Data Augmentation with Diversified Rephrasing for Low-Resource Neural Machine Translation

Yuan Gao y.gao1@massey.ac.nz
Feng Hou* f.hou@massey.ac.nz
Huia Jahnke h.t.jahnke@massey.ac.nz
Ruili Wang ruili.wang@massey.ac.nz
SMCS, Massey University, Auckland, 0632, New Zealand

Abstract

Data augmentation is an effective way to enhance the performance of neural machine translation models, especially for low-resource languages. Existing data augmentation methods are either at a token level or a sentence level. The data augmented using token level methods lack syntactic diversity and may alter original meanings. Sentence level methods usually generate low-quality source sentences that are not semantically paired with the original target sentences. In this paper, we propose a novel data augmentation method to generate diverse, high-quality and meaning-preserved new instances. Our method leverages high-quality translation models trained with high-resource languages to rephrase an original sentence by translating it into an intermediate language and then back to the original language. Through this process, the high-performing translation models guarantee the quality of the rephrased sentences, and the syntactic knowledge from the intermediate language can bring syntactic diversity to the rephrased sentences. Experimental results show our method can enhance the performance in various low-resource machine translation tasks. Moreover, by combining our method with other techniques that facilitate NMT, we can yield even better results.

1 Introduction

Current neural machine translation (NMT) (Ng et al., 2019; Wang et al., 2021; Wei et al., 2022; Shao and Feng, 2022) systems, especially those based on Transformer (Vaswani et al., 2017), have achieved human-level performance in translation quality (Hassan et al., 2018; Popel et al., 2020). These systems are trained using hundreds of millions of sentence pairs to ensure that they can generalize to unseen instances. However, large-scale parallel data is scarce and only available for a few high-resource language pairs (Lample et al., 2018; Haddow et al., 2022). Thus, the generalization of low-resource NMT models is far below an acceptable standard.

Recently, data augmentation (Sennrich et al., 2016a; Gao et al., 2019; Provilkov et al., 2020; Nguyen et al., 2020; Wei et al., 2022) has shown to be an effective way to improve the generalization of NMT models, especially for low-resource languages (Currey et al., 2017). Existing data augmentation methods for NMT can be categorized into token level or sentence level methods. Token level methods randomly replace words with rare words in both source and target sides to enhance the translation of rare words (Fadaee et al., 2017), or introduce

*Corresponding author

Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track, pages 35–47
September 4–8, 2023, Macau SAR, China.
©2023 The authors. This article is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

token level noises in the source side (Sennrich et al., 2016a; Lample et al., 2018; Artetxe et al., 2018; Wang et al., 2018; Gao et al., 2019; Provilkov et al., 2020) to improve the robustness of models (Khayrallah and Koehn, 2018). Sentence level methods are mainly based on back-translation (Sennrich et al., 2016b; Edunov et al., 2018), which uses target side monolingual data to synthesize pseudo-parallel data. Variants of back-translation include iterative back-translation (Hoang et al., 2018; Sánchez-Martínez et al., 2020), data diversification (Nguyen et al., 2020) and meta back-translation (Pham et al., 2021).

We argue that existing data augmentation methods for low-resource translations have two major limitations: (i) Token level methods perform token level manipulations (e.g., drop, reorder, replace) to generate new training data; thus, the generated sentences lack syntactic diversity; moreover, the token level manipulations may change the original meanings (Wei et al., 2022); (ii) Sentence level methods take natural sentences as input and generate synthetic corresponding translations using pre-trained low-quality models that are susceptible to errors (Edunov et al., 2018; Kambhatla et al., 2022), hence the augmented sentences often struggle to capture the complete semantics in the original sentences, resulting in the failure to semantically align with the target sentences. Pham et al. (2021) also noted the importance of the quality of augmented sentences.

In this paper, we propose a simple yet effective data augmentation method, **Bi**directional **T**ranslation-based **D**ata **A**ugmentation (BiTDA), to generate meaning-preserved and syntactic-diverse new training data for NMT. BiTDA uses pairs of high-quality translation models to rephrase the original sentences for low-resource translation. For example, for the Māori⇌English translation, the original English translation/sentence of a Māori sentence is first translated into an intermediate high-resource language (e.g., German or French) and then translated back into English. In this way, we obtain one more English translation for the Māori sentence. Instead of applying the translation models trained on the original low-resource data as back-translation does, we use the high-quality translation models trained with high-resource languages to generate new sentences. High-resource models generally yield higher-quality translations compared to low-resource translation models, leading to an enhancement in the quality of generated sentences. On the other hand, the knowledge of an intermediate language learned by the high-resource models can be injected into the generated sentences and resulting in syntactic diversity.

To evaluate the effectiveness of BiTDA, we conduct experiments on eight low-resource translation tasks. Experimental results show that our method significantly and consistently improves the translation performance for low-resource machine translation. We further combine our proposed method with other techniques that facilitate NMT, and the results demonstrate that BiTDA works well with the other techniques that facilitate NMT and achieves better results.

2 Methodology

2.1 BiTDA

Let $\mathcal{D} = (\mathcal{S}, \mathcal{T})$ be the original parallel training data for a low-resource translation, where $\mathcal{S}$ and $\mathcal{T}$ denotes the source and target side data, respectively; $\mathcal{M}_{\mathcal{S}\to\mathcal{I}}$ is a pre-trained translation model, which is used to translate sentences from source language $\mathcal{L}_{\mathcal{S}}$ to an intermediate high-resource language $\mathcal{L}_{\mathcal{I}}$. Given the source side data $\mathcal{S}$ from the training data and a pre-trained translation model $\mathcal{M}_{\mathcal{S}\to\mathcal{I}}$, we can obtain the translated sentences $\mathcal{I}$ in an intermediate language. This process introduces the linguistic knowledge of the intermediate language, and $\mathcal{I}$ exhibits a syntactic structure that is biased towards the intermediate language. Such diverse syntactic variants are beneficial for improving generalization.

Then, we use a reverse model $\mathcal{M}_{\mathcal{I}\to\mathcal{S}}$ to translate $\mathcal{I}$ back to the source language, the generated data is denoted as $\hat{\mathcal{S}}$. Although the generated sentences are still in language $\mathcal{L}_{\mathcal{S}}$ and

Algorithm 1 BiTDA

Inputs: Original dataset $\mathcal{D} = (\mathcal{S}, \mathcal{T})$,
Pre-trained translation models $\mathcal{M} \in \{\dots, \mathcal{M}_{\mathcal{S}\to\mathcal{I}_i}, \mathcal{M}_{\mathcal{I}_i\to\mathcal{S}}, \dots\}$

Output: A new training set $\hat{\mathcal{D}}$

procedure BiTDA($\mathcal{D} = (\mathcal{S}, \mathcal{T}), \mathcal{M}$)
  $\mathcal{D}_0 \leftarrow \mathcal{D}$
  for each $i \in 1, \dots, N$ **do**
    $\mathcal{I}_i \leftarrow$ Inference($\mathcal{M}_{\mathcal{S}\to\mathcal{I}_i}, \mathcal{S}$) ▷ Translate $\mathcal{S}$ to an intermediate language $\mathcal{L}_{\mathcal{I}i}$
    $\hat{\mathcal{S}}_i \leftarrow$ Inference($\mathcal{M}_{\mathcal{I}_i\to\mathcal{S}}, \mathcal{I}_i$) ▷ Translate $\mathcal{I}_i$ back to the source language $\mathcal{L}_{\mathcal{S}}$
    $\mathcal{D}_x \leftarrow \mathcal{D}_{x-1} \cup \left(\hat{\mathcal{S}}_i, \mathcal{T}\right)$ ▷ Merge original data and augmented data
  end for
  return $\hat{\mathcal{D}} \leftarrow \mathcal{D}_x$

largely hold the same meaning, the linguistic knowledge learned by translation models $\mathcal{M}_{\mathcal{S}\to\mathcal{I}}$ and $\mathcal{M}_{\mathcal{I}\to\mathcal{S}}$ have been injected into, and the rephrased sentences $\hat{\mathcal{S}}$ show syntactic diversity following the intermediate language. To describe our method clearly, we summarize the overall process in Algorithm 1.

As a result, we obtain multiple source sentences for one target sentence in this case. These rephrased sentences are directly paired with the corresponding target sentences from the original training data, and then we combine the synthetic data $(\hat{\mathcal{S}}, \mathcal{T})$ with the original training data as a larger training set to train our final translation model. The combined training set allows the model to learn from both the original data and the rephrased data, and the increased diversity provides the translation model with powerful generalization capabilities that can be applied to accurately translate a wider range of (unseen) sentences.

Our method can utilize multiple paired translation models with different intermediate languages to produce a more diverse set of augmented data. In practice, we only rephrase the sentences in English for low-resource translation tasks since the performance of low-resource translation models is consistently inadequate. In our research, we employ two high-resource languages, German and French, as intermediate languages to implement our method. As for the pre-trained translation models, we use the checkpoints shared by Facebook (Ng et al., 2019) instead of training them from scratch.

2.2 Relations with Existing Methods

Back Translation (BT) and **Data Diversification** BT is a widely used data augmentation method that generates new parallel data from monolingual data of the target side language using a backward translation model (i.e., target-to-source translation). Data diversification (Nguyen et al., 2020) generates a diverse set of synthetic training data from both lingual sides (in the parallel data) using multiple models trained for both forward and backward translation tasks. Similar to data diversification, our method uses the original bilingual data and multiple auxiliary translation models to generate sentence level new examples. However, data diversification is still based on back-translation and the generated source side is of low-quality (Wei et al., 2022). In contrast, we use pre-trained translation models of high-resource languages to generate high-quality sentences without requiring any monolingual data.

Knowledge Distillation Knowledge distillation is a technique that is frequently used in resource-limited scenarios (Kim and Rush, 2016; Wang et al., 2021). It uses the predictions of a pre-trained complex teacher model as soft targets to train a simple student model.

As a result, the student model is able to achieve comparable performance to the teacher model under limited resources. In our method, we use pre-trained models of high-resource languages to generate diverse training data that enhances the robustness of low-resource models. The knowledge acquired by the pre-trained models is also distilled into the augmented data.

Pivot Translation Pivot translation is particularly useful in scenarios where direct translation between the source and target languages is challenging due to limited training data. It works by incorporating a (relatively) high-resource *pivot* language to establish a bridge between the source and target languages and then translating sentences via the pivot language. Typically, the pivot language is required to be highly related to the low-resource side language and has a large amount of training data with the high-resource side language (Xia et al., 2019). Our method does not necessitate a strong relationship between the pivot language and the low-resource languages, making it more applicable to independent low-resource languages.

3 Experiments

In this section, we conduct experiments in a wide range of low-resource translation directions with different corpora sizes and languages to demonstrate the effectiveness of our method. In addition to the main experiments, we combine our method with other techniques to further improve the performance of translation models.

3.1 Datasets

To comprehensively evaluate BiTDA, we conduct experiments on both WMT and IWSLT tasks. For WMT* tasks, we conduct experiments on WMT2016 Romanian $\rightarrow$ English, Russian $\rightarrow$ English, WMT2017 Finnish $\rightarrow$ English, Latvian $\rightarrow$ English and WMT2018 Turkish $\rightarrow$ English. For IWSLT tasks†, we use IWSLT2014 Hebrew $\rightarrow$ English and IWSLT2015 Vietnamese $\rightarrow$ English. Besides, we also apply a tiny size dataset, Korean Parallel Dataset, from Google site‡. We use the officially provided training sets, development sets and test sets for all of these translation tasks.

Before performing translations, we use the standard Moses toolkit§ to preprocess all datasets and we use extra scripts from Sennrich et al. (2016a) to further process Romanian side data. To tackle unknown and rare words effectively, we use Byte Pair Encoding (BPE) (Sennrich et al., 2016c) to segment words with 4k merge operations for Vietnamese, Turkish and Korean $\rightarrow$ English. For Hebrew $\rightarrow$ English translation, we follow the set-up as Gao et al. (2019) with 10k merge operations; we also follow Sennrich et al. (2016a) which learns 89,500 merge operations for Romanian $\rightarrow$ English. As for Russian and Finnish $\rightarrow$ English, we adopt 40k merge operations. In our experiments, we build joint dictionaries for all tasks.

3.2 Training Settings

In our experiments, we adopt Transformer (Vaswani et al., 2017) as our translation model with a configuration that consists of 6 encoder and decoder layers with 4 attention heads. The dimensionalities of all sub-layers in the model are set to 512, and the inner layers of feed-forward networks have 1024 dimensions. Dropout is applied to all sub-layers, and the rate is set to 0.1. We train our models by using Adam (Kingma and Ba, 2015) as an optimizer with $(\beta_1, \beta_2) = (0.9, 0.98)$ and using cross-entropy as criterion with *label smoothing* $= 0.1$. The

*https://www.statmt.org/
†https://wit3.fbk.eu/
‡https://sites.google.com/site/koreanparalleldata
§https://github.com/moses-smt/mosesdecoder

	Vi→En	He→En	Tr→En	Ro→En
Baseline	31.64	36.52	21.86	34.08
+ WordDropout	31.62	36.67	21.92	34.16
+ Swap	31.63	36.56	21.94	34.22
+ SwitchOut	32.35	36.93	22.28	33.86
+ BPEDropout	32.73	37.66	22.95	34.83
+ BiTDA-de	32.33	37.20	22.72	**35.07**
+ BiTDA-fr	32.37	37.23	22.63	34.75
+ BiTDA-double	**32.96**	**37.72**	**23.56**	34.63
+ BiTDA-de + BPEDropout	**33.49**	**38.47**	23.40	**35.20**
+ BiTDA-de + MLS	33.19	37.38	22.90	34.38
	Ru→En	**Fi→En**	**Lv→En**	**Ko→En**
Baseline	28.69	28.01	17.20	5.26
+ WordDropout	28.15	28.12	17.32	5.46
+ Swap	28.92	28.31	17.52	5.37
+ SwitchOut	28.13	28.33	17.10	5.00
+ BPEDropout	28.94	27.55	17.61	5.86
+ BiTDA-de	**30.01**	**28.57**	**17.75**	5.63
+ BiTDA-fr	28.95	27.24	16.95	5.54
+ BiTDA-double	29.87	28.22	17.42	**5.89**
+ BiTDA-de + BPEDropout	29.98	**29.01**	**17.98**	**6.21**
+ BiTDA-de + MLS	29.64	28.74	17.82	5.02

Table 1: SacreBLEU scores on various translation tasks. The baseline denotes a Transformer model trained without any data augmentation.

initial learning rate is set to $1e^{-7}$, then gradually increases till $1e^{-4}$ within 4,000 warm-up updates. The batch size for a single GPU is set to 4k. During inference, we average the last five models before early stopping as the final model to decode where beam search is applied with the beam size 12. We calculate the BLEU (Papineni et al., 2002) score to evaluate the performance of models. Considering the discrepancy among different tokenization processes, we apply the SacreBLEU score (Post, 2018) for all experiments.

3.3 Results

The results are presented in Table 1. For our experiments, we utilize German and French as intermediate languages, and the methods employed with these languages are named BiTDA-de and BiTDA-fr, respectively. As we can see, for all translation tasks, our method consistently outperforms the baseline (Transformer without data augmentation) with up to +1.32 SacreBLEU points. In addition to using the data augmented by BiTDA-de and BiTDA-fr alone, we also combine the new training data obtained from both methods with the original data to train

Method	$\|\mathbf{D}\|$	*test2016*	*test2018*
Baseline	1×	20.53	21.86
+ BiTDA-de	2×	20.99	22.72
+ BT	11×	22.90	24.83
+ BT+ BiTDA-de	12×	23.44	25.17

Table 2: SacreBLEU scores in the Tr-En task with BT and BiTDA. $|\mathbf{D}|$ denotes the training sample size for each method

% of training data	AVG	*test2016*	*test2017*	*test2018*
0% BiTDA + 100% original	20.81	20.53	20.03	21.86
25% BiTDA + 75% original	20.58	20.42	19.73	21.58
50% BiTDA + 50% original	20.48	20.25	19.57	21.63
75% BiTDA + 25% original	20.35	20.02	19.56	21.46
100% BiTDA + 0% original	20.01	19.80	19.40	20.84

Table 3: SacreBLEU scores degradation as the proportion of synthetic data used.

translation models, named BiTDA-double. We find that the performance gains achieved by BiTDA-double are roughly equivalent to the combined performance gains achieved by BiTDA-de and BiTDA-fr when compared with the model trained only with natural text data. This shows that the improvements achieved through BiTDA-de and BiTDA-fr are largely independent of each other. Further, our finding encourages augmenting the training data with an intermediate language that has a distinctive syntactic structure from the target language.

Moreover, we compare our method with existing data augmentation methods, including WordDropout (Sennrich et al., 2016a), Swap (Lample et al., 2018), SwitchOut (Wang et al., 2018) and BPEDropout (Provilkov et al., 2020). For WordDropout and BPEDropout, we follow their (Sennrich et al., 2016a; Provilkov et al., 2020) configurations with a dropout rate of 0.1 and 0.1, respectively. We adopt a window size of 3 (Gao et al., 2019) to implement Swap. For SwitchOut, we reuse the hyperparameters in their repository[¶]. For all these methods, we merge the synthetic data with the original training set to train translation models together. Our proposed method also has demonstrated superior performance compared to the other data augmentation methods, which provides empirical evidence of the effectiveness of our method.

3.4 Analysis

Complements Existing Methods. We combine BiTDA with other methods that facilitate NMT, including BPEDropout (Provilkov et al., 2020) and MLS (Chen et al., 2022), which are data augmentation and label smoothing decoding techniques, respectively. BPEDropout works by randomly omitting some merge steps of BPE, which is able to generate diverse subword sequences and is a subword-level data augmentation method. MLS is a parameter-free label smoothing method, designed to ensure that soft probabilities are not assigned to words exclusive to the source side sentences during decoding. As shown in the bottom rows of Table 1, BiTDA-de demonstrates consistent improvements across 7 datasets when combined with each of the two methods separately. The results demonstrate the potential of synergising our

[¶]https://github.com/nsapru/SwitchOut

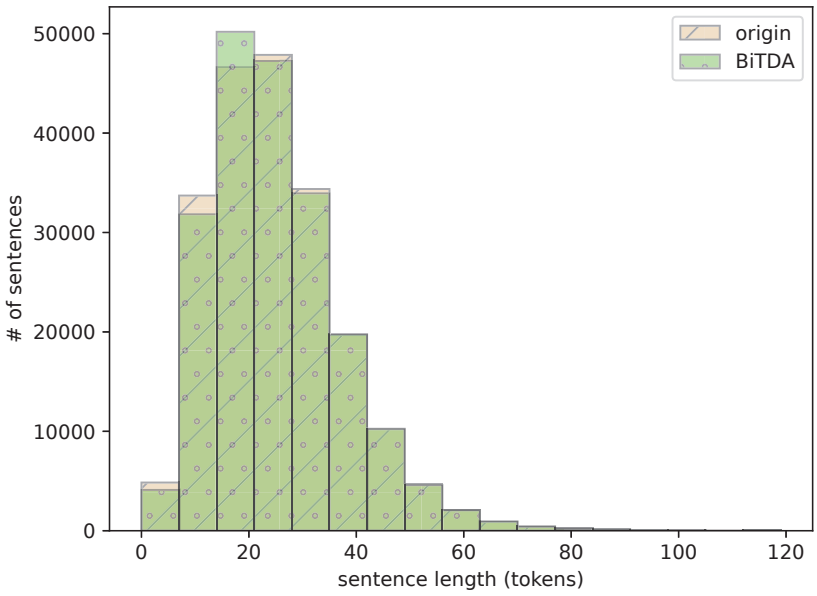


Figure 1: Distributions of sentence lengths in the English part of the original training set and the augmented training set in WMT2018 Tr-En.

	Baseline	**BiTDA-de**	**BiTDA-de**	**BiTDA-double**
Tr-orig	17.17	17.96	17.80	18.51
En-orig	25.90	27.57	27.23	28.21

Table 4: SacreBLEU scores for WMT18 Tr-En. Test sets are divided by their original source language.

method with others to further improve the performance of NMT models in partial translation directions.

Complements Back-Translation. We also combine our method with back-translation and find out the performance when they work together. To implement BT, we select WMT2018 Turkish $\rightarrow$ English (which contains 206K sentence pairs) as an example and extract 2,000,000 monolingual English sentences from News Crawl 2010. Thus, we obtain around 11 times more training examples after implementing back-translation. We conduct experiments on two test sets, *newstest2016* and *newstest2018*, both of which contain around $3,000$ sentence pairs. As shown in Table 2, BT outperforms baseline with 2.37 and 2.97 BLEU points on two sets, respectively. While BT has already achieved significant gains in performance, integrating the data generated by BiTDA results in an additional improvement of 0.34-0.54 points. The results demonstrate that BiTDA complements well with BT. It is worth noting that BiTDA does not utilize external monolingual data like BT, but rather relies solely on the original training data. Therefore, a direct comparison between BiTDA and BT based on the same amount of data was

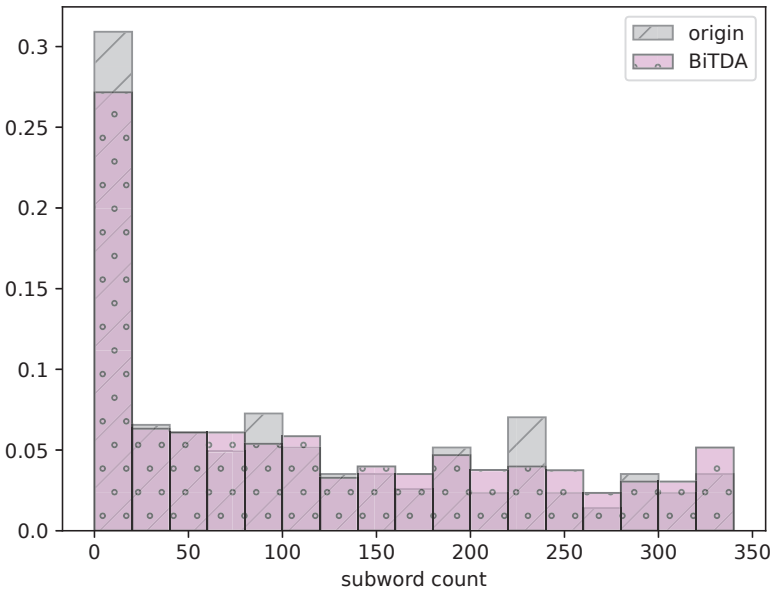
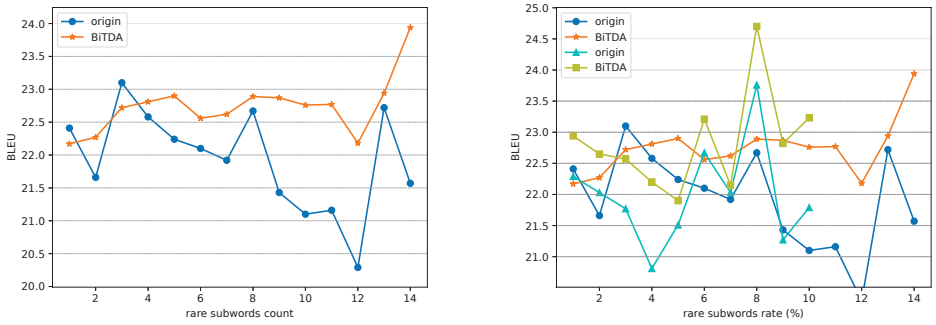


Figure 2: Distributions of top rare tokens, only 10% of the rarest words are shown. The range of numbers from 0 to 350 represents the count of subwords that appear in the whole set.

not conducted.

No Translationese Effects. Recently, Edunov et al. (2020) reveal that BT has the drawback of *translationese effect* (Gellerstam, 1986), i.e., an NMT model trained with back-translated data performs better on translated texts (simpler and shorter) than on natural texts (Marie et al., 2020). Thus, we conduct experiments to verify whether our method also suffers from this *translationese effect*. We first replace the original training data with the syntactic data in various proportions to train a translation model from scratch. We conduct experiments on the WMT2018 Turkish $\rightarrow$ English translation and present the results in Table 3. The results show that using the synthetic data as a part of training data can not directly improve the translation quality of a translation model and even does not impact the quality seriously (SacreBLEU only drops 0.8 on average when using 100% synthetic data). We then plot the distribution of sentence lengths in the English part of the original training set and the augmented training set in Figure 1. Note that the sentence lengths are counted in tokens instead of subwords from BPE encoding. As we can see, the lengths of the two sets show almost identical distributions. This finding supports the previously mentioned experimental results and underscores that our method can generate high-quality paraphrases that closely resemble natural sentences. We also follow the work of Freitag et al. (2019) in splitting each test set according to its original language. As illustrated in Table 4, BiTDA improves both the Tr-orig and En-orig test sets, further confirming our analysis.

Effect on Rare Subwords We conjecture that reducing the impact of rare subwords (encoded by BPE) is one of the reasons why BiTDA performs well. We argue that the syntactic diversity of the synthetic sentences provides a more comprehensive context for rare words, which can



(a) (b)

Figure 3: SacreBLEU score for sentences containing rare subwords. The range of numbers from 1 to 14 in (a) represents the count of rare subwords in a single sentence, and the range of numbers from 1 to 10 in (b) represents the proportion of rare subwords in a single sentence.

effectively enhance the model's ability to understand rare words. To verify this, we select the 10% rarest subwords as samples to illustrate the distribution of word frequencies. Specifically, we have selected the Turkish → English translation dataset from WMT2018 as an illustrative example. Figure 2 displays the distributions of subword frequencies in both the original set and the synthetic set by BiTDA (contains the same number of sentences as the original set). Comparing the subword distributions of the original set and the synthetic set, we observe that the synthetic set contains fewer rare subwords and increases the number of relatively common subwords. In other words, the number of partially rare subwords is increased, which enables more information to be shared between sentences. This advantage is crucial in contexts with limited resources. To provide a more intuitive demonstration of the enhanced performance of the BiTDA-augmented model, we have organized the sentences containing rare subwords and evaluated them separately. Two grouping methods have been employed in this study: grouping by the number of rare subwords in a single sentence, and grouping by the proportion of rare subwords in a single sentence. It is important to note that we have excluded results from groups with extremely small sample sizes, such as those with a proportion of rare words exceeding 10%. The results are presented in Figure 3. The model augmented by BiTDA exhibits superior performance when it comes to sentences containing rare subwords, providing further support for our conjecture.

3.5 Case Study

We present several examples generated by BiTDA in Table 5. We observe that BiTDA can reasonably adjust the syntactic structure of the original sentences, and some words are replaced with contextually appropriate alternatives. While word replacement is not the primary objective of our method, it does provide additional benefits for training NMT models.

4 Limitations

The limitations of our method are as follows: (i) It is restricted to the high-resource language side (e.g., English) of low-resource parallel data. While it is possible to use pairs of pre-trained low-resource translation models like BT can rephrase Non-English sentences, the quality of the generated sentences would be too low. (ii) It can be affected by domain shift (Deheeger et al., 2022) of the translation models we use. As seen in Table 1, using French translation models can

Original:	Ten years ago, when a local bank launched its first credit card, only one shop in Bucharest's downtown was able to accept electronic payments.
BiTDA-de:	When a local bank introduced its first credit card ten years ago, only one shop in downtown Bucharest could accept electronic payments.
BiTDA-fr:	Ten years ago, when a local bank started its first credit card, a single store in Bucharest, in the center-city was able to accept electronic payments.
Original:	Some foresee a growth of up to 500 per cent by the end of the year for transactions originating in Romania.
BiTDA-de:	Some expect up to 500 percent growth in transactions originating in Romania by the end of the year.
BiTDA-fr:	Some are forecasting a growth rate of up to 500%, at the end of the year for transactions from Romania.

Table 5: A case study on BiTDA.

be much worse than using German translation models. We conjecture that domain shift causes the sentences generated by French models to be of relatively low quality. Using high-resource translation models trained on multi-domain large-scale datasets would be better. (iii) With the same consideration as mentioned in (i), it cannot be used for the direct translation between two low-resource languages, e.g., Māori⇌Tongan.

5 Conclusion

In this work, we proposed BiTDA, a simple yet effective data augmentation method for low-resource NMT. Our method rephrases the original sentences using pairs of pre-trained high-resource translation models in opposite directions. Experiments validate the consistent effectiveness of our method across various low-resource translation tasks. Further experiments and analysis show that our method complements existing methods well.

In future work, we will explore using more pre-trained high-resource translation models and exploiting similarities (Mikolov et al., 2013) between the intermediate language and the language to be augmented.

Ethics Statement

We use public datasets and models that permit academic research. The preprocessing tools and model training toolkit are open-sourced without copyright conflicts.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments. This work is supported by the 2020 Catalyst: Strategic New Zealand - Singapore Data Science Research Programme Fund by Ministry of Business, Innovation and Employment (MBIE), New Zealand.

References

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2018). Unsupervised neural machine translation. In *International Conference on Learning Representations*.

Chen, L., Xu, R., and Chang, B. (2022). Focus on the target's vocabulary: Masked label smoothing for machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Currey, A., Miceli-Barone, A. V., and Heafield, K. (2017). Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*.

Deheeger, F., MOUGEOT, M., Vayatis, N., et al. (2022). Discrepancy-based active learning for domain adaptation. In *International Conference on Learning Representations*.

Edunov, S., Ott, M., Auli, M., and Grangier, D. (2018). Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Edunov, S., Ott, M., Ranzato, M., and Auli, M. (2020). On the evaluation of machine translation systems trained with back-translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Fadaee, M., Bisazza, A., and Monz, C. (2017). Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Freitag, M., Caswell, I., and Roy, S. (2019). Ape at scale and its implications on mt evaluation biases. In *Proceedings of the Fourth Conference on Machine Translation*.

Gao, F., Zhu, J., Wu, L., Xia, Y., Qin, T., Cheng, X., Zhou, W., and Liu, T.-Y. (2019). Soft contextual data augmentation for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Gellerstam, M. (1986). Translationese in swedish novels translated from english. *Translation studies in Scandinavia*.

Haddow, B., Bawden, R., Barone, A. V. M., Helcl, J., and Birch, A. (2022). Survey of low-resource machine translation. *Computational Linguistics*.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., et al. (2018). Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*.

Hoang, V. C. D., Koehn, P., Haffari, G., and Cohn, T. (2018). Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*.

Kambhatla, N., Born, L., and Sarkar, A. (2022). Cipherdaug: Ciphertext based data augmentation for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics)*.

Khayrallah, H. and Koehn, P. (2018). On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*.

Kim, Y. and Rush, A. M. (2016). Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations*.

Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. (2018). Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Marie, B., Rubino, R., and Fujita, A. (2020). Tagged back-translation revisited: Why does it really work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013). Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Ng, N., Yee, K., Baevski, A., Ott, M., Auli, M., and Edunov, S. (2019). Facebook fair's wmt19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation*.

Nguyen, X.-P., Joty, S., Kui, W., and Aw, A. T. (2020). Data diversification: a simple strategy for neural machine translation. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*.

Pham, H., Wang, X., Yang, Y., and Neubig, G. (2021). Meta back-translation. In *International Conference on Learning Representations*.

Popel, M., Tomkova, M., Tomek, J., Kaiser, Ł., Uszkoreit, J., Bojar, O., and Žabokrtskỳ, Z. (2020). Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature communications*.

Post, M. (2018). A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*.

Provilkov, I., Emelianenko, D., and Voita, E. (2020). Bpe-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Sánchez-Martínez, F., Sánchez-Cartagena, V. M., Pérez-Ortiz, J. A., Forcada, M. L., Espla-Gomis, M., Secker, A., Coleman, S., and Wall, J. (2020). An english-swahili parallel corpus and its use for neural machine translation in the news domain. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*.

Sennrich, R., Haddow, B., and Birch, A. (2016a). Edinburgh neural machine translation systems for wmt 16. In *Proceedings of the First Conference on Machine Translation*.

Sennrich, R., Haddow, B., and Birch, A. (2016b). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Sennrich, R., Haddow, B., and Birch, A. (2016c). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Shao, C. and Feng, Y. (2022). Overcoming catastrophic forgetting beyond continual learning: Balanced training for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*.

Wang, F., Yan, J., Meng, F., and Zhou, J. (2021). Selective knowledge distillation for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Wang, X., Pham, H., Dai, Z., and Neubig, G. (2018). Switchout: an efficient data augmentation algorithm for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Wei, X., Yu, H., Hu, Y., Weng, R., Luo, W., and Jin, R. (2022). Learning to generalize to more: Continuous semantic augmentation for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Xia, M., Kong, X., Anastasopoulos, A., and Neubig, G. (2019). Generalized data augmentation for low-resource translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.