Text Style Transfer Back-Translation

Daimeng Wei*, Zhanglin Wu*, Hengchao Shang, Zongyao Li,
Minghan Wang, Jiaxin Guo, Xiaoyu Chen, Zhengzhe Yu, Hao Yang
Huawei Translation Service Center, Beijing, China
{weidaimeng,wuzhanglin2,shanghengchao,lizongyao,wangminghan,
guojiaxin1,chenxiaoyu35,yuzhengzhe,yanghao30}@huawei.com

Abstract

Back Translation (BT) is widely used in the field of machine translation, as it has been proved effective for enhancing translation quality. However, BT mainly improves the translation of inputs that share a similar style (to be more specific, translation-like inputs), since the source side of BT data is machine-translated. For natural inputs, BT brings only slight improvements and sometimes even adverse effects. To address this issue, we propose Text Style Transfer Back Translation (TST BT), which uses a style transfer model to modify the source side of BT data. By making the style of source-side text more natural, we aim to improve the translation of natural inputs. Our experiments on various language pairs, including both high-resource and low-resource ones, demonstrate that TST BT significantly improves translation performance against popular BT benchmarks. In addition, TST BT is proved to be effective in domain adaptation so this strategy can be regarded as a general data augmentation method. Our training code and text style transfer model are open-sourced.[1]

1 Introduction

Works in neural machine translation (NMT) (Sutskever et al., 2014; Bahdanau et al., 2016; Wu and et.al, 2016; Vaswani et al., 2017) greatly improve translation quality. However, current methods generally require large amount of bilingual training data, which is a challenging and sometimes impossible task. As obtaining monolingual data is much easier, researchers have long exploited methods to enhance model performances using monolingual data, for example, language model fusion for phrase-based (Brants et al., 2007; Koehn, 2009) and neural machine translation (Gulcehre et al., 2015, 2017), back translation (Sennrich et al., 2016), and dual learning (Cheng et al., 2016; He et al., 2016; Xia et al., 2017). The combination of such monolingual methods can further improve model performances.

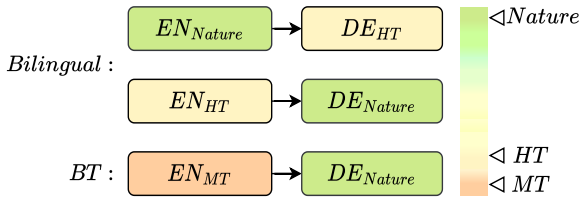

Figure 1: Bilingual and BT data used for English → German training. $Nature$ indicates data generated by native speakers; HT indicates data generated by human translators from another language, and MT indicates machine translation results. MT and HT styles are close, but far from $Nature$.

Back Translation (BT), a data augmentation method to generate synthetic parallel data by translating content from target language back to source language, is widely used in the field of machine translation. BT has many variants (Sennrich et al., 2016; Edunov et al., 2018; Caswell et al., 2019) and each has own merits.

In terms of text style, models that use BT are usually trained on three types of data. Real parallel data constitutes the first two types: natural source with human-translated target ($Nature \rightarrow HT$) or human-translated source with natural target ($HT \rightarrow Nature$). Back translation data constitutes the third type: machine-translated source with natural target ($MT \rightarrow Nature$), as shown in Figure 1.

Inspired by van der Werff et al. (2022), who find that a classifier can distinguish MT data from HT data, we train a similar classifier to classify $Nature$ and MT data and find that a high percentage of original text is marked as $Nature$ by the classifier. However, the percentage of $Nature$ content is low in human-translated data and even lower in machine-translated data. In general, hu-

*These authors contributed equally to this work.
[1] https://github.com/FrxxzHL/ssebt

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 7944–7959

Metrics	Method	Original	Reverse	All
BLEU	Bitext	**46.3**	34.9	42.2
	BT	41.8	**42.6**	**42.7**
COMET	Bitext	**58.7**	64.9	**61.8**
	BT	53.5	**69.8**	61.6

Table 1: English→German BLEU and COMET scores for models trained on WMT 2018 bitext (Bitext) and 24M BT data, measured on WMT 2018 $EN_{Nature} \rightarrow DE_{HT}$ (Original) and $EN_{HT} \rightarrow DE_{Nature}$ (Reverse) test sets.

man and machine translated data are similar, but far different from original text.

We find that when the input style is close to $Nature$, the output is biased towards HT; and when the input style is closed to HT, the output is biased towards $Nature$ (for details, see Section 6.1). Since the input used to generate BT data is $Nature$, the output is close to HT. So BT mainly improves the translation of translation-like inputs. For natural inputs, BT brings only slight improvements and sometimes even adverse effects. However, in practical use, most inputs sent to NMT models are natural language written by native speakers, rather than translation-like content.

We use one original test set ($Nature \rightarrow HT$) and one reverse test set ($HT \rightarrow Nature$) to measure BT performance respectively. As shown in Table 1, BLEU (Post, 2018) and COMET (Rei et al., 2020a) scores increase on the reserve test set but decrease on the original test set after BT.

Based on the finding, this paper aims to explore a method to enhance translation of $Nature$ input on basis of BT, while maintaining its effectiveness in translating translation-like content. Since BT connects translation-like input with $Nature$ target, we assume that if we could connect $Nature$ input with $Nature$ target, translation of $Nature$ input could be further enhanced.

Therefore, we propose Text Style Transfer Back Translation (TST BT), aiming to turn $MT \rightarrow Nature$ data into $Nature \rightarrow Nature$ data to enhance the translation of $Nature$ input. However, transferring translation-like text to a natural style is a zero-shot issue, because we can hardly obtain parallel data with the same meaning but different styles (MT and $Nature$). We propose two unsupervised methods. Our experiments on high-resource and low-resource language pairs demonstrate that TST BT can significantly enhance translation of $Nature$ input on basis of BT variants while brings no adverse effect on HT inputs. We also find that TST BT is effective in domain adaptation, demonstrating generalizability of our method.

Our contributions are as follows:

- We analyze the style of BT text and rationalize its ineffectiveness on $Nature$ input. We herein propose TST BT to solve this issue.
- TST BT combines Text Style Transfer with BT data to further improve translation of $Nature$ inputs in high and low resource, as well as in-domain and out-of-domain scenarios against various BT baselines.
- Our experiment results show that TST BT is effective in domain adaptation as well, which further improves model performance on basis of BT augmentation.

2 Related Work

2.1 Back Translation

Back Translation is first proposed by Bertoldi and Federico (2009); Bojar and Tamchyna (2011) for phrase-based systems, and then applied to neural systems by Sennrich et al. (2016).

In general, the standard BT adopts beam search for output generation, so in this paper, we denote it as Beam BT. The following are some BT variants:

- Sampling BT (Edunov et al., 2018): randomly samples translation results based on the probability of each word during decoding, thus largely increases BT data diversity.
- Noised BT (Edunov et al., 2018): adds three types of noise to the one-best hypothesis produced by beam search.
- Tagged BT(Caswell et al., 2019): adds an extra token to synthetic data to distinguish it from genuine bitext.

In our experiment, we use the above four variants as baselines. Other BT variants include Meta BT (Pham et al., 2021), a cascaded method to supervise synthetic data generation by using bitext information, aiming at generating more usable training data.

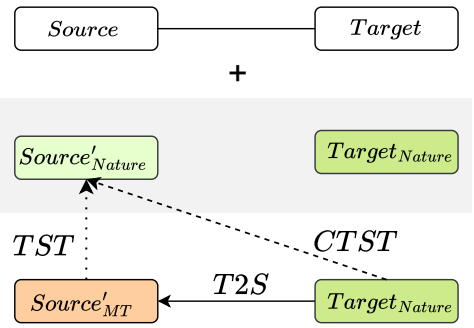

Figure 2: Direct and Cascaded methods for TST BT. Source and Target with white color means bilingual data, others mean BT data.

2.2 Unsupervised Text Style Transfer

Text Style Transfer (TST) (Fu et al., 2018; Jin et al., 2022), aiming to control attributes (e.g. politeness) of text, is an important task in the area of natural language generation. Three criteria are used to measure TST: transferred style strength, semantic preservation, and fluency.

As TST training data is difficult to obtain, unsupervised approaches (Dai and Liang, 2019; Yang et al., 2018; Krishna et al., 2020; Luo et al., 2019) are widely used. Among those, two particular approaches are closely related to machine translation and style transfer. Riley et al. (2020) propose using a classifier + tagging approach to make natural input be translated more naturally. This method is similar to the task of our paper, but it has high requirements on bilingual data size and cannot ensure a stable improvement. Freitag et al. (2019) propose training an Automatic Post-Editing (APE) model with large-scale target-side monolingual data. The APE model can also be considered as a natural style transfer.

We design our TST model by referring to the APE approach. The biggest difference between TST and APE is that APE lacks the ability to improve translation overall quality in some cases, while TST, which combines the advantages of style transfer and back translation, can achieve stable improvements on basis of standard BT.

3 Method

We propose cascaded and direct approaches (see Figure 2) to transfer the style of source-side BT data.

3.1 A Cascaded Approach

The cascaded approach generates standard BT data first and then modifies the style of the source-side BT data. However, modifying translation-like text to natural text is a zero-shot issue. To address this, we first train a Source to Target ($S2T$) model and a Target to Source ($T2S$) model. We use the reverse model ($T2S$) to generate BT data $\{Source'_{MT}, Target_{Nature}\}$. To generate TST training data, we employ Round Trip Translation (RTT) as shown in formula 1 and Figure 3(a).

$$Source' = T2S(S2T(Source_{Nature})) \quad (1)$$

We use $\{Source', Source_{Nature}\}$ to train the TST model, which uses an encoder-decoder architecture, and apply the model to the source-side BT data $Source'_{MT}$ to get $Nature \rightarrow Nature$ data, as shown in formula 2.

$$Source'_{Nature} = TST(Source') \quad (2)$$

The final training data is denoted as:

$$\{(Source, Target),$$

$$(Source'_{Nature}, Target_{Nature})\}$$

3.2 A Direct Approach

Directly translating $Nature$ data into $Nature$ outputs is also a zero-shot issue (Riley et al., 2020). In order to make $Nature$ input be translated more naturally, and avoid the data size limitations mentioned by Riley et al. (2020), we adopt a two-step training strategy, which is inspired by Zhang et al. (2021), as shown in Figure 3(b).

We first use source and target side monolingual data to generate $Source_{Nature}$ to $Target_{MT}$ and $Source_{MT}$ to $Target_{Nature}$ data respectively. We use only $Source_{Nature}$ to $Target_{MT}$ data to train the translation model and perform incremental training with $Source_{MT}$ to $Target_{Nature}$ data. During incremental training, we freeze all parameters in the encoder so the model only learns decoder parameters.

By using the two-step strategy, we aim to let the translation model learn how to produce $Nature \rightarrow Nature$ data. We consider this approach as a Conditional Text Style Transfer (CTST) method.

4 Experimental Setup

4.1 Data

Our main experiments are conducted on WMT18 EnDe, WMT17 ZhEn, and WMT16 EnRo news

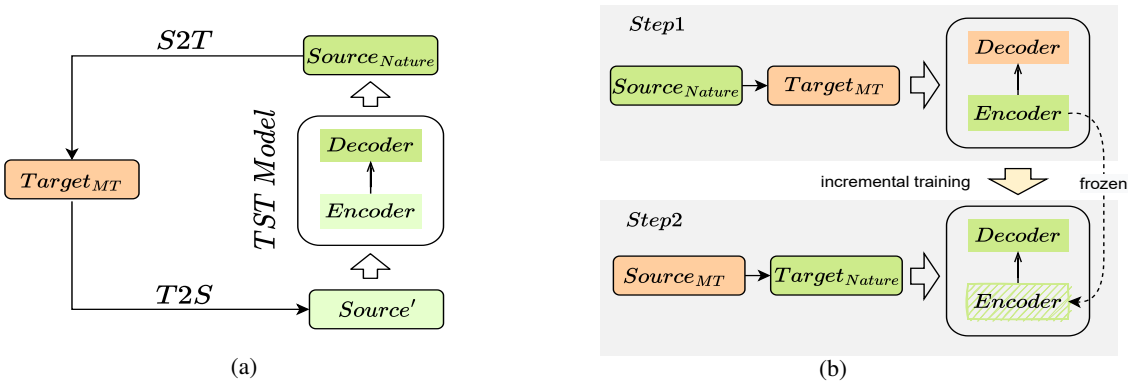

Figure 3: Left: TST Model and the process of training data generation. Right: our proposed two-step CTST training scheme.

translation data. For EnDe, we use 5.2M bilingual data except ParaCraw corpus to train the baseline model, and 226.2M German monolingual data from NewsCrawl 2007-2017 for back translation. For ZhEn, we use 19.1M bilingual data to train the baseline model, and 20.4M English monolingual data from News Crawl 2016 for back translation. For EnRo, we use 0.6M bilingual data to train the baseline model and 2.2M Romanian monolingual data from News Crawl 2015 for back translation.

Training the TST model requires source-side monolingual data. we use 24M English monolingual data from NewsCrawl 2007-2017 for EnDe and EnRo, and 24M Chinese monolingual data for ZhEn.

4.2 Evaluation

We use metrics including BLEU (Papineni et al., 2002), ChrF (Popović, 2015), COMET (Rei et al., 2020b) and BLEURT (Sellam et al., 2020) to evaluate models performances on test sets. Among them, BLEU and ChrF are calculated using SacreBLEU[2](Post, 2018), COMET using wmt20-comet-da[3], and BLEURT using BLEURT-20[4]. Based on the xlm-roberta-base[5] pre-training model, we use simpletransformers[6] to train a binary classifier to classify $Nature$ and MT text for subsequent experiments. The training data includes 10M natural monolingual data and 10M machine-translated monolingual data.

[2]https://github.com/mjpost/sacrebleu
[3]https://github.com/Unbabel/COMET
[4]https://github.com/google-research/bleurt
[5]https://huggingface.co/xlm-roberta-base
[6]https://github.com/ThilinaRajapakse/simpletransformers

BT type	Example sentence
Beam	Raise the child, love the child.
Sampling	Lift the child, love the child.
Noised	Raise child _ love child, the.
Tagged	<T> Raise the child, love the child.

Table 2: The source text of synthetic corpus for different BT methods

4.3 Architecture

We train our NMT models and TST models with Transformer (Vaswani et al., 2017) and fairseq (Ott et al., 2019), and employ FP16 to accelerate training under a joint source and target language vocabulary setting. Specifically, EnDE, ZhEn, and the TST models use the Transformer-big structure with a vocabulary size of 32K, while EnRo models use the Transformer-base structure with a vocabulary size 16K. The dropout rate for EnDe baseline model and TST model is 0.3, and 0.1 for other models. Other settings are as follows: batch size as 4096, learning rate as 7e-4, warmup steps as 4000, label-smoothing as 0.1 (Szegedy et al., 2016; Pereyra et al., 2017), Adam $\beta 1$ as 0.9, and $\beta 2$ as 0.98 (Kingma and Ba, 2017). For each training task, we select the best model according to the perplexities measured on the dev set.

5 Result

TST can be combined with popular BT strategies. Our strategy can be seen as a universal data argumentation method on basis of BT. To better verify the effectiveness of our method, Beam BT, Sampling BT, Noised BT, and Tagged BT are selected for comparative experiments (see Section 2.1).

Table 2 is an example of synthetic source sentences generated by four BT strategies. For Noised BT, noise is added after TST is performed. While

	BLEU			ChrF			COMET			BLEURT		
	All	O	R	All	O	R	All	O	R	All	O	R
Bitext	32.9	35.2	28.9	60.8	62.1	59.1	54.8	50.1	59.7	73.6	71.8	75.6
+Beam BT	32.1	28.5	36.4	59.2	55.0	65.0	45.9	28.0	65.4	71.7	66.2	77.8
+TST$_{Direct}$	33.3	31.4	34.8	60.8	58.4	64.1	53.3	42.2	65.4	73.9	70.3	77.8
+TST$_{Cascade}$	**35.3**	**33.0**	**37.7**	**62.8**	**60.6**	**65.8**	**59.3**	**51.6**	**67.6**	**75.8**	**73.1**	**78.7**
+Sampling BT	**36.0**	**32.7**	**40.2**	**63.0**	60.2	**66.9**	61.7	54.5	69.5	76.9	74.2	79.7
+TST	35.8	32.6	39.9	**63.0**	**60.3**	66.8	**62.5**	**55.9**	**69.6**	**77.2**	**74.7**	**79.8**
+Noised BT	36.6	36.2	36.4	63.6	62.6	65.0	59.8	53.4	66.9	75.7	73.0	78.5
+TST	**37.0**	**36.5**	**37.1**	**64.1**	**63.1**	**65.5**	**62.3**	**57.1**	**67.9**	**76.5**	**74.4**	**78.9**
+Tagged BT	37.0	36.6	**36.7**	63.9	63.1	**64.9**	61.6	56.0	**67.6**	76.2	73.9	78.6
+TST	**37.4**	**37.4**	36.5	**64.3**	**63.8**	**64.9**	**62.2**	**57.2**	**67.6**	**76.6**	**74.4**	**78.9**
+FT	33.6	36.4	28.9	61.5	63.1	59.3	56.3	52.4	60.4	74.1	72.5	75.8
+Beam BT	37.3	37.6	36.1	64.3	63.8	64.9	60.4	54.7	66.4	75.6	73.3	78.0
+TST	**37.8**	**37.7**	**37.2**	**64.6**	**64.0**	**65.5**	**61.3**	**55.4**	**67.6**	**76.1**	**73.8**	**78.6**

Table 3: English→German models trained on WMT 2018 bitext (Bitext) with four BT variants (Beam, Sampling, Noised and Tagged BT). Their averaged TST results on Original test set (O), Reverse test set (R) and the combined test sets (All) from WMT 2014-2018.

for other BT methods, we directly modify the source side of BT data using our TST model.

To prove the effectiveness of TST BT, We perform experiments on high-resource (EnDe and ZhEn) and low-resource (EnRo) languages, as well as domain adaptation.

5.1 TST BT for EnDe

We believe that when we add $Nature$ to $Nature$ BT data, the translation of $Nature$ input can be improved. However, the target side of original test set is human-translated, which could influences the scores measured by over-lapping metrics, such as BLEU and ChrF. For the purpose of fair evaluation, we report multiple metric scores, including BLEU, ChrF, COMET, and BLEURT. The final scores are averaged based on WMT14-18 test sets, as shown in Table 3. The detail results are shown in Appendix A.

All BT methods enhance model performance over baselines. It has greater effect on reverse test sets than original ones. Particularly, all metrics on original test set decline after Beam BT is applied. This result is consistent with our findings that merely adding BT data MT→$Nature$ deteriorates translation of $Nature$ input.

We try the two style transfer approaches mentioned above on basis of Beam BT. The result shows that both cascaded and direct approaches bring significant improvements but the cascaded approach is better. So we use the cascaded approach by default in following experiments.

In general, TST BT mainly brings improvement on original test sets while maintains standard BT's effectiveness on reverse test sets. Although BLEU and ChrF scores are fluctuated, we observe steady increase of COMET and BLEURT scores after TST BT is applied. We observe similar improvements against other BT baselines, with an average improvement of 1.0+ COMET score.

According to the experiment results, TST is a supplement to BT that further enhances the effectiveness of BT.

5.1.1 Ablation Experiment

Although TST BT does not directly use additional source text but the transfer model is trained with source data. So we perform forward translation (FT) or self-training (Zhang and Zong, 2016) with the same data and compare the FT, FT+BT (Wu et al., 2019), and FT + TST BT strategies, as shown in Table 3.

FT enhancement is considerable on the original test set but slight on the reverse test set. FT + BT brings significant improvement on the reverse and original test sets. When we perform TST BT on such a strong baseline, we observe further 0.7 and 1.2 COMET score increases on original and reverse sets respectively.

Although FT and TST use the same data, their mechanisms are different and the two methods can be used together. He et al. (2020) believe dropout is the key for FT while TST BT focuses on style transfer to create $Nature$ to $Nature$ data, which

	BLEU			ChrF			COMET			BLEURT		
	All	O	R	All	O	R	All	O	R	All	O	R
Bitext	24.7	23.8	26.1	53.4	53.0	54.2	43.5	34.2	55.0	68.0	65.9	70.6
+Beam BT	26.4	**23.7**	30.9	**55.1**	**53.5**	58.1	46.4	36.2	59.2	69.1	66.6	72.2
+TST	**26.6**	23.5	**31.8**	54.9	53.1	**58.4**	**47.8**	**37.4**	**60.5**	**69.5**	**67.0**	**72.7**

Table 4: Chinese→English models trained on WMT 2017 Bitext. The Beam BT and the averaged TST results on Original test set (O), Reverse test set (R) and the combined test set (All) from WMT 2017-2019.

	BLEU			ChrF			COMET			BLEURT		
	All	O	R	All	O	R	All	O	R	All	O	R
Bitext	28.7	28.8	28.6	56.0	54.1	57.9	52.5	28.8	76.3	71.6	64.7	78.5
+Beam BT	**32.3**	**29.0**	35.8	**59.0**	**54.8**	**63.5**	63.5	38.9	88.1	74.0	66.9	81.0
+TST$_{EnDe}$	31.7	27.8	35.6	58.6	54.1	63.3	**66.9**	**43.1**	**90.7**	**75.2**	**68.2**	**82.1**
+TST$_{EnRo}$	31.9	27.8	**36.1**	58.6	54.0	**63.5**	65.0	39.9	90.2	74.5	66.9	**82.1**

Table 5: English→Romanian models trained on WMT 2016 bitext (Bitext). Beam BT and TST results on each Original test set (O), Reverse test set (R) and the combined test set (All) from WMT 2016.

further improves the translation of $Nature$ input.

5.2 TST BT for ZhEn

The size of ZhEn bilingual data is 20M, four times that of EnDe. We perform TST on this language pair to see whether TST BT is effective when applied to a even larger data size and to a language from a different family. We use 20M English monolingual data to ensure the ratio of bilingual and monolingual data is 1:1. See overall results in Table 4 and detailed results in Appendix B.

The overall result is similar to that of EnDE. We observe significant increase of COMET and BLEURT scores after applying TST BT, although the BLEU and ChrF scores fluctuate. TST BT achieves 1.4 COMET score increase on average on basis of Beam BT. We observe significant increase on both original and reverse test sets.

Our experiments also show that TST BT achieves similar improvements against other BT baselines in addition to Beam BT on ZhEn. The result is different from the EnDe experiment, where the improvement brought by TST against Beam BT is much greater than other BT baselines. We assume that a larger bilingual data size and a different data ratio may be the reason.

It should be noted that the ZhEn baseline is already very strong considering the data size, and even stronger after adding the standard BT data. However, TST BT achieves further enhancement against such strong baselines.

5.2.1 Human Evaluation

We also perform human evaluation on ZhEn to verify the enhancement brought by TST BT. We randomly sample 300 sentences from WMT17-19 original and reverse test sets respectively. We follow the evaluation scheme mentioned by Callison-Burch et al. (2007), and 8 professional annotators are recruited to rate adequacy and fluency of three MT results on a 5-point scale, given source text and reference.

The result is listed in Table 6. TST improves adequacy and fluency on both original and reverse test sets. The result is consistent with COMET and BLEURT scores in Table 4. The human evaluation result again proves the effectiveness of our method. Both automatic metrics and human evaluations demonstrate that TST BT mainly brings enhancement on the original test set, indicating that TST BT improves the translation of $Nature$ input.

5.3 TST BT for EnRo

We further perform experiments in low-resource scenario to test the generalizability of TST BT. We use WMT16 EnRo bilingual data (0.6M bilingual) for the experiment. Table 5 presents the results.

In this experiment, we compare the effectiveness of two TST models: one is trained with EnRo models, and the other, used for our EnDe experiment, is trained with EnDe models. The style transfer model trained with EnRo data improves performance against BT baselines (by 1.5 COMET score and 0.5 BLEURT score).

Another interesting finding is that the TST model for the EnDe task also enhances the EnRo

	Original		Reverse	
	Adequacy	Fluency	Adequacy	Fluency
Bitext	3.89	4.52	4.57	4.81
+Beam BT	3.98	4.51	4.67	4.79
+TST	**4.03**	**4.52**	**4.69**	**4.82**

Table 6: Averaged Adequacy and Fluency results by human annotators on Original and Reverse test sets.

	BLEU	ChrF	COMET	BLEURT
Bitext	26.1	57.1	50.4	71.0
+Beam BT$_{Med}$	28.6	60.9	56.4	72.6
+TST$_{Med}$	**30.3**	**61.0**	**57.8**	**73.3**

Table 7: Metric scores of German→English models trained on WMT 2018 Bitext. Biomedical Beam BT (Beam BT$_{Med}$) and the biomedical TST (TST$_{Med}$) results measured on WMT 2018 biomedical test set.

model performance (by 3.4 COMET score and 1.2 BLUERT score), which is even greater than that of the TST$_{EnRo}$ model. The result indicates that it is possible to build a universal pre-trained model for sytle transfer. This result demonstrates that the style transfer model is universal and can be applied to other language pairs.

5.4 Domain Augmentation

We observe that the translation of in-domain natural inputs improve significantly after applying TST BT. We also found that TST BT still improve translation of out-of-domain natural inputs (like IWSLT14 and Flores (Goyal and Gao, 2022)) test set (for details, see Appendix Table 19).

Domain adaptation is a critical application of BT. BT can improve in-domain performance given in-domain monolingual data and an out-of-domain translation model (Edunov et al., 2018). If we train a TST model to modify the source-side text generated by BT to an in-domain style, we assume in-domain translation can be further improved.

To test our hypothesis, we train an out-of-domain DeEn model using WMT18 news bilingual data, and perform BT on 12M biomedical English monolingual data. 2.5M biomedical German monolingual data is used to train the in-domain TST model. The result is shown in Table 7.

We observe significant improvement brought by BT and more surprisingly, further significant improvement after we apply TST, with an increase of 1.4 COMET score and 0.7 BLEURT score. We believe the reason for such enhancement is the same as that on Flores and IWSLT test sets mentioned above: making the input style biased towards in-domain or $Nature$ text to augment the effectiveness of BT. The experiment again demonstrates the generalizability of TST BT.

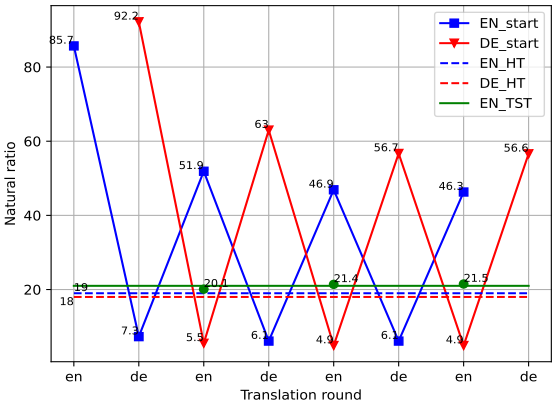

Figure 4: The $Nature$ ratio of each round of translation results starting with EN_{Nature} and DE_{Nature} (EN_start, DE_start). The dotted line indicates the $Nature$ ratio of English or German human translations (EN_HT, DE_HT). The green line represents the averaged $Nature$ ratio (EN_TST) of English data after style transfer.

6 Analysis

6.1 Style Tide

As shown in Figure 1, bilingual data can be divided into $Nature$ to HT or HT to $Nature$. By learning such data, the model inclines to generate translation-like output when the input is $Nature$, and vice versa. To illustrate the phenomenon, we perform several rounds of translation on EN_{Nature} and DE_{Nature} data from WMT18 EnDe test set. We calculate the proportion of $Nature$ text marked by the classifier after each round of translation.

As shown in Figure 4, the percentage of $Nature$ sentences fluctuates regularly after each round of translation, no matter the translation starts from De or En. For English original data, the percentage of $Nature$ data is 85.7% before translation. The percentage drops to 7.3% after the first round of translation into German, and then bounces back to 51.9 after the second round of translation back into English. As we analyzed above, the style of input determines the style of output.

In general, the composition of bilingual data, as well as the difference between $Nature$ and HT style, makes the source-side BT text significantly different from $Nature$ text. As a result, the translation of $Nature$ input can hardly be improved by

	Equal	TST better	MT better
Annotator 1	220	27	53
Annotator 2	212	23	65
Annotator 3	218	24	58

Table 8: Evaluation results by professional annotators on 300 randomly selected MT and TST sentences.

standard BT.

6.2 Style and Quality of TST

To understand what changes a TST model makes to the source-side text, we analyze the style and quality difference before and after applying TST to the source-side text.

Taking EnDe data as an example, we analyze the style of English text before and after TST, and compare the quality through human evaluation.

As shown in Figure 4, after TST, the percentage of $Nature$ text increases from 5.5 to 20.1. The improvement is significant, reaching the same level of $Nature$ as human-translated data, but there is still a certain gap with the real natural text.

In addition, to analyze the impact of TST on text quality, we randomly select 300 sentences from WMT14 test set and assess the quality of standard BT data and TST data against references. We invite three professional annotators to complete the assessment. We use relative ranking and classify the results into three categories: equal, TST better or MT better. The result is shown in Table 8, which is different from Freitag et al. (2019). APE can further improve translation quality but TST cannot.

Based on above analysis, we find that TST does not improve the overall quality of source-side BT data. Instead, it modifies the text towards a more natural style, thus overcomes the weakness of standard BT. In addition, TST BT still maintains BT's tolerance (Bogoychev and Sennrich, 2019) of data quality to make up the performance deterioration caused by TST.

6.3 Style Transfer and BT Effects

In order to analyze the relationship between style transfer results and final improvement on translation quality, we compare the improvements brought by TST BT data that is generated via two different approaches (cascaded/direct as we motioned above) on EnDe and EnRo.

We use Strength of Style Transfer (ACC) and Semantic Preservation (Seman) to measure style transfer results. Taking EnDe as an example, we perform BT on the DE_{nature} data from the reverse test set $\{EN_{HT}, DE_{nature}\}$, and calculate Seman (measured by BLEURT) against reference EN_{HT}. We then use the original test set $\{EN_{Nature}, DE_{HT}\}$ to measure the improvement of TST BT on the translation of $Nature$ input. The result shows that although the direct approach leads to higher ACC and Seman scores, the cascaded approach brings greater enhancement to the final translation performance. The results are shown in Table 9.

	TST		**TST BT**
	ACC	Seman	COMET
EnDe Beam BT	5.5	71.1	35.3
+TST$_{Direct}$	**25.2**	**70.8**	51.1
+TST$_{Cascade}$	20.1	69.9	**59.3**
EnRo Beam BT	25.4	65.6	38.9
+TST$_{EnDe}$	**55.4**	**65.1**	**43.1**
+TST$_{EnRo}$	52.2	65.7	39.9

Table 9: Style transfer performances of different TST models on WMT 2018 English→German and WMT 2016 English→Romanian translation tasks.

For EnRo, we compare style transfer models trained on EnRo and EnDe data as we stated before. Data modified by the TST$_{EnDe}$ achieves higher ACC and Seman scores, and lead to greater enhancement to the overall translation quality. The result is different from our EnDe experiment.

Therefore, the relationship between style transfer and the effect of BT enhancement can not be drawn and more researches are required.

7 Conclusion

This paper proposes Text Style Transfer Back Translation (TST BT) to further enhance BT effectiveness. We make a detailed analysis of training data styles and find that BT hardly improves translation of $Natural$ inputs, which are the main inputs in practical use. Our method simply modifies the style of source-side BT data, which brings significant improvements on translation quality, both in high-resource and low-resource language scenarios. Further experiment finds that TST BT is also effective in domain adaptation, which can further expand the application of our method. The generalizability of TST BT is thus proved.

8 Limitations

TST BT is simple and straightforward, which brings great improvements against BT baselines. However, comparing with standard BT, TST BT requires an additional style transfer model and additional time to process generated BT data.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

Nicola Bertoldi and Marcello Federico. 2009. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 182–189, Athens, Greece. Association for Computational Linguistics.

Nikolay Bogoychev and Rico Sennrich. 2019. Domain, translationese and noise in synthetic data for neural machine translation. *arXiv: Computation and Language*.

Ondřej Bojar and Aleš Tamchyna. 2011. Improving translation model by monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 330–336, Edinburgh, Scotland. Association for Computational Linguistics.

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 858–867, Prague, Czech Republic. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Shaw Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. *Workshop on Statistical Machine Translation*.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy. Association for Computational Linguistics.

Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Semi-supervised learning for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1965–1974, Berlin, Germany. Association for Computational Linguistics.

Ning Dai and Jianze et.al Liang. 2019. Style transformer: Unpaired text style transfer without disentangled latent representation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5997–6007, Florence, Italy. 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.

Markus Freitag, Isaac Caswell, and Scott Roy. 2019. APE at scale and its implications on MT evaluation biases. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 34–44, Florence, Italy. Association for Computational Linguistics.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. AAAI Press.

Naman Goyal and et.al Gao. 2022. The Flores-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, and Yoshua Bengio. 2017. On integrating a language model into neural machine translation. *Computer Speech Language*, 45:137–148.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *NIPS*, pages 820–828.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2020. Revisiting self-training for neural sequence generation. In *International Conference on Learning Representations*.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep Learning for Text Style Transfer: A Survey. *Computational Linguistics*, 48(1):155–205.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Philipp Koehn. 2009. *Statistical Machine Translation*. Cambridge University Press.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762, Online. Association for Computational Linguistics.

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Xu Sun, and Zhifang Sui. 2019. A dual reinforcement learning framework for unsupervised text style transfer. In *IJCAI*, pages 5116–5122.

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.

Gabriel Pereyra, George Tucker, Jan Chorowski, ukasz Kaiser, and Geoffrey Hinton. 2017. Regularizing neural networks by penalizing confident output distributions.

Hieu Pham, Xinyi Wang, Yiming Yang, and Graham Neubig. 2021. Meta back-translation. In *International Conference on Learning Representations*.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

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.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020a. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020b. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. 2020. Translationese as a language in "multilingual" NMT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7737–7746, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826.

Tobias van der Werff, Rik van Noord, and Antonio Toral. 2022. Automatic discrimination of human and neural machine translation: A study with multiple pre-trained models and longer context. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 161–170, Ghent, Belgium. European Association for Machine Translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Advances in neural information processing systems.

Lijun Wu, Yiren Wang, Yingce Xia, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2019. Exploiting monolingual data at scale for neural machine translation. 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 4207–4216, Hong Kong, China. Association for Computational Linguistics.

Yonghui Wu and Mike Schuster et.al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation.

Yingce Xia, Tao Qin, Wei Chen, Jiang Bian, Nenghai Yu, and Tie-Yan Liu. 2017. Dual Supervised Learning. *arXiv e-prints*, page arXiv:1707.00415.

Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. 2018. Unsupervised text style transfer using language models as discriminators. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Austin, Texas. Association for Computational Linguistics.

Meng Zhang, Liangyou Li, and Qun Liu. 2021. Two parents, one child: Dual transfer for low-resource neural machine translation. In *ACL/IJCNLP (Findings)*, pages 2726–2738.

A Experiment Details for EnDe

	2014			2015			2016			2017			2018		
	All	O	R	All	O	R	All	O	R	All	O	R	All	O	R
Bitext	28.2	28.2	28.3	30.8	32.8	26.1	34.6	37.6	29.8	28.7	31.1	25.2	42.2	46.3	34.9
+Beam BT	28.8	23.7	35.2	28.9	27.6	30.6	33.2	28.7	39.3	29.2	26.1	32.3	40.2	36.2	44.5
+Direct-TST	30.1	26.5	34.2	30.2	29.9	29.8	34.7	32.5	37.1	29.4	27.8	30.4	42.1	40.5	42.6
+Cascade-TST	**32.1**	**28.5**	**36.7**	**32.3**	**31.9**	**32.5**	**36.4**	**33.7**	**40.0**	**31.5**	**28.9**	**33.8**	**44.0**	**42.0**	**45.4**
+Sampling BT	33.7	28.8	**39.4**	**33.8**	**32.4**	35.7	**36.5**	**32.9**	**41.7**	**32.0**	**28.3**	**36.4**	**44.0**	40.9	**47.8**
+TST	**33.8**	**29.3**	38.9	33.5	32.0	**35.9**	36.4	32.6	**41.7**	31.7	28.1	36.0	43.8	**41.0**	47.1
+Noised BT	32.5	29.8	36.1	33.4	34.3	31.5	38.4	38.0	38.4	31.9	31.5	31.8	**46.6**	**47.2**	44.1
+TST	**33.0**	**30.4**	**36.4**	**34.2**	**34.9**	**32.8**	**38.8**	**38.1**	**39.4**	**32.6**	**32.0**	**32.7**	46.5	47.1	**44.3**
+Tagged BT	32.7	30.0	36.0	34.1	34.4	**32.3**	38.7	38.6	**38.7**	**32.9**	32.6	**32.3**	46.8	47.6	**44.4**
+TST	**33.0**	**30.6**	**36.1**	**34.5**	**35.6**	31.8	**39.3**	**39.8**	38.3	**32.9**	**32.7**	32.2	**47.3**	**48.5**	44.2
+FT	28.7	29.2	28.1	31.5	33.8	26.1	35.6	38.8	30.3	29.5	32.5	25.3	42.9	47.7	34.9
+Beam BT	32.3	**30.1**	35.2	33.7	**35.1**	30.5	39.8	40.4	38.5	32.7	32.6	32.0	47.9	49.6	44.2
+TST	**32.7**	30.0	**36.0**	**34.2**	**35.1**	**31.9**	**40.5**	**40.5**	**40.0**	**33.4**	**33.2**	**33.0**	**48.3**	**49.7**	**45.0**

Table 10: English→German BLEU scores on WMT 2014-2018 test sets.

	2014			2015			2016			2017			2018		
	All	O	R	All	O	R	All	O	R	All	O	R	All	O	R
Bitext	58.7	58.4	59.0	58.9	60.0	56.7	61.9	63.3	60.2	57.8	59.2	56.2	66.8	69.6	63.2
+Beam BT	57.4	51.8	64.9	56.4	54.1	61.2	60.1	54.9	66.9	57.4	53.8	61.9	64.6	60.4	70.1
+Direct-TST	59.2	55.4	64.3	58.3	57.2	60.7	61.8	58.9	65.7	58.0	56.0	60.6	66.6	64.7	69.0
+Cascade-TST	**61.4**	**58.2**	**65.8**	**60.3**	**59.3**	**62.4**	**63.7**	**60.9**	**67.5**	**60.1**	**58.1**	**62.7**	**68.4**	**66.6**	**70.8**
+Sampling BT	61.9	58.3	**66.9**	**61.0**	**59.7**	**63.7**	63.6	**60.3**	68.1	**60.3**	**57.4**	**63.9**	**68.1**	65.4	**71.8**
+TST	**62.1**	**58.8**	66.7	60.8	59.4	**63.7**	**63.7**	**60.3**	**68.4**	60.2	**57.4**	63.8	**68.1**	**65.6**	71.4
+Noised BT	62.0	59.5	65.3	61.0	60.8	61.5	64.9	63.7	66.5	60.3	59.3	61.6	69.9	69.8	70.1
+TST	**62.4**	**60.1**	**65.6**	**61.8**	**61.5**	**62.4**	**65.2**	**63.9**	**67.0**	**61.1**	**60.1**	**62.2**	**70.0**	**69.9**	**70.2**
+Tagged BT	62.0	59.8	65.0	61.2	61.0	**61.7**	65.1	64.1	**66.3**	60.9	60.4	61.6	70.2	70.4	**70.0**
+TST	**62.3**	**60.3**	**65.1**	**61.7**	**61.8**	61.5	**65.6**	**65.1**	**66.3**	**61.2**	**60.8**	**61.7**	**70.5**	**70.9**	69.9
+FT	59.3	59.3	59.4	59.5	60.9	56.6	62.7	64.2	60.5	58.6	60.2	56.5	67.6	70.9	63.3
+Beam BT	62.0	**59.9**	65.0	61.4	61.5	61.2	65.9	65.5	66.4	61.0	60.4	61.7	71.0	71.7	70.0
+TST	**62.2**	59.8	**65.4**	**61.7**	**61.7**	**61.9**	**66.4**	**65.6**	**67.5**	**61.5**	**60.9**	**62.3**	**71.3**	**71.9**	**70.4**

Table 11: English→German ChrF scores on WMT 2014-2018 test sets.

	2014			2015			2016			2017			2018		
	All	O	R	All	O	R	All	O	R	All	O	R	All	O	R
Bitext	54.4	51.2	57.6	51.4	49.8	54.4	54.5	46.2	62.7	51.9	44.7	59.1	61.8	58.7	64.9
+Beam BT	44.3	23.8	64.8	41.0	32.3	57.9	46.0	22.5	69.5	44.8	25.9	63.7	53.3	35.3	71.3
+Direct-TST	52.4	39.6	65.2	49.4	44.6	58.6	53.1	37.8	68.3	50.4	37.8	63.1	61.4	51.1	71.8
+Cascade-TST	**59.0**	**52.0**	**66.1**	**55.9**	**52.6**	**62.5**	**59.1**	**47.4**	**70.8**	**56.3**	**46.8**	**65.7**	**66.1**	**59.3**	**72.9**
+Sampling BT	61.6	55.0	**68.4**	58.7	55.8	64.4	61.5	51.0	72.0	59.0	50.0	**68.0**	67.6	60.7	**74.6**
+TST	**62.7**	**57.3**	68.0	**59.5**	**56.6**	**65.1**	**62.3**	**52.1**	**72.4**	**59.7**	**51.4**	**68.0**	**68.3**	**62.2**	74.5
+Noised BT	60.1	54.5	65.6	55.7	53.0	61.1	60.1	50.2	70.0	56.1	47.1	65.1	67.2	62.0	72.5
+TST	**62.0**	**57.6**	**66.4**	**57.8**	**55.7**	**62.1**	**62.8**	**54.2**	**71.4**	**59.7**	**53.0**	**66.3**	**69.1**	**64.8**	**73.3**
+Tagged BT	61.4	56.2	**66.7**	57.8	55.6	**62.0**	61.7	52.9	**70.6**	58.6	51.6	65.7	68.3	63.9	72.8
+TST	**61.8**	**57.4**	66.3	**58.2**	**56.2**	**62.0**	**62.4**	**54.2**	**70.6**	**59.3**	**52.6**	**66.0**	**69.2**	**65.4**	**73.1**
+FT	56.5	53.5	59.6	52.8	51.9	54.6	55.7	48.3	63.0	53.2	47.0	59.4	63.3	61.3	65.4
+Beam BT	60.6	**55.7**	65.5	56.7	54.7	60.5	60.1	51.0	69.2	57.2	49.2	65.2	67.3	63.0	71.5
+TST	**60.8**	55.4	**66.1**	**57.6**	**55.5**	**61.8**	**61.1**	**51.5**	**70.7**	**58.3**	**50.4**	**66.2**	**68.7**	**64.2**	**73.2**

Table 12: English→German COMET scores on WMT 2014-2018 test sets.

	2014			2015			2016			2017			2018		
	All	O	R	All	O	R	All	O	R	All	O	R	All	O	R
Bitext	73.5	71.9	75.2	72.7	72.2	73.7	73.5	70.5	76.5	72.7	69.8	75.6	75.8	74.5	77.2
+Beam BT	71.5	65.3	77.7	69.9	67.3	75.1	72.2	65.2	79.1	71.3	65.0	77.5	73.8	68.3	79.4
+Direct-TST	73.8	69.6	78.0	72.5	71.2	75.1	74.0	69.3	78.7	73.0	68.6	77.3	76.3	72.9	79.7
+Cascade-TST	**75.7**	**72.9**	**78.5**	**74.5**	**73.5**	**76.5**	**75.9**	**72.1**	**79.7**	**74.9**	**71.4**	**78.5**	**77.8**	**75.4**	**80.3**
+Sampling BT	77.1	74.4	**79.7**	75.7	74.7	77.6	76.8	73.1	80.5	76.0	72.7	79.3	78.7	76.1	**81.4**
+TST	**77.4**	**75.1**	**79.7**	**76.0**	**75.2**	**77.8**	**77.2**	**73.7**	**80.7**	**76.3**	**73.1**	**79.5**	**78.9**	**76.5**	81.3
+Noised BT	75.9	73.4	78.4	74.2	73.1	76.3	75.8	72.2	79.4	74.6	70.9	78.2	77.8	75.5	80.1
+TST	**76.5**	**74.3**	**78.7**	**75.0**	**74.2**	**76.6**	**76.8**	**73.6**	**80.1**	**75.7**	**72.9**	**78.6**	**78.7**	**76.8**	**80.6**
+Tagged BT	76.2	73.7	78.6	74.8	74.0	76.3	76.3	72.9	**79.7**	75.3	72.2	78.3	78.4	76.5	80.3
+TST	**76.5**	**74.3**	**78.8**	**75.2**	**74.5**	**76.6**	**76.7**	**73.7**	**79.7**	**75.6**	**72.6**	**78.6**	**78.8**	**76.9**	**80.6**
+FT	74.3	72.8	75.8	73.0	72.7	73.6	73.9	71.3	76.5	73.1	70.5	75.7	76.2	75.3	77.2
+Beam BT	75.8	73.5	78.1	74.2	73.5	75.4	75.5	72.1	79.0	74.6	71.2	77.9	77.9	76.0	79.7
+TST	**76.2**	**73.8**	**78.6**	**74.8**	**74.0**	**76.2**	**76.1**	**72.5**	**79.6**	**75.1**	**71.9**	**78.4**	**78.5**	**76.6**	**80.3**

Table 13: English→German BLEURT scores on WMT 2014-2018 test sets.

B Experiment Details for ZhEn

	2017			2018			2019			Average		
	All	O	R	All	O	R	All	O	R	All	O	R
Bitext	24.4	24.5	24.2	24.8	23.0	28.3	24.8	24.0	25.7	24.7	23.8	26.1
+Beam BT	25.7	23.2	29.1	26.5	23.3	32.8	**27.1**	**24.5**	30.8	26.4	**23.7**	30.9
+TST	**26.1**	**23.3**	**30.1**	**26.9**	**23.6**	**33.8**	26.7	23.7	**31.5**	**26.6**	23.5	**31.8**
+Sampling BT	**26.2**	**22.7**	**31.2**	26.6	22.9	**34.1**	**26.9**	**23.5**	**32.3**	**26.6**	23.0	**32.5**
+TST	**26.2**	22.6	31.1	**26.8**	**23.4**	33.9	26.8	**23.5**	32.2	**26.6**	**23.2**	32.4
+Noised BT	**26.3**	**24.4**	**28.9**	26.6	23.7	**32.5**	**27.0**	24.7	**30.6**	26.6	24.3	**30.7**
+TST	26.1	24.2	28.6	**26.9**	**24.1**	**32.5**	**27.0**	**24.8**	30.5	**26.7**	**24.4**	30.5
+Tagged BT	**26.3**	**24.3**	**29.0**	26.6	23.7	32.5	**27.1**	**24.5**	**31.0**	**26.7**	**24.2**	**30.8**
+TST	25.7	23.6	28.7	**27.0**	**24.2**	**32.7**	27.0	**24.5**	**31.0**	26.6	24.1	**30.8**

Table 14: Chinese→English BLEU scores on WMT 2017-2019 test sets.

	2017			2018			2019			Average		
	All	O	R	All	O	R	All	O	R	All	O	R
Bitext	53.4	54.0	52.5	53.2	52.0	55.7	53.5	53.1	54.3	53.4	53.0	54.2
+Beam BT	**54.7**	**53.6**	56.2	54.8	**52.6**	59.7	**55.8**	**54.3**	58.5	**55.1**	**53.5**	58.1
+TST	54.3	52.9	**56.4**	**54.9**	**52.6**	**60.0**	55.6	53.8	**58.7**	54.9	53.1	**58.4**
+Sampling BT	**54.4**	**52.4**	**57.2**	**54.5**	52.0	**60.2**	**55.0**	**52.7**	**59.0**	**54.6**	**52.4**	**58.8**
+TST	54.1	51.9	57.1	**54.5**	**52.1**	59.9	54.8	**52.7**	58.7	54.5	52.2	58.6
+Noised BT	**54.7**	**53.9**	**55.7**	54.6	52.7	**58.9**	**55.3**	**53.8**	**58.1**	**54.9**	**53.5**	**57.6**
+TST	54.4	53.6	55.5	**54.7**	**52.9**	**58.9**	55.1	53.7	57.8	54.7	53.4	57.4
+Tagged BT	**54.7**	**53.9**	**55.8**	54.5	52.5	**59.0**	**55.2**	**53.6**	**58.2**	**54.8**	**53.3**	**57.7**
+TST	54.3	53.3	55.7	**54.8**	**52.9**	**59.0**	55.1	53.4	58.1	54.7	53.2	57.6

Table 15: Chinese→English ChrF scores on WMT 2017-2019 test sets.

	2017			2018			2019			Average		
	All	O	R	All	O	R	All	O	R	All	O	R
Bitext	46.6	39.7	53.4	39.5	29.2	56.4	44.4	33.7	55.1	43.5	34.2	55.0
+Beam BT	48.9	40.0	57.8	42.4	31.0	61.3	48.0	37.6	58.4	46.4	36.2	59.2
+TST	**49.7**	**40.3**	**59.0**	**44.3**	**33.1**	**62.9**	**49.3**	**38.8**	**59.7**	**47.8**	**37.4**	**60.5**
+Sampling BT	48.6	37.0	60.1	42.4	30.4	62.3	47.8	34.7	**60.9**	46.3	34.0	61.1
+TST	**49.9**	**39.2**	**60.5**	**43.5**	**31.8**	**62.9**	**48.6**	**36.5**	60.6	**47.3**	**35.8**	**61.3**
+Noised BT	49.7	41.1	**58.4**	43.3	32.3	**61.5**	**48.8**	37.6	**60.1**	47.3	37.0	**60.0**
+TST	**50.1**	**41.9**	58.2	**43.8**	**33.3**	61.1	48.7	**38.1**	59.8	**47.5**	**37.8**	59.7
+Tagged BT	**49.7**	**41.1**	58.2	43.1	31.9	**61.7**	48.3	36.4	**60.2**	47.0	36.5	**60.0**
+TST	49.3	40.1	**58.4**	**44.1**	**33.8**	61.4	**48.8**	**37.7**	60.0	**47.4**	**37.2**	59.9

Table 16: Chinese→English COMET scores on WMT 2017-2019 test sets.

	2017			2018			2019			Average		
	All	O	R	All	O	R	All	O	R	All	O	R
Bitext	67.7	65.9	53.4	67.6	65.2	71.6	68.8	66.7	70.8	68.0	65.9	70.6
+Beam BT	68.6	66.2	57.8	68.7	65.9	73.3	70.0	67.7	72.3	69.1	66.6	72.2
+TST	**68.9**	**66.3**	**59.0**	**69.3**	**66.6**	**73.8**	**70.4**	**68.0**	**72.8**	**69.5**	**67.0**	**72.7**
+Sampling BT	68.7	65.5	60.1	68.7	65.5	73.9	70.0	67.0	73.1	69.1	66.0	73.0
+TST	**69.0**	**66.1**	**60.5**	**69.1**	**66.1**	**74.0**	**70.3**	**67.3**	**73.3**	**69.5**	**66.5**	**73.1**
+Noised BT	**68.9**	**66.5**	**58.4**	68.9	66.2	**73.3**	70.1	67.6	**72.7**	**69.3**	66.8	**72.4**
+TST	68.8	**66.5**	58.2	**69.0**	**66.4**	**73.3**	**70.2**	**67.8**	72.6	**69.3**	**66.9**	72.3
+Tagged BT	**68.9**	**66.6**	58.2	68.8	66.0	**73.5**	70.1	67.4	**72.8**	69.3	66.7	**72.5**
+TST	68.7	66.2	**58.4**	**69.1**	**66.5**	**73.5**	**70.3**	**67.8**	**72.8**	**69.4**	**66.8**	**72.5**

Table 17: Chinese→English BLEURT scores on WMT 2017-2019 test sets.

C Experiment Details for EnRo

	BLEU			ChrF			COMET			BLEURT		
	All	O	R	All	O	R	All	O	R	All	O	R
Bitext	28.7	28.8	28.6	56.0	54.1	57.9	52.5	28.8	76.3	71.6	64.7	78.5
+Beam BT	**32.3**	**29.0**	35.8	**59.0**	**54.8**	**63.5**	63.5	38.9	88.1	74.0	66.9	81.0
+TST$_{EnDe}$	31.7	27.8	35.6	58.6	54.1	63.3	**66.9**	**43.1**	**90.7**	**75.2**	**68.2**	**82.1**
+TST$_{EnRo}$	31.9	27.8	**36.1**	58.6	54.0	**63.5**	65.0	39.9	90.2	74.5	66.9	**82.1**
+Sampling BT	**32.6**	**29.3**	**35.9**	**59.0**	**54.8**	**63.5**	66.0	42.1	90.1	75.1	68.0	82.2
+TST$_{EnDe}$	31.9	28.2	35.7	58.5	54.1	63.2	**66.9**	**42.7**	**91.2**	**75.4**	**68.2**	**82.5**
+TST$_{EnRo}$	32.1	28.4	**35.9**	58.5	54.2	63.1	65.4	40.8	90.0	75.2	67.9	**82.5**
+Noised BT	32.2	30.8	**33.7**	58.8	55.9	**62.0**	66.8	45.0	88.5	75.3	68.7	81.9
+TST$_{EnDe}$	32.3	31.6	33.1	**59.2**	**56.8**	61.7	**68.7**	**47.5**	**90.0**	**76.2**	**70.0**	**82.4**
+TST$_{EnRo}$	**32.7**	**31.9**	**33.7**	59.1	56.5	61.8	67.3	45.9	88.7	75.7	69.1	**82.4**
+Tagged BT	32.5	31.7	33.2	58.9	56.5	61.5	67.7	45.8	89.6	75.7	69.1	82.2
+TST$_{EnDe}$	**33.1**	**32.6**	**33.7**	**59.3**	**57.1**	**61.7**	**70.3**	**49.8**	**90.7**	**76.8**	**70.9**	**82.8**
+TST$_{EnRo}$	32.6	32.0	33.3	59.0	56.7	61.5	68.2	46.5	89.8	76.1	69.5	82.7

Table 18: WMT 2016 English→Romanian results on the WMT 2016 test set.

D TST BT on OOD

	Flores				IWSLT-2014				Med-2021			
	BLEU	ChrF	COMET	BLEURT	BLEU	ChrF	COMET	BLEURT	BLEU	ChrF	COMET	BLEURT
Bitext	34.5	61.8	55.3	74.2	28.5	55.8	34.8	68.7	23.5	54.8	46.8	71.1
Beam BT	33.7	60.9	53.2	73.5	24.5	50.6	18.5	64.6	25.9	**57.6**	49.4	71.7
+TST	**34.9**	**62.2**	**59.6**	**75.8**	**27.9**	**55.2**	**38.0**	**69.7**	**26.3**	57.0	**49.9**	**72.2**
Tagged BT	37.2	63.6	61.1	76.0	29.7	56.7	40.4	70.0	25.5	56.1	49.1	71.9
+TST	**38.2**	**64.0**	**61.8**	**76.4**	**30.0**	**57.1**	**41.3**	**70.4**	**25.8**	**56.6**	**50.6**	**72.3**

Table 19: WMT 2018 English→German results on out-of-domain (Flores, IWSLT and Medical) original test set.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
8

☑ A2. Did you discuss any potential risks of your work?
6

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

4

☑ B1. Did you cite the creators of artifacts you used?
4

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
4

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
4

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
4

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
4

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
4

C ☑ Did you run computational experiments?

4

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No response.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

6.2

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.