Data Augmentation Methods for Low-Resource Machine Translation

Abstract

Over the years, multiple data augmentation methods have been proposed for low-resource machine translation, with most of them focusing on leveraging a larger monolingual corpus, either on the source side or the target side. This project focuses on the scenario where there are very few (6k) parallel examples for the low-resource language pair (Azerbaijani and English), but there exists a larger parallel corpus (182k) from a related language pair (Turkish and English) from the same domain. At the same time, the project also assumes limited monolingual data for the languages under study and explores different availabilities of such monolingual data from a mismatched domain (no monolingual data; monolingual data available on either the source side or the target side; monolingual data available for only the low-resource language or only the high-resource language). Under these assumptions, three different data augmentation methods (back-translation, self-training, and BART pre-training) were compared against each other. The results show that when some target-side monolingual data (300k) is available, back-translation is the most effective method, reaching a 14.55 BLEU score for the Azerbaijani to English direction and an 8.67 BLEU score for the English to Azerbaijani direction. Unfortunately, combining different approaches fails to obtain a higher BLEU score, indicating that these methods are not necessarily orthogonal. Future work is needed to understand the effect of combining different data augmentation methods with limited monolingual data.

1 Introduction and Related Work

Neural machine translation (NMT) powered by an encoder-decoder architecture (Bahdanau et al., 2015), combined with the recent development of the Transformer (Vaswani et al., 2017) architecture, has achieved state-of-the-art performance on multiple high-resource languages such as English-French (Liu et al., 2020a) and English-German (Takase and Kiyono, 2021). However, when there aren't enough parallel training examples, baseline NMT systems do not perform well and sometimes perform even worse than traditional phrase-based statistical machine translation (Koehn and Knowles, 2017; Sennrich and Zhang, 2019). Therefore, there is a general need for in-depth study on boosting the translation quality for low-resource language pairs.

Recent developments of unsupervised machine translation (Artetxe et al., 2018b,a; Lample et al., 2018a,b; Artetxe et al., 2019) that leverages only non-parallel source and target monolingual data would seem to be an interesting direction for low-resource machine translation. These unsupervised NMT frameworks usually only work best on high-resource monolingual corpora such as English-French/English-German, where they split the corresponding parallel corpora with millions of examples into non-parallel source side and target side data. When applied to low-resource scenarios, the BLEU scores on non-parallel monolingual data from dissimilar/diverse domains suffer (Marchisio et al., 2020; Kim et al., 2020). Marchisio et al. (2020) found that the stochasticity during word embedding training across dissimilar/diverse domains affects downstream translation results. Kim et al. (2020) reached similar conclusions about the dramatic failure of unsupervised NMT in low-resource scenarios, mismatched domains, and dissimilar language pairs. They also found that the inclusion of 20k parallel translation pairs is enough to beat the performance of completely unsupervised NMT. Therefore, instead of focusing on a truly unsupervised setting, the project instead assumes some availability of parallel data for the low-resource language but keeps such availability to a bare minimum (6k parallel examples only).

One approach that helps with low-resource machine translation is to train the system with one or more other languages. This involves training a single model from multiple source languages into multiple target languages. Aharoni et al. (2019) found that a many-to-many shared multilingual model performs the best in low-resource settings but does not uniformly out-perform the many-to-one/one-to-many models in the high-resource settings. This shows that when the model capacity is fixed, there could be a trade-off between the number of languages and translation performance. Wang and Neubig (2019) experimented on the same low-resource corpus used by Aharoni et al. (2019), and found that without their proposed target conditioned sampling, which are some similarity measures based on language model probabilities or vocabulary overlaps, a multilingual many-to-one model that utilizes all language pairs does not outperform the model that utilizes only the related higher-resource language pair. Guzmán et al. (2019) also reported that additional parallel data in English-Hindi, a related language pair for English-Nepali, further improve Nepali's translation quality for both the supervised and the unsupervised settings. Therefore, this project focuses on developing low-resource translation models that leverage the parallel examples from only one higher-resource language pair.

While the parallel examples from a higher resource language significantly improve the translation quality for the low-resource pair, the number of parallel examples overall can still be somewhat limited. For example, in this project, the parallel corpus (190k) from a related language pair (Turkish and English) is still below 200k, and the model may still suffer from data scarcity issues. Fortunately, usually, there exists some additional monolingual data from either the source side or the target side. Back-translation operates in a semi-supervised setup that assumes additional monolingual data in the target language is available. It first trains an initial model in the reverse direction, translating the target monolingual data into the source language. With synthetic parallel examples added to the genuine parallel examples, a final system is trained that translates from the source to the target language. Edunov et al. (2018) performed an extensive study on different settings (synthetic data generation methods, low resource vs. high resource setup, size and domain of monolingual data, e.t.c). Relevant findings to this project include the fact that sampling and beam+noise generation methods are too detrimental for low-resource settings, while pure beam search provides better results. On the other hand, self-training starts from a base model trained with parallel data for the designated translation direction and applies the trained model to obtain predictions for monolingual instances on the source side. It then trains a new model from scratch using only the pseudo parallel data generated by the current model, followed by finetuning the pre-trained model using real parallel data. He et al. (2020) explained that the secrets behind self-training for machine translation include the application of beam search, dropout during model training, and further input perturbation of the self-generated synthetic data, or noisy ST. They report that noisy ST can improve the performance over baseline for both the synthetic WMT100k English-German set and the real low-resource FloRes (Guzmán et al., 2019) English-Nepali pair under domain mismatch. Recently, mBART (Liu et al., 2020b), a multilingual sequence-to-sequence denoising auto-encoder, has been proposed. As in BART (Lewis et al., 2020), two types of noise (removing spans of text and replacing them with a mask token and sentence permuting) were applied to the monolingual inputs. They reported that initializing with the pre-trained mBART weights using 25 languages shows gains on all the low and medium resource pair compared with randomly initialized baselines. The mBART models trained on only the source and target languages can also achieve over 20 BLEU with only 10K training examples.

This project applies back-translation, self-training, and mBART for the many-to-one/one-to-many multilingual NMT models under the extremely low-resource scenario, with only one related higher resource language pair available. In addition to the limits on parallel examples (6k and 182k), all monolingual data from each language are relatively limited (300k) and come from a relatively mismatched domain than the parallel training and test examples. The project studies different availabilities of monolingual data as well, including no availability at all (for studying back-translation only), limited on the source side or the target side (for comparing back-translation with self-training and mBART), as well as mixing in only the low-resource or the high-resource synthetic examples (for simulating a lack of low-resource monolin-

gual data as well as different ablation studies on back-translation and self-training). Some simple combinations of different methods are also studied in this project to check for orthogonality between different data augmentation methods, and the inclusion of 600k more monolingual data from each language compares mBART and back-translation under further different data availability scenarios.

2 Experiments and Results

2.1 Experimental setup

The project takes the Azerbaijani (AZE) /English (ENG) pair from the multilingual TED corpus (Qi et al., 2018) as the low-resource language pair, which contains less than 6k parallel examples. Qi et al. (2018) showed that their best standard 1-layer encoder-decoder model with attention could only achieve a 2.1 BLEU score for the AZE to ENG direction. As another set of baselines, the project additionally exploits the Turkish (TUR) /English (ENG) pair in the TED corpus as the related higher-resource language pair, which consists of around 182k examples, as previous work on the same corpus (Qi et al., 2018; Wang and Neubig, 2019) usually found multilingual training with a related high resource pair lend itself to be immensely useful.

For each of the experiments, all texts from each language are encoded individually with a Byte Pair Encoding (BPE) model with a vocabulary size of 8000, trained on the training set using *sentencepiece* (Sennrich et al., 2016; Kudo and Richardson, 2018). All models are trained in *fairseq* (Ott et al., 2019) and use the same Transformer architecture, with 6 encoder layers and 6 decoder layers, a model dimension of 512, an FFN dimension of 1024, and 4 attention heads. For all the experiments, all the tokens from the source and target side languages are combined to form a shared dictionary, and a shared embedding layer is used for the encoder and the decoder. Unless otherwise specified, all models are trained with the same set of hyper-parameters, with a learning rate of 2e-4, a dropout rate of 0.3, 4000 warmup steps, and a label-smoothing rate of 0.1. The baseline AZE-ENG models are trained for 80 epochs, while all other multilingual models that utilize parallel TUR-ENG examples are trained for 40 epochs. The best checkpoint is recorded on the development set. We evaluate all models on the parallel AZE-ENG test set with 903 examples under the BLEU metric (Papineni et al., 2002) with standard *SacreBLEU* scripts (Post, 2018).

2.2 Objective 1: Establishing baselines

As the first step, Table 1 shows the baselines for both the AZE to ENG direction and the ENG to AZE direction, with or without additional parallel examples from the higher resource language pair (TUR-ENG). From the table, one could see that with only 6k parallel examples constructed from noisy TED talk transcripts, the standard NMT systems only achieve mediocre BLEU scores in both directions. As reported multiple times in other work, leveraging the parallel examples from a high-resource related language substantially improves the performance of the low-resource NMT systems for both directions. The ENG→AZE direction still suffers more from data scarcity issues than the AZE→ENG direction, and previous work (Johnson et al., 2017; Aharoni et al., 2019) has noticed the same issue.

AZE→ENG	ENG→AZE
2.67	1.64
AZE→ENG (Aug with TUR)	ENG→AZE (Aug with TUR)
12.33	5.88

Table 1: Baseline BLEU scores

2.3 Objective 2: Back-translation on the training set only

Under this setting, one assumes no availability for any additional monolingual data. This means applying data-augmentation methods to the original parallel examples from the low-resource and the higher resource language. Here, we only study the effect of back-translation on the training set. For augmenting the ENG→AZE direction, the original training examples in AZE and TUR are back-translated into English, while for augmenting the AZE→ENG direction, the original training examples in ENG are back-translated into both AZE and TUR. The main concern of this method is that the ENG→AZE model used for generating synthetic data is still weak at this point. Surprisingly, even under this less-than-ideal scenario, the results in Table 2 show that the BLEU scores on both translation directions improve by a significant margin compared to the HRL+LRL baseline. Furthermore, this set of experiments does not apply either iterative or on-the-fly back-translation (which in theory could be more beneficial); rather, after synthetic parallel data is collected, a new model is re-trained

on the combined pseudo-parallel corpus from random initialization. The results show that even when limited to the given parallel examples from a low-resource pair and a higher resource pair, knowledge distillation from back-translation still helps.

AZE→ENG (Aug with TUR)	ENG→AZE (Aug with TUR)
12.33	5.88
AZE→ENG (Aug with TUR +BT on training set)	ENG→AZE (Aug with TUR +BT on training set)
13.44	7.06

Table 2: BLEU scores with back-translation on the training set

2.4 Objective 3: Back-translation with target-side monolingual data from a mismatched domain

In this set of experiments, the project collects monolingual instances for all three languages by randomly sampling from their corresponding Wikipedia distribution. However, unlike most previous work, which assumes millions of available monolingual instances for each language, all results in this section only assume that 300k unlabeled examples exist for each language. Furthermore, as the multilingual TED corpus is constructed from talk transcripts, a moderate domain mismatch exists between the real parallel examples and the monolingual data leveraged for data augmentation. Furthermore, as the project operates in an LRL+HRL scenario, it explores different scenarios and settings regarding the availability of target-side monolingual data and different ways of constructing a pseudo-parallel corpus. For the back-translation experiments in this section, in addition to the two default settings where:

1) For augmenting the ENG→AZE direction, both monolingual AZE and TUR instances are back-translated into English;

2) For augmenting the AZE→ENG direction, monolingual ENG instances are back-translated into both AZE and TUR;

four additional scenarios are studied to gain a better picture of back-translation in the LRL+HRL setting:

3) For augmenting the ENG→AZE direction, only monolingual TUR instances are back-translated into English;

4) For augmenting the ENG→AZE direction, only monolingual AZE instances are back-translated into English;

5) For augmenting the AZE→ENG direction, monolingual ENG instances are back-translated into only TUR;

6) For augmenting the AZE→ENG direction, monolingual ENG instances are back-translated into only AZE.

The two models from 2.3 are used to generate synthetic back-translated examples, as they are the best models available up to this point.

AZE→ENG (Aug with TUR)	ENG→AZE (Aug with TUR)
12.33	5.88
AZE→ENG (Aug with TUR +BT on training set)	ENG→AZE (Aug with TUR +BT on training set)
13.44	7.06
AZE→ENG (Aug with TUR +BT on training set +BT with ENG→AZE)	ENG→AZE (Aug with TUR +BT on training set +BT with AZE→ENG)
14.03	8.04
AZE→ENG (Aug with TUR +BT on training set +BT with ENG→TUR)	ENG→AZE (Aug with TUR +BT on training set +BT with TUR→ENG)
10.73	6.03
AZE→ENG (Aug with TUR +BT on training set +BT with ENG→TUR +BT with ENG→AZE)	ENG→AZE (Aug with TUR +BT on training set +BT with TUR→ENG +BT with AZE→ENG)
14.55	8.67

Table 3: BLEU scores for different settings under back-translation

Setting 3 assumes a realistic scenario with no availability of monolingual low-resource instances. All other settings are for exploring the best way to build the pseudo-parallel corpus for back-translation. To ensure the results are comparable to the results in 2.3, the original back-translated examples on the training set from 2.3 are reused. Table 3 displays the results for all six settings, and the previous baselines are kept for comparison. The observations are as follows:

1) For the ENG→AZE direction, back-translated examples using monolingual data from the related high resource language alone do not help improve the BLEU score at all. The scores even fall behind the baseline that only applies back-translation on

the training set. When monolingual data from the low-resource language is available and utilized for constructing the pseudo-parallel corpus, the model achieves a significant improvement over the baseline. However, combining pseudo-parallel examples from both the target low-resource language and the high-resource related language achieves the highest BLEU score.

2) For the AZE→ENG direction, the results are strikingly similar, even though the same set of monolingual ENG is used across all relevant settings: when monolingual ENG is only back-translated to TUR and added to the synthetic corpus, the BLEU score falls behind the baseline that applies back-translation on the training set. When monolingual ENG is only back-translated to AZE, it significantly outperforms the baseline. Adding back the synthetic TUR-ENG subset obtained earlier further improves the performance.

The observations above indicate that for augmenting the ENG→AZE model, even though the two languages of study (AZE and TUR) are highly related, simply utilizing the monolingual data from the higher resource language does not help improve low-resource translation performance. Synthetic parallel corpora for both translation directions need to include the target low-resource synthetic examples to be useful. Only then could synthetic examples from the high resource language further boost the model performance.

2.5 Objective 4: Self-training with target-side monolingual data from a mismatched domain

For studying self-training, the project uses the same 300k unlabeled examples from Wikipedia for each language. Unlike back-translation that uses target-side monolingual data and a model trained under the reverse direction, source-side monolingual data is used, and synthetic parallel examples are generated using the model for the current translation direction. As the project operates in an LRL+HRL scenario, it also explores different scenarios and settings regarding the availability of source-side monolingual data and different ways of constructing a pseudo-parallel corpus for self-training. Therefore, like the experiments in 2.4, this subsection explores six different scenarios:
1) For augmenting the ENG→AZE direction, monolingual ENG instances are translated into both AZE and TUR;
1) For augmenting the AZE→ENG direction, both monolingual AZE and TUR instances are translated into English;
3) For augmenting the ENG→AZE direction, monolingual ENG instances are translated into only TUR;
4) For augmenting the ENG→AZE direction, monolingual ENG instances are translated into only AZE;
5) For augmenting the AZE→ENG direction, only monolingual TUR instances are back-translated into English;
6) For augmenting the AZE→ENG direction, only monolingual AZE instances are back-translated into English.
The two models from 2.3 are again used to generate synthetic examples for self-training to compare the results with those obtained using back-translation in 2.4. To make sure the results are comparable to the results in 2.3, the original back-translated examples on the training set from 2.3 are reused, and a new model is trained from random initialization on the combined pseudo-parallel corpus. Unlike He et al. (2020), no input perturbation is applied to the self-translated monolingual examples (unlike the original paper) as it is less effective. However, beam search and a large dropout rate during re-training still skew the self-translation distribution away from the true model distribution, allowing self-distillation. Table 4 displays the results for all six settings, along with baselines copied from previous sections.

Setting 6 assumes a realistic scenario with no availability of monolingual low-resource instances. All other settings are for exploring the best way to build the pseudo-parallel corpus for self-training. The observations are as follows:

1) For the AZE→ENG direction, self-training on monolingual data from the related high resource language alone does not help improve the BLEU score at all. Self-training on monolingual data from the low-resource language alone still does not beat the baseline established with only back-translation on the training set. However, the final model achieves a moderate improvement over the results in 2.3 when pseudo-examples from the high-resource related language are added back to the synthetic parallel corpus.

2) For the ENG→AZE direction, the results are very similar: when monolingual ENG is only translated to TUR and added to the synthetic corpus,

AZE→ENG (Aug with TUR)	ENG→AZE (Aug with TUR)
12.33	5.88
AZE→ENG (Aug with TUR +BT on training set)	ENG→AZE (Aug with TUR +BT on training set)
13.44	7.06
AZE→ENG (Aug with TUR +BT on training set +ST with AZE→ENG)	ENG→AZE (Aug with TUR +BT on training set +ST with ENG→AZE)
13.30	7.31
AZE→ENG (Aug with TUR +BT on training set +ST with TUR→ENG)	ENG→AZE (Aug with TUR +BT on training set +ST with ENG→TUR)
10.51	5.67
AZE→ENG (Aug with TUR +BT on training set +ST with TUR→ENG +ST with AZE→ENG)	ENG→AZE (Aug with TUR +BT on training set +ST with ENG→TUR +ST with ENG→AZE)
13.76	7.73

Table 4: BLEU scores for different settings under self-training

the BLEU score falls behind the baseline in 2.2. When monolingual ENG is only translated to AZE, it offers some moderate improvement over the results in 2.3, and adding back the synthetic subset from monolingual ENG to TUR further improves the performance.

The observations above indicate that one needs to combine the synthetic corpus from both the low-resource and the related high resource language pair for self-training to be useful. For both directions, using synthetic parallel examples from only the related high-resource pairs hurts the performance greatly. Overall, self-training does not lend itself to be as useful as back-translation in 2.4 for both translation directions. This is surprising for the AZE→ENG direction, as back-translation should have generated synthetic examples of lower quality than self-training. This indicates that the synthetic data distribution needs to be different enough from the current model distribution for more successful data augmentation.

2.6 Objective 5: Combining self-training with back-translation

To study whether the improvement brought by self-training and back-translation are orthogonal to each other, this experiment uses the two models obtained in Settings 1) and 2) of 3 to translate source-side monolingual data and append the newly-generated synthetic corpora to the original synthetic corpora used in those two experiments. The two models are then again trained from scratch. Table5 shows the results for two translation directions.

AZE→ENG (Aug with TUR +BT on training set +BT with ENG→TUR +BT with ENG→AZE +ST with TUR→ENG +ST with AZE→ENG)	ENG→AZE (Aug with TUR +BT on training set +BT with TUR→ENG + BT with AZE→ENG +ST with ENG→TUR +ST with ENG→AZE)
14.12	8.35

Table 5: BLEU scores for back-translation + self-training

Unfortunately, compared to the best back-translation model in 2.4, adding more self-training augmentation with source-side monolingual data does not further improve the translation quality for both directions, indicating that while both data augmentation methods can improve with only limited additional monolingual data significantly, their benefits are not necessarily orthogonal.

2.7 Objective 6: mBART pre-training with monolingual data from all three languages

This set of experiments explores applying the mBART objective to the LRL+HRL setting, where the encoder-decoder architecture is pre-trained on the 300k monolingual instances from AZE, TUR, and ENG under the mBART denoising objective. Following Liu et al. (2020b), 35% of the words are masked by random sampling a span length according to a Poisson distribution (λ= 3.5). Additionally, the order of sentences is permuted within each example (if multiple sentences exist in the example). A language id symbol ¡LID¿ is also used as the initial token to predict the sentence. The final NMT models for both directions are fine-tuned from the pre-trained model checkpoint, first using only the original LRL+HRL parallel corpus and then using the original corpus plus the same back-translated monolingual examples used in Settings 1) and 2) in 2.4. Table 6 displays the results. From the results, it seems that the effect of mBART pre-training with limited monolingual data is not significant: while starting from a pre-trained checkpoint offers a large improvement over starting from random initialization when only the original LRL+HRL parallel cor-

AZE→ENG (Random init +Aug with TUR)	ENG→AZE (Random init +Aug with TUR)
12.33	5.88
AZE→ENG (mBART init +Aug with TUR)	ENG→AZE (mBART init +Aug with TUR)
13.86	6.82
AZE→ENG (Random init +Aug with TUR +BT on training set +BT with ENG→TUR +BT with ENG→AZE)	ENG→AZE (Random init +Aug with TUR +BT on training set +BT with TUR→ENG +BT with AZE→ENG)
14.55	8.67
AZE→ENG (mBART init +Aug with TUR +BT on training set +BT with ENG→TUR +BT with ENG→AZE)	ENG→AZE (mBART init +Aug with TUR +BT on training set +BT with TUR→ENG +BT with AZE→ENG)
13.77	8.67

Table 6: BLEU scores for starting from a pre-trained mBART checkpoint

pus is used, the benefit disappears completely when back-translated monolingual examples are added to the training set. Surprisingly, for the AZE→ENG direction, when starting from an mBART checkpoint, the BLEU score with back-translated examples added is even lower than the case without. For the ENG→AZE direction, with back-translated examples added, starting from an mBART checkpoint achieves exactly the same BLEU score as starting from random initialization. Unlike what was shown in Liu et al. (2020b), the mBART objective is again not necessarily orthogonal with back-translation, at least in this ad-hoc HRL+LRL case when monolingual data is also somewhat limited and reused for back-translation. This is, however, not surprising. One reason is that pre-training methods usually tend to be very data-hungry. Even some of the low-resource languages in Liu et al. (2020b) still contain several gigabytes of monolingual data. Another reason is that while both multilingual denoising and translation learn a source side language model and a target side language model, they are still very different tasks. Thus, when monolingual data is somewhat limited, it is best to use it for creating a synthetic parallel corpus rather than using it for denoising pre-training.

The comparison of mBART + back-translation is repeated with 900k monolingual data. While mBART again helps under the case of no back-translation (test BLEU scores of 14.23 for AZE→ENG and 7.20 for ENG→AZE), when back-translation is applied, mBART again offers no additional gain(test BLEU scores of 15.83 for AZE→ENG and 9.23 for ENG→AZE, with or without mBART).

2.8 Objective 7: Comparing BLEU score with BERT-SCORE

Recently, BERT-SCORE (Zhang et al., 2020) was proposed as an automatic evaluation metric for text generation. Given a reference sentence $x = \langle x_1, \ldots, x_k \rangle$ and a hypothesis $\hat{x} = \langle \hat{x}_1, \ldots, \hat{x}_l \rangle$, the tokens are represented using contextualized embeddings, and pairwise cosine similarity was computed for matching. The score is then optionally weighted with inverse document frequency scores. In more details, the precision, recall, and F-1 BERT-SCORE can be calculated as

$$R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j \tag{1}$$

$$P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^\top \hat{\mathbf{x}}_j \tag{2}$$

$$F_{\text{BERT}} = 2 \frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}} \tag{3}$$

while idf re-weighting works as follows (for recall; the precision measure follows the same modification)

$$R_{\text{BERT}} = \frac{\sum_{x_i \in x} idf\,(x_i) \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j}{\sum_{x_i \in x} idf\,(x_i)} \tag{4}$$

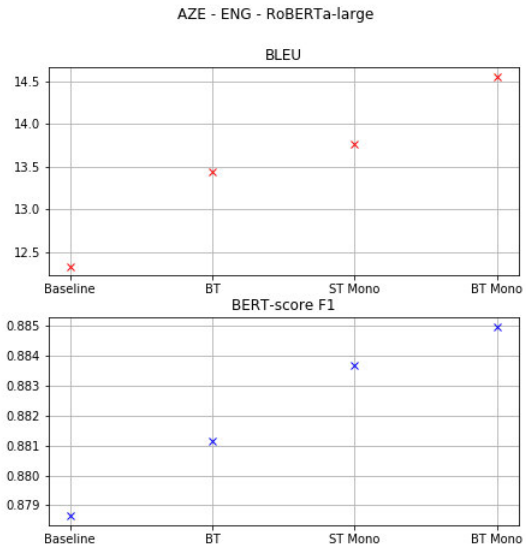

Figure 1: BERT-SCORE vs BLEU score for AZE→ENG

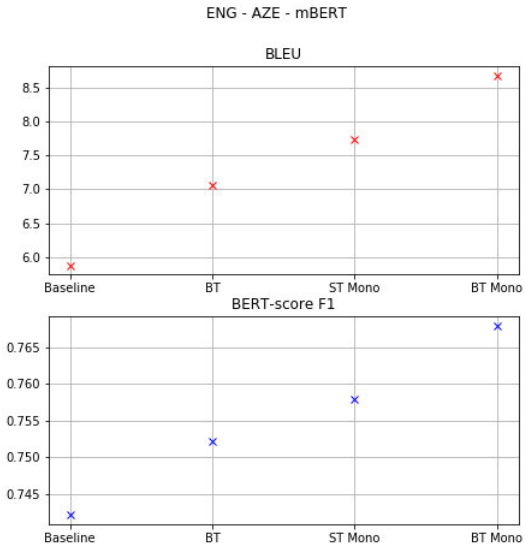

Figure 2: BERT-SCORE vs BLEU score for ENG→AZE

Here the idf-reweighted F-1 score is calculated by treating each test example in the parallel test set as a document. The calculation is only carried out on the baseline LRL+HRL models in 2.2 (named Baseline in the figures), the two models from 2.3 that back-translates the training set only (named BT in the figures), and the four models in Settings 1) and 2) from 2.4 and 2.5 (named BT Mono and ST Mono in the figures). The AZE→ENG BERT-SCOREs are calculated using a RoBERTa-large model (Liu et al., 2019), while the ENG→AZE BERT-SCOREs are calculated using the multilingual BERT (Devlin et al., 2019). Figures 1 and 2 indicates that the BLEU scores' improvements correlate well with improvements in contextualized token representations.

3 Conclusion

The project explores back-translation, self-training, and mBART denoising pre-training for a low-resource language pair, with additional parallel examples available for another related higher-resource language to train a many-to-one and a one-to-many multilingual NMT model. The project experiments with a wide array of monolingual data availability scenarios and synthetic corpus construction settings but kept the total amount of monolingual data to a bare minimum. The results show that even without additional monolingual data, back-translation on the HRL+LRL training set alone offers a considerable improvement over the HRL+LRL baseline and with additional monolingual data from a different domain at the target side available, back-translation gives an even greater boost. If instead source-side monolingual data is utilized, using the current model to perform self-training still offers some, albeit more moderate, improvements. Both back-translation and self-training, however, fail if only additional monolingual data is available for the higher resource language, or if additional monolingual English is translated into the higher resource language only, and works the best when both the low-resource and the higher resource language pairs are included in the synthetic pseudo parallel corpus. Unfortunately, a simple combination of synthetic corpora generated by back-translation and self-training offers no real benefit than the best back-translation model. While multilingual denoising pre-training on the monolingual data offers a better starting point, it also shows no benefit when back-translation is applied, which indicates that such pre-training does not help when there isn't an abundant amount of monolingual data. These results also indicate that all methods are ultimately limited by the amount of data available in current experimental settings. This calls for future work to explore the best set of strategies further to apply when monolingual data is not as abundant as assumed in previous work, which could be the case for some of the dying languages in the world.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, Brussels, Belgium. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018b. Unsupervised neural machine translation. In *Proceedings of the Sixth International Conference on Learning Representations*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. 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 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2020. Revisiting self-training for neural sequence generation. *ArXiv*, abs/1909.13788.

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

Yunsu Kim, Miguel Graça, and Hermann Ney. 2020. When and why is unsupervised neural machine translation useless? In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 35–44, Lisboa, Portugal. European Association for Machine Translation.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

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

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. *ArXiv*, abs/1711.00043.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Xiaodong Liu, Kevin Duh, Liyuan Liu, and Jianfeng Gao. 2020a. Very deep transformers for neural machine translation. *ArXiv*, abs/2008.07772.

Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Yinhan Liu, Jiatao Gu, Naman Goyal, X. Li, Sergey Edunov, Marjan Ghazvininejad, M. Lewis, and Luke

Zettlemoyer. 2020b. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Kelly Marchisio, Kevin Duh, and Philipp Koehn. 2020. When does unsupervised machine translation work? In *Proceedings of the Fifth Conference on Machine Translation*, pages 571–583, Online. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

S. Takase and Shun Kiyono. 2021. Lessons on parameter sharing across layers in transformers. *ArXiv*, abs/2104.06022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Xinyi Wang and Graham Neubig. 2019. Target conditioned sampling: Optimizing data selection for multilingual neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5823–5828, Florence, Italy. Association for Computational Linguistics.

Tianyi Zhang, V. Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.