Auxiliary Subword Segmentations as Related Languages for Low Resource Multilingual Translation

Nishant Kambhatla **Logan Born** **Anoop Sarkar**
School of Computing Science
Simon Fraser University
8888 University Drive, Burnaby BC, Canada
{nkambhat, loborn, anoop}@sfu.ca

Abstract

We propose a novel technique of combining multiple subword tokenizations of a single source-target language pair for use with multilingual neural translation training methods. These alternate segmentations function like related languages in multilingual translation, improving translation accuracy for low-resource languages and producing translations that are lexically diverse and morphologically rich. We also introduce a cross-teaching technique which yields further improvements in translation accuracy and cross-lingual transfer between high- and low-resource language pairs. Compared to other strong multilingual baselines, our approach yields average gains of +1.7 BLEU across the four low-resource datasets from the multilingual TED-talks dataset. Our technique does not require additional training data and is a drop-in improvement for any existing neural translation system.

1 Introduction

Multilingual neural machine translation (NMT, Dong et al. 2015; Johnson et al. 2017) models are capable of translating from multiple source and target languages. Besides allowing efficient parameter sharing (Aharoni et al., 2019) these models facilitate inherent transfer learning (Zoph et al., 2016; Firat et al., 2016) that can especially benefit low resource languages (Nguyen and Chiang, 2017; Gu et al., 2018; Neubig and Hu, 2018; Tan et al., 2019). A common technique to address lexical sharing and complex morphology in multilingual NMT is to decompose longer words into shorter subword units (Sennrich et al., 2016). Since subword units are produced using heuristic methods, not all subwords are created equally. This can put low- and extremely low-resource languages at a disadvantage, even when these languages are paired with a suitable high resource language. To diminish the impact of rare subwords in NMT, Kambhatla et al. (2022) leverage ciphertexts to augment the training data by constructing multiple-views of the source text. "Soft" decomposition methods based on transfer learning (Wang et al., 2018) address the problem of sub-optimal word segmentation with shared character-level lexical and sentence representations across multiple source languages (Gu et al., 2018). Wang et al. (2021) addressed this problem with a multiview-subword regularization technique that also improves the effectiveness of cross-lingual transfer in pretrained multilingual representations by simultaneously finetuning on different input segmentations from a heuristic and a probabilistic tokenizer. While subword-regularization methods (Kudo, 2018; Provilkov et al., 2020) have been widely explored in NMT, this work is the first to study them together with multilingual training methods.

Concretely, we construct pairs of "related languages" by segmenting an input corpus twice, each time with a different vocabulary size and algorithm for finding subwords; we use these "languages" (really, views of the same language) for multilingual training of an NMT model. We propose *Multi-Sub training*, a method that combines multilingual NMT training methods with a diverse set of auxiliary subword segmentations which func-

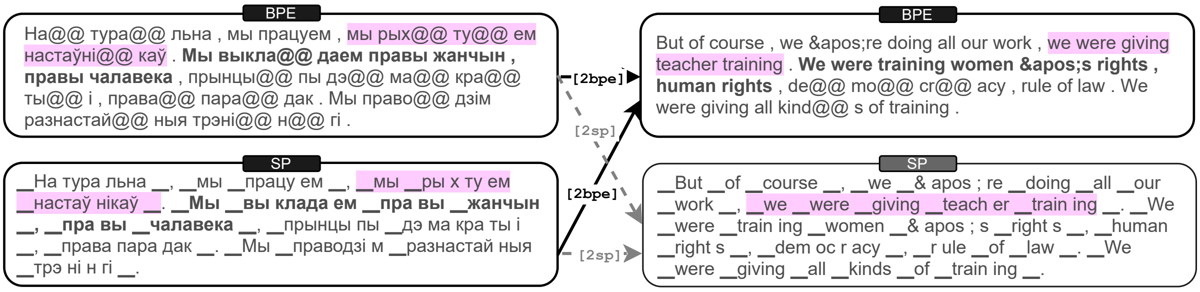

BPE
На@@ тура@@ льна , мы працуем , мы рых@@ ту@@ ем настаўні@@ каў . Мы выкла@@ даем правы жанчын , правы чалавека , прынцы@@ пы дэ@@ ма@@ кра@@ ты@@ і , права@@ пара@@ дак . Мы право@@ дзім разнастай@@ ныя трэні@@ н@@ гі .
SP
__На тура льна __, __мы __працу ем __, __мы __ры х ту ем __настаў нікаў __. __Мы __вы клада ем __пра вы __жанчын __, __пра вы __чалавека __, __прынцы пы __дэ ма кра ты і __, __права пара дак __. __Мы __праводзі м __разнастай ныя __трэ ні н гі __.
[2bpe]
[2sp]
[2bpe]
[2sp]
BPE
But of course , we &apos;re doing all our work , we were giving teacher training . We were training women &apos;s rights , human rights , de@@ mo@@ cr@@ acy , rule of law . We were giving all kind@@ s of training .
SP
__But __of __course __, __we __& apos ; re __doing __all __our __work __, __we __were __giving __teach er __train ing __. __We __were __train ing __women __& apos ; s __right s __, __human __right s __, __dem oc r acy __, __r ule __of __law __. __We __were __giving __all __kinds __of __train ing __.

Figure 1: An illustration of the interaction between the primary (BPE) and auxiliary (SP) subwords for the same sample from the be-en dev set where each type of segmentation is treated as a separate language. The model is taught to translate into a specific segmentation via multilingual training using the target "language" tags [2bpe] and [2sp]. The sentence in bold type font shows both variants of the source sentence translating to the same target sentence. The colored spans show different segmentations of the same word(s) in source/target.

tion like related languages in a multilingual setting since they have distinct but partially-overlapping vocabularies and share the same underlying lexical and grammatical features. Our model is able to transfer information between segmentations analogous to the way information is transferred between typologically similar languages.

We also introduce a *cross-teaching* technique in which a model is trained to translate source sentences from one subword tokenization into target sentences from a different subword tokenization. By using Multi-Sub training together with cross-teaching, we obtain strong results on four low-resource languages in the multilingual TED talks dataset outperforming strong multilingual baselines, with the most significant improvements in the lowest-resource languages. In addition to improving the BLEU scores, our technique captures word compositionality better leading to improved lexical diversity and morphological richness in the target language. Multi-Sub with cross-teaching is better at clustering different languages in the sentence embedding space than previous methods including Multi-Sub without cross-teaching.

2 Auxiliary Segmentation as a Related Language

Pairing related languages is common in multilingual NMT[1]: Nguyen and Chiang (2017) combine Uzbek/Turkish and Uzbek/Uyghur; Johnson et al. (2017) study multilingual translation to and from English with pairs such as Spanish/Portuguese or Japanese/Korean. Neubig and Hu (2018) pair low resource languages like Azerbaijani with a related "helper" language like Turkish.

We take these techniques as motivation for the present work. Our principal contribution is to rethink what it means to use "related" languages in a multilingual translation model. Beyond simply employing *other* languages from the same family, or those with high lexical overlap, we show that a model trained on different segmentations *of the same language* can produce improvements in translation quality.

Rather than segmenting a corpus with a single tokenizer prior to training a translation model, we produce multiple segmentations using different tokenizers. Consider the example sentences in Figure 1. On both the source and target sides, the same sentence is represented using both Byte-pair Encodings (BPEs, Sennrich et al. 2016, with a "@@" separator) and in parallel as sentencepieces (SP, Kudo 2018, with a "_" separator). Each segmentation uses a different vocabulary size, which guarantees that their subword sequences are to some extent distinct. The two tokenizations still resemble one other in many ways: (i) they have a non-trivial degree of lexical overlap (mostly between subwords which do not fall along word boundaries); (ii) they share the same grammatical structure, as both represent the same underlying language; and (iii) both sequences have the same semantic interpretation. We thus refer to the two segmentations as a pair of "related languages".

Applying two segmentations to a parallel corpus yields a total of four "languages": the source and target represented as BPE subwords, and the same represented using SP subwords. We obtain two source "languages" (each containing data from both high and low resource languages) and two target "languages". Using this four way configuration, we train a model following a common multi-

[1] Here we do not distinguish between languages which are related in the linguistic sense (having some genetic affiliation) and those which are related in a more pragmatic sense of having high lexical overlap.

lingual training method (Johnson et al., 2017): depending on the segmentation we want to translate into, we prepend a target token **[2bpe]** or **[2sp]** to the source side. We explore two different multilingual training configurations:

[BPE+SP]: In this setting, a source sentence in a particular segmentation is translated into the target with the same segmentation. Specifically, this model is trained multilingually on the pairs

$$\text{BPE [src]} \rightarrow \text{BPE [tgt]}$$

$$\text{SP [src]} \rightarrow \text{SP [tgt]}$$

Cross-teaching: In addition to [BPE+SP], in this setting, each source sentence with a particular segmentation is translated into the target with alternate segmentation. This multilingual model is therefore trained on the following pairs:

$$\text{BPE [src]} \rightarrow \text{SP [tgt]}$$

$$\text{SP [src]} \rightarrow \text{BPE [tgt]}$$

Using multilingual training, our model is able to *transfer* information between BPE and SP segmentations in much the same way that conventional multilingual models transfer information between languages with a shared linguistic affiliation. Unlike data augmentation techniques which generate synthetic training data, Multi-Sub training uses only the content of the original training corpus. Furthermore, contrary to other works which employ multiple segmentations (Wang et al., 2018; Wu et al., 2020), Multi-Sub training and cross-teaching do not affect model architecture and do not require specialised training. Thus Multi-Sub training can be used as a simple, drop-in improvement to an existing neural translation model.

3 Experiments

3.1 Experimental Setup

Data Following prior work on low-resource and multilingual NMT (Neubig and Hu, 2018; Wang et al., 2018) we use the multilingual Ted talks dataset (Qi et al., 2018). We use four low resource languages (LRL): Azerbaijani (az), Belarusian (be), Galician (gl) and Slovak (sk), and four high resource languages (HRL): Turkish (tr), Russian (ru), Brazilian-Portuguese (pt), and Czech (cs). In all experiments and baselines, each LRL is paired with the related HRL and English is the target language.

Table 1 shows general statistics for each dataset. Based on the size of the training data, we consider az, be and gl as extremely low-resource while sk is a slightly higher-resource dataset.

LRL	#train	#dev	#test	HRL	#train
az	5.9k	671	903	tr	182k
be	4.5k	248	664	ru	208k
gl	10.0k	682	1007	pt	185k
sk	61.5k	2271	2445	cs	103k

Table 1: Statistics from our low resource language (LRL) and high resource language (HRL) datasets.

Model Details Our model comprises a single bi-directional LSTM as encoder and decoder, with 128-dimensional word embeddings and 512-dimensional hidden states. We are careful to keep this configuration consistent with our baseline model (Neubig and Hu, 2018) to ensure a fair comparison. We use `fairseq`[2] to implement the baseline as well as our proposed models. We set dropout probability to 0.3, and use an adam optimizer with a learning rate of 0.001. In practice, we train a Multi-Sub model until convergence, and then use this model to continue training on cross-teaching data until convergence. For inference, we use beam size 5 with length penalty. We use `sacrebleu`[3] (Post, 2018) to report BLEU (Papineni et al., 2002) scores on the detokenized translations. We perform statistical significance tests for our results based on bootstrap resampling (Koehn, 2004) using `compare-mt` toolkit[4].

For fair comparison with prior work, we use BPE (Subword-nmt, Sennrich et al. 2016) as our primary segmentation toolkit and sentencepiece (SP, Kudo 2018) as our auxiliary tokenizer. We only use the BPE segmentations to tune our model via validation. In other words, while we train on both BPE and SP, we save model checkpoints that are optimized for BPE tokenized inputs[5].

Following Neubig and Hu (2018), we separately learn 8k BPE subwords on each of the source and target languages. When combining an LRL and a HRL, we take the union of the vocabulary on the source side and the target side separately. We use the same procedure with the SP tokenizer using a subword vocabulary size of 4k. To train BPE and SP together, we take the union of the vocabularies

[2] `https://github.com/pytorch/fairseq`

[3] SacreBLEU signature: BLEU+CASE.MIXED+NUMREFS.1 +SMOOTH.EXP+TOK.13A+VERSION.1.4.14

[4] `https://github.com/neulab/compare-mt`

[5] Our model can handle sentencepiece inputs as well. For a model that performs *equally* well on BPE and SP, construct a validation set with equal number of source sentences with both segmentations and save the checkpoints optimized for the validation metric. We chose BPE segments for validation to be comparable with previous work.

Lex Unit	**Model**	**tr/az**	**ru/be**	**pt/gl**	**cs/sk**
Word	Lookup	7.66	13.03	28.65	25.24
Sub-joint	Lookup	9.40	11.72	22.67	24.97
Sub-sep	UniEnc (Gu et al., 2018)	4.80	8.13	14.58	12.09
Sub-sep	Lookup (Neubig and Hu, 2018)[6]	10.8	16.2	27.7	28.4
Sub-sep	Adaptation (All→Bi) (ibid.)	11.7	18.3	28.8	28.2
Word	SDE (Wang et al., 2018)	11.82	18.71	**30.30**	28.77
Sub-sep	SDE (ibid.)	12.35	16.30	28.94	28.35
Multi-Sub	Lookup [BPE + SP] (Ours)	12.0*	18.5**	28.6*	**28.8**†
(BPE 8k + SP 4k)	Lookup + Cross-teaching (Ours)	**12.7****	**18.8****	29.6**	28.6†

Table 2: All models are trained on a LRL and a related HRL with English as the target language with LSTMs. BLEU scores are reported on the test set of the LRL. The sub-sep lookup model (Neubig and Hu, 2018) is our primary baseline (shaded in grey). Our best results compared to the baseline are underlined. Bolding indicates best overall results on the datasets. We indicate statistical significance w.r.t primary baseline with † ($p < 0.05$), $*$ ($p < 0.001$) and $**$ ($p < 0.0001$).

of the source and target sides separately, resulting in a vocabulary which is union of the BPE and SP subword vocabularies of each side.

3.2 Main results

We compare the results of our *Multi-Sub* models against various baselines in Table 2. *Sub-sep* models use a union of subword vocabularies learned separately for each of the source and target languages; the union is performed separately for the source and target sides yielding two separate vocabularies. *Sub-joint* refers to subword vocabularies learned jointly on the concatenation of all of the source and target languages. Such models consistently perform worse than their *sub-sep* counterparts for all datasets, as the HRL tends to occupy a larger share of the vocabulary and leaves the LRL with both a smaller vocabulary as well as smaller subwords. Our reimplementation of the *sub-sep* model (Neubig and Hu, 2018) mitigates this by (separately) learning the same number of subwords for the HRL and LRL. Using words instead of subwords performs on par with the *sub-sep* model for gl → en but worse for other languages.

We see that our model, Multi-Sub, handily outperforms all of these baselines. Compared to the *de-facto* sub-sep model (highlighted in grey, and used as the baseline in the rest of the paper), Multi-Sub without cross-teaching gains +1.2 BLEU points on az and be, and +0.9 on gl. The improvement on cs is not large, but is significant at +0.4 BLEU.

We also compare our approach against more sophisticated models, such as soft decoupled encoding (SDE, Wang et al. 2018) which shares lexical and latent semantic representations across multiple source languages. Our modest Multi-Sub model with cross-teaching outperforms SDE (with *words* as lexical units) on three out of four languages, with the largest gain being +0.9 BLEU on az → en. Multi-Sub consistently and significantly outperforms *subword*-level SDE on all language pairs with gains ranging from +0.4 BLEU to +2.5 BLEU. Note that although Multi-Sub is -0.7 BLEU behind *word-level* SDE on gl, it outperforms sub-sep by +2.6 BLEU and *subword-level* SDE by +2.5 BLEU.

Overall, our models are consistently better than the sub-sep baseline. For most languages, substantial improvements over the baseline come when the Multi-Sub model is combined with cross-teaching.

3.3 Comparison with Subword Regularization

Table 3 contrasts Multi-Sub against BPE-dropout (Provilkov et al., 2020), a subword regularization technique.[7] For comparison we report results from the baseline sub-sep model with and without subword regularization. Our implementation applies BPE-dropout to the training data with probability $p = 0.1$, and the model and training are otherwise identical to sub-sep.

	tr/az	**ru/be**	**pt/gl**	**cs/sk**
Sub-sep	10.8	16.2	27.7	28.4
+ SR	11.0	16.6	28.4	28.2
Multi-sub	**12.7**	**18.8**	**29.6**	**28.8**

Table 3: Comparing subword regularization (SR) with our best results. We use BPE-dropout (Provilkov et al., 2020) at $p = 0.1$.

[7] Using only one tokenizer (either BPE or SP) with different subword sizes closely resembles subword regularization. Using SP and BPE, on the other hand, results in different word-boundary markers that makes our technique distinct.

Although subword regularization improves upon the baseline model, the difference is small, likely because of the small amount of data available for the LRLs. By contrast our Multi-Sub technique yields much larger gains.

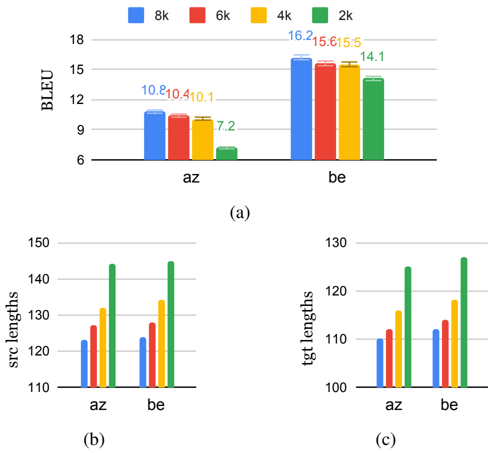
Discussion BPE-dropout (Provilkov et al., 2020) is a subword regularization technique that exposes the model to learn better word compositionalities by probabilistically producing multiple segmentations for each word. Multi-Sub, on the other hand, uses a secondary subword segmentation of lower vocabulary size and leverages its compositionalities as a related language to learn better representations. In Multi-Sub with cross-teaching, the model learns to produce four way translations on the same source and target languages: BPE [src] → {BPE [tgt] , SP [tgt]} and SP [src] → {BPE [tgt] , SP [tgt]}. Although this method is deterministic, and the model learns from only two unique subword sequences instead of one (e.g. sub-sep), this inter-segmentation interaction through multilingual training helps the model learn better compositionalities and morphology. See Section 4.2 for a discussion on the linguistic complexity of the output translations.

3.4 Choice of Auxiliary Subwords

Our primary subword tokenizer is BPE with 8000 subwords; we use sentencepiece (SP) as our auxiliary subword tokenizer. To choose the right auxiliary subword vocabulary size, we experiment with three different sizes (6k, 4k and 2k) on `tr/az` and `ru/be` datasets. To determine the optimal vocabulary size, we focus on two key aspects of the candidate segmentations: translation quality and average sentence length. Figure 2 presents a summary of our results.

On both datasets, subword vocabularies of sizes 6k and 4k yield slightly lower BLEU scores than the baseline with 8k subwords; the drop is minimal (`az`: 10.4 vs. 10.1, `be`: 15.6 vs. 15.5 for 6k and 4k). Performance is substantially worse on the same datasets with 2k subwords (7.2 for `az` and 14.1 for `be`) so we reject the 2k setting.

Next, we compare the average sentence lengths in the subword-tokenized training data (both source and target sides) across different subword vocabulary sizes. At a vocabulary size of 6k, sentence length does not vary substantially from the length found with 8k subwords (Figure 2(b, c)). 4k subwords yield a more significant increase in sentence length on both source (`tr/az`: +9, `ru/be`: +10) and target sides for both datasets. This is favourable since this guarantees as many new subwords as possible in the sentence without increasing its length dramatically. On the basis of these results, we have chosen 4k SP subwords for our auxiliary segmentations.



(a)

(b)

(c)

Figure 2: Effect of auxiliary subword vocabulary size on BLEU (a) and sentence length (b, c) in `tr/az` and `ru/be`.

4 Analysis

4.1 Correlation to Data Availability

Using a secondary subword model as a related language yields different degrees of improvement in different languages. We investigate whether these variations correlate with the degree to which the LRL is "low-resource".

We report (Table 4) the amount of training data available for the LRL, the word-level vocabulary size of each LRL (v_{LRL}), and the ratio of this size to the vocabulary size of the corresponding HRL (v_{HRL}). The ratio v_{LRL}/v_{HRL} is directly pro-

	#train	$\mathbf{v_{LRL}}$	$\frac{\mathbf{v_{LRL}}}{\mathbf{v_{HRL}}}$	**BLEU Δ**
az	5.94k	13.1k	11.29	+1.90
be	4.50k	9.9k	11.43	+2.61
gl	10.03k	10.9k	27.69	+1.90
sk	61.50k	48.5k	80.01	+0.40

Table 4: Comparison of size of training data in LRL with the BLEU improvements. Column 4 shows the ratio of the word vocabularies of LRL (v_{LRL}) to HRL (v_{HRL}). The ratios are multiplied by 100 for readability.

[7] The numbers are from our reimplementation of Neubig and Hu (2018). Original BLEU scores on this dataset were az: 10.9, be: 15.8, gl: 27.3, sk: 25.5 while a reimplementation by Wang et al. (2018) yields az: 10.9, be: 16.17, gl: 28.1, sk: 28.5. Our implementation matches the performance on all test sets except for gl where we lag by 0.5 points.

	Model	BLEU	TTR	RTTR	LTTR	MTTR ↓	HD-D	MTLD	MTLD-A	MTLD-Bi	Yule's K ↓
Az→En	Reference	–	0.1845	22.98	0.8248	0.0417	0.8738	106.60	108.47	108.17	80.68
1	Base	10.8	0.0855	10.9615	0.7466	0.0600	0.7750	33.9342	38.3466	38.1259	170.4321
2	BPE 8k + SP 4k	12.0	0.0971	12.2866	0.7591	0.0572	0.7936	40.0937	44.7958	44.8005	152.0778
3	2 + Cross-teach	12.7	**0.0993**	**12.4746**	**0.7610**	**0.0569**	**0.7961**	**41.3529**	**45.4622**	**45.3590**	**149.4563**
Be→En	Reference	–	0.1863	20.83	0.8219	0.0434	0.8687	102.95	104.44	104.3692	85.73
1	Base	16.2	0.1149	13.0503	0.7714	0.0556	0.8045	51.1452	52.4293	52.6571	139.7345
2	BPE 8k + SP 4k	18.5	0.1225	13.7806	0.7777	0.0542	0.8017	51.9363	52.9719	53.0382	147.5613
3	2 + Cross-teach	18.8	**0.1249**	**14.0746**	**0.7799**	**0.0536**	**0.8071**	**54.8368**	**55.6391**	**55.7884**	**142.6042**
Gl→En	Reference	–	0.1484	19.45	0.8043	0.0462	0.8643	91.22	94.81	94.67	87.92
1	Base	27.7	0.1329	17.1629	0.7924	0.0492	0.8312	72.9798	73.9316	73.8523	120.5782
2	BPE 8k + SP 4k	28.6	0.1365	17.6551	0.7952	0.0485	0.8328	**76.0790**	**75.5915**	**75.5815**	119.1850
3	2 + Cross-teach	29.6	**0.1366**	**17.7624**	**0.7955**	**0.0484**	**0.8307**	74.6902	73.7315	73.7201	**112.5075**
Sk→En	Reference	–	0.1253	25.5328	0.8047	0.0423	0.8689	95.38	102.52	102.24	86.20
1	Base	28.4	0.0935	18.9185	0.7769	0.0484	0.8383	72.7529	74.8386	74.9117	112.8484
2	BPE 8k + SP 4k	28.8	**0.0954**	19.3010	**0.7787**	**0.0480**	**0.8411**	**74.5821**	**76.1596**	**76.2799**	**110.8807**
3	2 + Cross-teach	28.6	0.0947	**19.3118**	0.7784	**0.0480**	0.8379	72.8657	74.7803	74.8770	114.8330

Table 5: Lexical diversity of the reference human translations vs. model outputs in different settings for each LRL.

portional to the number of training samples in the LRLs. This ratio has a generally *negative* correlation to the BLEU gains in our models—the more training data is available, the smaller the improvements. This strongly suggests that using auxiliary subwords as a foreign language is a technique best suited to low resource languages.

4.2 Linguistic Complexity

While estimating linguistic complexity is a multifarious task, lexical and morphological diversity are two of its major components. In this section we perform an exhaustive assessment of our models' translations using lexical diversity metrics (Section 4.2.1) and morphological inflectional diversity metrics (Section 4.2.2).

4.2.1 Lexical Richness

We use several metrics to quantify lexical diversity across translations from different models[8]. The metrics include type-token ratio (TTR) and its variants—Root TTR (RTTR, Guiraud 1960), Log TTR (LTTR), and (MATTR, Covington and McFall 2010)—hypergeometric distribution D (HDD, McCarthy and Jarvis 2007), measure of textual, lexical diversity (MTLD, McCarthy 2005) and Yule's K (Yule, 2014). The scores for these measures are presented in Table 5 for our model outputs and for the reference human translations.

On average, Multi-Sub training with cross-teaching significantly improves the lexical diversity of the generated translations. Improvements in lexical diversity correlate with BLEU scores in all languages (which need not be the case, cf. Vanmassenhove et al. 2021), implying that our methods produce translations which are not only more accurate, but also richer and more varied in terms of vocabulary. These effects are most pronounced in the lowest-resource languages, `az` and `be`, where cross-teaching yields improvements in every metric relative to both the baseline and Multi-Sub training without cross-teaching. In `gl`, cross-teaching yields improvements in all metrics except MTLD and its variants, which are optimized by Multi-Sub training without cross-teaching. `Sk` is unique in that the greatest improvements for most metrics come from Multi-Sub training without cross-teaching. This parallels the pattern observed in the BLEU scores (Table 4), and confirms our earlier claim that cross-teaching is most effective in cases of extreme data scarcity, while Multi-Sub training without cross-teaching works better for high resource languages.

4.2.2 Morphological Richness

To examine the morphological complexity of the translations produced by our models, we averaged the inflectional diversity of the lemmas. Following Vanmassenhove et al. (2021), we used the Spacy-udpipe lemmatizer to retrieve all lemmas[9].

Shannon Entropy (H, Shannon 1948) is used to measure the variety of inflected forms associated with a given lemma (higher entropy means more variation). Entropy is averaged across each lemma in the model outputs.

[8] The intent of this section is not to claim that LD metrics are potential indicators of proficiency, quality or sophistication; they simply represent qualities which may be desirable for certain applications, cf. Vanmassenhove et al. (2021)

[9] `https://github.com/TakeLab/spacy-udpipe`

	Model	BLEU	H ↑	D ↓
Az→En	Reference	–	69.26	54.75
1	Base	10.8	64.12	59.14
2	BPE 8k + SP 4k	12.0	63.67	59.67
3	2 + Cross-teach	12.7	**65.62**	**57.97**
Be→En	Reference	–	71.24	53.97
1	Base	16.2	64.12	59.14
2	BPE 8k + SP 4k	18.5	67.32	67.78
3	2 + Cross-teach	18.8	**67.78**	**57.52**
Gl→En	Reference	–	68.27	55.88
1	Base	27.7	66.64	56.95
2	BPE 8k + SP 4k	28.6	**66.93**	56.95
3	2 + Cross-teach	29.6	66.20	**56.92**
Sk→En	Reference	–	69.03	55.41
1	Base	28.4	62.96	59.18
2	BPE 8k + SP 4k	28.8	**63.41**	58.91
3	2 + Cross-teach	28.6	62.50	**59.37**

Table 6: Morphological diversity measures comparing our model outputs against the human references.

Simpson's Diversity Index (D, Simpson 1949) measures the probability that two randomly-sampled items have the same label; large values imply homogeneity (most items belong to the same category). We measure morphological diversity by computing the probability that two instances of a given lemma represent the same inflected form.

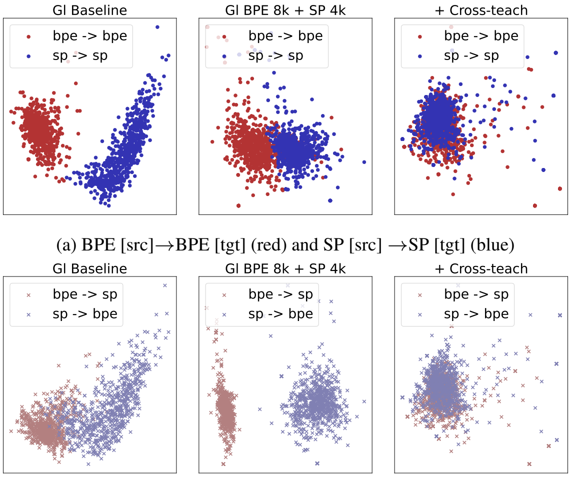
The results in Table 6 parallel the lexical diversity evaluation: in the extremely low-resource languages `az` and `be`, cross-teaching yields a clear improvement in both the entropy and diversity index of the output translations. The model thus employs a greater variety of inflectional forms, which provides more choices to the decoder (Vanmassenhove et al., 2021) (c.f. Fig. 8). In slightly higher-resource languages like `sk`, the impact of cross-teaching is less pronounced: the best diversity index is in `gl`, but Multi-Sub training without cross-teaching yields the best entropy. Multi-Sub training without cross-teaching also yields the greatest degree of morphological diversity in `sk`.

Model	**gl**	**sk**
Base	0.39	0.11
Multi-Sub/Cross-teaching	**0.51***†	**0.12**†

Table 7: F1 scores on zero-shot NER in `sk` and `gl`. † means the best result comes from cross-teaching; ∗ means the best result comes without cross-teaching.

4.3 Improved Cross-lingual Transfer

Downstream Task: NER Multi-Sub training improves the usefulness of subword embeddings for downstream tasks. We train NER models on `pt` and `cs` using the pre-trained embeddings from our translation models; then, following Sharoff 2017, we evaluate each of these models on the corresponding LRL.[10] Since the NER models are never trained on LRL data, this is a zero-shot evaluation where model performance should reflect the degree of multilinguality in the pre-trained embeddings. Table 7 reports F1 scores for this task. We observe that Multi-Sub training on its own can yield significant performance improvements (as in `gl`), but cross-teaching is sometimes required to obtain optimal results (as in `sk`). Together with the results in Figure 3, this suggests that cross-teaching can play a crucial role in facilitating cross-lingual transfer.



(a) BPE [src]→BPE [tgt] (red) and SP [src] →SP [tgt] (blue)

(b) BPE [src]→SP [tgt] (red) and SP [src] →BPE [tgt] (blue)

Figure 3: PCA decomposition of Galician sentence representations in the baseline (left), Multi-Sub (center), and cross-teaching (right) settings. Multi-Sub training can reduce separation between tokenizations, while the addition of cross-teaching eliminates separation entirely.

Visualizations of Sentence Embeddings We find that cross-teaching significantly reduces the separation between different tokenizations in the sentence representations of certain languages. Figure 3 shows the distribution of sentence representations produced by our two tokenizers. In the baseline, BPE-tokenized sentences are clearly separated from (parallel) SP-tokenized sentences; in the Multi-Sub setting we observe less separation, although distinct clusters of BPE and SP inputs are still clearly visible. By contrast, in the cross-teaching setting, there is significant overlap between the representations of BPE and SP inputs.

[10] `cs` training data taken from Sevcíková et al. 2007, `sk` test data from Piskorski et al. 2017, and `pt/gl` training and test data from Garcia and Gamallo 2014

gl (src)	en (ref.)	sub-sep	SDE	multi-sub+cross-teach
Se queres saber sobre o clima, preguntas a un climatólogo.	If you want to know about climate, you ask a climatologist.	If you want to know about climate, you're asking a college friend.	If you want to know about climate, they ask for a weather.	If you want to know about the climat, you ask a climatologist.

Table 8: Example of translations of the same source sentence from `gl`→`en` test set with different models.

This suggests that cross-teaching serves to eliminate "monolingual" subspaces (that is, subspaces representing a single tokenization) in favor of representing all input languages in the same joint space. On the basis of this result, we argue that cross-teaching is an effective technique for increasing the degree of multilinguality in a translation model.[11]

5 Qualitative Analysis

We list translations for the baseline *sub-sep* and SDE models along with our Multi-Sub model in Table 8. While sub-sep results in an entirely unrelated translation of the `gl` word climatólogo, SDE produces a related word weather. Multi-Sub, however, produces an accurate translation of the word which is climatologist.

6 Related Work

Several techniques have been proposed to improve lexical representations for multilingual machine translation. Zoph et al. (2016) propose to first train a HRL parent model, then transfer some of the learned parameters to the LRL child model to initialize and constrain training. Similarly, Nguyen and Chiang (2017) pair related languages together and transfer source word embeddings from parent-HRL words to their child-LRL equivalents. Johnson et al. (2017); Neubig and Hu (2018), on the other hand, learn a joint vocabulary over several languages and train a single NMT model on the concatenated data. Gu et al. (2018) introduce a latent embedding space shared by all languages to enhance parameter sharing in lexical representation. Wang et al. (2018); Gao et al. (2020) use a similar idea but use character n-gram encodings (SDE) instead of the conventional subword/word embeddings. By contrast Multi-Sub does not involve any architectural changes and improves the representation of low-resource languages by training on multiple segmentations of the same corpus.

Subword-regularization methods (Kudo, 2018; Provilkov et al., 2020) share the motivation of alleviating sub-optimal subwords by exposing a model to multiple segmentations of the same word. However, our method is substantially different in that (i) we use two completely different subword algorithms with different vocabulary sizes (*contra* Wang et al. 2021), and (ii) we do not rely on expensive sampling procedures (*contra* Kudo 2018) or additional data to learn an LM. Especially for low-resource languages, our method not only improves translation quality but also enhances a model's cross-lingual transfer capabilities. Finally, this simple architecture-agnostic technique can act as drop-in improvement for existing methods.

7 Conclusion

This work introduces Multi-Sub training with cross-teaching—a novel technique that combines multiple alternative subword tokenizations of a source-target language pair—to improve the representation of low-resource languages. Our proposed methods obtain significant gains on low-resource datasets from multilingual TED-talks. We performed exhaustive analysis to show that our methods also increase the lexical and morphological diversity of the output translations, and produce better multilingual representations which we demonstrate by performing zero-shot NER by exploiting representations from a high resource language. Multi-Sub training and cross-teaching are simple architecture-agnostic steps which can be easily applied to existing single or multilingual neural machine translation models and do not require any external data.

Acknowledgements

N.K would like to thank Kumar Abhishek for the numerous discussions that helped shape this paper. The research was partially supported by the Natural Sciences and Engineering Research

[11] In this respect, cross-teaching has a similar effect to BPE-dropout (Provilkov et al., 2020), which serves to eliminate monolingual subspaces at the level of subword embeddings (but recall our prior comments on the distinction between BPE-dropout and Multi-Sub in Section 3.3).

Council of Canada grants NSERC RGPIN-2018-06437 and RGPAS-2018-522574 and a Department of National Defence (DND) and NSERC grant DGDND-2018-00025 to the third author.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Michael A Covington and Joe D McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of quantitative linguistics*, 17(2):94–100.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Beijing, China.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Luyu Gao, Xinyi Wang, and Graham Neubig. 2020. Improving target-side lexical transfer in multilingual neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Marcos Garcia and Pablo Gamallo. 2014. Multilingual corpora with coreferential annotation of person entities. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor OK Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

P Guiraud. 1960. *Problèmes et Méthodes de la Statistique Linguistique.* Presses universitaires de France.

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

Nishant Kambhatla, Logan Born, and Anoop Sarkar. 2022. CipherDAug: Ciphertext Based Data Augmentation for Neural Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Philip M McCarthy. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph.D. thesis, The University of Memphis.

Philip M McCarthy and Scott Jarvis. 2007. vocd: A theoretical and empirical evaluation. *Language Testing*, 24(4):459–488.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Toan Q Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*.

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 As-*

sociation for Computational Linguistics, pages 311–318.

Jakub Piskorski, Lidia Pivovarova, Jan Šnajder, Josef Steinberger, and Roman Yangarber. 2017. The first cross-lingual challenge on recognition, normalization, and matching of named entities in slavic languages. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, Valencia, Spain.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, Belgium, Brussels. Association for Computational Linguistics.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. Bpe-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535.

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*, pages 1715–1725.

Magda Sevcíková, Zdenek Zabokrtský, and Oldrich Kruza. 2007. Named entities in czech: Annotating data and developing NE tagger. In *Text, Speech and Dialogue, 10th International Conference, TSD 2007, Pilsen, Czech Republic, September 3-7, 2007, Proceedings*, volume 4629 of *Lecture Notes in Computer Science*, pages 188–195. Springer.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(3):379–423.

Serge Sharoff. 2017. Toward pan-Slavic NLP: Some experiments with language adaptation. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, Valencia, Spain. Association for Computational Linguistics.

Edward H Simpson. 1949. Measurement of diversity. *nature*, 163(4148):688.

Xu Tan, Jiale Chen, Di He, Yingce Xia, Tao Qin, and Tie-Yan Liu. 2019. Multilingual neural machine translation with language clustering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 963–973.

Fiona J Tweedie and R Harald Baayen. 1998. How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities*, 32(5):323–352.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online..

Xinyi Wang, Hieu Pham, Philip Arthur, and Graham Neubig. 2018. Multilingual neural machine translation with soft decoupled encoding. In *International Conference on Learning Representations*.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2021. Multi-view subword regularization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 473–482, Online.

Lijun Wu, Shufang Xie, Yingce Xia, Yang Fan, Jian-Huang Lai, Tao Qin, and Tieyan Liu. 2020. Sequence generation with mixed representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10388–10398. PMLR.

C Udny Yule. 2014. *The statistical study of literary vocabulary*. Cambridge University Press.

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