Sebastián Q. (2008); Mihas (2011) and consists of stories, educational texts, and environmental laws. The **Aymara** training data consists mainly of news text from the GlobalVoices corpus (Prokopidis et al., 2016) as available through OPUS (Tiedemann, 2012). The **Bribri** training data is from six sources (Feldman and Coto-Solano, 2020; Margery, 2005; Jara Murillo, 2018a; Constenla et al., 2004; Jara Murillo and Segura, 2013; Jara Murillo, 2018b; Flores Solórzano, 2017) ranging from dictionaries and textbooks to story books. The **Guaraní** training data consists of blogs and web news sources collected by Chiruzzo et al. (2020). The **Nahuatl** training data comes from the Axolotl parallel corpus (Gutierrez-Vasques et al., 2016). The **Quechua** training data was created from the JW300 Corpus (Agić and Vulić, 2019), including Jehovah's Witnesses text and dictionary entries collected by Huarcaya Taquiri (2020). The **Rarámuri** training data consists of phrases from the Rarámuri dictionary (Brambila, 1976). The **Shipibo-Konibo** training data consists of translations of a subset of the Tatoeba dataset (Montoya et al., 2019), translations from bilingual education books (Galarreta et al., 2017), and dictionary entries (Loriot et al., 1993). The **Wixarika** training data consists of translated Hans Christian Andersen fairy tales from Mager et al. (2018).

No formal citation was given for the source of the **Hñähñu** training data (see Mager et al., 2021).

B Hyperparameter Details

Pre-training The character embeddings for our multilingual model are initialized by training CBOW (Mikolov et al., 2013) on the AmericasNLP training set for 32 epochs, with a window size of 5. Special tokens like <bos> that do not appear in the training corpus are randomly initialized. These pre-trained embeddings are not frozen during training.

We pre-train for 16,768 steps, using the Adam optimizer (Kingma and Ba, 2015). We apply a linear warmup for 1024 steps, and a linear decay afterward. We sweep eight learning rates on a grid of the interval $[0.0005, 0.0009]$ and encoder dropout values $\{12.5\%, 25\%\}$. A dropout rate of 6.25% is applied both to the embeddings before being passed to the encoder, and to the hidden-state and start-symbol encodings input to the decoder (see Downey et al., 2021). Checkpoints are taken every 128 steps.

K'iche' Transfer Experiments Similar to the pre-trained model, character embeddings are initialized using CBOW on the given training set for 32 epochs with a window size of 5, and these embeddings are not frozen during training.

All models are trained using the Adam optimizer (Kingma and Ba, 2015) for 8192 steps on all but the two smallest sizes, which are trained for 4096 steps. A linear warmup is used for the first 1024 steps (512 for the smallest sets), followed by linear decay. We set the maximum segment length to 10. A dropout rate of 6.25% is applied to the input embeddings, plus h and the start-symbol for the decoder. Checkpoints are taken every 64 steps for sizes 256 and 512, and every 128 steps for every other size.

For all training set sizes, we sweep 5 learning rates and 3 encoder dropout rates, but the swept set is different for each. For size 256, we sweep learning rates {5e-5, 7.5e-5, 1e-4, 2.5e-4, 5e-4} and (encoder) dropout rates {12.5%, 25%, 50%}. For size 2048, we sweep learning rates {1e-4, 2.5e-4, 5e-4, 7.5e-4, 1e-3} and dropouts {12.5%, 25%, 50%}. For the full training size, we sweep learning rates {1e-4, 2.5e-4, 5e-4, 7.5e-4, 1e-3} and dropouts {6.5%, 12.5%, 25%}.

Language	File	Lines	Total Tokens	Unique Tokens	Total Characters	Unique Characters	Mean Token Length
All	train.anlp	259,207	2,682,609	400,830	18,982,453	253	7.08
All	train_balanced.anlp	171,830	1,839,631	320,331	11,981,011	241	6.51
All	train_downsampled.anlp	120,145	1,284,440	255,392	8,365,710	221	6.51
Asháninka	train.cni	3,883	26,096	12,490	232,494	65	8.91
Asháninka	train_1.mono.cni	12,010	99,329	27,963	919,897	48	9.26
Asháninka	train_2.mono.cni	593	4,515	2,325	42,093	41	9.32
Aymara	train.aym	6,424	96,075	33,590	624,608	156	6.50
Bribri	train.bzd	7,508	41,141	7,858	167,531	65	4.07
Guaraní	train.gug	26,002	405,449	44,763	2,718,442	120	6.70
Hñähñu	train.oto	4,889	72,280	8,664	275,696	90	3.81
Nahuatl	train.nah	16,684	351,702	53,743	1,984,685	102	5.64
Quechua	train.quy	120,145	1,158,273	145,899	9,621,816	114	8.31
Quechua	train_downsampled.quy	32,768	315,295	64,148	2,620,374	95	8.31
Rarámuri	train.tar	14,720	103,745	15,691	398,898	74	3.84
Shipibo Konibo	train.shp	14,592	62,850	17,642	397,510	56	6.32
Shipibo Konibo	train_1.mono.shp	22,029	205,866	29,534	1,226,760	61	5.96
Shipibo Konibo	train_2.mono.shp	780	6,424	2,618	39,894	39	6.21
Wixarika	train.hch	8,948	48,864	17,357	332,129	67	6.80

Table 3: Composition of the AmericasNLP 2021 training sets

Language	File	Lines	Total Tokens	Unique Tokens	Total Characters	Unique Characters	Mean Token Length
All	dev.anlp	9,122	79,901	27,597	485,179	105	6.07
Asháninka	dev.cni	883	6,070	3,100	53,401	63	8.80
Aymara	dev.aym	996	7,080	3,908	53,852	64	7.61
Bribri	dev.bzd	996	12,974	2,502	50,573	73	3.90
Guaraní	dev.gug	995	7,191	3,181	48,516	70	6.75
Hñähñu	dev.oto	599	5,069	1,595	22,712	69	4.48
Nahuatl	dev.nah	672	4,300	1,839	31,338	56	7.29
Quechua	dev.quy	996	7,406	3,826	58,005	62	7.83
Rarámuri	dev.tar	995	10,377	2,964	55,644	48	5.36
Shipibo Konibo	dev.shp	996	9,138	3,296	54,996	65	6.02
Wixarika	dev.hch	994	10,296	3,895	56,142	62	5.45

Table 4: Composition of the AmericasNLP 2021 validation sets