Monolingual Transfer Learning via Bilingual Translators for Style-Sensitive Paraphrase Generation

Tomoyuki Kajiwara,[1] **Biwa Miura,**[2] **Yuki Arase**[3]
[1]Institute for Datability Science, Osaka University, [2]AI Samurai Inc.,
[3]Graduate School of Information Science and Technology, Osaka University
kajiwara@ids.osaka-u.ac.jp, miura@aisamurai.co.jp, arase@ist.osaka-u.ac.jp

Abstract

We tackle the low-resource problem in style transfer by employing transfer learning that utilizes abundantly available raw corpora. Our method consists of two steps: pre-training learns to generate a semantically equivalent sentence with an input assured grammaticality, and fine-tuning learns to add a desired style. Pre-training has two options, auto-encoding and machine translation based methods. Pre-training based on AutoEncoder is a simple way to learn these from a raw corpus. If machine translators are available, the model can learn more diverse paraphrasing via roundtrip translation. After these, fine-tuning achieves high-quality paraphrase generation even in situations where only 1k sentence pairs of the parallel corpus for style transfer is available. Experimental results of formality style transfer indicated the effectiveness of both pre-training methods and the method based on roundtrip translation achieves state-of-the-art performance.

1 Introduction

Style transfer is a class of tasks that generate paraphrases by controlling information other than the meaning in a sentence. Style transfer is beneficial for various applications. For example, it allows us to simplify text by changing difficult expressions to easier ones for language learning support (Petersen and Ostendorf 2007; Belder and Moens 2010). Furthermore, it is useful as preprocessing for information extraction and machine translation (Evans 2011; Štajner and Popovic 2016). In this study, we focus on formality and simplicity, denoted as *style* herein.

Paraphrase generation, including style transfer, can be formalized as a monolingual machine translation problem (Specia 2010; Xu et al. 2012). Generally, machine translation requires the availability of million-scale parallel corpora to train statistical or neural models. Bilingual texts are produced and accumulated in daily life. Meanwhile, monolingual parallel sentences of a specific style are difficult to collect because they are unlikely to be produced naturally. Hence, only small-scale monolingual parallel corpora with hundreds of thousands of sentence pairs (Zhang and Lapata 2017; Rao and Tetreault 2018) are available for style transfer. From such small datasets, only a limited amount of rewrite rules can be acquired. This results in a conservative style transfer model that rewrites only a few phrases in an input sentence (Niu, Rao, and Carpuat 2018).

To train a high-quality style transfer model that conducts active rewrites, rule-based data augmentation (Rao and Tetreault 2018) and multitask learning with style-sensitive machine translation (Niu, Rao, and Carpuat 2018) are proposed. However, these are high-cost methods that rely on manual rules or special datasets; they cannot be extended easily to other styles.

We address the low-resource problem in style transfer by transfer learning. Our method is independent of any manual process for data augmentation and hence widely applicable to various types of styles. As a key contribution, our method allows to train a style transfer model with just a thousand parallel sentence pairs. In style transfer, a successful paraphrase should reserve meaning equivalence, grammaticality, and style fidelity (Rao and Tetreault 2018; Niu, Rao, and Carpuat 2018). Our key concept is that styles should be learned from the monolingual parallel corpus, but meaning equivalence and grammaticality can be ensured by utilizing other corpora of larger sizes. Specifically, we pre-train the paraphrase generation model to ensure that grammatical sentences of equivalent meaning are generated using a style-independent raw corpus. Subsequently, we fine-tune the model to learn styles using the monolingual parallel corpus specialized for a target style.

We propose two methods for pre-training using AutoEncoder and machine translators. AutoEncoder is a simpler approach, where the paraphrase generator learns to generate exactly the same sentence as the input using any raw corpus. The pre-trained model that learns various expressions using a large-scale raw corpus becomes a high-quality paraphrase generator via fine-tuning. If reliable machine translators are available, we can pre-train the model with more diverse paraphrasal expressions on pseudo-parallel corpus generated by roundtrip translation. Fine-tuning on the truly-parallel corpus fits the model to generate paraphrases with the target style distilling a number of rewriting patterns acquired in pre-training.

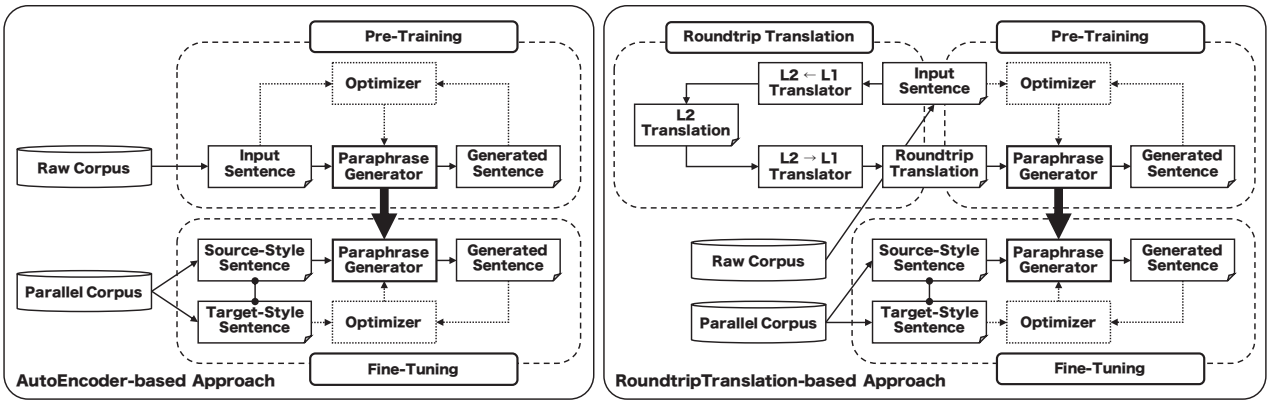

Figure 1: Pre-training based on auto-encoding or roundtrip translation followed by fine-tuning with a small-scale parallel corpus.

Input Sentence	Roundtrip Translation
I love watching the show.	I love to see the show.
Thanks for asking the question.	Thank you for the question.
The key to a successful relationship is good communication.	Good communication is the key to a successful relationship.

Table 1: Example sentence pairs of roundtrip translation that replaces phrases and changes syntactic structures while preserving the meaning equivalence.

Evaluation using two datasets empirically confirm that our approach allows an effective transfer learning, and achieves state-of-the-art performance without a costly data augmentation process.

2 Proposed Method

We propose a method of transfer learning to solve the low-resource problem in style transfer. In style transfer, appropriate sentences must be generated from three perspectives: meaning equivalence between the input and output sentences, grammaticality of the output sentence, and style fidelity of the output sentence (Rao and Tetreault 2018). However, it is difficult to learn these three features simultaneously from only a small-scale parallel corpus. Therefore, we train the paraphrase generation model in two steps: pre-training learns to generate a semantically equivalent sentence with an input assuring grammaticality, and fine-tuning learns to add a desired style. Our pre-training uses a style-independent raw corpus such that it can be easily applied to any style, unlike previous studies.

Pre-Training

To efficiently learn styles from a small monolingual parallel corpus in the fine-tuning step, we propose to first conduct pre-training to acquire the ability to generate a grammatical paraphrase with meaning equivalence. In this study, we pre-train a paraphrase generation model using the AutoEncoder-based method and roundtrip translation-based method, as shown in Figure 1. The former approach utilizes a raw corpus to train a paraphrase generator as an AutoEncoder. In the latter approach, a paraphrase generator as a denoising AutoEncoder is trained using a pseudo-paraphrase corpus constructed by roundtrip translating a raw corpus. As these pre-training methods use only style-independent raw corpora and translators, they can be applied to any style.

AutoEncoder-based Approach In this pre-training, we perform text-to-text generation as AutoEncoder (AE) that outputs the input sentence as it is. Any raw corpus can be used for AE-based pre-training. This method satisfies both conditions of retaining meaning and grammatical correctness in a simple manner.

RoundtripTranslation-based Approach As shown in Figure 1, machine translators in each direction of L1 (target language for style transfer) $\rightarrow$ L2 (another language that bilingual corpora with L1 are available) and L2 $\rightarrow$ L1 are prepared, and the entire raw corpus is translated and back-translated. Roundtrip translation (RT) inevitably adds noise to an input sentence, thus resulting in style-insensitive paraphrases, as shown in Table 1. Of course, not all translations can be expected to be paraphrasal, *e.g.*, sentences of different meanings and agrammatical sentences can be generated owing to translation errors. Despite this limitation, we utilize the advantage of generating pseudo-paraphrases from only the raw corpus.

We pre-train a paraphrase generation model using a denoising AutoEncoder that reproduces an input sentence from roundtrip translation. By the denoising feature, we expect that the model learns to actively rewrite sentences while preserving the semantic equivalency. As mentioned, roundtrip translations produce not only pseudo-paraphrases but also sentence pairs with translation errors. Hence, we input the roundtrip translations to the generator to avoid training it to generate noisy sentences.

		Informal → Formal		Formal → Informal	
	Train	Dev	Test	Dev	Test
Entertainment & Music (E&M)	52,595	2,877	1,416	2,356	1,082
Family & Relationships (F&R)	51,967	2,788	1,332	2,247	1,019

Table 2: Number of sentence pairs of GYAFC dataset.

Fine-Tuning

Fine-tuning primarily learns to add a desired style to an input sentence. As shown in Figure 1, a pre-trained paraphrase generation model is simply fine-tuned by supervised learning using the monolingual parallel corpus of the target style.

In pre-training via roundtrip translation, the model should have learned both paraphrases and noisy translation errors from a pseudo-parallel corpus. As truly-parallel corpora can be used in fine-tuning, we expect to distill paraphrases suitable for the target style from the pre-trained model.

3 Experimental Settings

We evaluate the performance of the proposed methods with the GYAFC (Rao and Tetreault 2018), as shown in Table 2.

Setup for Style Transfer

For experiments of style transfer, we used the GYAFC corpus (Rao and Tetreault 2018) after normalization and tokenization using the Moses toolkit.[1] The GYAFC corpus is a monolingual parallel corpus consisting of formal and informal English sentences. These sentences were extracted from Entertainment & Music (E&M) and Family & Relationships (F&R) domains of the Yahoo Answers L6 corpus.[2] URLs, question sentences, and sentences that are shorter than 5 words or longer than 25 words were discarded during the preprocessing. Its development (Dev) and test (Test) sets are multi-referenced, and each source sentence contains 4 references with a target style *i.e.*, formal or informal.

Additionally, a raw corpus for pre-training was constructed from the Yahoo Answers L6 corpus. After the same preprocessing as the GYAFC corpus, we extracted 3 million sentences as the training set and 3,000 sentences of the development set for each domain. For both pre-training and fine-tuning, we used byte-pair encoding[3] to limit the number of tokens to 32,000 per domain.

As a paraphrase generation model, we constructed the recurrent neural network (RNN), convolutional neural network (CNN), and self-attention network (SAN) models using the Sockeye toolkit (Hieber et al. 2017).[4] Our RNN model uses a 4-layer long short-term memory of 1,024 hidden dimensions for both the encoder and decoder, and multi-layer perceptron attention with a layer size of 1,024. Our CNN model uses 8 layers in the encoder and decoder, where the hidden dimensions were set to 512. Its convolutional kernel size was set to 3. Our SAN model uses a 6-layer transformer with a model size of 512 and 8 attention heads. We used word embeddings in 512 dimensions tying the source, target, and the output layer's weight matrix. We added dropout to all embeddings and hidden layers. In addition, we applied layer-normalization and label-smoothing as regularization. All models were optimized using the Adam optimizer. The batch size was 4,096 tokens. We created a checkpoint for the model at every 200 updates. The training stopped after 32 checkpoints without improvement in the validation perplexity. All the hyper-parameter settings omitted here are the same as Sockeye's `arxiv_1217` branch.[5]

Setup for Roundtrip Translation

For roundtrip translation in pre-training, we used the SAN model with the same setting as the paraphrase generation model. We chose German as the pivot language because large-scale bilingual corpora are freely available. We used the dataset of WMT-2017 En-De translation task (Bojar et al. 2017) for our machine translators. For the Train set, we used approximately 4.5 million sentence pairs from the News Commentary, Europarl, and Common Crawl corpora. For the Dev and Test sets, we used 2,999 sentence pairs of newstest-2016 and 3,004 sentence pairs of newstest-2017, respectively. Each translator achieved 27.6 and 33.8 test-set BLEU on En → De and De → En, respectively. They outperformed the single model of the WMT-2017 winning team (Sennrich et al. 2017) in both directions.

Method Comparison

We compare the proposed method to the previous methods for formality style transfer. R&T-PBMT (Rao and Tetreault 2018) is a phrase-based statistical machine translation (PBMT) model (Koehn et al. 2007) trained on the GYAFC corpus with data augmentation using a rule-based method. R&T-NMT (Rao and Tetreault 2018) is a neural machine translation (NMT) model (Jhamtani et al. 2017) trained on the GYAFC corpus with data augmentation using R&T-PBMT. BiFT-Single and BiFT-Ensemble (Niu, Rao, and Carpuat 2018) trained bi-directional paraphrases of both Informal → Formal and Formal → Informal in a single model in the manner of multilingual translation (Johnson et al. 2017). MultiTask (Niu, Rao, and Carpuat 2018) is a model based on the multitask learning of style-sensitive machine translation and formality style transfer. Note that BiFT-Ensemble and MultiTask are ensemble models combining 4 models with different seeds.

[1] https://github.com/moses-smt/mosesdecoder

[2] https://webscope.sandbox.yahoo.com

[3] https://github.com/rsennrich/subword-nmt

[4] https://github.com/awslabs/sockeye

[5] https://github.com/awslabs/sockeye/tree/arxiv_1217/arxiv

	Informal → Formal		Formal → Informal	
	E&M	F&R	E&M	F&R
Source	49.09 (100.00)	51.03 (100.00)	29.85 (100.00)	29.85 (100.00)
Reference	100.00 (27.88)	100.00 (29.77)	100.00 (15.05)	100.00 (15.64)
R&T-PBMT	68.22 (51.62)	72.94 (51.56)	33.54 (61.53)	32.64 (74.01)
R&T-NMT	68.41 (54.16)	74.22 (54.66)	33.56 (52.95)	35.03 (59.57)
BiFT-Single	69.20 (n/a)	73.52 (n/a)	35.44 (n/a)	37.72 (n/a)
Ours (RNN-RT)	**71.14** (**49.07**)	**75.73** (**50.82**)	**38.51** (**47.85**)	**39.79** (**51.73**)
BiFT-Ensemble	71.36 (55.86)	74.49 (59.48)	36.18 (61.21)	38.34 (63.60)
MultiTask	72.13 (54.55)	75.37 (58.11)	38.04 (55.47)	39.09 (58.02)
Ours (RNN-RT)	**72.41** (**48.62**)	**76.40** (**51.28**)	**39.22** (**48.42**)	**39.31** (**52.68**)

Table 3: BLEU scores of formality style transfer in GYAFC dataset. Parentheses are BLEU scores between input and output sentences (BLEU_{IO}) where lower values mean that the model is actively rewriting.

4 Experimental Results

Following the previous studies (Rao and Tetreault 2018; Niu, Rao, and Carpuat 2018), BLEU[6] evaluates the performance of each model based on the phrasal match rate between model outputs and reference sentences (BLEU_{OR}). Style transfer models tend to conduct conservative paraphrases that yield only a small number of rewrites. To evaluate if rewriting has been actively performed, we calculated the BLEU between input and output sentences (BLEU_{IO}). A lower BLEU_{IO} implies that the output sentence is rewritten significantly. Hence, an ideal style transfer should achieve a higher BLEU_{OR} and a lower BLEU_{IO}.

Comparison to Previous Methods

Table 3 shows the comparison results, *i.e.*, the BLEU_{OR} scores of the proposed and comparative methods with BLEU_{IO} scores in parentheses. Herein, we only present the results of RNN-RT for brevity, which has been confirmed to achieve the best performance. The first two rows indicate the performances when the source sentence itself or the reference sentence is regarded as a paraphrase instead of a model output, thus setting the standard for score interpretation. The BLEU_{OR} score of Source corresponds to the most conservative model that conducts no rewriting at all. Similarly, the BLEU_{IO} score of Reference implies the upper bound of BLEU_{IO}, where all the rewritings in references are conducted. In the middle rows, we compare our method with the previous single models. As the system outputs of BIFT-Single are not available, we borrowed the BLEU_{OR} score from Niu, Rao, and Carpuat (2018). In the bottom rows, we compare our method with the previous ensemble models. Here, following Niu, Rao, and Carpuat (2018), we combine the Train sets of two domains (E&M+F&R) and train a single model on it. For a comparison with ensemble models, we conducted model ensembling by combining four models of different seeds.

The experimental results in Table 3 indicate that the proposed method consistently achieves the highest BLEU_{OR} and lowest BLEU_{IO}. These results indicate that our method allows to generate paraphrases closer to references across styles or domains. Simultaneously, the lowest BLEU_{IO} indicates that our method conducts more active rewriting than previous methods. These evaluation results demonstrate that our transfer learning has successfully benefited from the pre-training based on roundtrip translation.

[6]https://github.com/mjpost/sacreBLEU

Effects of Transfer Learning

Table 4 shows the BLEU_{OR} of our methods with different combinations of the pre-training methods and model architectures, as well as baselines that were trained using the parallel corpus without pre-training. On any models, performance of the style transfer improved consistently and significantly for four tasks owing to the RT-based pre-training. Specifically, for the Informal → Formal task, improvements on BLEU_{OR} range from 6.67 to 16.93 points compared to the baselines. For the Formal → Informal task, those are 3.92 to 10.68 points.

The AE-based method improved the performance of the paraphrase generation in many cases (10 out of 12). However, the improvement was smaller compared to the RT-based method. This is because the paraphrase generation model learns more diverse synonymous expressions than AE-based methods, as shown in Table 1 via the RT-based pre-training. These results indicated that by adding bilingual noise to the pre-training corpus via roundtrip translation, transfer learning can be performed more effectively.

Qualitative Analysis

Table 5 shows examples of model outputs. In R&T's method, informal expressions appear in the output sentence in the Informal → Formal task, because the informal expressions out of its rules failed to be formalized and remain in data augmentation. As shown by BLEU_{IO} in Table 3, MultiTask tends to conduct conservative rewriting. This tendency is obvious here, where informal expressions remained in the output. In the Formal → Informal task, the R&T-PBMT failed to generate a grammatical output. Although outputs by R&T-NMT and MultiTask are fluent, they failed to preserve the meaning of the input sentence. Nonetheless, our model

	Informal → Formal				Formal → Informal			
	E&M		F&R		E&M		F&R	
	pre-train	fine-tune	pre-train	fine-tune	pre-train	fine-tune	pre-train	fine-tune
RNN-Base	57.76		66.45		27.68		33.86	
RNN-AE	48.55	64.58	50.51	67.38	29.26	32.19	29.36	32.99
RNN-RT	46.84	**69.03**	47.96	**74.28**	32.03	**38.36**	31.86	**39.00**
CNN-Base	48.06		60.58		24.85		30.93	
CNN-AE	48.91	57.18	50.64	51.41	29.30	28.12	29.49	31.67
CNN-RT	46.96	**64.99**	47.94	**69.36**	31.33	**34.17**	31.10	**36.12**
SAN-Base	60.57		67.52		30.09		34.64	
SAN-AE	48.54	65.57	50.55	70.34	29.25	33.45	29.36	34.95
SAN-RT	46.82	**69.58**	47.97	**74.19**	31.50	**38.84**	31.89	**38.56**

Table 4: BLEU scores for each method. -Base is a model trained on parallel corpus only. -AE is a proposed model that pre-trains based on AutoEncoder. -RT is a proposed model that pre-trains based on roundtrip tranlation.

	E&M: Informal → Formal	E&M: Formal → Informal
Source	I LOOOOOVVVVVVVEEE this song SOOO Much!!!!!!	I thoroughly enjoy the hair bands of the 1980s.
R&T-PBMT	I loovvvvvvveee this song very Much.	I love the hair bands are THOROUGHLY + the 1980S.
R&T-NMT	I loovvvvvvveee this song so Much.	I just like the hair of the brids.
MultiTask	I really enjoy VVVVVVVVEEE this song.	I love the 80's hair.
Ours (RNN-RT)	I love this song very much.	I love the hair bands of the 80's.

Table 5: Example model outputs in E&M domain.

outputs were both semantically and grammatically correct, and successfully performed style transfer.

5 Analysis

In this section, we conduct detailed analyses of our method using the SAN model for its computational efficiency.

Detailed Automatic Evaluation

For a more detailed evaluation of each model, we employed three automatic evaluation metrics that evaluate: meaning equivalence, grammaticality, and style fidelity. Following the success of embedding-based metrics in machine translation (Shimanaka, Kajiwara, and Komachi 2018; 2019), we fine-tuned the state-of-the-art sentence encoder XLNet (Yang et al. 2019)[7] for each criterion. We trained the meaning evaluator with the Semantic Textual Similarity Benchmark (STS-B),[8] the grammar evaluator with the Grammatical versus UnGrammatical (GUG) dataset,[9] and the style evaluator with the formality corpus,[10] same as Rao and Tetreault (2018). Table 6 shows the statistics of each dataset and Pearson correlations between outputs of trained XLNet models and human labels.[11] The results show that these XLNet-based evaluators are reliable.

	Train	Dev	Test	Label	Pearson r
Meaning	$5,749$	$1,500$	$1,379$	$[0, 5]$	0.859
Grammar	$1,518$	747	754	$[1, 4]$	0.695
Style	$5,297$	$1,000$	$4,977$	$[-3, 3]$	0.704

Table 6: Details of datasets used to train the XLNet model to estimate meaning equivalence, grammaticality, and style fidelity and its correlation to human labels.

Table 7 shows the results of the XLNet-based evaluation. These are the average scores of the evaluation results for each sentence of the model output in the GYAFC test set. Note that each score was converted to $[0, 1]$ using min-max normalization, as the range of scores varies by criteria. Our method has the highest evaluation in most settings. Especially in the style fidelity, the proposed method consistently achieved the best scores. Whereas the MultiTask (Niu, Rao, and Carpuat 2018) uses a large-scale additional corpus consisting of sentences with the target style, we have successfully transformed the style without extra style-dependent corpus. Our method effectively learns target styles from a small-scale parallel corpus by first learning to generate grammatical sentences that represent the equivalent senses with the inputs (pre-training) then learning to mimic target styles (fine-tuning).

[7] We used the pre-trained XLNet-Base model available at https://github.com/zihangdai/xlnet

[8] http://ixa2.si.ehu.es/stswiki/images/4/48/Stsbenchmark.tar.gz

[9] https://github.com/EducationalTestingService/gug-data

[10] https://www.seas.upenn.edu/~nlp/resources/formality-corpus.tgz

[11] For the formality corpus, we used the Yahoo Answers domain for Test set and the other domains for Train/Dev set.

	Informal → Formal						Formal → Informal					
	E&M			F&R			E&M			F&R		
	M	G	S	M	G	S	M	G	S	M	G	S
Reference	n/a	0.692	0.344	n/a	0.710	0.356	n/a	0.720	0.623	n/a	0.734	0.614
R&T-NMT	0.756	0.682	0.322	0.743	0.705	0.339	0.602	0.758	0.578	0.601	0.774	0.573
MultiTask	**0.766**	0.689	0.321	0.745	0.709	0.337	**0.639**	**0.775**	0.575	0.620	0.783	0.570
Ours (RNN-RT)	0.764	**0.691**	**0.325**	**0.750**	**0.712**	**0.340**	**0.639**	0.771	**0.582**	**0.621**	**0.788**	**0.576**

Table 7: Results of XLNet-based automated evaluation for three criteria: **M**eaning equivalence between the reference and output sentences, **G**rammaticality of the output sentence, and **S**tyle fidelity of the output sentence.

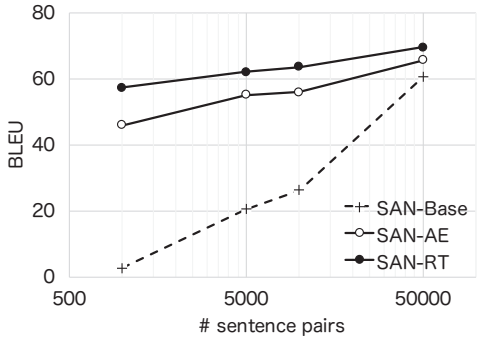

Figure 2: Learning curve in → Formal E&M task.

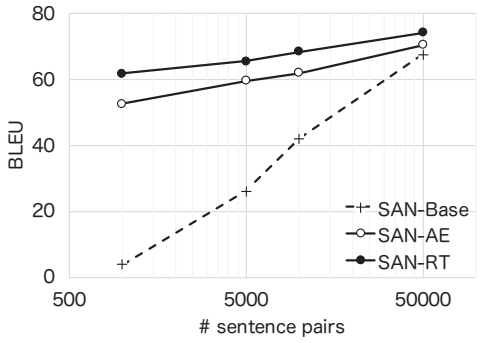

Figure 3: Learning curve in → Formal F&R task.

Effectiveness in Lower Resource Setting

We evaluate the effectiveness of the proposed method in situations where significantly less monolingual parallel corpora are available for fine-tuning. Figures 2 and 3 show the performance changes when the monolingual parallel corpus for fine-tuning is reduced to 10k, 5k, and 1k sentence pairs in Informal → Formal tasks. These experimental results indicate that our pre-training allows for high-quality paraphrase generation even in situations where only a small monolingual parallel corpus, such as just 1k sentence pairs, is available. Notably, RT-based pre-training exhibits the equivalent performance to fully supervised learning using 50k parallel corpus after fine-tuning only 5k parallel corpus.

Back- vs. Roundtrip Translation

As a relevant technique to roundtrip translation, back-translation allows to generate pseudo-paraphrasal sentence pairs. Wieting and Gimpel (2018) back-translated the Czech side of the English–Czech parallel corpus and constructed an English pseudo-paraphrase corpus.

We compare the effects of back- and roundtrip translation on style transfer in this section. We back-translated the German side of the English–German parallel corpus of Section 3 and used it for pre-training (BT). As with AE and RT, we used 3 million sentences for BT-based pre-training.

Table 8 shows the experimental results in the Informal → Formal tasks. BT-based pre-training demonstrated a higher performance than AE-based one. Same as RT-based pre-training, BT-based pre-training performs text-to-text generation as a denoising AutoEncoder using a pseudo-parallel corpus including both paraphrases and translation errors. Because fine-tuning allows one to distill paraphrases suitable for the target style, denoising the bilingual noise is effective for our transfer learning.

RT-based pre-training further outperformed BT-based one. This is because roundtrip translation allows for pseudo-paraphrases to be generated in the same domain from a monolingual raw corpus. Back-translation can be a substitute for roundtrip translation when only a limited amount of raw corpora in the target domain is available.

Effectiveness in Other Style

Finally, we evaluate the effectiveness of the proposed method in other styles. Text simplification, which rewrites a complex sentence to a simpler sentence, is a style transfer task. For text simplification, only a small monolingual parallel corpus similar in size to the GYAFC corpus exists. We used 1.2 million sentences of Simple English Wikipedia[12] for pre-training. In addition, we used 88,837 sentence pairs of WikiSmall[13] (Zhang and Lapata 2017) and 296,402 sentence pairs of WikiLarge[13] (Zhang and Lapata 2017) for fine-tuning. For evaluation, we calculated the SARI (Xu et al. 2016) on multi-reference dataset.[14] SARI is an automatic evaluation metric that correlates with manual evaluation for meaning equivalence, grammaticality, and simplicity.

[12]https://dumps.wikimedia.org/simplewiki/20181201

[13]https://github.com/louismartin/dress-data

[14]https://github.com/cocoxu/simplification

	E&M	F&R
SAN-Base	60.57	67.52
SAN-AE	65.57	70.34
SAN-BT	69.06	73.39
SAN-RT	**69.58**	**74.19**

Table 8: Comparison with pre-training based on back-translation (BT) in Informal → Formal tasks.

Table 9 shows the results. Our method outperformed the baseline that conducts supervised learning using the parallel corpus. Note that these scores are not comparable to the state-of-the-art text simplification. We do not intend to show that our method outperforms previous studies dedicated to text simplification tasks. Rather, this additional experimental result indicates that our method is effective for style transfer tasks other than formal–informal transformation.

6 Related Work

As it is difficult to train a high-quality style transfer model with only a small monolingual parallel corpus, methods for mitigating the low-resource problem have been proposed. Rao and Tetreault (2018) performed data augmentation based on the rule-based method and trained a copy-enriched NMT model (Jhamtani et al. 2017). Niu, Rao, and Carpuat (2018) further improved by the multitask learning of style-sensitive machine translation (French → formal English and French → Informal English) and formality style transfer. However, the former approach is costly because paraphrasing rules must be developed manually for each style, and the latter large-scale bilingual corpus with style labels is unlikely to be available in practice. In contrast to these methods, our transfer learning allows to utilize a raw corpus; it is free from any human efforts for data augmentation nor availability of style labels. Although it requires a large-scale bilingual corpus to train translators for roundtrip translation, such a bilingual corpus is widely available for major languages.

The general conditions for style transfer targeted in this study are summarized as follows:

1. Small parallel corpora containing hundreds of thousands of sentences are available depending on the target style.
2. Large parallel corpora (with millions of sentence pairs) for other types of styles are inaccessible.
3. Abundant raw and/or bilingual corpus are available.

Although domain adaptation in machine translation (Chu and Wang 2018) and transfer learning in dialogue response generation (Akama et al. 2017) are closely related, they assume that out-of-domain parallel corpora (that correspond to corpora of other types of styles for us) are sufficiently available. However, the second condition hinders us from using the domain adaptation approach. Instead, we proposed pre-training methods that does not rely on a parallel corpus.

There are related studies that work on style transfer without any monolingual parallel corpus (Hu et al. 2017; Shen et al. 2017; Fu et al. 2018; Prabhumoye et al. 2018).

	WikiSmall	WikiLarge
SAN-Base	34.19	34.58
SAN-RT	**35.46**	**35.99**

Table 9: SARI scores on text simplification task.

However, their targets are transferring a sentiment in a sentence, which inevitably changes the meaning of the original sentence. We regard that such tasks are different from ours that aims to preserve the original meaning.

Among the studies on sentiment transfer, Prabhumoye et al. (2018) use a common approach with us that employ bilingual translators. They generate representations of an input sentence using bilingual translators assuming that such representations preserve the core meaning of the input. Then they conduct adversarial training on the decoder to generate a sentence with the target sentiment using a classifier that discriminates sentiments. Their method requires a large-scale style-specific corpus to build the classifier. Furthermore, careful hyper-parameter tuning is required to balance training losses of the decoder and classifier.

7 Conclusion

To address the low-resource problem in style transfer, we proposed a transfer learning method comprising two steps: pre-training that learned to generate a semantically equivalent sentence with an input assured grammaticality, and fine-tuning that learned to add a desired style. It was noteworthy that the model could learn paraphrase generation via pre-training that added bilingual noise to the raw corpus using machine translators and denoising it. Our proposed method did not rely on manual annotation or a special dataset; therefore, it is a low-cost and style-independent method that achieves state-of-the-art performance.

Our detailed analysis indicated that even in situations where only 1k sentence pairs of monolingual parallel corpus was available, high-quality paraphrase generation could be achieved by the proposed method. In addition, our method was effective for styles other than formality. Roundtrip translation enabled a monolingual parallel corpus to be created from a raw corpus of any domains or styles in a cost-effective manner. Hitherto, paraphrase generation tasks in minor languages, domains, and styles have been poor due to lack of large-scale parallel corpora that are mandatory to train supervised learning models. Our method allows for high-quality paraphrases of desired styles to be generated with only a small parallel corpus and a raw corpus, and an even better quality is assured if reliable machine translators are available.

Acknowledgments

This work was supported by JST, ACT-I Grant Number JPMJPR18UB, Japan.

References

Akama, R.; Inada, K.; Inoue, N.; Kobayashi, S.; and Inui, K. 2017. Generating Stylistically Consistent Dialog Responses

with Transfer Learning. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, 408–412.

Belder, J. D., and Moens, M.-F. 2010. Text Simplification for Children. In *Proceedings of the SIGIR Workshop on Accessible Search Systems*, 19–26.

Bojar, O.; Chatterjee, R.; Federmann, C.; Graham, Y.; Haddow, B.; Huang, S.; Huck, M.; Koehn, P.; Liu, Q.; Logacheva, V.; Monz, C.; Negri, M.; Post, M.; Rubino, R.; Specia, L.; and Turchi, M. 2017. Findings of the 2017 Conference on Machine Translation. In *Proceedings of the Second Conference on Machine Translation*, 169–214.

Chu, C., and Wang, R. 2018. A Survey of Domain Adaptation for Neural Machine Translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, 1304–1319.

Evans, R. J. 2011. Comparing Methods for the Syntactic Simplification of Sentences in Information Extraction. *Literary and Linguistic Computing* 26(4):371–388.

Fu, Z.; Tan, X.; Peng, N.; Zhao, D.; and Yan, R. 2018. Style Transfer in Text: Exploration and Evaluation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 663–670.

Hieber, F.; Domhan, T.; Denkowski, M.; Vilar, D.; Sokolov, A.; Clifton, A.; and Post, M. 2017. Sockeye: A Toolkit for Neural Machine Translation. *arXiv:1712.05690*.

Hu, Z.; Yang, Z.; Liang, X.; Salakhutdinov, R.; and Xing, E. P. 2017. Toward Controlled Generation of Text. In *Proceedings of the 34th International Conference on Machine Learning*, 1587–1596.

Jhamtani, H.; Gangal, V.; Hovy, E.; and Nyberg, E. 2017. Shakespearizing Modern Language Using Copy-Enriched Sequence to Sequence Models. In *Proceedings of the Workshop on Stylistic Variation*, 10–19.

Johnson, M.; Schuster, M.; Le, Q. V.; Krikun, M.; Wu, Y.; Chen, Z.; Thorat, N.; Viégas, F.; Wattenberg, M.; Corrado, G.; Hughes, M.; and Dean, J. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics* 5:339–351.

Koehn, P.; Hoang, H.; Birch, A.; Callison-Burch, C.; Federico, M.; Bertoldi, N.; Cowan, B.; Shen, W.; Moran, C.; Zens, R.; Dyer, C.; Bojar, O.; Constantin, A.; and Herbst, E. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, 177–180.

Niu, X.; Rao, S.; and Carpuat, M. 2018. Multi-Task Neural Models for Translating Between Styles Within and Across Languages. In *Proceedings of the 27th International Conference on Computational Linguistics*, 1008–1021.

Petersen, S. E., and Ostendorf, M. 2007. Text Simplification for Language Learners: A Corpus Analysis. In *Proceedings of the Speech and Language Technology in Education Workshop*, 69–72.

Prabhumoye, S.; Tsvetkov, Y.; Salakhutdinov, R.; and Black, A. W. 2018. Style Transfer Through Back-Translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 866–876.

Rao, S., and Tetreault, J. 2018. Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 129–140.

Sennrich, R.; Birch, A.; Currey, A.; Germann, U.; Haddow, B.; Heafield, K.; Miceli Barone, A. V.; and Williams, P. 2017. The University of Edinburgh's Neural MT Systems for WMT17. In *Proceedings of the Second Conference on Machine Translation*, 389–399.

Shen, T.; Lei, T.; Barzilay, R.; and Jaakkola, T. 2017. Style Transfer from Non-Parallel Text by Cross-Alignment. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, 6830–6841.

Shimanaka, H.; Kajiwara, T.; and Komachi, M. 2018. RUSE: Regressor Using Sentence Embeddings for Automatic Machine Translation Evaluation. In *Proceedings of the Third Conference on Machine Translation*, 751–758.

Shimanaka, H.; Kajiwara, T.; and Komachi, M. 2019. Machine Translation Evaluation with BERT Regressor. *arXiv:1907.12679*.

Specia, L. 2010. Translating from Complex to Simplified Sentences. In *Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language*, 30–39.

Štajner, S., and Popovic, M. 2016. Can Text Simplification Help Machine Translation? *Baltic Journal of Modern Computing* 4(2):230–242.

Wieting, J., and Gimpel, K. 2018. ParaNMT-50M: Pushing the Limits of Paraphrastic Sentence Embeddings with Millions of Machine Translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 451–462.

Xu, W.; Ritter, A.; Dolan, B.; Grishman, R.; and Cherry, C. 2012. Paraphrasing for Style. In *Proceedings of the 24th International Conference on Computational Linguistics*, 2899–2914.

Xu, W.; Napoles, C.; Pavlick, E.; Chen, Q.; and Callison-Burch, C. 2016. Optimizing Statistical Machine Translation for Text Simplification. *Transactions of the Association for Computational Linguistics* 4:401–415.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; and Le, Q. V. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv:1906.08237*.

Zhang, X., and Lapata, M. 2017. Sentence Simplification with Deep Reinforcement Learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 595–605.