Measuring the Impact of Data Augmentation Methods for Extremely Low-Resource NMT

Zeyneb Kaya [1], **Annie K. Lamar**[2]
[1] Saratoga High School, Saratoga, CA
zeynebnahidekaya@gmail.com
[2] Department of Classics and Stanford Data Science
Stanford University, Stanford, CA
kalamar@stanford.edu

Abstract

Data augmentation (DA) is a popular strategy to boost performance on neural machine translation tasks. The impact of data augmentation in low-resource environments, particularly for diverse and scarce languages, is understudied. In this paper, we introduce a simple yet novel metric to measure the impact of several different data augmentation strategies. This metric, which we call Data Augmentation Advantage (DAA), quantifies how many true data pairs a synthetic data pair is worth in a particular experimental context. We demonstrate the utility of this metric by training models for several linguistically-varied datasets using the data augmentation methods of back-translation, SwitchOut, and sentence concatenation. In lower-resource tasks, DAA is an especially valuable metric for comparing DA performance as it provides a more effective way to quantify gains when BLEU scores are especially small and results across diverse languages are more divergent and difficult to assess.

1 Introduction

Neural Machine Translation (NMT) has been established as the dominant approach for developing state-of-the-art Machine Translation (MT) systems. The neural network-based architecture enables effective translation without expert linguistic knowledge while better capturing contextual information. However, many NMT systems are data-inefficient and are dependent on large amounts of parallel data pairs in order to attain reliable performance, limiting their applicability in low-resource tasks. This paper is particularly interested in applicability of DA methods in the preservation of low-resource and scarce languages. There is therefore a significant performance gap in NMT for low-resource language pairs (Zoph et al., 2016).

One way that this gap is addressed is the generation of synthetic data through unsupervised Data Augmentation (DA). DA has been largely used in other deep learning modalities like image- and tabular-based data (Yang et al., 2022; Shorten et al., 2021). Multilingual text DA, in particular, has been the frontier of DA research. Sennrich et al. (2016a) proposed the backtranslation of sentences from monolingual data to generate bitext for a pseudo-parallel corpora. Recently, many more new DA approaches have been presented in order to improve NMT systems.

As opposed to approaches for low-resource NMT exploiting auxiliary languages through transfer learning, which rely heavily on the availability of data on a rich-resourced and linguistically similar language, DA in particular has potential to expand language technologies further by addressing that many low-resource and indigenous languages tend to be the most specialized and idiosyncratic, and are often part of smaller language families that are endangered as a whole (Sennrich et al. (2016a)). DA thus is especially relevant in the preservation of low-resource and scarce languages.

However, DA methods often do not exhibit consistent improvement across translation tasks (Li et al., 2019). In the case of low-resource languages, the effectiveness of DA may be even more irregular. Synthetic pairs based on very limited amounts of data may have compromised quality,

Proceedings of the The Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023), pages 101–109
May 6, 2023 ©2023 Association for Computational Linguistics

and the generalizability of these methods for scarce and orthographically diverse languages is understudied.

In this paper, we propose a method to measure the impact of DA on machine translation tasks in a low-resource environment. We then use this metric to assess the performance of three DA methods–back translation, switch-out, and sentence concatenation–on a machine translation task. We first measure the impact of DA on variously-sized subsets of high-resource language datasets, including English-Italian, English-Turkish, and German-English, to assess the generalizability and consistency of the selected DA methods. We then demonstrate how DA methods can be employed and measured in truly scarce linguistic environments by measuring the impact of DA on a machine translation task for the language pairs English-Romany, English-Māori, English-Uyghur, and English-Kabyle.

2 Background

We implement and investigate three multilingual DA approaches in our analyses. Each of these approaches have been shown to improve performance in high-resource environments, underscoring the importance of measuring the impact of such approaches in low-resource settings as well.

2.1 DA Methods for NMT

Back-translation: The augmentation procedure of back-translation (Sennrich et al., 2016a) uses monolingual data to generate more training data for a machine translation task. A backward intermediate model is trained on the available parallel corpora and then used to generate synthetic source-side translations from a target-side monolingual language corpus. Synthetic and true pairs are mixed together in the training data and not distinguished during model training.

Back-translation has shown promising results for neural machine translation tasks, particularly for large datasets. Sugiyama and Yoshinaga (2019) show that back translation has a significant positive impact on context-aware large-scale NMT tasks. Several iterations of previous work (Jin et al. 2022, Aji & Heafield 2020, Li & Specia 2019) show that back-translation can supplement other data augmentation techniques to improve performance in neural translation tasks. Such work emphasizes the need to better understand the impact of back-translation in low-resource environments so that such work can keep pace with work in high-resource settings.

SwitchOut: SwitchOut (Wang et al., 2018) independently replaces words in both the source and target sentences with words randomly sampled from their respective vocabularies to encourage smoothness and diversity. Wang et al. treat DA as an optimization problem and use hamming distance sampling to sample data pairs. Wang et al. find that these 'switches' in combination with their sampling strategy yield an improvement of 0.5 BLEU on multilingual datasets. They also find that the performance gain from SwitchOut is more significant than the gain from back translation. Notably, all the datasets used by Wang et al. are high-resource languages, including English, German, and Vietnamese.

SwitchOut has been used in combination with other DA methods in other low-resource investigations, namely that of Maimaiti et al. (2021). Maimaiti et al. compare their own, novel method of constrained sampling for machine translation to the results achieved by other DA methods, including SwitchOut, and conclude that their method is state-of-the-art. As above, such work emphasizes the need for a straightforward evaluation framework for foundational DA methods.

Sentence Concatenation: The sentence concatenation (Kondo et al., 2021) method is straightforward: sentence pairs are selected at random from the parallel corpora and concatenated with a separator token, <SEP>, in between. Notably, this method was developed with low-resource datasets in mind. Konda et al.'s method prioritizes performance on longer sentences, which are more common in low-resource datasets. Notably, Konda et al. find that their method is even more effective when combined with back-translation. For the purposes of our study, we do not combine the two methods.

2.2 Low-Resource and Scarce Languages

To evaluate the utility of the DA methods in low-resource language pairs across linguistically diverse languages from various regions, we perform our experiments on a range of low-resource languages from the Tatoeba Dataset (Tiedemann, 2020), which contains parallel data for translation systems of ranging sizes. We test DA methods for four language pairs, including English-Romany, English-Māori, English-Uyghur, and English-Kabyle.

Romany is a Balkan language classified as “definitely endangered” of the Indo-Aryan language family (New et al., 2017). It is spoken by small groups in various countries but is stateless and a minority, with a history of persecution and suppression. Availability of Romany resources is very small, with limited access to books and computers. Projects to support and Romany have arisen to help preserve the language and prevent its loss. The dataset contains English-Romany pairs, with 24K parallel sentences.

Māori, spoken in the indigenous population of New Zealand, is an endangered Eastern Polynesian language (Love, 1983). Māori is an analytical language and marks many grammatical categories. It became a minority language and English became increasingly powerful, and has since had several movements towards its revitalization. The dataset contains English-Māori pairs, with 221K parallel sentences.

Uyghur is the Turkic language spoken in the Xinjiang region of Western China (Imin et al., 2021). Primarily Muslim, the Uyghur people have been targeted by the Chinese on the basis of ethnic and religious identity. With ongoing crimes against the minority community, recognised as a genocide, teaching of the Uighur language has been banned in schools and the culture suppressed. The dataset contains English-Uyghur pairs, with 143K parallel sentences.

Kabyle in the Afro-Asiatic language of the Berbers (Rousan et al., 2018), the indigenous people of north Africa. The language has a history of brutal suppression, and today, most Berber varieties are endangered or extinct. It has limited official status, as French and Arabic are primarily used. The dataset contains English-Kabyle pairs, with 84K parallel sentences.

These languages are a selection of extremely low-resource languages from around the world with diverse linguistic features. For these languages, the development of effective NMT systems have potential to support both preservation and promotion. They are only a sample of the languages that could benefit from such technologies, and demonstrate the implications of DA towards advancing linguistic vitality and cultural preservation.

3 Datasets

In this paper, we use two groups of data. Both groups of data are from the Tatoeba Dataset (Tiedemann, 2020). The training data for the Tatoeba Dataset was obtained from OPUS’ parallel corpora (Tiedemann, 2012), which is made up of translated texts from the web. First, we determine the generalizability and consistency of different DA algorithms by measuring their performance in a simulated low-resource environment, that is, using small samples of high-resource languages.

Dataset	**5%**	**10%**	**20%**	**50%**	**100%**
English-Italian	50K	100K	200K	500K	1M
English-Turkish	50K	100K	200K	500K	1M
German-English	50K	100K	200K	500K	1M
		25%	**50%**	**75%**	**100%**
English-Romany		6K	12K	18K	24K
English-Māori		55K	111K	166K	221K
English-Uyghur		36K	72K	107K	143K
English-Kabyle		21K	42K	63K	84K

Table 1: Datasets and sampling sizes for simulated and true low-resource experiments.

Second, we measure the efficacy of DA algorithms for true low-resource environments using scarce language datasets.

To simulate a low-resource setting we randomly sample 1M pairs each from the English-Italian, English-Turkish, and German-English Tatoeba Dataset training data. We train multiple models by sampling the data in increments of 5%, 10%, 20%, 50%, and 100%. We sample from the unused portions of the dataset for use in the augmentation methods requiring monolingual data. We report results on the 2021 test sets. Second, we test DA methods for four truly scarce language pairs, including English-Romany (24K pairs), English-

Māori (221K pairs), English-Uyghur (143K pairs), and English-Kabyle (84K pairs) (see section 2.2 above). We train multiple models by sampling the data in increments of 25%, 50%, 75%, and 100%.

4 Data Augmentation Advantage

Across DA methods, synthetic pairs contribute different amounts of value to the training data. In some cases, synthetic pairs have the same impact as a true pair in the training pair, while in other cases, synthetic pairs seem to have no value or even negative value within the training dataset. In this section, we offer a simple yet novel metric to quantify how many true data pairs a synthetic data pair is worth. We call this metric Data Augmentation Advantage (DAA). We calculate DAA as follows.

First, we perform linear interpolation for the baseline model, where *x* is the number of training pairs and *y* is a BLEU score. Then for a point *y* we calculate the interpolant as in Equation 1 below. Note that $y_a < y < y_b$ and $x_a < x < x_b$.

$$y = y_a + (y_b - y_a)\frac{x - x_a}{x_b - x_a} \qquad (1)$$

For a specified target BLEU score *b* on the linear interpolation described above, let x_t be the number of training pairs needed to achieve *b* in the current experiment, and let x_b be the number of training pairs needed to achieve *b* in the baseline model. We can then calculate x_{adv} as in Equation 1 below.

$$x_{adv} = x_b - x_t \qquad (2)$$

Using x_{adv}, we calculate Data Augmentation Advantage (*DAA*) as in Equation 3. This process is summarized in Figure 1.

$$DAA = \frac{x_{adv}}{x_t} \qquad (3)$$

DAA represents the number of true data points that each synthetic data point is worth. For example, if DAA is 0.5, then the addition of DA is comparable to having 50% more data and a synthetic data point is worth 0.5 true data points. In the results section below, the overall DAA values of a DA method are obtained by averaging the values across the language tasks.

5 Experiments

For all the experiments, we use the OpenNMT-py toolkit (Klein et al., 2017) for the translation models. The NMT system is a 4-layer attention-based encoder-decoder model (Luong et al., 2015). This system estimates the probability distribution of a sentence in the target languages given a sentence in the source language. An encoder recurrent neural network (RNN) maps each source word to a word vector; the word vectors are then mapped to a set of hidden vectors. The decoder RNN decodes the source-side hidden vectors to predict the next word in the target languages. Note that target-side decoder used is aware of the previously generated words. We train the model with hidden dimension 1024 and batch size 64. We use a dropout probability of 0.3. We employ early stopping to train until convergence in order to control for the role of training time in performance changes. The settings used in training the models are the same for each language pair.

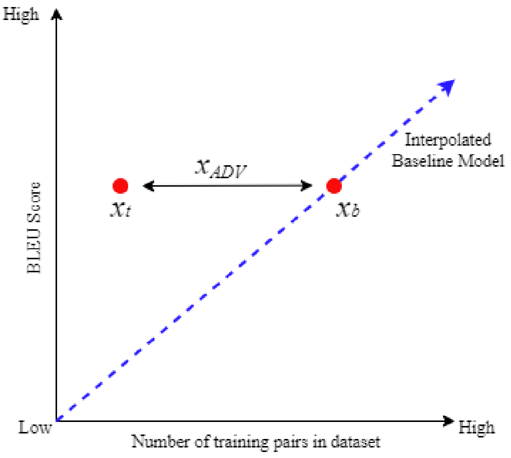
High
BLEU Score
x_{ADV}
x_t
x_b
Interpolated Baseline Model
Low
Number of training pairs in dataset
High

Figure 1

Across the experiments we process the source and target language sentences with Byte-pair encoding (BPE) (Sennrich et al., 2016b) based on the SentencePiece subword model (Kudo & Richardson, 2018) with a vocabulary size of 8K. BPE is a method for segmenting words into subword units based on their frequency of occurrence. It enables better coverage and generalization in handling rare and out-of-vocabulary words by breaking them down, and is especially relevant in languages with complex morphology such as Turkish. SentencePiece is a powerful and flexible method for unsupervised tokenization and subword segmentation, and provides an implementation of the BPE algorithm. For models with augmentation, BPE is applied after DA, and for augmentation methods with an intermediate model, BPE is applied for each. For all DA methods, we generate synthetic data with a 1:1 ratio.

In the experiments, we compare a baseline model with no augmentation to models trained with the original training data in addition to the synthetic data obtained through each DA method. We additionally ran control experiments by simply duplicating the data to verify that any results were due to DA, and observed no improvement from the baselines. We run a model for each of the subsets of data and report our final results for each. The translation quality is measured by a single reference BLEU score (Papineni et al., 2002). Three language pairs, English-Italian, English-Turkish, and German-English, are used to assess the generalizability and consistency of the methods.

6 Results & Discussion

6.1 Simulated Low-Resource Environment

DA can obtain different benefits across different sizes of available data and examine the trends and limits as data grows smaller. In this section, we examine the trends in the performance of the three DA methods across decreasing amounts of initial language pairs on multiple translation tasks. Table 2 shows the BLEU scores achieved by the various models demonstrating the performance of the various DA methods across dataset sizes and languages. Table 3 shows the calculated DAA for each DA method and dataset size.

Training Pairs	**50K**	**100K**	**200K**	**500K**	**1M**
		eng → ita			
Baseline	11.1	23.9	27.8	29.6	30.4
Ba-Trans	19.9↑	20.9	28.4	28.4	29.0
Sw-Out	12.1	24.8	27.5	27.6	28.2
Sen-Con	13.6	24.8↑	28.8↑	30.7↑	31.0↑
		eng → tur			
Baseline	5.5	9.8	14.4	15.7	16.5
Ba-Trans	7.1 ↑	13.2↑	15.6↑	16.4	17.0
Sw-Out	6.1	10.4	13.0	14.1	14.2
Sen-Con	5.6	9.8	14.5	16.7↑	17.6↑
		deu → eng			
Baseline	11.1	17.7	21.0	21.5	23.1
Ba-Trans	15.1↑	19.3↑	20.7	20.9	21.1
Sw-Out	13.0	18.1	20.4	20.8	21.6
Sen-Con	11.8	18.2	21.0	23.3↑	23.0

Table 2: BLEU scores for three datasets (English-Italian, English-Turkish, and German-English) for a baseline model, back-translation (Ba-Trans), SwitchOut (Sw-Out) and sentence concatenation (Sen-Con).

Training Pairs	**50K**	**100K**	**200K**	**500K**	**1M**
		eng → ita			
Ba-Trans	0.69	-0.12	0.50	-0.40	-0.60
Sw-Out	0.08	0.23	-0.04	-0.61	-0.75
Sen-Con	0.20	0.23	0.83	1.38	0.38
		eng → tur			
Ba-Trans	0.37	0.74	1.38	0.87	0.31
Sw-Out	0.14	0.13	-0.15	-0.61	-0.80
Sen-Con	0.02	0.0	0.12	1.25	0.69
		deu → eng			
Ba-Trans	0.60	0.48	-0.05	-0.61	-0.74
Sw-Out	0.29	0.12	-0.09	-0.61	-0.47
Sen-Con	0.11	0.15	0.0	1.12	-0.03

Table 3: DAA scores for three datasets (English-Italian, English-Turkish, and German-English) for back-translation (Ba-Trans), SwitchOut (Sw-Out) and sentence concatenation (Sen-Con).

As expected, baseline model performance increases significantly with greater data sizes, and as the number of language pairs is less, its impact on model performance is greater. Although intuitively, DA performance might be expected to decrease with less available data with the limited quality of the generated synthetic data, it is observed that the improvement from DA increases with fewer initial pairs. DA thus shows potential and value for the development of low-resource NMT systems.

There is no single consistently best DA method across the configurations. SC shows improvement in nearly all runs. However, while BT primarily shows the best improvement, its gains are not always consistent. SO follows a similar trend, harming performance in larger data sizes, but providing near the highest gains in smaller data sizes. In multiple cases, such as in both the 200K and 500K data size models, application of SC and BT attain results that can exceed or perform competitively with the results of baseline models trained on datasets with up to over twice as many pairs. Here, synthetic data provides as much value to the models as a true pair.

The most effective methods are based on introducing lexical and syntactic diversity to the datasets, presenting potentially important characteristics of effective DA methods and opening paths for future development. Overall, the results demonstrate trends that present the limits of DA and show surprising potential for its application in low-resource conditions.

6.2 True Low-Resource Environment

Many methods presenting studies on DA and low-resource NMT have often applied methods to simulated low-resource settings, like in the previous section, to enable certain analyses (Fadaee et al., 2017; Li et al., 2020). However, with the unique linguistic characteristics of truly low-resource languages as well as the varying quality of the sentences available in such data, it is also important to understand the effectiveness of the application of DA methods in a true low-resource environment. In this section, we assess the capabilities of DA methods in developing translation systems for truly low-resource languages and demonstrate the potential of such methods in advancing language technologies for supporting these communities. We evaluate the performance of the NMT systems with and without the generated augmentations on the low-resource languages. The models used follow the architecture of those described in the simulated low-resource conditions in the Experiments section.[1]

Training Pairs	**25%**	**50%**	**75%**	**100%**
eng → rom				
Baseline	0.4	0.5	0.5	0.7
Ba-Trans	0.4	0.5	0.6	1.0↑
Sw-Out	0.7↑	0.7↑	0.9↑	0.6
Sen-Con	0.6	0.5	0.5	0.6
eng → mri				
Baseline	5.5	10.2	10.8	12.0
Ba-Trans	8.5↑	6.4	10.0	12.4
Sw-Out	7.1	10.4↑	11.4	12.6↑
Sen-Con	5.3	9.0	11.5↑	12.4
eng → uig				
Baseline	0.5	0.6	0.5	0.6
Ba-Trans	0.4	0.4	0.6	0.4
Sw-Out	0.7↑	0.5	0.7↑	0.6
Sen-Con	0.6	0.6	0.6	0.7↑
eng → kab				
Baseline	1.3	1.5	1.4	1.6
Ba-Trans	1.0	1.2	1.3	1.4
Sw-Out	1.1	1.5	1.7	1.4
Sen-Con	1.1	1.3	1.8↑	1.5

Table 4: BLEU scores for four datasets (English-Romany, English-Māori, English-Uyghur, and English-Kabyle) for a baseline model, back-translation (Ba-Trans), SwitchOut (Sw-Out) and sentence concatenation (Sen-Con).

Overall, BLEU scores (Table 4) are, as expected, lower than those in the simulated low-resource conditions, even comparing datasets of similar sizes. One reason for this may be because of the lower quality of the data. The training data for the Tatoeba Dataset was obtained from OPUS' parallel corpora (Tiedemann, 2012), which is made up of translated texts from the web. The resources on these languages are limited, and therefore not only is there less data, but also the sentences are potentially less diverse and clean. Another reason is that many of these languages that are low-resource are indigenous languages, which often have more unique linguistic characteristics. Linguistic similarity between the source and target is an influential factor in the performance of NMT systems (Subramanian & Sundararaman, 2021).

These differences also create some shifts in the performance of the DA methods. As in Section 2, the best augmentation methods are also not consistent, yet here SO seems to perform better with respect to the other augmentation methods than it did in the simulated low-resource. SC also continues to show gains in the BLEU scores as well. Across the tasks, DA was able to considerably improve results, even with extremely limited available data, establishing the value DA. In fact, comparing our models' performances to the results reported on the OPUS-MT leaderboard[2] for the 2021 Tatoeba test sets shows that in two of the tasks, eng→uig and eng→mri, the top DA methods' performances surpassed the previous best OPUS-MT scores by 0.4 BLEU points each. The previous best OPUS-MT model was trained with multilingual training, and our models' comparably better performance may demonstrate that DA is more suitable in low-resource NMT to address the differences that the condition presents. This is consistent with previous findings demonstrating the limitations of utilizing auxiliary languages in low-resource NMT (Eo et al., 2021). The results

[1] Note that for back-translation, we sample from the monolingual data presented in the Tatoeba dataset.

[2] https://opus.nlpl.eu/leaderboard/

show the potential of DA to be furthered towards multilingual NLP systems and language technologies enabling inclusivity.

The advantage provided by DA is especially significant in lower-resource settings, and generated synthetic data can provide up to as much value as a true source-target pair. Comparing the DA methods, BT provides the most considerable value, and SC is the most consistently beneficial. The DAA values for BT show a greater advantage in the eng→tur task, and SC is more effective in the eng→ita, while SO has less variation across languages. These analyses of the generalizability of the methods go beyond the information presented by the BLEU scores, where the greatest net gains are not consistent with these trends.

DAA enables further insights into the performance of DA methods that BLEU scores do not capture. The absolute gains are not comparable across languages and dataset sizes. Observing the BLEU scores of the SO method in the 11K and 221K eng→mri datasets, while there are greater net gains in the larger of the two, the DAA values show that DA has a far greater impact in the smaller dataset. DAA accounts for the non-linear variation in the worth of true data with larger corpora. DAA shows the performance of BLEU gains between language tasks; for instance, the 11K eng→mri task and the 12K eng→rom task, although shows having similar dataset sizes and BLEU gains with SO, eng→rom experiences a greater impact.

Training Pairs	**25%**	**50%**	**75%**	**100%**
eng → rom				
Ba-Trans	0.0	0.0	0.17	0.38
Sw-Out	3.0	1.0	0.67	-0.13
Sen-Con	2.5	0.0	0.0	-0.13
eng → mri				
Ba-Trans	0.64	-0.40	-0.35	0.08
Sw-Out	0.34	0.17	0.17	0.13
Sen-Con	-0.04	-0.13	0.19	0.08
eng → uig				
Ba-Trans	-0.2	-0.6	0.33	-0.8
Sw-Out	1.0	-0.5	0.33	0.0
Sen-Con	1.0	0.0	0.33	1.0
eng → kab				
Ba-Trans	-0.23	-0.54	-0.67	-0.25
Sw-Out	-0.15	0.0	1.0	-0.5
Sen-Con	-0.15	-0.5	1.66	-0.5

Table 5: DAA scores for four datasets (English-Romany, English-Māori, English-Uyghur, and English-Kabyle) for a baseline model, back-translation (Ba-Trans), SwitchOut (Sw-Out) and sentence concatenation (Sen-Con).

7 Conclusion & Future Work

Data augmentation (DA) is a popular strategy to boost performance on neural machine translation tasks. The impact of data augmentation in low-resource environments, particularly for diverse and scarce languages, is understudied. In this paper, we introduce a simple yet novel metric to measure the impact of several different data augmentation strategies. This metric, which we call Data Augmentation Advantage (DAA), quantifies how many true data pairs a synthetic data pair is worth in a particular experimental context.

Because DAA provides a consistent measure comparable across the results, we are able to determine that SwitchOut and sentence concatenation show the greatest language task generalizability, providing more consistent DAA. In general, SwitchOut is especially advantageous with less available data, most evident in the increasing DAA in eng→mri and eng→rom, while back-translation has more limitations with regards to the minimum amount of data required for best performance. In particular, in lower-resource tasks, DAA is an especially valuable metric for comparing DA performance as it provides a more effective way to quantify gains when BLEU scores are especially small and results across diverse languages are more divergent and difficult to assess.

Limitations

DA demonstrates promising gains in low-resource NMT. However, in the current exploration there are certain limitations. Our experiments use the Tatoeba Dataset, which contains varying, and sometimes quite high, levels of noise; this can impact the quality of translations, the extent of overfitting, and the effectiveness of generated synthetic data. Furthermore, such data can also affect the sensitivity of evaluations to minor changes in outputs, affecting the significance of performance changes.

Additionally, DA has limitations in its effectiveness. The quality of the data generated by augmentation is inconsistent, and can degrade model performance. DA is a tradeoff between noise vs. knowledge injection (Li et al., 2019), so it is important to understand the effects that DA can have. Although we experiment with a diverse range of languages and DA methods, the study is a limited yet promising analysis of the impact of DA for low-resource NMT.

Finally, this paper does not engage in discussion regarding the value of experiments performed with true low-resource datasets vs. simulated ones. This issue is topical and requires further investigation in an expanded work.

Ethics Statement

Global linguistic diversity is currently fragile with the rapid loss of languages. The current overlap between these fading languages and emerging technologies, natural language processing tools are especially critical towards supporting diverse languages. However, globalization has only furthered English domination across the web and available language resources as NLP advancements grow in high-data tasks, and minority languages have been unrepresented in NLP literature and technologies, leaving many behind. Language technologies are a valuable aspect of supporting minority languages, yet the low-data setting has made it difficult to fully take advantage of this critical era. The application of effective multilingual DA methods in NMT systems for these languages is valuable for greater materials in accessibility, promotion, education, and connection.

References

Alham Fikri Aji and Kenneth Heafield. 2020. Fully synthetic data improves neural machine translation with knowledge distillation. CoRR, abs/2012.15455.

Eo, S., Park, C., Moon, H., Seo, J., & Lim, H. 2021. Dealing with the Paradox of Quality Estimation. *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages* (*LoResMT2021*), 1–10.

Fadaee, M., Bisazza, A., & Monz, C. 2017. Data Augmentation for Low-Resource Neural Machine Translation. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers), 567–573.

Imin, G., Ablimit, M., & Hamdulla, A. 2021. A Review of Morphological Analysis Methods on Uyghur Language. *2021 International Conference on Asian Language Processing* (*IALP*), 310–315.

Jin, C., Qiu, S., Xiao, N., & Jia, H. 2022. AdMix: A Mixed Sample Data Augmentation Method for Neural Machine Translation (arXiv:2205.04686).

Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *Proceedings of ACL 2017, System Demonstrations*, 67–72.

Kondo, S., Hotate, K., Hirasawa, T., Kaneko, M., & Komachi, M. 2021. Sentence Concatenation Approach to Data Augmentation for Neural Machine Translation. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, 143–149.

Kudo, T., & Richardson, J. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (System Demonstrations)*, 66–71.

Li, G., Liu, L., Huang, G., Zhu, C., & Zhao, T. 2019. Understanding Data Augmentation in Neural Machine Translation: Two Perspectives towards Generalization. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (*EMNLP-IJCNLP*), 5689–5695.

Li, J., Liu, L., Li, H., Li, G., Huang, G., & Shi, S. 2020. Evaluating Explanation Methods for Neural Machine Translation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 365–375.

Li, Z., & Specia, L. 2019. Improving Neural Machine Translation Robustness via Data Augmentation: Beyond Back-Translation. *Proceedings of the 5th Workshop on Noisy User-Generated Text* (*W-NUT 2019*), 328–336.

Love, P. A. (1983). The Maori language in New Zealand: A case study of language shift.

Luong, M.-T., Pham, H., & Manning, C. D. 2015. Effective Approaches to Attention-based Neural Machine Translation. Stanford NLP Group.

Maimaiti, M., Liu, Y., Luan, H., & Sun, M. 2022. Data augmentation for low-resource languages: NMT guided by constrained sampling. *International Journal of Intelligent Systems*, 37(1), 30–51.

New, W., Kyuchukov, H., & Villiers, J. de. 2017. 'We don't talk Gypsy here': Minority Language Policies in Europe. *Journal of Language and Cultural Education*, 5(2), 1–24.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. 2002. BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 311–318.

Rousan, R. A., & Ibrir, L. 2018. Language Change and Stability in Algeria: A Case Study of Mzabi and Kabyle. *Jordan Journal of Modern Languages and Literature*, 10(2), 177-198.

Sennrich, R., Haddow, B., & Birch, A. 2016a. Neural Machine Translation of Rare Words with Subword Units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1715-1725.

Sennrich, R., Haddow, B., & Birch, A. 2016b. Improving Neural Machine Translation Models with Monolingual Data. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 86–96.

Shorten, C., Khoshgoftaar, T. M., & Furht, B. 2021. Text Data Augmentation for Deep Learning. Journal of Big Data, 8(1), 101.

Subramanian, V., & Sundararaman, D. 2021. How do lexical semantics affect translation? An empirical study. arXiv preprint arXiv:2201.00075.

Sugiyama, A., & Yoshinaga, N. (2019). Data augmentation using back-translation for context-aware neural machine translation. *Proceedings of the Fourth Workshop on Discourse in Machine Translation* (*DiscoMT 2019*), 35–44.

Tiedemann, J. 2012. Parallel Data, Tools and Interfaces in OPUS. *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (*LREC'12*), 2214–2218.

Tiedemann, J. 2020. The Tatoeba Translation Challenge – Realistic Data Sets for Low Resource and Multilingual MT. *Proceedings of the Fifth Conference on Machine Translation*, 1174–1182.

Wang, X., Pham, H., Dai, Z., & Neubig, G. 2018. SwitchOut: An Efficient Data Augmentation Algorithm for Neural Machine Translation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 856–861.

Whalen, D. H., & Simons, G. F. 2012. Endangered Language Families. *Language*, 88(1), 155–173.

Yang, S., Xiao, W., Zhang, M., Guo, S., Zhao, J., & Shen, F. 2022. Image Data Augmentation for Deep Learning: A Survey (arXiv:2204.08610).

Zoph, B., Yuret, D., May, J., & Knight, K. 2016. Transfer Learning for Low-Resource Neural Machine Translation (arXiv:1604.02201).