ST^2: Small-data Text Style Transfer via Multi-task Meta-Learning

Xiwen Chen, Kenny Q. Zhu
Advanced Data and Programming Technologies Lab
Shanghai Jiao Tong University
{victoria-x@sjtu, kzhu@cs.sjtu}.edu.cn

Abstract

Text style transfer aims to paraphrase a sentence in one style into another style while preserving content. Due to lack of parallel training data, state-of-art methods are unsupervised and rely on large datasets that share content. Furthermore, existing methods have been applied on very limited categories of styles such as positive/negative and formal/informal. In this work, we develop a meta-learning framework to transfer between any kind of text styles, including personal writing styles that are more fine-grained, share less content and have much smaller training data. While state-of-art models fail in the few-shot style transfer task, our framework effectively utilizes information from other styles to improve both language fluency and style transfer accuracy.

1 Introduction

Text style transfer aims at rephrasing a given sentence in a desired style. It can be used to rewrite stylized literature works, generate different styles of journals or news (e.g., formal/informal), and to transfer educational texts with specialized knowledge for education with different levels.

Due to lack of parallel data for this task, previous works mainly focused on unsupervised learning of styles, usually assuming that there is a substantial amount of nonparallel corpora for each style, and that the contents of the two corpora do not differ significantly (Shen et al., 2017; John et al., 2018; Fu et al., 2018). Existing state-of-art models either attempt to disentangle style and content in the latent space (Shen et al., 2017; John et al., 2018; Fu et al., 2018), directly modifies the input sentence to remove stylized words (Li et al., 2018), or use reinforcement learning to control the generation of transferred sentences in terms of style and content (Wu et al., 2019a; Luo et al., 2019). However, most of the approaches fail on low-resource datasets based on our experiments. This calls for new few-shot style transfer techniques.

The general notion of style is not restricted to the heavily studied sentiment styles, but also writing styles of a person. However, even the most productive writer can't produce a fraction of the text corpora commonly used for unsupervised training of style transfer today. Meanwhile, in real world, there exists as many writing styles as you can imagine. Viewing the transfer between each pair of styles as a separate domain-specific task, we can thus formulate a multi-task learning problem, each task corresponding to a pair of styles. To this end, we apply a meta-learning scheme to take advantage of data from other domains, i.e., other styles to enhance the performance of few-shot style transfer (Finn et al., 2017).

Moreover, existing works mainly focus on a very limited range of styles. In this work, we take both personal writing styles and previously studied general styles, such as sentiment style, into account. We test our model and other state-of-the-art style transfer models on two datasets, each with several style transfer tasks with small training data, and verify that information from different style domains used by our model enhances the abilities in content preservation, style transfer accuracy, and language fluency.

Our contributions are listed as follows:

- We show that existing state-of-the-art style transfer models fail on small training data which naturally shares less content (see Section 3.3 and Section 3.4).
- We propose Multi-task Small-data Text Style Transfer (ST^2) algorithm, which adapts meta-learning framework to existing state-of-art models, and this is the first work that applies meta-learning on text style transfer to the best of our knowledge (see Section 2).

- The proposed algorithm substantially outperforms the state-of-the-art models in the few-shot text style transfer in terms of content preservation, transfer accuracy and language fluency (see Section 3).

- We create and release a literature writing style transfer dataset, which the first of its kind (see Section 3.1).

2 Approach

In this section, we first present two simple but effective style transfer models, namely *CrossAlign* (Shen et al., 2017) and *VAE* (John et al., 2018) as our base models, and then present a meta-learning framework called model-agnostic meta-learning (Finn et al., 2017) that incorporates the base models to solve the few-shot style transfer problem.

2.1 Preliminaries

Cross Align

The *CrossAlign* model architecture proposed by Shen et al. (2017) is shown in Figure 1. Let X and Y be two corpora with styles s_x and s_y, respectively. E and D are encoders and decoders that take both the sentence x or y, and their corresponding style labels s_x or s_y as inputs. Then the encoded sentences z_x and z_y, together with their labels are input to two different adversarial discriminators D_1 and D_2, which are trained to differentiate between logits generated by the concatenation of content embedding and the original/opposite style label.

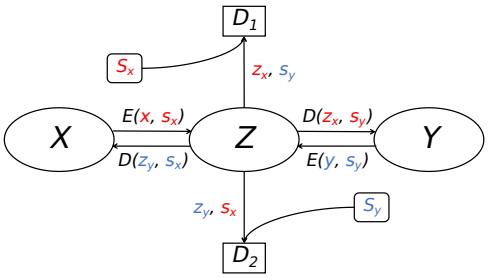

Figure 1: CrossAlign architecture

In training phase, the discriminators and the seq2seq model are trained jointly. The objective is to find

$$\theta^* = \arg\min_{\theta} \mathcal{L}_{\text{rec}}(\theta_E, \theta_D) + \mathcal{L}_{\text{adv}}(\theta_E, \theta_D),$$

where

$$\mathcal{L}_{\text{adv}}(\theta_E, \theta_D) = \mathbb{E}_{x\sim X}[-\log D(E(x, s_x))] + \mathbb{E}_{y\sim Y}[-\log(1 - D(E(y, s_y)))].$$

The discriminators are implemented as CNN classifiers (Kim, 2014).

VAE for Style Transfer

In order to disentangle style and content in the latent space, John et al. (2018) used variational autoencoder (VAE) and their specially designed style-oriented and content-oriented losses to guide the updates of the latent space distributions for the two components (Kingma and Welling, 2013).

The architecture of this model is shown in Figure 2. Given a corpus X with unknown latent style space and content space, an RNN encoder maps a sequence x into the latent space, which defines a distribution of style and content (Cho et al., 2014). Then style embedding and content embedding are sampled from their corresponding latent distributions and are concatenated as the training sentence embedding.

The two embeddings are used to calculate multi-task loss J_{mul} and adversarial loss J_{adv} for content and style to separate their information. Then this concatenated latent vector is used as a generative latent vector, and is concatenated to every step of the input sequence and fed into decoder D, which reconstructs the sentence x'. The final loss is the sum of these multi-task losses and the usual VAE reconstruction J_{rec} with KL divergence for both style embedding and content embedding (Kingma and Welling, 2013).

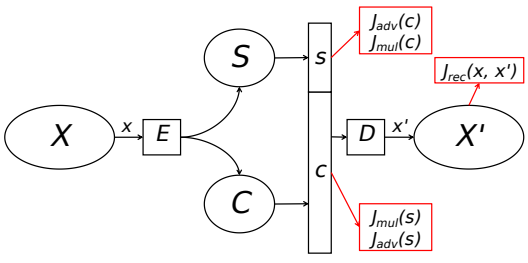

Figure 2: VAE architecture

The main designs of style- and content-oriented losses are as follows (John et al., 2018).

1. The style embedding should contain enough information to be discriminative. Therefore, a multitask discriminator is added to align

the predicted distribution and the ground-truth distribution of labels.

$$J_{\text{mul}}(\theta_E; \theta_{\text{mul(s)}}) = -\sum_{l \in \text{labels}} t_s(l) \log y_s(l),$$

where $t_s(l)$ is the distribution of ground-truth style labels, and $y_s(l)$ is the predicted output by the style discriminator.

2. The content embedding should not contain too much style information. Therefore, an adversarial discriminator is added, with loss of the discriminator and adversarial loss for the autoencoder given by

$$J_{\text{dis}(s)}(\theta_{\text{dis}(s)}) = -\sum_{l \in \text{labels}} t_s(l) \log y_s(l),$$
$$J_{\text{adv(s)}}(\theta_E) = -\sum_{l \in \text{labels}} y_s(l) \log y_s(l),$$

where $\theta_{\text{dis}(s)}$ contains the weights for a fully connected layer, and $t_c(l)$ is the predicted distribution of style labels when taking content embedding as an input.

3. The content embedding needs to be able to predict the information given by bag-of-words (BoW), which is defined as

$$t_c(w) := \frac{\sum_{i=1}^{N} \mathbb{I}\{w_i = w\}}{N},$$

for each word w in the vocabulary V with sentence length N (Wallach, 2006). Therefore, a multitask discriminator is added to align the predicted BoW distribution with ground-truth.

$$J_{\text{mul}(c)}(\theta_E; \theta_{\text{mul}(c)}) = -\sum_{w \in V} t_c(w) \log y_c(w),$$

where $t_c(w)$ is the distribution of true BoW representations, and $y_c(w)$ is the predicted output by the content discriminator.

4. The style embedding should not contain content information. Similar as before, an adversarial discriminator is trained to predict the BoW features from style embedding, with loss for discriminator and adversarial loss given by

$$J_{\text{dis}(c)}(\theta_{\text{dis}(c)}) = -\sum_{w \in V} t_c(w) \log y_c(w),$$
$$J_{\text{adv}(c)}(\theta_E) = -\sum_{w \in V} y_c(l) \log y_c(l).$$

In the training phase, the adversarial discriminators are trained together with other parts of the model, and the final loss of the autoencoder is given by the weighted sum of the loss from traditional VAE, the multitask losses for style and content, and the adversarial losses given by the style and content discriminators. Then in the inference phase, the style embedding is extracted from the latent space of a target domain, and the original style embedding is substituted by this target embedding in decoding.

2.2 Model-Agnostic Meta-learning (MAML)

Meta-learning is designed to help a model quickly adapt to a new tasks, given that it has been trained on several other similar tasks. Compared with other model-based meta-learning methods, model-agnostic meta-learning algorithm (MAML) utilizes only gradient information (Finn et al., 2017). Therefore, it can be easily applied to models based on gradient descent training.

Given a distribution of similar tasks $p(\mathcal{T})$, a task-specific loss function $\mathcal{L}_{\mathcal{T}_i}$ and shared parameters θ, we aim to jointly learn a model so that in fine-tuning with the new task, the parameters are well-initialized so that the model quickly converges with fewer epochs and a smaller dataset.

Figure 3 shows the architecture of MAML. We define the shared model with parameters θ as a meta-learner. The data for each task is divided in to a support set D_s and a query set D_q. Every update of the meta-learner's parameters consist of K-step updates for each of the N tasks. The support set for each task is used to update the N sub-tasks, and the query set is used to evaluate a query loss that is later used for meta-learner's updates.

In each sub-task training, the sub-learner is initialized with the parameters of the meta-learner. Then this parameter is updated K times using the support data for this specific task. After updating, the new parameter is θ_i' for the i-th task, and a loss $\mathcal{L}_{\mathcal{T}_i}(f_{\theta'})$ is evaluated using the query dataset for this sub-task. This sub-training process is performed for each sub-task, and losses from all sub-tasks are aggregated to obtain a loss for meta-training.

In our application, the sub-tasks contain different pairs of styles to be transferred. The meta-learner contains the transfer function $f_\theta : (x, s) \mapsto x'$, which takes a sentence x with its style label s, and outputs a sentence x' in the target style with sim-

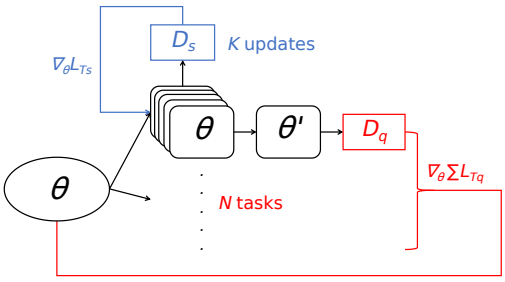

Figure 3: MAML architecture. For every update of meta-learner's parameters θ, we first update each sub-task on the support dataset D_s for K steps and obtain the new parameter θ'. Then we use the loss evaluated using this new parameter on the query set D_q, and sum up all losses from N tasks to update meta-learner's parameters.

Algorithm 1: ST^2

Input: a set of style pairs, $\{(s_{t,1}, s_{t,2}), \ldots\}$, where $t = 1, \ldots, N$, parameters α, β

Output: transfer function $f_\theta : (x, s) \mapsto y$, where s is the source style, x is the original sentence, y is the transferred sentence in target style

while *not done* **do**
  foreach *style pair* $(s_{t,1}, s_{t,2})$ **do**
    Initialize sub learner with $\theta_t = \theta$;
    for *step in 1, ..., K* **do**
      Sample batch data from support set of t;
      Update transfer function f_θ using $\theta_t = \theta_t - \alpha \nabla_{\theta_t} \mathcal{L}_t(f_{\theta_t})$;
    end
    Sample batch data from query set of t;
    Evaluate $\mathcal{L}_t(f_{\theta_t})$;
  end
  Update meta-learner with $\theta = \theta - \beta \nabla_\theta \sum_{t=1}^{T} \mathcal{L}_t(f_{\theta_t})$;
end

ilar content. This transfer function is shared by all pairs of styles in the meta-training phase. In addition, both our base models include adversarial functions for style disentanglement, the updates for the adversarial parameters are also included in the updates of meta-learner. Since the data size for each task with a single pair of styles is assumed to be small, the goal of MAML is to use information from other style pairs for a better initialization in the fine-tuning phase of a specific sub-task. The multi-task style transfer via meta-learning (ST^2) algorithm is described in Algorithm 1.

3 Experiments

In order to incorporate a more diverse range of styles, we gather two datasets for our experiments. The first is collected from literature translations with different writing styles, and the second is a grouped standard dataset used for existing style transfer works, which also contains different types of styles.

We test our ST^2 model and state-of-the-art models on these two datasets, and verify our model's effectiveness on few-shot style transfer scheme. By comparing our models with the pretrained base models, we verify that meta-learning framework improves the performance both in terms of language fluency and style transfer accuracy.

3.1 Datasets

Since we extend the definition of style to the general writing style of a person, we do not need to be limited to the widely used Yelp/Amazon review and GYAFC datasets. To model the real situation where we have different style pairs with not enough data for each style pair, we propose to use the literature translations dataset and a set of popular style transfer datasets with reduced sizes.

Literature Translations (LT)

Current state-of-the-art works on text style transfer require large datasets for training, and thus they are not able to be applied to personal writing styles. One reason is that personal writing styles are relatively difficult to learn, compared with more discriminative styles such as sentiment and formality. Furthermore, sources of data reflecting personal writing styles are quite limited.

For the reasons above, we consider literature translations dataset. Firstly, there are multiple versions of translation from the same source. Since it is possible to align these comparable sentences to construct ground-truth references, they are well-suited for our test data. Moreover, in addition to the common-source translated work, a translator has other written works, which can be used for our non-parallel training data.

We collect a set of writers with unknown different writing styles $\{s_1, \ldots, s_n\}$, with each writer has his/her own set of written works $\{c_{s_i,1}, \ldots, c_{s_i,n_i}\}$. In order to have a test set with ground-truth references, we used translated works from non-English sources[1], so that each writer in our set has at least one translated work that is from the same source as another writer. Namely, for each writing style s_i in the set, there exists another style s_j and $\exists\ k_1, k_2$ such that $src(c_{s_i,k_1}) = src(c_{s_j,k_2})$. In this dataset, each writer has approximately 10k

[1] Obtained from `http://gen.lib.rus.ec/`.

Common Source	Writer A	Writer B
Notre-Dame de Paris	Alban Kraisheimer	Isabel F. Hapgood
The Brothers Karamazov	Andrew R. MacAndrew	Richard Pevear
The Story of Stone	David Hawkes	Yang Xianyi
The Magic Mountain	John E. Woods	H. T. Lowe-Porter
The Iliad	Ian C. Johnston	Robert Fagles
Les Miserables	Isabel F. Hapgood	Julie Rose
Crime and Punishment	Michael R. Katz	Richard Pevear

Table 1: Literature translations dataset. The first column shows the name of translated works with common source for the two writers in the same row.

nonparallel sentences for training.

We used the aligned sentences for each style pair using the algorithm provided by Chen et al. (2019) for testing. The sentence pairs are extracted from the common translated work for each writer pair. The test data has approximately 1k sentences for each writer. More information is shown in Table 1.

Grouped Standard Datasets (GSD)

In our second set, we group popular datasets for style transfer. For large datasets, we use only a small portion of then in order to model our few-shot style transfer task. The datasets we use are listed in Table 2. For the standard/simple versions of Wikipedia, we use the aligned sentences by Hwang et al. (2015) for testing. For all datasets listed in the table, we use 10k sentences for training and 1k sentences for testing.

Dataset	Style
Yelp	(health) positive/negative
Amazon	(musical instrument) positive/negative
GYAFC	(relations)formal/informal
Wikipedia	standard/simple
Bible	standard/easy
Britannica	standard/simple
Shakespeare	original/modern

Table 2: Grouped dataset.

3.2 Metrics

BLEU for Content Preservation

To evaluate content preservation of transferred sentences, we use a multi-BLEU score between reference sentences and generated sentences (Papineni et al., 2002). When ground-truth sentences are available in the dataset, we calculate the BLEU scores between generated sentences and ground-truth sentences. When they are missing, we calculate self-BLEU scores based on the original sentences[2].

Perplexity (PPL)

Following the metrics used by John et al. (2018), we use a bigram Kneser-Key bigram language model to evaluate the fluency and naturalness of generated sentences (Kneser and Ney, 1995). The language models are trained in the target domain for each style pair. We use the training data before reduction to train the language model for GSD set.

Transfer Accuracy (ACC)

To evaluate the effectiveness of style transfer, we pretrain a TextCNN classifier proposed by Kim (2014). The transfer accuracy is the score output by the CNN classifier. Our classifier achieves accuracy of 80% on GSD and 77% even on LT dataset, which serves as a reasonable evaluator for transfer effectiveness.

Human Evaluation of Fluency and Content

We conduct an additional human evaluation, following Luo et al. (2019). Two native English speakers are required to score the generated sentences from 1 to 5 in terms of fluency, naturalness, and content preservation, respectively. Before annotation, the two evaluators are given the best and worst sentences generated so as to know the upper and lower bound, and thus score more linearly. The final score for each model is calculated as the average score given by the annotators. The kappa inter-judge agreement is 0.769, indicating significant agreement.

3.3 Multi-task Style Transfer

We compare the results with the state-of-the-art models for the style transfer task. All the baseline models are trained on the single style pair. The ST[2] model is trained on all the tasks for both LT and GSD sets, and then fine-tuned using a specific

[2]We use BLEU score provided by `multi-bleu.perl`

Model	LT					GSD				
	B-ref$^{\uparrow}$	**B-ori**	**PPL**$^{\downarrow}$	**ACC**$^{\uparrow}$	**Human**$^{\uparrow}$	**B-ref**$^{\uparrow}$	**B-ori**	**PPL**$^{\downarrow}$	**ACC**$^{\uparrow}$	**Human**$^{\uparrow}$
Template	**41.6**	81.48	**5.4**	0.31	**4.3** / **4.2**	**81.7**	88.8	**5.3**	0.42	4.2 / **4.2**
CrossAlign	2.2	2.1	1895.6	0.45	1.2 / 1.1	2.7	2.2	1049.7	0.36	1.0 / 1.0
DeleteRetrieve	**35.9**	41.6	63.3	0.33	1.0 / 1.0	20.5	21.4	28.8	0.41	2.1 / 1.3
DualRL	4.1	3.9	1400.7	0.49	1.2 / 1.2	25.4	27.5	171.0	0.41	2.9 / 1.5
VAE	13.5	16.3	8.5	0.49	3.5 / 1.7	12.4	26.4	21.5	0.45	**4.3** / 2.1
ST^2-CA (ours)	6.3	6.8	54.8	**0.65**	3.1 / **2.3**	**66.7**	73.2	21.4	0.42	3.6 / **3.8**
ST^2-VAE (ours)	20.5	15.1	**8.2**	0.62	**3.8** / 1.9	14.7	13.9	**10.9**	**0.71**	**4.3** / 2.7

Table 3: Results for multi-task style transfer. The larger$^{\uparrow}$/lower$^{\downarrow}$, the better. B-ref and B-ori means BLEU score and self-BLEU score, respectively. The human evaluation scores include language fluency/content preservation, respectively. Our base models are underlined.

style pair in the sets. The trained meta-learner is fine-tuned on each of the sub-tasks, and the scores are calculated as the average among all sub-tasks for both ST^2 models and baselines. The results are shown in Table 3.

We note that the BLEU and PPL scores for the template based model appear to be superior to those of other models. This is because it directly modifies the original sentence by changing a couple of words. So the modification is actually minimum. However, its transfer accuracy suffers, which is well expected. Thus its should only serve as a reference in our task.

For qualitative analysis, we randomly select sample sentences output by the baseline models, pretrained base models and our ST^2 models on the Translations dataset and Yelp positive/negative review dataset, which are shown in Table 4.

From the results, we notice that state-of-the-art models fail to achieve satisfying performances in our few-shot style transfer task, and many baseline models fail to generate syntactically or logically consistent sentences. In comparison, our methods are able to generate relatively fluent sentences both in terms of automatic evaluation and human evaluation, meanwhile achieving a higher transfer accuracy.

We might be tempted to conclude that this is simply because the ST^2 models learn better language models because they are trained on larger data, i.e., data from all styles rather than only a single pair of styles. Therefore, further experiments are required.

3.4 Pretrained Base Models

Based on the previous reasoning, we extract and pretrain the language model part in our base models (*CrossAlign* and *VAE*) on the union of data from all sub-tasks. Starting with a well-trained language model, we then fine-tune the models for the style transfer task. By comparing these models with our ST^2 model, we verify that meta-learning framework can improve the style transfer accuracy in addition to language fluency. We perform this experiment only on the GSD dataset, since they are enough for analysis purposes.

In addition, to examine the effect of pretraining combined with meta-learning, we also add a pretraining phase to our ST^2 model. The quantitative and qualitative results are included in Table 5 and Table 4 (on Yelp dataset for the pretrained base models).

By adding a pretraining phase, the models get a chance to see all the data and learn to generate fluent sentences via reconstruction. Therefore, it is not surprising that the content preservation measure (BLEU) and sentence naturalness measure (PPL) give significantly better results than before but at a cost of style transfer accuracy.

In effect, the models tend to reconstruct the original sentence and do not transfer the style. In comparison, our ST^2 model learn to generate reasonable sentences and transfer styles jointly in the training phase. Therefore, it is still superior in terms of style transfer accuracy. This verifies that the success of ST^2 is not merely resulted from a larger training dataset. The way that the model updates its knowledge is parallel, rather than sequential, which contributes to better language models and more effective style transfer.

Furthermore, we notice that the pretraining phase in our ST^2 model is not crucial, suggesting that it is the meta-learning framework that significant contributes to the model's improvements in generating fluent sentences and effectively transferring styles.

Original Sentence (Notre-Dame de Paris)	*in their handsome tunics of purple camlet , with big white crosses on the front .*
Template	*in their handsome tunics of purple camlet, with big white crosses on front.*
CrossAlign	*heel skilful skilful skilful skilful*
DeleteRetrieve	*the man , and the man , the man , the*
DualRL	*lyres lyres lyres*
VAE	*the gypsy girl had stirred up from the conflict*
ST^2-CrossAlign (ours)	*so the spectacle who prayed and half white streets ,*
ST^2-VAE (ours)	*all four were dressed in robes of white and were white locks from*
Original Sentence (Yelp positive)	*the staff is welcoming and professional .*
Template	*the staff is welcoming and professional .*
CrossAlign	*glad glad glad*
CrossAlign (pretrained)	*the staff is welcoming and professional .*
DeleteRetrieve	*the staff is a time .*
DualRL	*less expensive have working .*
VAE	*the staff is rude and rude*
VAE (pretrained)	*the staff is extremely welcoming and professional .*
ST^2-CrossAlign (ours)	*the staff is friendly and unprofessional*
ST^2-VAE (ours)	*the staff are rude and unprofessional .*
Original Sentence (Yelp negative)	*these people do not care about patients at all !*
Template	*these people wonderful about patients at all !*
CrossAlign	*glad glad glad*
CrossAlign (pretrained)	*these people do not care about patients at all !*
DeleteRetrieve	*i was n't be a a appointment and i have .*
DualRL	*and just like that it was over and i was .*
VAE	*these people do not care about patients or doctors*
VAE (pretrained)	*these guys do n't care about the patients at time*
ST^2-CrossAlign (ours)	*these people do not satisfied at all !*
ST^2-VAE (ours)	*i was so happy and i did n't consent*

Table 4: Randomly selected sample outputs for the Alban Kraisheimer/Isabel F. Hapgood pair in LT dataset and Yelp positive/negative review dataset.

Model	**BLEU**$^\uparrow$	**PPL**$^\downarrow$	**ACC**$^\uparrow$	**Human**$^\uparrow$
CA (pre.)	**70.4**	12.2	0.32	3.9
VAE (pre.)	17.2	22.4	0.48	4.0
ST^2-CA (pre.)	62.7	23.2	0.37	3.7
ST^2-VAE (pre.)	13.6	10.9	0.66	4.2
ST^2-CA (ours)	66.7	21.4	0.42	3.6
ST^2-VAE (ours)	14.7	**10.9**	**0.71**	**4.3**

Table 5: Results on GSD for pretrained (pre.) base models (CrossAlign abbreviated as CA) and ST^2.

3.5 Disentanglement of Style

Following the experiments adapted by John et al. (2018), we use t-SNE plots (shown in Figure 4) to analyze the effectiveness of disentanglement of style embedding and content embedding in the latent space (Maaten and Hinton, 2008). In particular, we compare the pretrained base models (*CrossAlign* and *VAE*) and our ST^2 models.

These two models, together with our ST^2 models attempt to disentangle style and content in latent space, and thus is well suited for this experiment, while it is unreasonable to treat hidden state vectors in other baseline models as content/style embedding. Therefore, they are excluded from this experiment.

As we can see from the figures, the content space learned by all models are relatively clustered, while the style spaces are more separated in our ST^2 models than the pretrained base models. This verifies that the improvements of meta-learning framework is not limited to a better language model, but also in terms of the disentanglement of styles.

4 Related Work

Fu et al. (2018) devised a multi-encoder and multi-embedding scheme to learn a style vector via adversarial training. Adapting a similar idea, Zhang et al. (2018) built a shared private encoder-decoder model to control the latent content and style space. Also based on a seq2seq model, Shen et al. (2017) proposed a cross-align algorithm to align the hidden states with a latent style vector from target domain using teacher-forcing. More recently, John et al. (2018) used well-defined style-oriented and

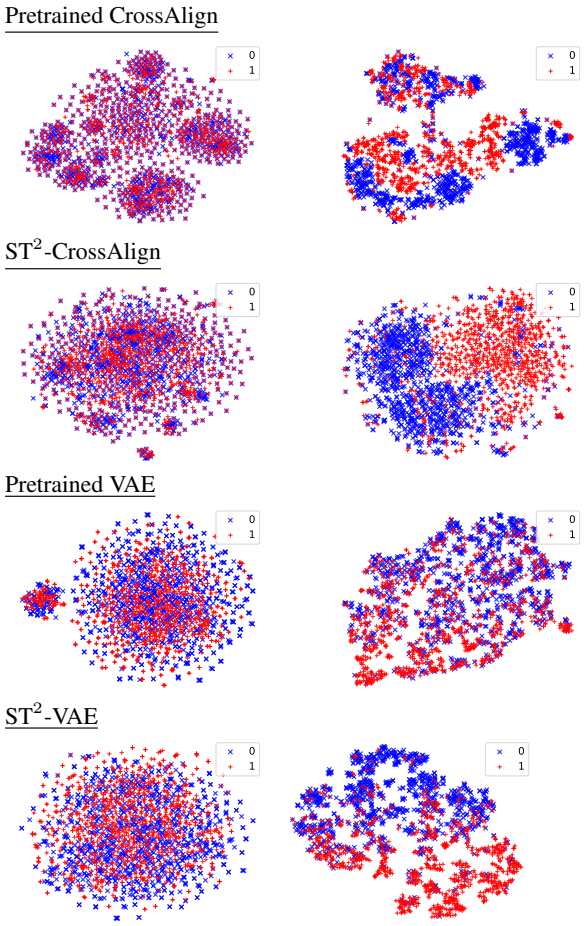

Figure 4: t-SNE plots for content(left) and style(right) embedding

content-oriented losses based on a variational autoencoder to separate style and content in latent space.

Li et al. (2018) directly removed style attribute words based on TF-IDF weights and trained a generative model that takes the remaining content words to construct the transferred sentence. Inspired by the recent achievements of masked language models, Wu et al. (2019b) used an attribute marker identifier to mask out the style words in source domain, and trained a "infill" model to generate sentences in target domain.

Based on reinforcement learning, Xu et al. (2018) proposed a cycled-RL scheme with two modules, one for removing emotional words (neutralization), and the other for adding sentiment words (emotionalization). Wu et al. (2019a) devised a hierarchical reinforced sequence operation method using a *point-then-operate* framework, with a high-level agent proposing the position in a sentence to operate on, and a low-level agent altering the proposed positions. Luo et al. (2019) proposed a dual reinforcement learning model to jointly train the transfer functions using explicit evaluations for style and content as a guidance. Although their methods work well in large datasets such as Yelp (Asghar, 2016) and GYAFC (Rao and Tetreault, 2018), it fails in our few-shot style transfer task.

Prabhumoye et al. (2018) adapted a back-translation scheme in an attempt to remove stylistic characteristics in some intermediate language domain, such as French.

There are also meta-learning applications on text generation tasks. Qian and Yu (2019) used the model agnostic meta-learning algorithm for domain adaptive dialogue generation. However, their task has paired data for training, which is different from our task. In order to enhance the content-preservation abilities, Li et al. (2019) proposed to first train an autoencoder on both source and target domain. But in addition to utilizing off-domain data, we are applying meta-learning method to enhance models' performance both in terms of language model and transfer abilities.

5 Conclusion

In this paper, we extend the concept of style to general writing styles, which naturally exist as many as possible but with a limited size of data. To tackle this new problem, we propose a multi-task style transfer (ST^2) framework, which is the first of its kind to apply meta-learning to small-data text style transfer. We use the literature translation dataset and the augmented standard dataset to evaluate the state-of-the-art models and our proposed model.

Both quantitative and qualitative results show that ST^2 outperforms the state-of-the-art baselines. Compared with state-of-the-art models, our model does not rely on a large dataset for each style pair, but is able to effectively use off-domain information to improve both language fluency and style transfer accuracy.

Noticing that baseline models might not be able to learn an effective language model from small datasets, which is a possible reason for their bad performances, we further eliminate the bias in our experiment by pretraining the base models using data from all tasks. From the results, we ascertain that the enhancement of meta-learning framework is substantial.

References

Nabiha Asghar. 2016. Yelp dataset challenge: Review rating prediction. *arXiv preprint arXiv:1605.05362*.

Xiwen Chen, Kenny Q. Zhu, and Mengxue Zhang. 2019. Aligning sentences between comparable texts of different styles. In *The 9th Joint International Semantic Technology Conference*.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

William Hwang, Hannaneh Hajishirzi, Mari Ostendorf, and Wei Wu. 2015. Aligning sentences from standard wikipedia to simple wikipedia. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 211–217.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2018. Disentangled representation learning for text style transfer. *arXiv preprint arXiv:1808.04339*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 181–184. IEEE.

Dianqi Li, Yizhe Zhang, Zhe Gan, Yu Cheng, Chris Brockett, Ming-Ting Sun, and Bill Dolan. 2019. Domain adaptive text style transfer. *arXiv preprint arXiv:1908.09395*.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: A simple approach to sentiment and style transfer. *arXiv preprint arXiv:1804.06437*.

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Zhifang Sui, and Xu Sun. 2019. A dual reinforcement learning framework for unsupervised text style transfer. *arXiv preprint arXiv:1905.10060*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. *arXiv preprint arXiv:1804.09000*.

Kun Qian and Zhou Yu. 2019. Domain adaptive dialog generation via meta learning. *arXiv preprint arXiv:1906.03520*.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. *arXiv preprint arXiv:1803.06535*.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in neural information processing systems*, pages 6830–6841.

Hanna M Wallach. 2006. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning*, pages 977–984. ACM.

Chen Wu, Xuancheng Ren, Fuli Luo, and Xu Sun. 2019a. A hierarchical reinforced sequence operation method for unsupervised text style transfer. *arXiv preprint arXiv:1906.01833*.

Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019b. " mask and infill": Applying masked language model to sentiment transfer. *arXiv preprint arXiv:1908.08039*.

Jingjing Xu, Xu Sun, Qi Zeng, Xuancheng Ren, Xiaodong Zhang, Houfeng Wang, and Wenjie Li. 2018. Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. *arXiv preprint arXiv:1805.05181*.

Ye Zhang, Nan Ding, and Radu Soricut. 2018. Shaped: Shared-private encoder-decoder for text style adaptation. *arXiv preprint arXiv:1804.04093*.