Improving Stylized Neural Machine Translation with Iterative Dual Knowledge Transfer

Xuanxuan Wu[1*] , **Jian Liu**[1*] , **Xinjie Li**[2] , **Jinan Xu**[1†] ,
Yufeng Chen[1] , **Yujie Zhang**[1] , **Hui Huang**[1]
[1]Beijing Jiaotong University, Beijing, China
[2]Global Tone Communication Technology Co., Ltd., Beijing, China
{19120414, jianliu}@bjtu.edu.cn, lixinjie@gtcom.com.cn,
{jaxu, cheny, yjzhang, 18112023}@bjtu.edu.cn

Abstract

Stylized neural machine translation (NMT) aims to translate sentences of one style into sentences of another style, which is essential for the application of machine translation in a real-world scenario. However, a major challenge in this task is the scarcity of high-quality parallel data which is stylized paired. To address this problem, we propose an iterative dual knowledge transfer framework that utilizes informal training data of machine translation and formality style transfer data to create large-scale stylized paired data, for the training of stylized machine translation model. Specifically, we perform bidirectional knowledge transfer between translation model and text style transfer model iteratively through knowledge distillation. Then, we further propose a data-refinement module to process the noisy synthetic parallel data generated during knowledge transfer. Experiment results demonstrate the effectiveness of our method, achieving an improvement over the existing best model by 5 BLEU points on MTFC dataset. Meanwhile, extensive analyses illustrate our method can also improve the accuracy of formality style transfer.

1 Introduction

Stylized neural machine translation aims to translate sentences of one style into sentences of another style, which is essential for a wide range of NLP applications, such as non-native speaker assistants and child education. Table 1 shows an example of stylized translation. Despite its importance, few efforts have been made to deal with translation with stylistic transformation. Recently, [Wu *et al.*, 2020] introduce a dataset, the Machine Translation Formality Corpus (MTFC), serving as a benchmark dataset for translation formality transfer task, which aims to translate an informal sentence to a formal one while keeping its semantics unchanged.

Informal Source Sentence:
哇(Wow)，我的(my)观察技术(observation skill)
真的(is really)很差……(so bad......)

Informal Translation: Wow , I am very dumb in my observation skills ... ...
Formal Translation: My observation skills are not very good.

Table 1: An example of stylized Chinese-to-English translation. The source sentence can be translated as an informal target sentence, with colloquial expressions such as ”Wow”, or more formally with correct grammar and formal terms.

However, a major challenge in stylized neural machine translation is the scarcity of stylized parallel training data, which makes training a sequence-to-sequence model directly impractical. Therefore, traditional pivot-based [Cohn and Lapata, 2007; Chen *et al.*, 2017] methods conduct stylized machine translation in a pipeline manner, with the decoding process approximated as two steps: the first step is to translate sentences from source language to target language, and the second step is to transfer the target sentences to certain style. Although the pivot-based method is a reasonable solution to this task, it suffers from error propagation problem. To deal with this problem, [Wu *et al.*, 2020] propose a teacher-student framework to build synthetic parallel data. This procedure allows parameters to be estimated in one model, avoiding the error cascading, thus improving translation formality by a large margin. But there are still some issues to be resolved. First, they only consider knowledge transfer from style transfer model to NMT model unidirectionally. Moreover, they only perform one-pass transfer, failing to consider the interactions between the two models. Second, as the synthetic parallel data is built by an imperfect formality style transfer model, it may contain noises, which can harm the performance of the NMT model.

To overcome limitations of previous methods, we propose an **I**terative **D**ual **K**nowledge **T**ransfer (**IDKT**) framework for stylized NMT, with a motivation that the bidirectional knowledge transfer between NMT model and formality style transfer model should be repeatedly performed to fully exploit the knowledge in different datasets. Particularly, first, to tackle the problem of lacking stylized translation pairs, we use a formality style transfer model, trained on style transfer data, as a teacher model to generate formal target sentences, where

[*]Equal contribution
[†]Corresponding author

the generated target sentences are then combined with the source sentences to train an NMT model. After that, we iteratively perform bidirectional formality knowledge transfer with knowledge distillation. In this way, both NMT model and formality style transfer model are expected to continually boost each other. Second, to address the noise in synthetic data, we further enhance the proposed framework by a **data-refinement** module to regenerate the target formal sentences with higher quality.

To verify the effectiveness of our method, we perform experiments on MTFC dataset with informal Chinese translated to formal English. Our method surpasses the existing best model by 5 BLEU points and gets the highest formality accuracy.

The contributions of this work are summarized as follows:

- We propose an iterative dual knowledge transfer framework for improving stylized NMT, which can iteratively perform bidirectional knowledge transfer between NMT model and text style transfer model to boost each other.
- We extend our framework with a data-refinement module to improve the quality of synthetic parallel data generated during knowledge transfer.
- We conduct experiments on two benchmark datasets, MTFC and GYAFC, and achieves state-of-the-art results in producing stylized translation sentences based on both automatic and human evaluation[1].

2 Related Work

Stylized Neural Machine Translation. Previously, the work been made on stylized NMT is limited. [Rabinovich *et al.*, 2016; Michel and Neubig, 2018] take an adaptation approach to personalize MT with gender-specific or speaker-specific data, [Niu *et al.*, 2018] uses multi-task learning to perform formality transfer and translation. Other work like [Niu and Carpuat, 2020] focuses on controlling the formality of NMT. However, the training process of these NMT models usually require translation pairs with the target-style, [Wu *et al.*, 2020] proposes the machine translation formality corpus by extending the Grammarly's Yahoo Answers Formality Corpus (GYAFC) [Rao and Tetreault, 2018], which makes it possible to benchmark stylized machine translation.

Text Style Transfer. Text style transfer is a more general task of changing the style of a sentence while preserving its content. Previous work explores this task as a sequence-to-sequence learning task using paired sentences in different styles, [Jhamtani *et al.*, 2017] adopts an end-to-end neural network with pointer network to transform Modern English text to Shakespearean style English. However, available paired stylized texts are very limited. Some works focus on learning style-independent content representation. Recent works [Li *et al.*, 2018; Sudhakar *et al.*, 2019] propose to separate style information and content representation by directly removing stylistic words, however, the separation is challenging due to the content and style interacting in subtle ways in natural language. [Shen *et al.*, 2017; Fu *et al.*, 2018] use adversarial network to ensure that the content representation is separated from style. Regarding the tasks, formality style transfer is one of the most wild direction of text style transfer, aiming to change the style of a given sentence from informal to formal. The research of formality style transfer is mainly motivated by the GYAFC dataset, released by [Rao and Tetreault, 2018]. [Xu *et al.*, 2019] uses CNN-based discriminators and cyclic reconstruction objective in a semi-supervised setting. Due to the small size of the parallel corpus, [Wang *et al.*, 2019; Chawla and Yang, 2020] incorporate a pre-trained neural network for formality style transfer, and [Zhang *et al.*, 2020] proposes several data augmentation methods to generate augmented data with various formality style transfer knowledge.

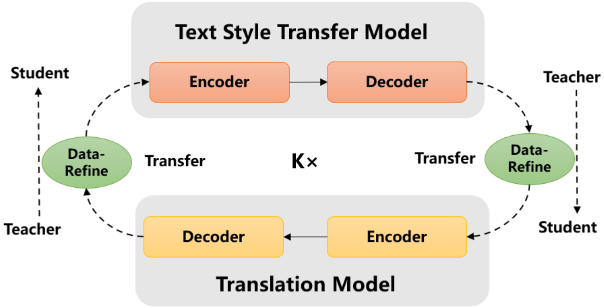

Figure 1: Overview of our framework, dotted-line refers to the directions of knowledge transfer between translation model and text style transfer model. Data-Refinement module is used to filter out noise during knowledge transfer.

3 Our Method

In this section, we first illustrate the overview of our iterative dual knowledge transfer framework, then describe the iterative knowledge transfer strategy and the architecture of data-refinement module.

3.1 Overview

As shown in Figure 1, our proposed framework consists of three sub-modules: (1) A stylized NMT model to translate sentences from informal to formal style. (2) A text style transfer model to regenerate the sentences from informal to formal style. (3) A data-refinement module aims to correct the mistakes produced by above two models. The NMT model and the style transfer model will iteratively teach each other to enhance their performance. The NMT model and text style transfer model in our study use the Transformer-based encoder and decoder architecture and the parameters of the text style transfer model are initialized using pre-trained BART [Lewis *et al.*, 2019].

In this study, we propose to build a formal NMT model without translation parallel sentences in the target style. Our method leverages two parts of data during training, one is machine translation corpus $D_{mt} = \{\langle x_1, y_1\rangle, \ldots, \langle x_n, y_n\rangle\}$, where n is the number of training samples. The other is text style transfer corpus $D_{style} = \{\langle y_1^s, y_1^t\rangle, \ldots, \langle y_m^s, y_m^t\rangle\}$, where y_i^s is the source-style sentence and y_i^t is the target-style sentence. Our stylized NMT model aims to translate

[1] Code and data are available at https://github.com/mt887/IDKT

sentences into target language while changing the style of the source sentences and preserving their style-independent content.

3.2 Iterative Dual Knowledge Transfer

Figure 1 illustrates our iterative dual knowledge transfer procedure between the translation model and the text style transfer model. At first, we train two sequence-to-sequence models individually via minimizing the negative likelihood on their training corpora D_{mt} and D_{style}, the NMT model θ_{mt}:$X \rightarrow \hat{Y}$ translates the source language sentence x into the target language sentence $\hat{y}$, while the style transfer model θ_{style}:$\hat{Y} \rightarrow Y$ transfers an informal sentence $\hat{y}$ into a formal sentence y.

Then, we perform iterative dual knowledge transfer between the NMT model and text style transfer model. We introduce sequence-level knowledge distillation [Kim and Rush, 2016] to transfer knowledge between the two models, which is a method for improving the performance of student model by imitating the behaviour of the teacher model. The sequence-level knowledge distillation involves three steps:

1. Randomly initialize a teacher model and train it until convergence on real-world training data.
2. Decode the source-side training data using the teacher model to produce source-target pairs.
3. Randomly initialize a student model and train it until convergence on the synthetic source-target pairs.

During the transfer from θ_{style} to θ_{mt}, the student model is optimized by minimizing the following loss function

$$\mathcal{L} = -\sum_{(x^s,y^s)\in\mathcal{D}} \sum_{y^t} q\left(y^t \mid y^s\right) \log P\left(y^t \mid x^s; \theta_{mt}\right) \quad (1)$$

where $q\left(y^t \mid y^s\right)$ represents the probability outputs of the previous best formality style transfer model. The student model is trained on the synthetic source-target pairs generated by the teacher model.

Although other knowledge transfer methods like fine-tuning can be effective in this task, it may cause catastrophic forgetting of previously learned knowledge. Sequence-level knowledge distillation is more effective and can alleviate the style difference between NMT model and text style transfer model. Despite the predicted results are not hundred-percent accurate and fluent, the output texts contain task-related knowledge which could be helpful for strengthening counterpart model. More importantly, despite the limitation of parallel data, monolingual data is readily accessible, thus refraining our model from the data-scarcity nature of style transfer task.

To better understand our proposed framework, we summarize the training procedure of the framework in Algorithm 1. The whole framework starts with pre-trained NMT model, style transfer model and data-refinement module (**Line** 1), then a joined training process is carried out to train the models iteratively (**Lines** 3-13). In each iteration, we first leverage text style transfer model to transfer the target informal sentences Y^s_{mt} to formal style $\hat{Y}^t_{mt}$. In this way, we obtain stylized paired translation synthetic data $X^* = \{\langle x^s_{mt}, \hat{y}^t_{mt}\rangle\}$, which can be applied to train the stylized NMT model (**Lines** 4-6). Further, a batch of parallel sentences $\left\{\left\langle x^s_p, y^s_p\right\rangle\right\}$ are sampled from D_{mt} (**Line** 7), then the stylized NMT model is adopted to decode the sampled sentences x^s_p into target-style translation $\hat{y}^t_p$. After that, we obtain the synthetic text style transfer data $Y^* = \left\{\left\langle y^s_p, \hat{y}^t_p\right\rangle\right\}$ to further improve text style transfer model (**Line** 8-10). During this process, we evaluate and save the best model parameters. Specifically, the synthetic parallel data produced by a teacher model will be regenerated by a data-refinement module to filter out the syntactic and semantic noise.

Algorithm 1 Iterative Dual Knowledge Transfer for Improving Stylized NMT

```
Input: Training corpora { $D_{mt} = \{\langle X^s_{mt}, Y^s_{mt}\rangle\}$, $D_{style} = \left\{\left\langle Y^s_{style}, Y^t_{style}\right\rangle\right\}$, development sets { $D^v_{mt}$, $D^v_{style}$} and the max iteration number $K$.
Output: Stylized NMT model $\theta^*_{mt}$
 1: $\theta^*_{mt} \leftarrow$ TrainModel($D_{mt}$) , $\theta^*_{style} \leftarrow$ TrainModel($D_{style}$)
    $\theta^*_{dr} \leftarrow$ TrainModel($D_{style}$ and $\hat{y}^t$)
 2: k=0
 3: for $k \leq K$ do
 4:     Use $\theta^*_{style}$ to build synthetic data $X^* = \{\langle x^s_{mt}, \hat{y}^t_{mt}\rangle\}$
 5:     Use $\theta^*_{dr}$ to refine $\hat{y}^t_{mt}$ and build $\bar{X} = \{\langle x^s_{mt}, y^t_{mt}\rangle\}$
 6:     $\theta^*_{mt} \leftarrow$ TransferKnowledge($\theta^*_{style}$,$\bar{X}$) based on Eq.1
 7:     Sample sentences $\left\{\left\langle x^s_p, y^s_p\right\rangle\right\}$ from $D_{mt}$
 8:     Use $\theta^*_{mt}$ to build synthetic data $Y^* = \left\{\left\langle y^s_p, \hat{y}^t_p\right\rangle\right\}$
 9:     Use $\theta^*_{dr}$ to refine $\hat{y}^t_p$ and build $\bar{Y} = \left\{\left\langle y^s_p, y^t_p\right\rangle\right\}$
10:     $\theta^*_{style} \leftarrow$ TransferKnowledge($\theta^*_{mt}$,$\bar{Y}$)
11:     Fine-tune $\theta^*_{dr}$ with $\left\{\left\langle y^s_p, \hat{y}^t_p, y^t_p\right\rangle\right\}$
12:     k=k+1
13: end for
```

In this way, we enable stylized NMT model to retain the previously learned translation knowledge and absorb the effective style transfer knowledge from text style transfer model. Similarly, we perform knowledge transfer in the inverse direction to enhance the performance of text style transfer model. During this procedure, the knowledge contained in two different models is thoroughly interacted with each other, leading to better single model with limited training data.

3.3 Data-Refinement Module

Since the synthetic data contains noise, we further propose a data-refinement module to improve its quality, which is motivated by the advance of Automatic post-editing (APE [Correia and Martins, 2019]). APE is to automatically correct the mistakes produced by an NMT model. We design our data-refinement module architecture, following the practices in APE. The module takes the source sentences and the imperfect style transferred sentences as inputs to regenerate target sentences with mistakes corrected.

To fit in our proposed framework, we adopt the Multi-source Transformer [Junczys-Dowmunt and Grundkiewicz, 2018] with an encoder-decoder structure for our data-refinement model, as shown in Figure 2. Since pre-trained

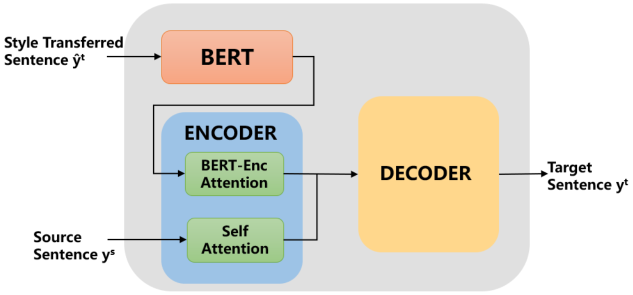

Figure 2: The architecture of the data-refinement module. For convenience, we omit some repeated components, such as feed forward layer and layer normalization. Here $(y^s, \hat{y}^t)$ is the inputs of the model, and y^t is the refined output sentence.

text representations like BERT are trained on a large corpus, beneficial to correct grammar errors. Following the practice in [Zhu *et al.*, 2020], we combine BERT [Devlin *et al.*, 2018] with a standard transformer to improve the performance of data-refinement module, in which we first employ BERT to extract representations for an input sentence, then the representations are combined with each layer of the encoder in transformer model through attention mechanisms. The data-refinement module is trained via minimizing the negative likelihood of triplets $\widehat{Y} = \{\langle y^s, \hat{y}^t, y^t\rangle\}$, containing the source sentence, the style transferred sentence, and the target sentence respectively.

$$\mathcal{L}(\theta) = \sum_{(y^s,\hat{y}^t,y^t)\in\widehat{Y}} -\log P\left(y^t \mid \hat{y}^t, y^s; \theta\right) \tag{2}$$

Given an style transferred sentence $\hat{y}^t = (y_1, \ldots, y_n)$, where n is its sentence length, y_i represent i-th token in $\hat{y}^t$. BERT is trained on a large-scale corpus to generate style-independent and robustness representation, so we first use BERT to encode $\hat{y}^t$ into representation $H_B = (b_1, \ldots, b_n)$. Then the encoder encodes the other input y^s into $H_E^l = (h_1^l, \ldots, h_m^l)$, where h_i^l is the i-th representation of the l-th layer in the encoder. Next, the representation of BERT and encoder are combined through attention mechanism. We calculate h_i^l as follows:

$$h_i^l = \frac{1}{2}\left(H_B^l + H_E^l\right) \tag{3}$$

$$H_E^l = \mathrm{Attn}_E\left(h_i^{l-1}, H_E^{l-1}, H_E^{l-1}\right) \tag{4}$$

$$H_B^l = \mathrm{Attn}_B\left(h_i^{l-1}, H_B, H_B\right) \tag{5}$$

where $Attn$ is a self-attention model, which works as follows, note that we leverage the transform in self-attention to ensure H_B^l and H_E^l are the same dimensions:

$$\mathrm{Attn}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

We utilize an attention strategy to incorporate the pre-trained model into the original transformer-based text style transfer model, which can adequately leverage the information of the source sentence and the style transferred sentence.

Dataset	Train	Valid	Test
GYAFC (E&M)	52k	2877	1416
GYAFC (F&R)	52k	2788	1432
MTFC	14280k	2877	1416

Table 2: The number of training sentences in each dataset.

4 Experiments

4.1 Setup

We use two datasets to evaluate our proposed method, the size of each training dataset is presented in Table 2.

MTFC. Machine Translation Formality Corpus (MTFC) is built by selecting bilingual subtitle parallel data and extending the GYAFC dataset. MTFC dataset contains 14 million informal Chinese-English sentence pairs.

GYAFC. We use Grammarly's Yahoo Corpus Dataset (GYAFC) as our parallel data for text style transfer model training.

Settings

We choose Transformer [Vaswani *et al.*, 2017] as the architecture of NMT model and style transfer model, the texts are preprocessed with Byte Pair Encoding (BPE) [Sennrich *et al.*, 2016] with a vocabulary size of 25,000. We used Fairseq [Ott *et al.*, 2019] library to implement our experiments. The dimensionality of all input and output layers is 1024, and that of FFN layer is 4096. Both the encoder and decoder have 6 layers with 8 attention heads. For formality transfer model (a teacher model), we follow the work of [Chawla and Yang, 2020], and use BART-large [Lewis *et al.*, 2019] model pre-trained on CNN-DM summarization [Nallapati *et al.*, 2016] data as our initialized parameters. The data-refinement module is pretrained on grammatical error correction (GEC) data and style transfer data. Besides, we expand these parallel data to triplets by adding the noise sentences generated by imperfect style transfer model or GEC model. We train our models with Adam [Kingma and Ba, 2015] optimizer using β_1=0.9 β_2=0.98 on 2 NVIDIA 2080Ti GPUs.

Baselines

We compare our stylized NMT model with the following models, which is proposed in [Wu *et al.*, 2020], namely:

- **Base-model**: This method directly uses a transformer-based NMT model to evaluate on stylized test sets.
- **Pivot-rule**: This method first translates informal Chinese sentences as informal English sentences with NMT model. Then, it rewrites the generation results with several effective rules (Rules include capitalization, lowercase words with all upper characters, etc.).
- **Pivot-model**: This method first trains an informal NMT model and a formality style transfer model, then the two models decode source sentences in a pipeline manner.
- **Teacher-student**: For this method, it first trains a formality style transfer model that serves as a teacher model, then the student model is trained on the data generated by the teacher model.

	MTFC			
Model	Formality	Fluency	BLEU	Human
Base-model	0.557	3.57	33.47	1.70
Pivot-rule	0.756	3.35	37.83	2.05
Pivot-model	0.853	3.51	38.75	1.83
T-S	0.846	3.34	40.07	2.22
BT	0.785	3.26	40.68	2.28
IDKT	0.865	3.64	43.67	2.25
IDKT+Refine	**0.897**	3.66	**45.58**	**2.32**

Table 3: The performance of NMT model on MTFC corpus evaluated using automatic and human evaluation, “Formality” shows the accuracy of formality style transfer, “BLEU” reflects the overall quality of the output. “ T-S” and “BT” denote Teacher-student and Back-Translation respectively. “+Refine” means incorporating data-refinement module in our framework.

- **Back-translation**: This method trains two back-directional models, including a target-style to source-style model and a model that translates source-style English sentences to source-style Chinese ones. Then the synthetic-parallel data are created by using a limited parallel corpus and a large-scale non-parallel corpus.

We also compare our text style transfer model with previous state-of-the-art works:

- **Hybrid** [Xu *et al.*, 2019]: This method uses classification feedback and reconstruction constraints to train the model.
- **Multi-Task** [Niu *et al.*, 2018]: This method combines the training data of GYAFC corpus with data from machine translation to train their model with a multi-task learning schema.
- **GPT-CAT** [Wang *et al.*, 2019]: This method uses a pre-trained GPT2 model and combines source sentences with rule-based processed sentences to train the model.
- **BART+tune** [Chawla and Yang, 2020]: This method finetunes BART on our data.
- **BART+LM** [Chawla and Yang, 2020]: This method trains a semi-supervised formality text style transfer model using language model discriminator and mutual information maximization.

4.2 Evaluation Metrics

To evaluate different models, we apply three automatic metrics, including Formality, Fluency and Overall Evaluation, mostly following [Wu *et al.*, 2020].

Formality: We train a BERT-based classifier using the training data of the GYAFC by regarding this problem as a binary classification problem.

Fluency: We evaluate the fluency of translation sentences using a language model, which is trained on GYAFC training data by Kenlm[2]. Each sentence is scored from 0 to 4 by the language model according to the syntactic correctness.

	E&M			
Model	Formality	Fluency	BLEU	Human
Hybrid	0.782	3.44	69.63	1.92
Multi-Task	0.720	3.61	72.13	1.85
GPT-CAT	0.794	3.49	72.70	2.01
BART+tune	-	-	74.66	-
BART+LM	0.818	3.48	76.52	2.13
IDKT	0.778	3.59	76.64	2.08
IDKT+Refine	**0.855**	3.60	**77.03**	**2.21**

Table 4: The performance of text style transfer model on E&M domain of GYAFC corpus.

	F&R			
Model	Formality	Fluency	BLEU	Human
Hybrid	0.779	3.53	74.43	1.88
Multi-Task	0.753	3.68	75.37	2.04
GPT-CAT	-	-	76.87	-
BART+tune	-	-	78.89	-
BART+LM	**0.798**	3.57	79.92	2.24
IDKT	0.773	3.74	79.43	2.16
IDKT+Refine	0.794	3.73	**80.34**	**2.28**

Table 5: The performance of text style transfer model on F&R domain of GYAFC corpus.

Overall Evaluation: We evaluate the overall quality of formality-transferred sentences with BLEU [Papineni *et al.*, 2002].

Human Evaluation: Similar to [Wu *et al.*, 2020], at first, we randomly sample 300 output sentences for human evaluation, then three human annotators are required to score the overall quality of all model outputs from -3 to 3, denoted as: -3: very informal, -2: informal, -1: somewhat informal, 0: neutral, 1: somewhat Formal, 2: formal and 3: very formal.

4.3 Results

We show the results on MTFC corpus in Table 3. As can be seen, our proposed method IDKT+Refine significantly surpasses all previous methods, achieving improvements by up to 5 BLEU points compared to previous best results and gaining best score 0.897 on formality accuracy. The poor performance of the Base-model on test data indicates that a general NMT model is not applicable in this specific task. Although the pivot-based method can significantly improve the accuracy of formality style transfer and overall performance, it still suffers from the problem of error propagation. The teacher-student method achieves high score on formality accuracy, because the NMT model learns style knowledge from the teacher model. Compared to the pivot-based method, our method absorbs the benefit of teacher-student method and generate stylized translation in one model, which can reduce the number of model parameters and accelerate the decoding speed. The back-translation method produces good results in terms of BLEU, but the noise in synthetic-data of back-translation may harm formality accuracy.

[2] https://github.com/kpu/kenlm

Model	Formality	Fluency	BLEU
IDKT+refine	**0.897**	**3.66**	**45.58**
w/o BART	0.842	3.55	43.41
w/o Data-refinement	0.865	3.64	43.67
w/o Iterative	0.873	3.62	42.83

Table 6: Ablation study of our framework by removing some key components. Automatic evaluation metrics scores are reported, "w/o BART" means the text style transfer model without pre-trained BART parameters, 'w/o data-refinement' means we do not use data-refinement module to improve the quality of synthetic parallel data, 'w/o Iterative' means we just perform one-pass knowledge transfer.

Besides, as shown in Table 4 and Table 5, by utilizing the knowledge from NMT model, our formality style transfer model achieves more than 2 BLEU points improvement compared to the BART+tune baseline. After refining the synthetic parallel data with the data-refinement module, our method achieves consistent improvement in terms of BLEU and formality accuracy. Notably, our method only needs the synthetic data generated by NMT model but can still outperform previous baselines, indicating that the knowledge in NMT model is beneficial to text style transfer model.

We also conduct human evaluation, which reflects the superior performance of our NMT model and text style transfer model. Our method gets the highest scores in human evaluation, which is consistent with BLEU and formality accuracy.

4.4 Discussion

Ablation Study. We implement an ablation study to check the contributions of the key components in our proposed framework. The performance of our method is reported in row (1) of Table 6. Comparing row (1) and row (2), we can see that if the text style transfer model is not initialized with pre-trained BART, the performance would suffer a degradation of 2.1 BLEU points, proving that the style transfer model contains more style information when the model has better performance. Moreover, when we remove the data-refinement module, as shown in row (3), the BLEU score drops by 1.9 points, which shows that the data refinement module can filter out the noise of synthetic data effectively. When only performing one-pass knowledge transfer from text style transfer model to NMT model, as shown in row (4), the performance of the stylized NMT model decreases from 45.58 to 42.83, which indicates that corpora of machine translation and formality style transfer should be repeatedly used to fully utilize the knowledge in these datasets.

Effect of Iteration Number K. We conduct experiments with different iteration number K, on small batches of data that are sampled from MTFC, as shown in Figure 3. It can be observed that the performance of NMT model increases steadily during the first three iterations. However, the model performance does not change significantly when the number of iterations exceeds 3. So we set K=3 in our experiments.

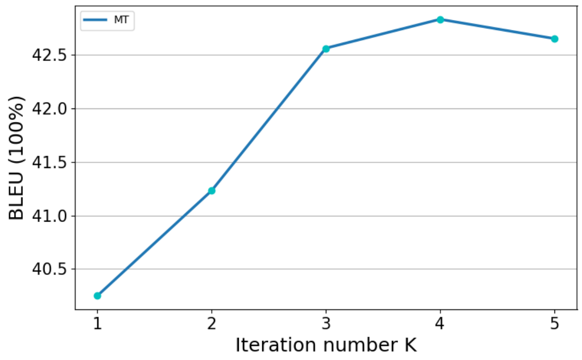

Figure 3: Effect of iteration number K

FST(BLEU)	Formality	Fluency	BLEU
66.4	0.745	3.52	41.32
69.3	0.793	3.48	42.46
72.3	0.855	3.62	43.67
75.6	0.863	3.60	45.21

Table 7: Results of formal NMT model using formality style transfer (FST) models that have different performance, the first column on the left is the BLEU scores of text style transfer model.

Effect of The Performance of Text Style Model. In this group of experiments, we investigate the impacts of text style transfer model on the performance of stylized NMT model. As shown in Table 7, the performance of the NMT model is related to the performance of the text style transfer model. When the performance of the latter increases, the performance of the former also increases. A better text style transfer model may contain more style knowledge, which enables the NMT model to achieve better performance after knowledge transfer. In other words, our bidirectional knowledge transfer method is beneficial to both the stylized NMT model and the text style transfer model.

5 Conclusion

In this paper, we propose an iterative dual knowledge transfer framework for producing coherent and stylized translations. Our method can perform bidirectional knowledge transfer between translation model and text style transfer model iteratively to fully exploit the knowledge and reinforce each other. We further leverage a data-refinement module to refine the low-quality synthetic data. Experiments on formality style translation demonstrate the effectiveness of our method, achieving obvious improvements over previous work.

In the future, we will explore how to effectively leverage unpaired monolingual data in our method. We also prepare to apply our framework to other text generation tasks.

Acknowledgments

The research work descried in this paper has been supported by the National Key R&D Program of China (2020AAA0108001) and the National Nature Science Foundation of China (No. 61976015, 61976016, 61876198 and 61370130). The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve this paper.

References

[Chawla and Yang, 2020] Kunal Chawla and Diyi Yang. Semi-supervised formality style transfer using language model discriminator and mutual information maximization. *In EMNLP*, 2020.

[Chen *et al.*, 2017] Y. Chen, Y. Liu, Y. Cheng, and V. Li. A teacher-student framework for zero-resource neural machine translation. *In ACL*, 2017.

[Cohn and Lapata, 2007] Trevor Cohn and Mirella Lapata. Machine translation by triangulation: Making effective use of multi-parallel corpora. *In ACL*, 2007.

[Correia and Martins, 2019] Gonalo M Correia and André F. T Martins. A simple and effective approach to automatic post-editing with transfer learning. *In ACL*, 2019.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *In NAACL*, 2018.

[Fu *et al.*, 2018] Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. Style transfer in text: Exploration and evaluation. *In AAAI*, 2018.

[Jhamtani *et al.*, 2017] Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. Shakespearizing modern language using copy-enriched sequence to sequence models. In *Proceedings of the Workshop on Stylistic Variation*, pages 10–19, September 2017.

[Junczys-Dowmunt and Grundkiewicz, 2018] Marcin Junczys-Dowmunt and Roman Grundkiewicz. Ms-uedin submission to the wmt2018 ape shared task: Dual-source transformer for automatic post-editing. *In WMT*, 2018.

[Kim and Rush, 2016] Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. *In EMNLP*, 2016.

[Kingma and Ba, 2015] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *In ICLR*, 2015.

[Lewis *et al.*, 2019] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *In ACL*, 2019.

[Li *et al.*, 2018] Juncen Li, Robin Jia, He He, and Percy Liang. Delete, retrieve, generate: A simple approach to sentiment and style transfer. *In NAACL*, 2018.

[Michel and Neubig, 2018] Paul Michel and Graham Neubig. Extreme adaptation for personalized neural machine translation. *In ACL*, 2018.

[Nallapati *et al.*, 2016] Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. Abstractive text summarization using sequence-to-sequence rnns and beyond. *In CoNLL*, 2016.

[Niu and Carpuat, 2020] Xing Niu and Marine Carpuat. Controlling neural machine translation formality with synthetic supervision. *In AAAI*, 2020.

[Niu *et al.*, 2018] Xing Niu, Sudha Rao, and Marine Carpuat. Multi-task neural models for translating between styles within and across languages. *In COLING*, 2018.

[Ott *et al.*, 2019] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. *In NAACL*, 2019.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002.

[Rabinovich *et al.*, 2016] Ella Rabinovich, Shachar Mirkin, Raj Nath Patel, Lucia Specia, and Shuly Wintner. Personalized machine translation: Preserving original author traits. *In EACL*, 2016.

[Rao and Tetreault, 2018] Sudha Rao and Joel Tetreault. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. *In NAACL*, 2018.

[Sennrich *et al.*, 2016] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *In ACL*, 2016.

[Shen *et al.*, 2017] Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. Style transfer from non-parallel text by cross-alignment. *In NIPS*, 2017.

[Sudhakar *et al.*, 2019] Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. Transforming delete, retrieve, generate approach for controlled text style transfer. *In EMNLP*, 2019.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, 2017.

[Wang *et al.*, 2019] Yunli Wang, Yu Wu, Lili Mou, Zhoujun Li, and Wenhan Chao. Harnessing pre-trained neural networks with rules for formality style transfer. *In EMNLP*, 2019.

[Wu *et al.*, 2020] Yu Wu, Yunli Wang, and Shujie Liu. A dataset for low-resource stylized sequence-to-sequence generation. *In AAAI*, 2020.

[Xu *et al.*, 2019] Ruochen Xu, Tao Ge, and Furu Wei. Formality style transfer with hybrid textual annotations. *arXiv preprint arXiv:1903.06353*, 2019.

[Zhang *et al.*, 2020] Yi Zhang, Tao Ge, and Xu Sun. Parallel data augmentation for formality style transfer. *In ACL*, 2020.

[Zhu *et al.*, 2020] Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. Incorporating bert into neural machine translation. *In ICLR*, 2020.