

# Generalized Data Augmentation for Low-Resource Translation

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, Graham Neubig  
Language Technologies Institute, Carnegie Mellon University  
{mengzhox, xiangk, aanastas, gneubig}@andrew.cmu.edu

## Abstract

Translation to or from low-resource languages (LRLs) poses challenges for machine translation in terms of both adequacy and fluency. Data augmentation utilizing large amounts of monolingual data is regarded as an effective way to alleviate these problems. In this paper, we propose a general framework for data augmentation in low-resource machine translation that not only uses target-side monolingual data, but also pivots through a related high-resource language (HRL). Specifically, we experiment with a two-step pivoting method to convert high-resource data to the LRL, making use of available resources to better approximate the true data distribution of the LRL. First, we inject LRL words into HRL sentences through an induced bilingual dictionary. Second, we further edit these modified sentences using a modified unsupervised machine translation framework. Extensive experiments on four low-resource datasets show that under extreme low-resource settings, our data augmentation techniques improve translation quality by up to 1.5 to 8 BLEU points compared to supervised back-translation baselines.

## 1 Introduction

The task of Machine Translation (MT) for low resource languages (LRLs) is notoriously hard due to the lack of the large parallel corpora needed to achieve adequate performance with current Neural Machine Translation (NMT) systems (Koehn and Knowles, 2017). A standard practice to improve training of models for an LRL of interest (e.g. Azerbaijani) is utilizing data from a related high-resource language (HRL, e.g. Turkish). Both transferring from HRL to LRL (Zoph et al., 2016; Nguyen and Chiang, 2017; Gu et al., 2018) and joint training on HRL and LRL parallel data (Johnson et al., 2017; Neubig and Hu, 2018) have shown

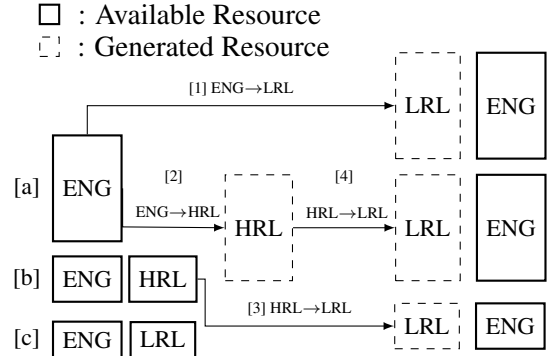


Figure 1: With a low-resource language (LRL) and a related high-resource language (HRL), typical data augmentation scenarios use any available parallel data [b] and [c] to back-translate English monolingual data [a] and generate parallel resources ([1] and [2]). We additionally propose scenarios [3] and [4], where we pivot through HRL in order to generate a LRL-ENG resource.

to be effective techniques for low-resource NMT. Incorporating data from other languages can be viewed as one form data augmentation, and particularly large improvements can be expected when the HRL shares vocabulary or is syntactically similar with the LRL (Lin et al., 2019). Simple joint training is still not ideal, though, considering that there will still be many words and possibly even syntactic structures that will not be shared between the most highly related languages. There are model-based methods that ameliorate the problem through more expressive source-side representations conducive to sharing (Gu et al., 2018; Wang et al., 2019), but they add significant computational and implementation complexity.

In this paper, we examine how to better share information between related LRL and HRLs through a framework of *generalized data augmentation* for low-resource MT. In our basic setting, we have access to parallel or monolingual data of an LRL of interest, its HRL, and the target language, which

we will assume is English. We propose methods to create pseudo-parallel LRL data in this setting. As illustrated in Figure 1, we augment parallel data via two main methods: 1) back-translating from ENG to LRL or HRL; 2) converting the HRL-ENG dataset to a pseudo LRL-ENG dataset.

In the first thread, we focus on creating new parallel sentences through back-translation. Back-translating from the target language to the source (Sennrich et al., 2016) is a common practice in data augmentation, but has also been shown to be less effective in low-resource settings where it is hard to train a good back-translation model (Currey et al., 2017). As a way to ameliorate this problem, we examine methods to instead translate from the target language to a highly-related HRL, which remains unexplored in the context of low-resource NMT. This pseudo-HRL-ENG dataset can then be used for joint training with the LRL-ENG dataset.

In the second thread, we focus on converting an HRL-ENG dataset to a pseudo-LRL-to-ENG dataset that better approximates the true LRL data. Converting between HRLs and LRLs also suffers from lack of resources, but because the LRL and HRL are related, this is an easier task that we argue can be done to some extent by simple (or unsupervised) methods.<sup>1</sup> In our proposed method, for the first step, we substitute HRL words on the source side of HRL parallel datasets with corresponding LRL words from an induced bilingual dictionary generated by mapping word embedding spaces (Xing et al., 2015; Lample et al., 2018b). In the second step, we further attempt translate the pseudo-LRL sentences to be closer to LRL ones utilizing an unsupervised machine translation framework.

In sum, our contributions are four fold:

1. We conduct a thorough empirical evaluation of data augmentation methods for low-resource translation that take advantage of all accessible data, across four language pairs.
2. We explore two methods for translating between related languages: word-by-word substitution using an induced dictionary, and unsupervised machine translation that further uses this word-by-word substituted data as input. These methods improve over simple unsupervised translation from HRL to LRL by more than 2 to 10 BLEU points.

<sup>1</sup>This sort of pseudo-corpus creation was examined in a different context of pivoting for SMT (De Gispert and Marino, 2006), but this was usually done with low-resource source-target language pairs with English as the pivot.

3. Our proposed data augmentation methods improve over standard supervised back-translation by 1.5 to 8 BLEU points, across all datasets, and an additional improvement of up to 1.1 BLEU points by augmenting from both ENG monolingual data, as well as HRL-ENG parallel data.

## 2 A Generalized Framework for Data Augmentation

In this section, we outline a generalized data augmentation framework for low-resource NMT.

### 2.1 Datasets and Notations

Given an LRL of interest and its corresponding HRL, with the goal of translating the LRL to English, we usually have access to 1) a limited-sized LRL-ENG parallel dataset  $\{\mathcal{S}_{LE}, \mathcal{T}_{LE}\}$ ; 2) a relatively high-resource HRL-ENG parallel dataset  $\{\mathcal{S}_{HE}, \mathcal{T}_{HE}\}$ ; 3) a limited-sized LRL-HRL parallel dataset  $\{\mathcal{S}_{HL}, \mathcal{T}_{HL}\}$ ; 4) large monolingual datasets in LRL  $\mathcal{M}_L$ , HRL  $\mathcal{M}_H$  and English  $\mathcal{M}_E$ .

To clarify notation, we use  $\mathcal{S}$  and  $\mathcal{T}$  to denote the source and target sides of parallel datasets, and  $\mathcal{M}$  for monolingual data. Created data will be referred to as  $\hat{\mathcal{S}}_{A \rightarrow B}^m$ . The superscript  $m$  denotes a particular augmentation approach (specified in Section 3). The subscripts denote the translation direction that is used to create the data, with the LRL, HRL, and ENG denoted with ‘L’, ‘H’, and ‘E’ respectively.

### 2.2 Augmentation from English

The first two options for data augmentation that we explore are typical back-translation approaches:

1. **ENG-LRL** We train an ENG-LRL system and back-translate English monolingual data to LRL, denoted by  $\{\hat{\mathcal{S}}_{E \rightarrow L}, \mathcal{M}_E\}$ .
2. **ENG-HRL** We train an ENG-HRL system and back-translate English monolingual data to HRL, denoted by  $\{\hat{\mathcal{S}}_{E \rightarrow H}, \mathcal{M}_E\}$ .

Since we have access to LRL-ENG and HRL-ENG parallel datasets, we can train these back-translation systems (Sennrich et al., 2016) in a supervised fashion. The first option is the common practice for data augmentation. However, in a low-resource scenario, the created LRL data can be of very low quality due to the limited size of training data, which in turn could deteriorate the LRL $\rightarrow$ ENG translation performance. As we show in Section 5, this is indeed the case.

The second direction, using HRL back-translated data for LRL $\rightarrow$ ENG translation, has not been explored in previous work. However, we suggest that in low-resource scenarios it has potential to be more effective than the first option because the quality of the generated HRL data will be higher, and the HRL is close enough to the LRL that joint training of a model on both languages will likely have a positive effect.

### 2.3 Augmentation via Pivoting

Using HRL-ENG data improves LRL-ENG translation because (1) adding extra ENG data improves the target-side language model, (2) it is possible to share vocabulary (or subwords) between languages, and (3) because the syntactically similar HRL and LRL can jointly learn parameters of the encoder. However, regardless of how close these related languages might be, there still is a mismatch between the vocabulary, and perhaps syntax, of the HRL and LRL. However, translating between HRL and LRL should be an easier task than translating from English, and we argue that this can be achieved by simple methods.

Hence, we propose ‘‘Augmentation via Pivoting’’ where we create an LRL-ENG dataset by translating the source side of HRL-ENG data, into the LRL. There are again two ways in which we can construct a new LRL-ENG dataset:

3. **HRL-LRL** We assume access to an HRL-ENG dataset. We then train an HRL-LRL system and convert the HRL side of  $\mathcal{S}_{\text{HE}}$  to LRL, creating a  $\{\hat{\mathcal{S}}_{\text{H}\rightarrow\text{L}}, \mathcal{T}_{\text{HE}}\}$  dataset.
4. **ENG-HRL-LRL** Exactly as before, except that the HRL-ENG dataset is the result of back-translation. That means that we have first converted English monolingual data  $\mathcal{M}_{\text{E}}$  to  $\hat{\mathcal{S}}_{\text{E}\rightarrow\text{H}}$ , and then we convert those to the LRL, creating a dataset  $\{\hat{\mathcal{S}}_{\text{E}\rightarrow\text{H}\rightarrow\text{L}}, \mathcal{M}_{\text{E}}\}$ .

Given a LRL-HRL dataset  $\{\mathcal{S}_{\text{LH}}, \mathcal{T}_{\text{LH}}\}$  one could also train supervised back-translation systems. But we still face the same problem of data scarcity, leading to poor quality of the augmented datasets. Based on the fact that an LRL and its corresponding HRL can be similar in morphology and word order, in the following sections, we propose methods to convert HRL to LRL for data augmentation in a more reliable way.

## 3 LRL-HRL Translation Methods

In this section, we introduce two methods for converting HRL to LRL for data augmentation.

### 3.1 Augmentation with Word Substitution

Mikolov et al. (2013) show that the word embedding spaces share similar innate structure over different languages, making it possible to induce bilingual dictionaries with a limited amount of or even without parallel data (Xing et al., 2015; Zhang et al., 2017; Lample et al., 2018b). Although the capacity of these methods is naturally constrained by the intrinsic properties of the two mapped languages, it’s more likely to create a high-quality bilingual dictionary for two highly-related languages. Given the induced dictionary, we can substitute HRL words with LRL ones and construct a word-by-word translated pseudo-LRL corpus.

**Dictionary Induction** We use a supervised method to obtain a bilingual dictionary between the two highly-related languages. Following Xing et al. (2015), we formulate the task of finding the optimal mapping between the source and target word embedding spaces as the Procrustes problem (Schönemann, 1966), which can be solved by singular value decomposition (SVD):

$$\min_W \|WX - Y\|_F^2 \text{ s.t. } W^T W = I,$$

where  $X$  and  $Y$  are the source and target word embedding spaces respectively.

As a seed dictionary to provide supervision, we simply exploit identical words from the two languages. With the learned mapping  $W$ , we compute the distance between mapped source and target words with the CSLS similarity measure (Lample et al., 2018b). Moreover, to ensure the quality of the dictionary, a word pair is only added to the dictionary if both words are each other’s closest neighbors. Adding an LRL word to the dictionary for every HRL word results in relatively poor performance due to noise as shown in Section 5.3.

**Corpus Construction** Given an HRL-ENG  $\{\mathcal{S}_{\text{HE}}, \mathcal{T}_{\text{HE}}\}$  or a back-translated  $\{\hat{\mathcal{S}}_{\text{E}\rightarrow\text{H}}, \mathcal{M}_{\text{E}}\}$  dataset, we substitute the words in  $\mathcal{S}_{\text{HE}}$  with the corresponding LRL ones using our induced dictionary. Words not in the dictionary are left untouched. By injecting LRL words, we convert the original or augmented HRL data into pseudo-LRL, which explicitly increases lexical overlap between the concatenated LRL and HRL data. The

created datasets are denoted by  $\{\hat{\mathcal{S}}_{H \rightarrow E}^w, \mathcal{T}_{HE}\}$  and  $\{\hat{\mathcal{S}}_{E \rightarrow H \rightarrow L}^w, \mathcal{M}_E\}$  where  $w$  denotes augmentation with word substitution.

### 3.2 Augmentation with Unsupervised MT

Although we assume LRL and HRL to be similar with regards to word morphology and word order, the simple word-by-word augmentation process will almost certainly be insufficient to completely replicate actual LRL data. A natural next step is to further convert the pseudo-LRL data into a version closer to the real LRL. In order to achieve this in our limited-resource setting, we propose to use unsupervised machine translation (UMT).

**UMT** Unsupervised Neural Machine Translation (Artetxe et al., 2018; Lample et al., 2018a,c) makes it possible to translate between languages without parallel data. This is done by coupling denoising auto-encoding, iterative back-translation, and shared representations of both encoders and decoders, making it possible for the model to extend the initial naive word-to-word mapping into learning to translate longer sentences.

Initial studies of UMT have focused on data-rich, morphologically simple languages like English and French. Applying the UMT framework to low-resource and morphologically rich languages is largely unexplored, with the exception of Neubig and Hu (2018) and Guzmán et al. (2019), showing that UMT performs exceptionally poorly between dissimilar language pairs with BLEU scores lower than 1. The problem is naturally harder for morphologically rich LRLs due to two reasons. First, morphologically rich languages have a higher proportions of infrequent words (Chahuneau et al., 2013). Second, even though still larger than the respective parallel datasets, the size of monolingual datasets in these languages is much smaller compared to HRLs.

**Modified Initialization** As pointed out in Lample et al. (2018c), a good initialization plays a critical role in training NMT in an unsupervised fashion. Previously explored initialization methods include: 1) word-for-word translation with an induced dictionary to create synthetic sentence pairs for initial training (Lample et al., 2018a; Artetxe et al., 2018); 2) joint Byte-Pair-Encoding (BPE) for both the source and target corpus sides as a pre-processing step. While the first method intends to give a reasonable prior for parameter search, the second method simply forces the source and target

languages to share the same subword vocabulary, which has been shown to be effective for translation between highly related languages.

Inspired by these two methods, we propose a new initialization method that uses our word substitution strategy (§3.1). Our initialization is comprised of a sequence of three steps:

1. First, we use an induced dictionary to substitute HRL words in  $\mathcal{M}_H$  to LRL ones, producing a pseudo-LRL monolingual dataset  $\hat{\mathcal{M}}_L$ .
2. Second, we learn a joint word segmentation model on both  $\mathcal{M}_L$  and  $\hat{\mathcal{M}}_L$  and apply it to both datasets.
3. Third, we train a NMT model in an unsupervised fashion between  $\mathcal{M}_L$  and  $\hat{\mathcal{M}}_L$ . The training objective  $\mathcal{L}$  is a weighted sum of two loss terms for denoising auto-encoding and iterative back-translation:

$$\begin{aligned} \mathcal{L} = & \lambda_1 (\mathbb{E}_{x \sim \mathcal{M}_L} - \log P_{s \rightarrow s}(x|C(x)) \\ & + \mathbb{E}_{y \sim \hat{\mathcal{M}}_L} - \log P_{t \rightarrow t}(y|C(y))) \\ & + \lambda_2 (\mathbb{E}_{x \sim \mathcal{M}_L} - \log P_{t \rightarrow s}(x|u^*(y|x)) \\ & + \mathbb{E}_{y \sim \hat{\mathcal{M}}_L} - \log P_{s \rightarrow t}(y|u^*(x|y))) \end{aligned}$$

where  $u^*$  denotes translations obtained with greedy decoding,  $C$  denotes a noisy manipulation over input including dropping and swapping words randomly,  $\lambda_1$  and  $\lambda_2$  denotes the weight of language modeling and back translation respectively.

In our method, we do not use any synthetic parallel data for initialization, expecting the model to learn the mappings between a true LRL distribution and a pseudo-LRL distribution. This takes advantage of the fact that the pseudo-LRL is naturally closer to the true LRL than the HRL is, as the injected LRL words increase vocabulary overlap.

**Corpus Construction** Given the word-level augmented datasets  $\{\hat{\mathcal{S}}_{H \rightarrow E}^w, \mathcal{T}_{HE}\}$  and  $\{\hat{\mathcal{S}}_{E \rightarrow H \rightarrow L}^w, \mathcal{M}_E\}$ , we use the UMT model trained with this method to translate the pseudo-LRL data from  $\hat{\mathcal{S}}_{H \rightarrow E}^w$  and from  $\hat{\mathcal{S}}_{E \rightarrow H \rightarrow L}^w$ . We obtain new parallel datasets  $\{\hat{\mathcal{S}}_{H \rightarrow E}^m, \mathcal{T}_{HE}\}$  and  $\{\hat{\mathcal{S}}_{E \rightarrow H \rightarrow L}^m, \mathcal{M}_E\}$  with superscript  $m$  denoting Modified UMT (M-UMT). We use superscript  $u$  for un-modified standard UMT.

### 3.3 Why Pivot for Back-Translation?

Pivoting through an HRL in order to convert English to LRL will be a better option compared to

Datasets	LRL (HRL)			
	AZE (TUR)	BEL (RUS)	GLG (POR)	SLK (CES)
$\mathcal{S}_{LE}, \mathcal{T}_{LE}$	5.9K	4.5K	10K	61K
$\mathcal{S}_{HE}, \mathcal{T}_{HE}$	182K	208K	185K	103K
$\mathcal{S}_{LH}, \mathcal{T}_{LH}$	5.7K	4.2K	3.8K	44K
$\mathcal{M}_L$	2.02M	1.95M	1.98M	2M
$\mathcal{M}_H$	2M	2M	2M	2M
$\mathcal{M}_E$	2M/ 200K			

Table 1: Statistics (number of sentences) of all datasets.

directly translating ENG to LRL under the following three conditions: 1) HRL and LRL are related enough to allow for the induction of a high-quality bilingual dictionary; 2) There exists a relatively high-resource HRL-ENG dataset; 3) A high-quality LRL-ENG dictionary is hard to acquire due to data scarcity or morphological distance.

Essentially, the direct ENG→LRL back-translation may suffer from both data scarcity and morphological differences between the two languages. Our proposal breaks the process into two easier steps: ENG→HRL translation is easier due to the availability of data, and HRL→LRL translation is easier because the two languages are related.

A good example is the agglutinative language of Azerbaijani, where each word may consist of several morphemes and each morpheme could possibly map to an English word itself. Correspondences to (also agglutinative) Turkish, however, are easier to uncover. To give a concrete example, the Azerbaijani word “düşüncələrim” can be fairly easily aligned to the Turkish word “düşüncelerim” while in English it corresponds to the phrase “my thoughts”, which is unlikely to be perfectly aligned.

## 4 Experimental Setup

### 4.1 Data

We use the multilingual TED corpus (Qi et al., 2018) as a test-bed for evaluating the efficacy of each augmentation method. We conduct extensive experiments over four low-resource languages: Azerbaijani (AZE), Belarusian (BEL), Galician (GLG), and Slovak (SLK), along with their highly related languages Turkish (TUR), Russian (RUS), Portuguese (POR), and Czech (CES) respectively. We also have small-sized LRL-HRL parallel datasets, and we download Wikipedia dumps to acquire monolingual datasets for all languages.

The statistics of the parallel datasets are shown in Table 1. For AZE, BEL and GLG, we use all available Wikipedia data, while for the rest of the languages we sample a similar-sized corpus. We sample 2M/200K English sentences from Wikipedia data, which are used for baseline UMT training and augmentation from English respectively.

### 4.2 Pre-processing

We train a joint sentencepiece<sup>2</sup> model for each LRL-HRL pair by concatenating the monolingual corpora of the two languages. The segmentation model for English is trained on English monolingual data only. We set the vocabulary size for each model to 20K. All data are then segmented by their respective segmentation model.

We use FastText<sup>3</sup> to train word embeddings using  $\mathcal{M}_L$  and  $\mathcal{M}_H$  with a dimension of 256 (used for the dictionary induction step). We also pre-train subword level embeddings on the segmented  $\mathcal{M}_L$ ,  $\hat{\mathcal{M}}_L$  and  $\mathcal{M}_H$  with the same dimension.

### 4.3 Model Architecture

**Supervised NMT** We use the self-attention Transformer model (Vaswani et al., 2017). We adapt the implementation from the open-source translation toolkit OpenNMT (Klein et al., 2017). Both encoder and decoder consist of 4 layers, with the word embedding and hidden unit dimensions set to 256.<sup>4</sup> We use a batch size of 8096 tokens.

**Unsupervised NMT** We train unsupervised Transformer models with the UnsupervisedMT toolkit.<sup>5</sup> Layer sizes and dimensions are the same as in the supervised NMT model. The parameters of the first three layers of the encoder and the decoder are shared. The embedding layers are initialized with the pre-trained subword embeddings from monolingual data. We set the weight parameters for autodenosing language modeling and iterative back translation as  $\lambda_1 = 1$  and  $\lambda_2 = 1$ .

### 4.4 Training and Model Selection

After data augmentation, we follow the pre-train and fine tune paradigm for learning (Zoph et al., 2016; Nguyen and Chiang, 2017). We first train a base NMT model on the concatenation of  $\{\mathcal{S}_{LE}, \mathcal{T}_{LE}\}$  and  $\{\mathcal{S}_{HE}, \mathcal{T}_{HE}\}$ . Then we adopt the

<sup>2</sup><https://github.com/google/sentencepiece>

<sup>3</sup><https://github.com/facebookresearch/fastText>

<sup>4</sup>We tuned on multiple settings to find the optimal parameters for our datasets.

<sup>5</sup><https://github.com/facebookresearch/UnsupervisedMT>

Training Data		BLEU for X→ENG			
		AZE (TUR)	BEL (RUS)	GLG (POR)	SLK (CES)
Results from Literature					
SDE (Wang et al., 2019)		12.89	18.71	31.16	29.16
many-to-many (Aharoni et al., 2019)		12.78	21.73	30.65	29.54
Standard NMT					
1	$\{\mathcal{S}_{LE}\mathcal{S}_{HE}, \mathcal{T}_{LE}\mathcal{T}_{HE}\}$ (supervised MT)	11.83	16.34	29.51	28.12
2	$\{\mathcal{M}_L, \mathcal{M}_E\}$ (unsupervised MT)	0.47	0.18	1.15	0.75
Standard Supervised Back-translation					
3	+ $\{\hat{\mathcal{S}}_{E→L}^s, \mathcal{M}_E\}$	11.84	15.72	29.19	29.79
4	+ $\{\hat{\mathcal{S}}_{E→H}^s, \mathcal{M}_E\}$	12.46	16.40	30.07	30.60
Augmentation from HRL-ENG					
5	+ $\{\hat{\mathcal{S}}_{H→L}^s, \mathcal{T}_{HE}\}$ (supervised MT)	11.92	15.79	29.91	28.52
6	+ $\{\hat{\mathcal{S}}_{H→L}^u, \mathcal{T}_{HE}\}$ (unsupervised MT)	11.86	13.83	29.80	28.69
7	+ $\{\hat{\mathcal{S}}_{H→L}^w, \mathcal{T}_{HE}\}$ (word subst.)	14.87	23.56	32.02	29.60
8	+ $\{\hat{\mathcal{S}}_{H→L}^m, \mathcal{T}_{HE}\}$ (modified UMT)	14.72	23.31	32.27	29.55
9	+ $\{\hat{\mathcal{S}}_{H→L}^w, \hat{\mathcal{S}}_{H→L}^m, \mathcal{T}_{HE}\mathcal{T}_{HE}\}$	15.24	<b>24.25</b>	32.30	30.00
Augmentation from ENG by pivoting					
10	+ $\{\hat{\mathcal{S}}_{E→H→L}^w, \mathcal{M}_E\}$ (word subst.)	14.18	21.74	31.72	30.90
11	+ $\{\hat{\mathcal{S}}_{E→H→L}^m, \mathcal{M}_E\}$ (modified UMT)	13.71	19.94	31.39	30.22
Combinations					
12	+ $\{\hat{\mathcal{S}}_{H→L}^w, \hat{\mathcal{S}}_{E→H→L}^w, \mathcal{T}_{HE}\mathcal{M}_E\}$ (word subst.)	<b>15.74</b>	<b>24.51</b>	<b>33.16</b>	<b>32.07</b>
13	+ $\{\hat{\mathcal{S}}_{H→L}^w, \hat{\mathcal{S}}_{H→L}^m, \mathcal{T}_{HE}\mathcal{T}_{HE}\}$ + $\{\hat{\mathcal{S}}_{E→H→L}^w, \hat{\mathcal{S}}_{E→H→L}^m, \mathcal{M}_E\mathcal{M}_E\}$	<b>15.91</b>	23.69	32.55	31.58

Table 2: Evaluation of translation performance over four language pairs. Rows 1 and 2 show pre-training BLEU scores. Rows 3–13 show scores after fine tuning. Statistically significantly best scores are **highlighted** ( $p < 0.05$ ).

mixed fine-tuning strategy of Chu et al. (2017), fine-tuning the base model on the concatenation of the base and augmented datasets. For each setting, we perform a sufficient number of updates to reach convergence in terms of development perplexity.

We use the performance on the development sets (as provided by the TED corpus) as our criterion for selecting the best model, both for augmentation and final model training.

## 5 Results and Analysis

A collection of our results with the baseline and our proposed methods is shown in Table 2.

### 5.1 Baselines

The performance of the base supervised model (row 1) varies from 11.8 to 29.5 BLEU points. Generally, the more distant the source language is from English, the worse the performance. A standard unsupervised MT model (row 2) achieves extremely low scores, confirming the results of Guzmán et al.

(2019), indicating the difficulties of directly translating between LRL and ENG in an unsupervised fashion. Rows 3 and 4 show that standard supervised back-translation from English at best yields very modest improvements. Notable is the exception of SLK-ENG, which has more parallel data for training than other settings. In the case of BEL and GLG, it even leads to worse performance. Across all four languages, supervised back-translation into the HRL helps more than into the LRL; data is insufficient for training a good LRL-ENG MT model.

### 5.2 Back-translation from HRL

**HRL-LRL** Rows 5–9 show the results when we create data using the HRL side of an HRL-ENG dataset. Both the low-resource supervised (row 5) and vanilla unsupervised (row 6) HRL→ENG translation do not lead to significant improvements. On the other hand, our simple word substitution approach (row 7) and the modified UMT approach (row 8) lead to improvements across

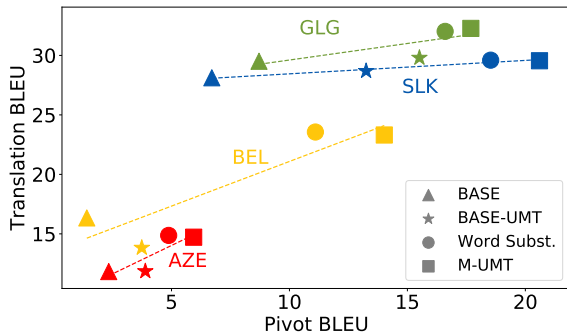


Figure 2: Correlation between HRL-LRL (augmentation) pivot BLEU and LRL-ENG translation BLEU.

the board: +3.0 BLEU points in AZE, +7.8 for BEL, +2.3 for GLG, +1.1 for SLK. These results are significant, demonstrating that the quality of the back-translated data is indeed important.

In addition, we find that combining the datasets produced by our word substitution and UMT models provide an additional improvement in all cases (row 9). Interestingly, this happens despite the fact that the ENG data are the exact same between rows 5–9.

**ENG-HRL-LRL** We also show that even in the absence of parallel HRL-LRL data, our pivoting method is still valuable. Rows 10 and 11 in Table 2 show the translation accuracy when the augmented data are the result of our two-step pivot back-translation. In both cases, monolingual ENG is first translated into HRL and then into LRL with either just word substitution (row 10) or modified UMT (row 11). Although these results are slightly worse than our one-step augmentation of a parallel HRL-LRL dataset, they still outperform the baseline standard back-translation (rows 3 and 4). An interesting note is that in this setting, word substitution is clearly preferable to UMT for the second translation pivoting step, which we explain in §5.3.

**Combinations** We obtain our best results by combining the two sources of data augmentation. Row 12 shows the result of using our simple word substitution technique on the HRL side of both a parallel and an artificially created (back-translated) HRL-ENG dataset. In this setting, we further improve not only the encoder side of our model, as before, but we also aid the decoder’s language modeling capabilities by providing ENG data from two distinct resources. This leads to improvements of 3.6 to 8.2 BLEU points over the base model and 0.3 to 2.1 over our best results from HRL-ENG

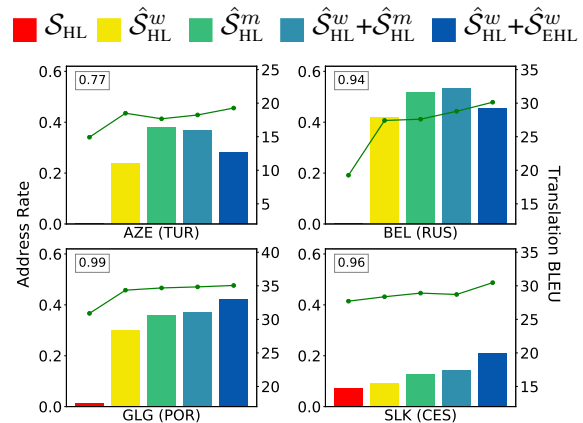


Figure 3: Rare word address rate (bars) and LRL-ENG BLEU scores (line plot) for each data augmentation method. The numbers in each upper left corner is the Pearson correlation coefficient.

augmentation.

Finally, row 13 shows our attempt to obtain further gains by combining the datasets from both word substitution and UMT, as we did in setting 7. This leads to a small improvement of 0.2 BLEU points in AZE, but also to a slight degradation on the other three datasets.

We also compare the results of our augmentation methods with other state-of-the-art methods that either perform improvements to modeling to improve the ability to do parameter sharing (Wang et al., 2019), or train on many different target languages simultaneously (Aharoni et al., 2019). The results demonstrate that the simple data augmentation strategies presented here improve significantly over these previous methods.

### 5.3 Analysis

In this section we focus on the quality of HRL→LRL translation, showing that our better M-UMT initialization method leads to significant improvements compared to standard UMT.

We use the dev sets of the HRL-LRL datasets to examine the performance of M-UMT between related languages. We calculate the pivot BLEU<sup>6</sup> score on the LRL side of each created dataset ( $\mathcal{S}_{HL}$ ,  $\hat{\mathcal{S}}_{H \rightarrow L}^w$ ,  $\hat{\mathcal{S}}_{H \rightarrow L}^u$ ,  $\hat{\mathcal{S}}_{H \rightarrow L}^m$ ). In Figure 2 we plot pivot HRL-LRL BLEU scores against the translation LRL-ENG BLEU ones. First, we observe that across all datasets, the pivot BLEU of our M-UMT method is higher than standard UMT (the squares are all further right than their corresponding stars). Vanilla

<sup>6</sup>We will refer to *pivot BLEU* in order to avoid confusion with translation BLEU scores from the previous sections.

Data	Example Sentence	Pivot BLEU
$\mathcal{S}_{LE}$ (GLG)	Pero con todo, veste obrigado a agardar nas mans dunha serie de estraños moi profesionais.	
$\mathcal{S}_{HE}$ (POR)	Em vez disso, somos obrigados a esperar nas mãos de uma série de estranhos muito profissionais.	0.09
$\hat{\mathcal{S}}_{H \rightarrow L}^w$	En vez disso, somos obrigados a esperar nas <b>mans</b> de <b>unha serie</b> de <b>estraños</b> moito <b>profesionais</b> .	0.18
$\hat{\mathcal{S}}_{H \rightarrow L}^m$	En vez diso, somos obrigados a esperar nas <b>mans dunha serie</b> de <b>estraños moi profesionais</b> .	<b>0.54</b>
$\mathcal{T}_{LE}$	But instead, you are forced there to wait in the hands of a series of very professional strangers.	

Table 3: A POR-GLG pivoting example with corresponding pivot BLEU scores. Edits by word substitution or M-UMT are highlighted.

UMT’s scores are 2 to 10 BLEU points worse than the M-UMT ones. This means that UMT across related languages significantly benefits from initial-izing with our simple word substitution method.

Second, as illustrated in Figure 2, the pivot BLEU score and the translation BLEU are im- perfectly correlated; even though M-UMT reaches the highest pivot BLEU, the resulting translation BLEU is comparable to using the simple word sub- stitution method (rows 7 and 8 in Table 2). The reason is that the quality of  $\{\hat{\mathcal{S}}_{H \rightarrow L}^m, \mathcal{T}_{HE}\}$  is natu- rally restricted by the  $\{\hat{\mathcal{S}}_{H \rightarrow L}^w, \mathcal{T}_{HE}\}$ , whose quality is in turn restricted by the induced dictionary. How- ever, by combining the augmented datasets from these two methods, we consistently improve the translation performance over using only word sub- stitution augmentation (compare Table 2 rows 7 and 9). This suggests that the two augmented sets improve LRL-ENG translation in an orthogonal way.

Additionally, we observe that augmentation from back-translated HRL data leads to generally worse results than augmentation from original HRL data (compare rows 7,8 with rows 10,11 in Table 2). We believe this to be the result of noise in the back-translated HRL, which is then compounded by further errors from the induced dictionary. There- fore, we suggest that the simple word substitution method should be preferred for the second pivoting step when augmenting back-translated HRL data.

Table 3 provides an example conversion of an HRL sentence to pseudo-LRL with the word sub- stitution strategy, and its translation with M-UMT. From  $\mathcal{S}_{HE}$  to  $\hat{\mathcal{S}}_{H \rightarrow L}^w$ , the word substitution strategy achieves very high unigram scores (0.50 in this case), largely narrowing the gap between two lan- guages. The M-UMT model then edits the pseudo- LRL sentence to convert all its words to LRL.

**Rare Word Coverage** Next, we quantitatively evaluate how our pivoting augmentation methods increase rare word coverage and the correlation

	AZE (TUR)	BEL (RUS)	GLG (POR)	SLK (CES)
WT-Bi	35K	42K	34K	51K
WT-Uni	211K	179K	89K	117K
WN-Bi	1.6M	2.5M	3.1M	2.0M
WN-Uni	2.9M	3.8M	3.8M	2.9M
BLEU-Bi	14.33	21.55	31.72	29.09
BLEU-Uni	14.10	21.86	30.51	28.58

Table 4: Injected word type (WT), injected word num- ber (WN) and BLEU score (BLEU) on low-resource translation with different induced dictionaries. **Bi** denotes **bidirectional** and **Uni** denotes **unidirectional** word induction.

with LRL-ENG translation quality. For each word in the tested set, we define a word as “rare” if it is in the training set’s lowest 10<sup>th</sup> frequency per- centile. This is particularly true for LRL test set words when using concatenated HRL-LRL training data, as the LRL data will be smaller. We further define rare words to be “addressed” if after adding augmented data the rare word is not in the lowest 10<sup>th</sup> frequency percentile anymore. Then, we de- fine the “address rate” of a test dataset as the ratio of the number of addressed words to the number of rare words. The address rate of each method, along with the corresponding translation BLEU score is shown in Figure 3. As indicated by the Pearson cor- relation coefficients, these two metrics are highly correlated, indicating that our augmentation meth- ods significantly mitigate problems caused by rare words, improving MT quality as a result.

**Dictionary Induction** We conduct experiments to compare two methods of dictionary induction from the mapped word embedding spaces: 1) Uni- directional: For each HRL word, we collect its clos- est LRL word to be added to the dictionary; 2) Bidi- rectional: We only add word pairs the two words



of which are each other’s closest neighbor to the dictionary.

In order to know how many LRL words are injected into the HRL corpus, we show the number of injected unique word types, number of injected words, and the corresponding BLEU score of models trained with bidirectional and unidirectional word induction in Table 4. It can be seen that the ratio of word numbers is higher than that of word types between bidirectional and unidirectional word induction, indicating that the injected words using the bidirectional method are of relatively high frequency. The BLEU scores show that bidirectional word induction performs better than unidirectional induction in most cases (except BEL). One explanation could be that adding each word’s closest neighbor as a pair into the dictionary introduces additional noise that might harm the low-resource translation to some extent.

## 6 Related Work

Our work is related to multilingual and unsupervised translation, bilingual dictionary induction, as well as approaches for triangulation (pivoting).

In a low-resource MT scenario, multilingual training that aims at sharing parameters by leveraging parallel datasets of multiple languages is a common practice. Some works target learning a universal representation for all languages either by leveraging semantic sharing between mapped word embeddings (Gu et al., 2018) or by using character n-gram embeddings (Wang et al., 2019) optimizing subword sharing. More related with data augmentation, Nishimura et al. (2018) fill in missing data with a multi-source setting to boost multilingual translation.

Unsupervised machine translation enables training NMT models without parallel data (Artetxe et al., 2018; Lample et al., 2018a,c). Recently, multiple methods have been proposed to further improve the framework. By incorporating a statistical MT system as posterior regularization, Ren et al. (2019) achieved state-of-the-art for en-fr and en-de MT. Besides MT, the framework has also been applied to other unsupervised tasks like non-parallel style transfer (Subramanian et al., 2019; Zhang et al., 2018).

Bilingual dictionaries learned in both supervised and unsupervised ways have been used in low-resource settings for tasks such as named entity recognition (Xie et al., 2018) or information re-

trieval (Litschko et al., 2018). Hassan et al. (2017) synthesized data with word embeddings for spoken dialect translation, with a process that requires a LRL-ENG as well as a HRL-LRL dictionary, while our work only uses a HRL-LRL dictionary.

Bridging source and target languages through a pivot language was originally proposed for phrase-based MT (De Gispert and Marino, 2006; Cohn and Lapata, 2007). It was later adapted for Neural MT (Levinboim and Chiang, 2015), and Cheng et al. (2017) proposed joint training for pivot-based NMT. Chen et al. (2017) proposed to use an existing pivot-target NMT model to guide the training of source-target model. Lakew et al. (2018) proposed an iterative procedure to realize zero-shot translation by pivoting on a third language.

## 7 Conclusion

We propose a generalized data augmentation framework for low-resource translation, making best use of all available resources. We propose an effective two-step pivoting augmentation method to convert HRL parallel data to LRL. In future work, we will explore methods for controlling the induced dictionary quality to improve word substitution as well as M-UMT. We will also attempt to create an end-to-end framework by jointly training M-UMT pivoting system and low-resource translation system in an iterative fashion in order to leverage more versions of augmented data.

## Acknowledgements

The authors thank Junjie Hu and Xinyi Wang for discussions on the paper. This material is based upon work supported in part by the Defense Advanced Research Projects Agency Information Innovation Office (I2O) Low Resource Languages for Emergent Incidents (LORELEI) program under Contract No. HR0011-15-C0114 and the National Science Foundation under grant 1761548. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

## References

- Roei 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*.
- Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *International Conference on Learning Representations*.
- Victor Chahuneau, Eva Schlinger, Noah A Smith, and Chris Dyer. 2013. Translating into morphologically rich languages with synthetic phrases. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1677–1687.
- Yun Chen, Yang Liu, Yong Cheng, and Victor OK Li. 2017. A teacher-student framework for zero-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1925–1935.
- Yong Cheng, Qian Yang, Yang Liu, Maosong Sun, and Wei Xu. 2017. Joint training for pivot-based neural machine translation. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 3974–3980.
- Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 385–391.
- Trevor Cohn and Mirella Lapata. 2007. Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 728–735.
- Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156.
- Adrià De Gispert and Jose B Marino. 2006. Catalan-english statistical machine translation without parallel corpus: bridging through spanish. In *Proc. of 5th International Conference on Language Resources and Evaluation (LREC)*, pages 65–68. Citeseer.
- 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, Volume 1 (Long Papers)*, volume 1, pages 344–354.
- Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc’Aurelio Ranzato. 2019. Two new evaluation datasets for low-resource machine translation: Nepali-English and Sinhala-English. *arXiv preprint arXiv:1902.01382*.
- Hany Hassan, Mostafa Elaraby, and Ahmed Tawfik. 2017. Synthetic data for neural machine translation of spoken-dialects. *arXiv preprint arXiv:1707.00079*.
- 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.
- Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of 55th Annual Meeting of the Association for Computational Linguistics, System Demonstrations*, pages 67–72.
- Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39.
- Surafel M Lakew, Quintino F Lotito, Matteo Negri, Marco Turchi, and Marcello Federico. 2018. Improving zero-shot translation of low-resource languages. *arXiv preprint arXiv:1811.01389*.
- Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.
- Guillaume Lample, Alexis Conneau, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018b. Word translation without parallel data. In *International Conference on Learning Representations*.
- Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, et al. 2018c. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049.
- Tomer Levinboim and David Chiang. 2015. Supervised phrase table triangulation with neural word embeddings for low-resource languages. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1079–1083, Lisbon, Portugal. 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 *The 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, Florence, Italy.
- Robert Litschko, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. 2018. Unsupervised cross-lingual information retrieval using monolingual data only. *arXiv preprint arXiv:1805.00879*.
- Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.
- Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Brussels, Belgium.
- Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proc. IJCNLP*, volume 2, pages 296–301.
- Yuta Nishimura, Katsuhito Sudoh, Graham Neubig, and Satoshi Nakamura. 2018. Multi-source neural machine translation with data augmentation. In *Proceedings of the First Workshop on Neural Machine Translation*.
- 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)*, volume 2, pages 529–535.
- Shuo Ren, Zhirui Zhang, Shujie Liu, Ming Zhou, and Shuai Ma. 2019. Unsupervised neural machine translation with SMT as posterior regularization. *arXiv preprint arXiv:1901.04112*.
- Peter H Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 86–96.
- Sandeep Subramanian, Guillaume Lample, Eric Michael Smith, Ludovic Denoyer, Marc’Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text style. In *International Conference on Learning Representations*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.
- Xinyi Wang, Hieu Pham, Philip Arthur, and Graham Neubig. 2019. Multilingual neural machine translation with soft decoupled encoding. In *International Conference on Learning Representations*.
- Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A Smith, and Jaime Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 369–379.
- Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, Denver, Colorado. Association for Computational Linguistics.
- Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Earth mover’s distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945.
- Zhirui Zhang, Shuo Ren, Shujie Liu, Jianyong Wang, Peng Chen, Mu Li, Ming Zhou, and Enhong Chen. 2018. Style transfer as unsupervised machine translation. *arXiv preprint arXiv:1808.07894*.
- 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.