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

Anonymous NAACL-HLT 2021 submission

Abstract

Text style transfer aims to rephrase a sentence in one style into another style while preserving its semantic content. Due to the lack of parallel training data, most state-of-the-art neural methods resort to an unsupervised paradigm and rely on large-scale data in each domain of interest to sufficiently train the networks. Furthermore, existing methods have been applied to very limited categories of styles such as sentiment and formality. In this work, we adapt meta-learning to transfer between any kind of text styles, including a newly constructed dataset regarding personal writing styles, that are more fine-grained, share less content and have much smaller training data. Experimental results demonstrate that, while state-of-the-art models fail in the few-shot style transfer task, our framework effectively utilizes information from other styles to improve the overall transfer effectiveness.

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 adapt educational texts with specialized knowledge to learners at various levels.

Due to the lack of parallel data for this task, previous work (Shen et al., 2017; John et al., 2018; Fu et al., 2018) mainly focused on unsupervised learning of styles, usually assuming that there is a substantial amount of non-parallel corpora for each style and that the contents of the two corpora do not differ significantly. Existing state-of-the-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., 2019; Luo et al., 2019). However, most of the approaches fail to generate fluent sentences with the desired style on low-resource datasets based on our experiments (Section 3.2).

Moreover, existing work has been limited to a small range of discrete styles such as sentiment polarity and textual formality, with no evidence to show that they can be generalized to more challenging out-of-domain transfer tasks. In real-world scenarios, the general notion of style is not restricted to the heavily studied discrete attribute labels, but also includes the writing style 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. In the real world, there exists as many writing styles as you can imagine, making it impossible to train style transfer models tailored for each task from scratch.

By viewing the transfer between each pair of styles as a separate domain-specific task, we propose to formulate a multi-task learning problem where each task corresponds to the transfer between a pair of styles. Based on the multi-task formulation, we further 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). To extend the scope of text style transfer beyond the coarse-grained styles, we take both personal writing styles and previously studied general styles, such as sentiment style, into account. We apply our framework to several state-of-the-art style transfer models on two collections of datasets, each with several style transfer tasks with small training data, and verify that information from different style domains can be effectively utilized to enhance the abilities in content preservation, style transfer accuracy, and language fluency.

Our contributions are listed as follows:

- We create and release a literature writing style

transfer dataset, which is the first of its kind that captures more fine-grained stylistic characteristics of text rather than discrete style label (e.g., sentiment).

- We propose the Multi-task Small-data Text Style Transfer (ST^2) framework, which adapts meta-learning to enable flexible plug-in of existing state-of-the-art models, and this is the first work that applies meta-learning on text style transfer to the best of our knowledge.

- Experimental results demonstrate that the proposed algorithm substantially improves its base models in the few-shot text style transfer task for both traditional and our newly created datasets, in terms of content preservation, transfer accuracy and language fluency.

2 Approach

We first present the small-data text style transfer (ST^2) framework, which adapts meta-learning scheme to effectively exploit relevant knowledge from other style pairs, then introduce a newly constructed literature translation dataset covering a board range of fine-grained personal writing styles.

Algorithm 1: ST^2

Input: a set of N style pairs, $\{(s_{t,src}, s_{t,tgt}), \ldots\}$, where $t = 1, \ldots, N$, step sizes α, β
Output: transfer function $f_\theta : (x, s_{src}) \mapsto y$, 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* $(s_{t,src}, s_{t,tgt})$ **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
    $\theta = \theta - \beta \nabla_\theta \sum_{t=1}^{N} \mathcal{L}_t(f_{\theta_t})$;
end

2.1 Small Data Text Style Transfer

In contrast to traditional single style pair transfer, in our application, the sub-tasks contain different pairs of styles to be transferred. The meta-learner contains the transfer function $f_\theta : (x, s_{src}) \mapsto x'$, which takes a sentence x with its style label s_{src}, and outputs a sentence x' in the target style with similar content. This transfer function is shared by all pairs of styles in the meta-training phase. In practice, the transfer function f_θ can be parameterized by any existing single style pair neural transfer model. In addition, for base models which 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 ST^2 is to transfer knowledge 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.

2.2 Literature Translation (LT) Dataset Construction

We leverage literature translations by different translators as a new challenging text style transfer task. Because there are multiple versions of translation from the same source and it is possible to align these comparable sentences to construct ground-truth references, they are well-suited for style transfer. Moreover, in addition to a popular book translated by several other translators, a translator may have other written works, which can be used as a non-parallel training corpus as in standard style transfer setting. We align 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 1k sentences for each writer. While it is difficult to characterize different writing styles using discrete representation, we report some statistics regarding each writer's translation. For example, Ian C. Johnston tends to use more concise expressions(13.4 v.s. 26.6 tokens per sentence) than Robert Fagles even they possess similar vocabulary(11168 v.s. 13521) when translating *The Iliad*. See table 1 for more detailed statistics.

We collect a set of writers (from 1 through n) with unknown writing styles $\{s_1, \ldots, s_n\}$, each writer having his/her own set of written works W_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

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

Statistics	Alban Kraisheime	Isabel F.Hapgood	Andrew R. MacAndrew	Richard Pevear	David Hawkes	Yang Xianyi	John E. Woods	H. T. Lowe-Porter	Ian C. Johnston	Robert Fagles	Julie Rose	Michael R. Katz
Vocab Size	16,810	17,205	11,814	14,831	15,121	11,436	21,869	20,819	11,168	13,521	18,020	13,908
Average Length	20.9	20.4	19.5	20.7	18.6	15.9	25.1	26.7	13.5	26.6	20.0	19.0
# of Adjectives	11,798	11,187	10,050	11,364	9,622	7,581	17,521	19,187	8,236	15,246	11,255	11,283
# of Adverbs	9,462	8,325	14,848	13,580	10,622	9,555	14,667	17,369	7,901	13,991	9,362	13,506
# of Conjunctions	22,649	20,673	16,640	19,675	17,086	13,762	25,837	26,789	13,916	27,801	20,112	16,650
Flesch Readability	68.7	66.7	71.1	67.4	73.0	79.6	61.3	59.6	79.8	70.2	70.6	66.4
Dale-Chall Readability	6.2	6.3	5.2	5.7	5.4	4.9	6.5	6.7	5.2	6.1	5.8	5.9

Table 1: Linguistic statistics of each writer. The higher the Flesch readability, the easier it is to read. The Dale-Chall readability score indicates the grade level required to understand the text. See https://en.wikipedia.org/wiki/Dale-Chall_readability_formula for full description.

the set, there exists another style s_j and $\exists\, w_m \in W_i$ and $\exists\, w_n \in W_j$ such that $src(w_m) = src(w_n)$. In this dataset, each writer has approximately 10k non-parallel sentences for training.

3 Experiments

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 standard dataset.

3.1 Setup

We use the Literature Translation Datasets (LT) (see Section 2.2) and a grouped standard dataset (GSD) (see Table 2) (Li et al., 2018; Sudhakar et al., 2020; Rao and Tetreault, 2018) in the following experiments. For all datasets listed in Table 2, we use 10k sentences for training and 1k sentences for testing.

We use the following evaluation metrics:

- **BLEU-3.** We report the BLEU-3 score (Papineni et al., 2002) between references and model outputs.
- **Perplexity (PPL).** We use a Kneser-Ney bigram language model as measurement of fluency (Kneser and Ney, 1995). The language models are trained in the target domain for each style pair before reduction.
- **Transfer Accuracy (ACC).** We pretrain a RoBERTa (Liu et al., 2019) classifier for each style pair. It achieves test accuracy of 85.0% on LT and 83.9% on GSD on average.
- **Overall Performance.** The geometric mean(**G3**) and harmonic mean(**H3**) of BLEU-3, $\frac{1}{\log PPL}$ and ACC.
- **Human Evaluation (HE).** For each model with each transfer direction, we randomly sample 25 sentences for human evaluation. Each annotator(two native English speakers) is asked to assess the overall transfer effectiveness of each output sentence at a 4-point scale by jointly considering content preservation, transfer strength and language fluency, given the source sentence and the target style. For LT collection, we additionally provide annotators with writer-specific statistics as auxiliary information. The final score for each model is calculated as the average score given by the annotators. The kappa inter-judge agreement is 0.769.

We adopt the following as our base models (**CrossAlign** (Shen et al., 2017), **VAE** (John et al., 2018) and **CP-VAE** (Xu et al., 2019)) against: **DeleteRetrieve** (Li et al., 2018), **DualRL** (Luo et al., 2019), **B-GST**(Sudhakar et al., 2020).

3.2 Main Results

The results are shown in Table 3. From the results, we notice that state-of-the-art models fail to achieve satisfying performances in few-shot style transfer tasks, and many baseline models fail to generate syntactically or logically consistent sentences. However, even without the help of large-scale pretrained language model (e.g., GPT in B-GST), models equipped with ST^2 are able to generate more fluent sentences both in terms of automatic evaluation and human evaluation, meanwhile achieving a higher transfer accuracy. By inspecting evaluation results of all base models in Table 3 before and after being equipped with ST^2, we show that ST^2 is robust to the choice of base models as all evaluation metrics regarding transfer effectiveness obtain consistent improvement after applying ST^2 to the base models. It is worth noting that both CrossAlign and VAE are significantly inferior to the two state-of-the-art models B-GST and CP-VAE in terms

Model	LT						GSD					
	BLEU-3$^{\uparrow}$	PPL$^{\downarrow}$	ACC$^{\uparrow}$	G3$^{\uparrow}$	H3$^{\uparrow}$	HE$^{\uparrow}$	BLEU-3$^{\uparrow}$	PPL$^{\downarrow}$	ACC$^{\uparrow}$	G3$^{\uparrow}$	H3$^{\uparrow}$	HE$^{\uparrow}$
DeleteRetrieve	0.27	63.3	0.33	1.3	0.3	1.9	0.71	28.8	0.41	1.8	0.4	2.9
DualRL	0.01	1400.7	0.49	0.3	0.1	1.9	5.80	171.0	0.41	3.4	0.4	2.4
B-GST	0.56	24.4	0.50	1.8	0.4	1.9	15.62	31.1	0.36	4.8	0.6	1.9
CrossAlign	0.0	1895.6	0.45	0.0	0.0	1.8	0.0	1049.7	0.36	0.0	0.0	1.8
ST^2-CA	0.26	54.8	0.54	1.3	0.3	2.4	**17.6**	21.4	0.45	**5.3**	0.8	3.3
VAE	0.11	8.5	0.49	1.2	0.2	1.8	0.35	21.5	0.45	1.5	0.4	2.9
ST^2-VAE	0.34	8.2	**0.62**	1.9	0.5	3.3	0.80	10.9	**0.71**	2.4	0.6	3.2
CP-VAE	0.57	11.4	0.41	1.8	0.5	2.4	2.52	8.3	0.64	3.8	0.8	3.1
ST^2-CP-VAE	**0.71**	**8.1**	0.49	**2.3**	**0.7**	**3.4**	2.87	**4.6**	0.66	5.0	**1.6**	**3.5**

Table 3: Results for multi-task style transfer. The larger$^{\uparrow}$/lower$^{\downarrow}$, the better. Our base models are underlined.

of overall performance before the enhancement by ST^2. By incorporating related tasks into a unified transfer model, the learned parameters can be better adapted to one specific task with minimal amount of data. While ST^2-VAE yields better transfer strength and ST^2-CA achieves higher BLEU score on GSD, ST^2-CP-VAE learns to strike a balance and obtains the best overall performance. A similar phenomenon is also observed on LT. For qualitative analysis, we randomly select transferred sentences by baseline models, pretrained base models and ST^2 models and show them in the Appendix.

We might be tempted to conclude that this is simply because the ST^2 models learn better language models. Therefore, further experiments are required.

3.3 Pretrained Base Models

Model	BLEU-3$^{\uparrow}$	PPL$^{\downarrow}$	ACC$^{\uparrow}$	HE$^{\uparrow}$
CA*	**24.21**	12.2	0.32	1.9
VAE*	1.84	22.4	0.48	2.0
ST^2-CA*	14.62	23.2	0.37	2.2
ST^2-VAE*	0.95	10.9	0.66	2.9
ST^2-CA	17.60	21.4	0.45	3.3
ST^2-VAE	0.80	**10.9**	**0.71**	3.2

Table 4: Results on GSD for pretrained (*) base models and ST^2. HE means human evaluation score.

Based on the previous reasoning, we extract and pretrain the language model module in two of our base models (CrossAlign and VAE) on the union of training data from all sub-tasks. Starting with a well-trained language model, we then fine-tune the models for each style transfer task. By comparing pretrained base models with our ST^2 models, we verify that meta-learning framework can improve the style transfer accuracy in addition to language fluency.

In addition, to examine the effect of pretraining combined with meta-learning, we also add a pretraining phase to our ST^2 model. The results are included in Table 4. 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 BLEU and PPL gives 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 learns 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 has 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. We also notice that the pretraining is not crucial to ST^2, suggesting that it is the meta-learning framework that significantly contributes to the model's improvements in generating fluent sentences and efficacy in style transfers.

4 Conclusion

We extend the concept of text style to general writing styles with limited training data for style transfer. To tackle this new challenging 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 grouped standard dataset to evaluate the state-of-the-art models and our proposed framework. Unlike previous state-of-the-art models that are resource-demanding to impart rich knowledge into the networks, ST^2 is able to effectively utilize off-domain information to improve both language fluency and style transfer accuracy in a way that conventional pretrained language models fall short.

References

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*.

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*.

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

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.

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*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

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*.

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.

Sudha Rao and Joel Tetreault. 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, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2020. Transforming delete, retrieve, generate approach for controlled text style transfer. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*.

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

Peng Xu, Yanshuai Cao, and Jackie Chi Kit Cheung. 2019. Unsupervised controllable text generation with global variation discovery and disentanglement. *CoRR*, abs/1905.11975.