A Dual Reinforcement Learning Framework for Unsupervised Text Style Transfer

Fuli Luo[1] , **Peng Li**[2] , **Jie Zhou**[2] , **Pengcheng Yang**[1,3] , **Baobao Chang**[1,4] , **Xu Sun**[1] , **Zhifang Sui**[1,4]

[1]Key Lab of Computational Linguistics, School of EECS, Peking University
[2]Pattern Recognition Center, WeChat AI, Tencent Inc, China
[3]Deep Learning Lab, Beijing Institute of Big Data Research, Peking University
[4]Peng Cheng Laboratory, China
luofuli@pku.edu.cn, {patrickpli,withtomzhou}@tencent.com, {yang_pc,chbb,szf,xusun}@pku.edu.cn

Abstract

Unsupervised text style transfer aims to transfer the underlying style of text but keep its main content unchanged without parallel data. Most existing methods typically follow ***two steps***: first separating the content from the original style, and then fusing the content with the desired style. However, the separation in the first step is challenging because the content and style interact in subtle ways in natural language. Therefore, in this paper, we propose a dual reinforcement learning framework to directly transfer the style of the text via a ***one-step*** mapping model, without any separation of content and style. Specifically, we consider the learning of the source-to-target and target-to-source mappings as a dual task, and two rewards are designed based on such a dual structure to reflect the style accuracy and content preservation, respectively. In this way, the two one-step mapping models can be trained via reinforcement learning, without any use of parallel data. Automatic evaluations show that our model outperforms the state-of-the-art systems by a large margin, especially with more than 8 BLEU points improvement averaged on two benchmark datasets. Human evaluations also validate the effectiveness of our model in terms of style accuracy, content preservation and fluency. Our code and data, including outputs of all baselines and our model are available at https://github.com/luofuli/DualRL. [1]

1 Introduction

Text style transfer aims to rephrase the input text in the desired style while preserving its original content. It has various application scenarios such as sentiment transformation (transferring a positive review to a negative one) and formality modification (revising an informal text into a formal one). As parallel data, i.e., aligned sentences with the same content but different style, is hard to collect for this task, previous works mainly focus on unsupervised text style transfer.

Most existing methods of unsupervised text style transfer follow a two-step process: first separating the content from the original style and then fusing the content with the desired style. One line of research [Shen *et al.*, 2017; Fu *et al.*, 2018; Hu *et al.*, 2017; Tsvetkov *et al.*, 2018] learns a style-independent content representation vector via adversarial training, and then passes it to a style-dependent decoder for rephrasing. Another line of research [Li *et al.*, 2018; Xu *et al.*, 2018] directly removes the specific style attribute words in the input, and then feeds the neutralized sequence which only contains content words to a style-dependent generation model. However, each line has its own drawback.

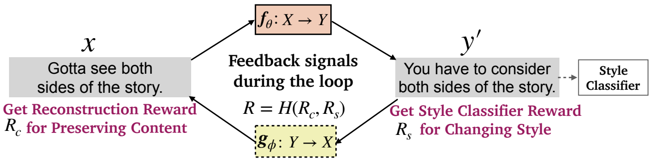
x
Gotta see both sides of the story.
Get Reconstruction Reward
R_c for Preserving Content
$f_\theta: X \to Y$
Feedback signals during the loop
$R = H(R_c, R_s)$
$g_\phi: Y \to X$
y'
You have to consider both sides of the story.
Style Classifier
Get Style Classifier Reward
R_s for Changing Style

Figure 1: The proposed DualRL framework for unsupervised text style transfer with an informal-to-formal text example, where both $\boldsymbol{f}_\theta$ and $\boldsymbol{g}_\phi$ are a sequence-to-sequence mapping model.

The former line of research tends to only change the style but fail in keeping the content, since it is hard to get a style-independent content vector without parallel data [Xu *et al.*, 2018; Lample *et al.*, 2019]. For example, on the sentiment transfer task, given "*The food is delicious*" as input, the model may generate "*The movie is bad*" instead of "*The food is awful*". Thus, the latter line of research focuses on improving content preservation in a more direct way by neutralizing the text in the discrete token space, other than the continuous vector space. However, these models have a limited range of applications, since they are challenged by the examples like "*The only thing I was offered was a free dessert!!!*", whose negative sentiment is implicitly expressed such that there is no specific emotional style word.

To alleviate the above problems caused by the two-step process, we propose to directly learn a one-step *mapping* model between the source corpora and the target corpora of different styles. More importantly, due to the lack of parallel data, we consider the learning of the source-to-target and target-to-source mapping models as a dual task, and propose a dual reinforcement learning algorithm **DualRL** to train them. Taking Figure 1 for example, the forward one-step mapping model $\boldsymbol{f}$ transfers an informal sentence $\boldsymbol{x}$ into a formal sentence $\boldsymbol{y}'$, while the backward one-step mapping model

[1]Joint work between WeChat AI and Peking University.

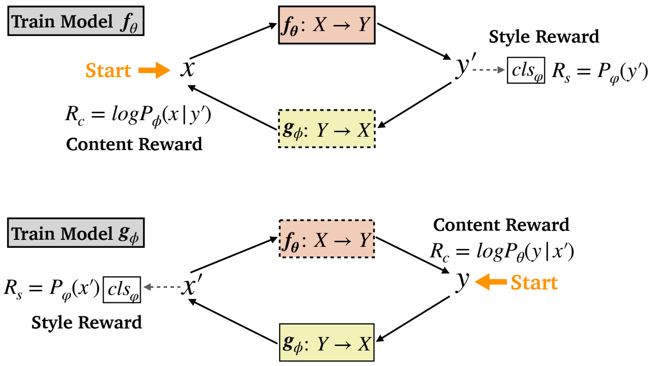

Figure 2: Training process of DualRL. We alternately train the two mapping models $\boldsymbol{f_\theta}$ and $\boldsymbol{g_\phi}$.

$\boldsymbol{g}$ transfers a formal sentence $\boldsymbol{y}$ into an informal sentence $\boldsymbol{x}'$. Since the two models can form a closed loop, we let them to teach each other interactively via two elaborately designed quality feedback signals to ensure the success of style transfer: changing style while preserving content. Specially, the two signals are combined as a reward for the reinforcement learning (RL) method to alternately train the model $\boldsymbol{f}$ and $\boldsymbol{g}$ (Section 2.1). Furthermore, in order to better adapt DualRL to the unsupervised scenario, we propose an *annealing pseudo teacher-forcing* algorithm to construct pseudo-parallel data on-the-fly via back-translation to warm up RL training and gradually shift to pure RL training (Section 2.2). The proposed framework is simple and generic and can be potentially adapted to other sequence-to-sequence generation tasks that lack parallel data. Experiments on two benchmark datasets show that our model outperforms the state-of-the-art systems by a large margin in both automatic and human evaluation.

2 Dual Reinforcement Learning for Unsupervised Text Style Transfer

Given two corpora $\mathcal{D}_X = \{\boldsymbol{x}^{(i)}\}_{i=1}^n$ and $\mathcal{D}_Y = \{\boldsymbol{y}^{(j)}\}_{j=1}^m$ with two different styles s_x and s_y, the goal of text style transfer task is to generate a sentence of the target style while preserving the content of the source input sentence. In general, the two corpora are non-parallel such that the gold pair $(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(j)})$ that describes the same content but expresses the different style is unavailable.

2.1 DualRL: Dual Reinforcement Learning

In this paper, we directly learn two *one-step* mappings (as style transfer models) between the two corpora of different styles. Formally, the forward model $\boldsymbol{f_\theta}: \mathcal{X} \rightarrow \mathcal{Y}$ transfers the sequence $\boldsymbol{x}$ with style s_x into a sequence $\boldsymbol{y}'$ with style s_y, while the backward model $\boldsymbol{g_\phi}: \mathcal{Y} \rightarrow \mathcal{X}$ transfers the sequence $\boldsymbol{y}$ with style s_y into a sequence $\boldsymbol{x}'$ with style s_x.

Due to the lack of parallel data, the two transfer models can not be trained in a supervised way as usual. Fortunately, since text style transfer task always happens in dual directions, we loop the two transfer models of the two directions and the loop process can provide quality feedbacks to guide the training of the two style transfer models even using non-parallel data only. In order to encourage changing style but preserving content, we design two corresponding quality feedbacks. For the former, a style classifier is adopted to assess how well the transferred sentence $\boldsymbol{y}'$ matches the target style. For the latter, the probability of feeding $\boldsymbol{y}'$ to the backward model $\boldsymbol{g}$ to reconstruct $\boldsymbol{x}$ can reflect how much content information preserved in the source sentence $\boldsymbol{x}$. Because of the discrete connection $\boldsymbol{y}'$ of the two models, the loss function is no longer differentiable w.r.t. to the parameters of the forward model. Therefore, we treat the two quality feedbacks as rewards and train the model via RL.

Algorithm 1 The dual reinforcement learning algorithm for unsupervised text style transfer.

```
Pre-train text style transfer models $f_\theta$ and $g_\phi$ using pseudo-parallel sentence pairs from corpora $\mathcal{D}_X$ and $\mathcal{D}_Y$
Pre-train a binary style classifier $cls_\varphi$
for each iteration $i = 1, 2, ..., M$ do
                                        ▷ Start to train model $f_\theta$
    Sample sentence $x$ from $\mathcal{D}_X$
    Generate sentence $y'$ of opposite style via model $f_\theta$
    Compute style reward $R_s$ based on Eq. 1
    Compute content reward $R_c$ based on Eq. 2
    Compute total reward $R$ based on Eq. 3
    Update $\theta$ using reward $R$ based on Eq. 4
    Update $\theta$ using annealing teacher-forcing via MLE
                                        ▷ Start to train model $g_\phi$
    Sample sentence $y$ from $\mathcal{D}_Y$
    Generate sentence $x'$ of opposite style via model $g_\phi$
    Compute style reward $R_s$ similar to Eq. 1
    Compute content reward $R_c$ similar to Eq. 2
    Compute total reward $R$ based on Eq. 3
    Update $\phi$ using reward $R$ similar to Eq. 4
    Update $\phi$ using annealing teacher-forcing via MLE
end for
```

In order to enable the two models to boost each other, we propose a dual training algorithm **DualRL** to train the two models simultaneously, inspired by [He *et al.*, 2016]. As Figure 2 shows, starting from sampling a sequence $\boldsymbol{x}$ from corpus $\mathcal{D}_X$, model $\boldsymbol{f}$ will be trained based on two rewards provided by the style classifier cls_φ and model $\boldsymbol{g}$. Meanwhile, starting from sampling a sequence $\boldsymbol{y}$ from $\mathcal{D}_Y$, model $\boldsymbol{g}$ can be trained based on the two rewards provided by the style classifier cls_φ and model $\boldsymbol{f}$. The overview of DualRL is shown in Algorithm 1. The definitions of the two rewards and the gradients for model $\boldsymbol{f}$ are introduced as follows, and those for model $\boldsymbol{g}$ is computed in a similar way.

Reward

Since the gold transferred result of input $\boldsymbol{x}$ is unavailable, the quality of the generated sentence $\boldsymbol{y}'$ can not be directly evaluated. Therefore, we design two rewards that can assess the style accuracy and the content preservation, respectively.

Reward for changing style. A pre-trained binary style classifier [Kim, 2014] is used to evaluate how well the transferred sentence $\boldsymbol{y}'$ matches the target style. Formally, the style classifier reward is formulated as

$$R_s = P(s_y|\boldsymbol{y}'; \varphi) \quad (1)$$

where φ is the parameter of the classifier and is fixed during the training process.

Reward for preserving content. Intuitively, if the two transfer models are well-trained, it is easy to reconstruct the source sequence via back transferring. Therefore, we can estimate how much the content preserved in $\boldsymbol{y}'$ by means of the probability that the model $\boldsymbol{g}$ reconstructs $\boldsymbol{x}$ when taking $\boldsymbol{y'}$ as input. Formally, the corresponding reconstruction reward is formulated as

$$R_c = \log(P(\boldsymbol{x}|\boldsymbol{y'};\boldsymbol{\phi})) \tag{2}$$

where $\boldsymbol{\phi}$ is the parameter of model $\boldsymbol{g}$. Another intuitive way to measure the content preservation is to calculate the BLEU score [Papineni *et al.*, 2002] of $\boldsymbol{x''}$ with the input $\boldsymbol{x}$ as reference, where $\boldsymbol{x''}$ is the output of the backward model $\boldsymbol{g}$ when taking $\boldsymbol{y'}$ as input [Xu *et al.*, 2018]. However, primary experiments show that this method exhibits poor performance in our framework.

Overall reward. To encourage the model to improve both the content preservation and the style accuracy, the final reward is the harmonic mean of the above two rewards

$$R = (1+\beta^2)\frac{R_c \cdot R_s}{(\beta^2 \cdot R_c) + R_s} \tag{3}$$

where β is a harmonic weight aiming to control the trade-off between the two rewards.

Policy Gradient Training

The policy gradient algorithm [Williams, 1992] is used to maximize the expected reward $\mathbb{E}[R]$ of the generated sequence $\boldsymbol{y'}$, whose gradient w.r.t. the parameter $\boldsymbol{\theta}$ of the forward model $\boldsymbol{f}$ is estimated by sampling as

$$\begin{aligned}
\nabla_\theta \mathbb{E}[R] &= \nabla_\theta \sum_k P(\boldsymbol{y'_k}|\boldsymbol{x};\theta) R_k \\
&= \sum_k P(\boldsymbol{y'_k}|\boldsymbol{x};\theta) R_k \nabla_\theta \log(P(\boldsymbol{y'_k}|\boldsymbol{x};\theta)) \\
&\simeq \frac{1}{K}\sum_{k=1}^{K} R_k \nabla_\theta \log(P(\boldsymbol{y'_k}|\boldsymbol{x};\theta))
\end{aligned} \tag{4}$$

where R_k is the reward of the k_{th} sampled sequence $\boldsymbol{y'_k}$ from model $\boldsymbol{f}$, and K is the sample size.

2.2 DualRL for Unsupervised Task

When applied to the field of text generation, RL faces two ingrained challenges: 1) the RL framework needs to be well pre-trained to provide a warm-start, and 2) the RL method may find an unexpected way to achieve a high reward but fail to guarantee the fluency or readability of the generated text [Ranzato *et al.*, 2016; Pasunuru and Bansal, 2018]. An effective solution to these two challenges in supervised tasks is to expose the parallel data to the model and train it via MLE (Maximum Likelihood Estimation) [Ranzato *et al.*, 2016; Paulus *et al.*, 2017; Li *et al.*, 2017]. However, due to the lack of parallel data, these two challenges become intractable on unsupervised scenarios. In this paper, we tackle these two challenges via pseudo-parallel data. Specifically, in order to pre-train our Seq2Seq mapping models, we exploit pseudo-parallel data generated by a simple template-based baseline [Li *et al.*, 2018] to train via MLE; in order to enhance the quality of the generated text, we propose a annealing pseudo teacher-forcing algorithm.

Algorithm 2 The annealing pseudo teacher-forcing algorithm for dual reinforcement learning.

```
Initialize the iteration interval $p$
for each iteration $i = 1, 2, ..., M$ do
                                        ▷ Start to train model $\boldsymbol{f_\theta}$
    Update parameter $\boldsymbol{\theta}$ via RL based on Eq. 4
    if $i$ % $p = 0$ then                ▷ Pseudo Teacher-Forcing
        Generate a pair of data $(\boldsymbol{x}'_i, \boldsymbol{y}_i)$, where $\boldsymbol{x}'_i = \boldsymbol{g}(\boldsymbol{y}_i)$
        Update $\boldsymbol{\theta}$ using data $(\boldsymbol{x}'_i, \boldsymbol{y}_i)$ via MLE
    end if
                                        ▷ Start to train model $\boldsymbol{g_\phi}$
    Update parameter $\boldsymbol{\phi}$ via RL similar to Eq. 4
    if $i$ % $p = 0$ then                ▷ Pseudo Teacher-Forcing
        Generate a pair of data $(\boldsymbol{y}'_i, \boldsymbol{x}_i)$, where $\boldsymbol{y}'_i = \boldsymbol{f}(\boldsymbol{x}_i)$
        Update $\boldsymbol{\phi}$ using data $(\boldsymbol{y}'_i, \boldsymbol{x}_i)$ via MLE
    end if
    Exponential increase in $p$ based on Eq. 5
end for
```

Annealing Pseudo Teacher-Forcing

Teacher-forcing is the strategy that feeds the parallel data $(\boldsymbol{x}, \boldsymbol{y})$ into the Seq2Seq model, and then either 1) train the model by optimizing a weighted sum of RL and MLE loss [Paulus *et al.*, 2017], or 2) alternately update the model using the RL and the MLE objective [Li *et al.*, 2017]. An intuitive but not ideal solution is to utilize the pseudo-parallel data which was used during pre-training. However, we have done primary experiments which show that the quality of the pseudo-parallel data is not acceptable for the later iterations of training. Inspired by back-translation in unsupervised machine translation [Lample *et al.*, 2018a; Lample *et al.*, 2018b], we leverage the latest version of model $\boldsymbol{f_\theta}/\boldsymbol{g_\phi}$ at previous iteration $i-1$ to generate a *higher* quality of pseudo-parallel data $(\boldsymbol{x}'_i, \boldsymbol{y}_i)/(\boldsymbol{y}'_i, \boldsymbol{x}_i)$ [2] than those used during pre-training on-the-fly to update model $\boldsymbol{g_\phi}/\boldsymbol{f_\theta}$ via MLE at iteration i, respectively.

As long as the model gets better during training, the generated pseudo-parallel data can become more closer to real parallel data. However, there still exists a gap between the distribution of the generated pseudo-parallel data and real parallel data during training. Moreover, models trained via MLE often exhibit "exposure bias" problem [Ranzato *et al.*, 2016]. Therefore, in order to get rid of the dependence of pseudo-parallel data, we propose an *annealing* strategy of teacher-forcing, as shown in Algorithm 2. More specifically, we enlarge the training interval of teacher-forcing to decay its frequency of updating parameters via MLE. Formally, at iteration i, we adopt an exponential increase in the interval of teacher-forcing p

$$p = \min(p_0 \times r^{\frac{i}{d}}, p_{max}) \tag{5}$$

where p_0 is the initial iteration interval, p_{max} is the max iteration interval, r is the increase rate ($r > 1$) and d is the increase gap. A deep study of the influence of teacher forcing (trained via MLE) will be given in Section 3.7 and Figure 3.

[2] $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ denote the original data and $\boldsymbol{y}'_i$ and $\boldsymbol{x}'_i$ denote the corresponding generated pseudo-parallel data at i-th iteration.

	YELP				GYAFC			
	ACC	BLEU	G2	H2	ACC	BLEU	G2	H2
Retri [Li *et al.*, 2018]	**96.0**	2.9	16.7	5.7	**91.3**	0.4	6.0	0.8
BackTrans [Tsvetkov *et al.*, 2018]	95.4	5.0	21.9	9.6	70.2	0.9	8.1	1.9
StyleEmbed [Fu *et al.*, 2018]	8.7	42.3	19.2	14.4	22.7	7.9	13.4	11.7
MultiDec [Fu *et al.*, 2018]	50.2	27.9	37.4	35.9	17.9	12.3	14.8	14.6
CrossAlign [Shen *et al.*, 2017]	75.3	17.9	36.7	28.9	70.5	3.6	15.9	6.8
Unpaired [Xu *et al.*, 2018]	64.9	37.0	49.0	47.1	79.5	2.0	12.6	3.9
Del [Li *et al.*, 2018]	85.3	29.0	49.7	43.3	18.8	29.2	23.4	22.9
DelRetri [Li *et al.*, 2018]	89.0	31.1	52.6	46.1	55.2	21.2	34.2	30.6
Template [Li *et al.*, 2018]	81.8	45.5	61.0	58.5	52.9	35.2	43.1	42.3
UnsuperMT [Zhang *et al.*, 2018b]	95.4	44.5	65.1	60.7	70.8	33.4	48.6	45.4
DualRL	85.6	**55.2**	**68.7**	**67.1**	71.1	**41.9**	**54.6**	**52.7**
Human	74.0	100.0	86.0	85.1	84.3	100.0	91.8	91.5

Table 1: Automatic evaluation results on the YELP and GYAFC datasets. "ACC" shows the accuracy of output labeled as the target style by a pre-trained style classifier. "BLEU" measures content similarity between the output and the four human references. G2 and H2 are geometric mean and harmonic mean of ACC and BLEU. **Bold** denotes the best results and underline denotes the best overall scores.

3 Experiments

3.1 Dataset

We evaluate our model on two instances of style transfer task.

Sentiment Transfer. The representative **YELP** restaurant reviews dataset is selected for this task. Following common practice, reviews with rating above 3 are considered as positive, and those below 3 as negative. This dataset is widely used by previous work and the train, dev and test split is the same as [Li *et al.*, 2018].

Formality Transfer. A newly released dataset **GYAFC** (Grammarly's Yahoo Answers Formality Corpus) [Rao and Tetreault, 2018] is used for this task. And we choose the family and relationships domain. Although it is a parallel dataset, the alignments are only used for evaluation but not training.

3.2 Human References

While four human references are provided for each test sentence in the GYAFC dataset, only one reference is provided for each test sentence in the YELP dataset, which makes the automatic evaluation less reliable. Therefore, we hired crowdworkers on CrowdFlower to write three more human references for each test sentence in the YELP dataset. All these references and the generated results of all the involved models in this paper will be released for reproducibility, hopefully to enable more reliable empirical comparisons in future work.

3.3 Training Details

The hyper-parameters are tuned on the development set. Both $\boldsymbol{f}$ and $\boldsymbol{g}$ are implemented as a basic LSTM-based encoder-decoder model with 256 hidden size [Bahdanau *et al.*, 2015]. The word embeddings of 300 dimension are learned from scratch. The optimizer is Adam [Kingma and Ba, 2014] with 10^{-3} initial learning rate for pre-training and 10^{-5} for dual learning. The batch size is set to 32 for pre-training and 128 for dual learning. Harmonic weight β in Eq. 3 is 0.5. For annealing teacher forcing (Eq. 5), the initial gap p_0 is 1, the max gap p_{max} is 100, increase rate r is 1.1, and increase gap d is 1000. Before dual learning, model $\boldsymbol{f}$ and $\boldsymbol{g}$ are pre-trained for 5 epochs. During dual learning, training runs for up to 20 epochs with early stopping if the development set performance does not improve within last one epoch.

3.4 Baselines

We compare our proposed method with the following state-of-the-art systems: StyleEmbed and MultiDec [Fu *et al.*, 2018]; CrossAlign [Shen *et al.*, 2017]; BackTrans [Tsvetkov *et al.*, 2018]; Template, Retri, Del and DelRetri [Li *et al.*, 2018]; Unpaired [Xu *et al.*, 2018]. Moreover, a most recent and representative work UnsuperMT [Zhang *et al.*, 2018b] which treats style transfer as unsupervised machine translation is also considered.

3.5 Evaluation Metrics

We conduct both automatic and human evaluation.

Automatic Evaluation. Following previous work [Li *et al.*, 2018; Zhang *et al.*, 2018b], we adopt the following metrics to evaluate each system. A pre-trained binary style classifier TextCNN [Kim, 2014] is used to evaluate the style accuracy of the outputs. The classifier can achieve the accuracy of 95% and 89% on the two datasets respectively. The BLEU score [Papineni *et al.*, 2002] [3] between the outputs and the four human references is used to evaluate the content preservation performance. In order to evaluate the overall performance, we report the geometric mean and harmonic mean of the two metrics [Xu *et al.*, 2018].

Human Evaluation. We distribute the outputs of different systems to three annotators with linguistic background and the annotators have no knowledge in advance about which model the generated text comes from. They are required to score the generated text from 1 to 5 in terms of three criteria: the accuracy of the target style, the preservation of the original content and the fluency. Finally, following [Li *et al.*, 2018], a transferred text is considered to be "*successful*" if it is rated 4 or 5 on all three criteria.

[3] The BLEU score is computed using `multi-bleu.perl`.

	YELP					GYAFC				
	Sty	Con	Flu	**Avg**	**Suc**	Sty	Con	Flu	**Avg**	**Suc**
MultiDec [Fu *et al.*, 2018]	2.14	3.02	3.27	2.81	5%	2.21	1.95	2.54	2.23	4%
CrossAlign [Shen *et al.*, 2017]	2.88	2.79	3.40	3.02	14%	2.96	1.33	3.27	2.52	3%
Unpaired [Xu *et al.*, 2018]	2.93	3.38	3.44	3.25	17%	2.69	1.19	2.38	2.09	2%
Template [Li *et al.*, 2018]	3.12	3.71	3.42	3.42	23%	2.74	3.60	3.43	3.26	9%
DelRetri [Li *et al.*, 2018]	3.39	3.49	3.71	3.53	28%	2.47	2.57	2.67	2.57	5%
UnsuperMT [Zhang *et al.*, 2018b]	3.82	3.90	3.93	3.95	40%	3.27	3.54	3.76	3.52	21%
DualRL	**4.11**	**4.33**	**4.31**	**4.25**	**54%**	**3.65**	**3.62**	**3.80**	**3.69**	**28%**

Table 2: Human evaluation results on two datasets. We show human ratings for and target style accuracy (Sty), content preservation (Con), fluency (Flu) on a 1 to 5 Likert scale. We also calculate the average ratings (Avg) and success rate (Suc) as overall scores.

Automatic	ACC	BLEU		G2	H2
Human	Sty	Con	Flu	Avg	
YELP	0.89*	0.96*	0.72	0.93*	0.89*
GYAFC	0.68	0.99*	0.76	0.96*	0.94*

Table 3: Pearson correlation between automatic evaluation and human evaluation. Scores marked with * denotes $p < 0.01$.

	Sty	Con	Flu	**Avg**	**Suc**
RL+MLE	4.11	**4.33**	**4.31**	**4.25**	**54%**
RL	**4.29**	4.08	3.73	4.03	43%
MLE	3.45	4.19	**4.31**	3.98	41%

Table 4: Human evaluation results on full model (RL+MLE) and ablated models on the YELP dataset.

3.6 Results and Discussions

Table 1 shows the automatic evaluation results of the systems. We can observe that our model DualRL achieves the best overall performance (G2, H2). More specifically, our model significantly outperforms the other systems by over 8 BLEU points averaged on two datasets.

It is worth mentioning that, our model does not get the best style classifier accuracy (ACC), so does the human reference. The reasons are in two-folds. First, the Retri system, which directly retrieves a similar sentence from the training dataset of the *target style*, can naturally achieve an accuracy close to the training dataset which lets the classifier performs best. However, most of systems including Retri only show good results either in ACC or BLEU, implying that they tend to sacrifice one for the other. Second, since both the generated sentence and human reference sometimes only change a few words, which can be adversarial examples [Iyyer *et al.*, 2018] and mislead the classifier.

Table 2 shows the human evaluation results of several well-performed systems in the automatic evaluation. We find that our model achieves the best average score (Avg). Moreover, our system can generate more than 10% successfully (Suc) transferred instances, averaged on two datasets. And all systems show better results on YELP than GYAFC, revealing that text formality is more challenging than sentiment transfer.

Furthermore, Table 3 shows the system-level Pearson correlation between automatic evaluation metrics and human evaluation results. We find that: (1) BLEU score significantly correlates with content preservation, but not fluency. (2) The correlation between automatic calculated accuracy (ACC) and the human ratings of style accuracy varies between datasets. (3) Both the automatic overall metrics G2 and H2 well correlate with the human average ratings.

3.7 Ablation Study

In this section, we give a deep analysis of the key components of our model. Figure 3 and Table 4 show the learning curves and human evaluation results of the full model and models which ablated RL and MLE training on the YELP dataset. It shows if we only train the model based on RL, the ACC (Sty) will increase, while the BLEU (Con, Flu) will decline. The reason is that the RL training may encourage the model to generate tricky sentences which can get the high reward but fails in quality and readability (measured by BLEU). For example, given a negative review "*We sit down and got some really slow and lazy service*" as the input, the model may generate "*We sit down and got some really great and great service*". This output sentence is not fluent but it can get a high style classifier reward and content preservation reward, thus leading the model to train towards a bad direction. In contrast, the Seq2Seq model using MLE objective is essentially a conditional language model which can ensure readability of the output, thus showing higher BLEU (Con, Flu) score than RL. However, the ACC (Sty) of MLE declines, since there is no specific mechanism in MLE to directly control the style accuracy. Finally, the combination of RL and MLE can get best BLEU score without compromising ACC, with over 3.5 points absolute improvement in H2/G2 score and 13% more success rate (Suc) than model trained only based on MLE.

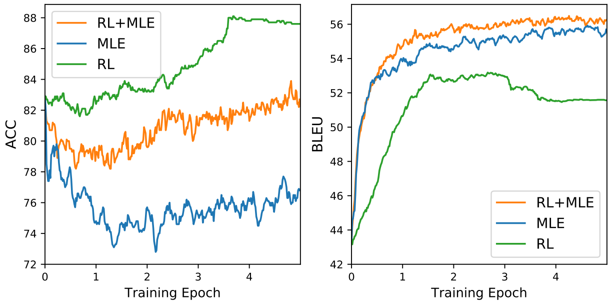

Figure 3: Learning curves of the full model (RL+MLE) and the ablated models (RL, MLE) on the YELP dataset.

	From negative to positive (YELP)	From informal to formal (GYAFC)
Source	Moving past the shape, they were dry and truly tasteless.	(That's what i called it) .. but, why?
CrossAlign	Everyone on the fish, they were fresh and filling.	And i know what this helps me.
Template	Moving past the shape, they a wonderful truly.	(That's what it is called it) .. but, why?
Del-Retri	Moving past the shape is awesome, and they will definitely be back!	(That's what i you it you but why, you?
UnsuperMT	Moving moving the shape, they were juicy and truly delicious.	(That's what i said it) but that is why you were doing.)
DualRL	Moving past the shape, they were tasty and truly delicious.	It is what i called it, but why?

Table 5: Example outputs on the YELP and GYAFC datasets. Improperly generated words and grammar errors are colored.

3.8 Case Study

In this section, we present one randomly sampled example of representative systems and analyze the strengths and weaknesses of them. Table 5 shows the example outputs on the YELP and GYAFC datasets. We can observe that: (1) The CrossAlign system, which learns a style-independent content representation vector via adversarial training, tends to sacrifice the content preservation. (2) The Template and Del-Retri systems, which directly removes the specific style attribute words in the input, can better preserve the content. However, these two systems may fail when the style is implicitly expressed in the input (See the informal-to-formal example). (3) Promisingly, our model achieves a better balance among preserving content, changing the style and improving fluency.

3.9 Error Analysis

Although the proposed method outperforms the state-of-the-art systems, we also observe a few failure cases. The typical type of failure cases is that the analogy or metaphor of the style (sentiment). A representative example is "*over cooked so badly that it was the consistency of canned tuna fish*". The *canned tuna fish* does not represent its literal content but just an analogy of *"over cooked"*. However, it is really hard for our system as well as other existing methods to balance between preserving the original content and transferring the style when encountering such analogy examples.

4 Related Work

Recently, increasing efforts have been devoted to unsupervised text style transfer. Despite the increasing efforts devoted to the text style transfer in recent years, the lack of parallel data is still the major challenge for this task.

To relief from the need of parallel data, early works generally learn style-independent content representation. In this way, they can train the style rendering model using non-parallel data only. [Fu *et al.*, 2018] leverages the adversarial network to make sure that the content representation does not include style representation. [Shen *et al.*, 2017; Hu *et al.*, 2017; Yang *et al.*, 2018b] combine Variational Auto-encoder with a style discriminator. Besides, [Tsvetkov *et al.*, 2018] strives to get a style-independent content representation through the English-to-French translation models. However, some recent works [Li *et al.*, 2017; Lample *et al.*, 2019] argue that it is often easy to fool the discriminator without actually removing the style information. In other words, the style-independent content representation in latent space may indeed not be able to be achieved in practice, thus causing bad content preservation. On the contrary, [Li *et al.*, 2018; Zhang *et al.*, 2018a; Xu *et al.*, 2018] propose to separate content and style by directly removing the style words. The former takes advantage of the prior knowledge that style words only localized in corresponding corpora, while the latter skillfully exploits the self-attention mechanism. However, this explicit separation is not suitable for the text whose style can only be expressed as a whole.

Another way to relief from the need of parallel data is to construct pseudo-parallel data via back-translation, which achieves promising results in unsupervised machine translation [Artetxe *et al.*, 2018; Lample *et al.*, 2018a; Lample *et al.*, 2018b]. There are also two most recent works [Zhang *et al.*, 2018b; Lample *et al.*, 2019] directly adopt unsupervised machine translation methods to this task. However, learning from pseudo-parallel data inevitably accompanies with the data quality problem, thus further influence the control of the preservation of content and accuracy of style. In contrast, we adopt the reinforcement learning algorithm with specifically designed rewards, which directly ensures the two aims of style transfer (Section 2.1). Meanwhile, the proposed annealing pseudo teacher-forcing algorithm (Section 2.2) not only benefits our model from pseudo-parallel data at the beginning of training, but also gradually gets rid of it in the latter stage of training when the model is completely warmed up and is suitable for training mainly based on DualRL.

5 Conclusion

In this work, we aim at solving text style transfer by learning a direct *one-step* mapping model for the source-to-target style transfer and a dual mapping model for the target-to-source style transfer. Due to the lack of parallel data, we propose a dual reinforcement learning algorithm DualRL in order to train the two mapping models solely based on the automatically generated supervision signals. In this way, we do not need to do any explicit separation of content and style, which is hard to achieve in practice even with parallel data. Experimental results on the sentiment transfer and formality transfer datasets show that our model significantly outperforms the previous approaches, empirically demonstrating the effectiveness of learning two *one-step* mapping models and the proposed DualRL training algorithm.

Acknowledgments

This paper is supported by NSFC project 61751201, 61876004 and 61772040. The contact authors are Baobao Chang and Zhifang Sui.

References

[Artetxe *et al.*, 2018] Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. In *Proceedings of ICLR*, 2018.

[Bahdanau *et al.*, 2015] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR*, 2015.

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

[He *et al.*, 2016] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. Dual learning for machine translation. In *Proceedings of NIPS*, 2016.

[Hu *et al.*, 2017] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. Toward controlled generation of text. In *Proceedings of ICML*, 2017.

[Iyyer *et al.*, 2018] Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of NAACL*, 2018.

[Kim, 2014] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of EMNLP*, 2014.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of ICLR*, 2014.

[Lample *et al.*, 2018a] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. In *Proceedings of ICLR*, 2018.

[Lample *et al.*, 2018b] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. In *Proceedings of EMNLP*, 2018.

[Lample *et al.*, 2019] Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. Multiple-attribute text rewriting. In *Proceedings of ICLR*, 2019.

[Li *et al.*, 2017] Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. Adversarial learning for neural dialogue generation. In *Proceedings of EMNLP*, 2017.

[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 *Proceedings of NAACL*, 2018.

[Liu *et al.*, 2017] Tianyu Liu, Kexiang Wang, Baobao Chang, and Zhifang Sui. A soft-label method for noise-tolerant distantly supervised relation extraction. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the EMNLP 2017*, 2017.

[Liu *et al.*, 2018] Tianyu Liu, Kexiang Wang, Lei Sha, Baobao Chang, and Zhifang Sui. Table-to-text generation by structure-aware seq2seq learning. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI, (AAAI-18), 2018*, 2018.

[Ma *et al.*, 2018] Shuming Ma, Lei Cui, Damai Dai, Furu Wei, and Xu Sun. Livebot: Generating live video comments based on visual and textual contexts. *CoRR*, abs/1809.04938, 2018.

[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 ACL*, 2002.

[Pasunuru and Bansal, 2018] Ramakanth Pasunuru and Mohit Bansal. Multi-reward reinforced summarization with saliency and entailment. In *Proceedings of NAACL*, 2018.

[Paulus *et al.*, 2017] Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization. In *Proceedings of ICLR*, 2017.

[Ranzato *et al.*, 2016] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *Proceedings of ICLR*, 2016.

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

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

[Tsvetkov *et al.*, 2018] Yulia Tsvetkov, Alan W. Black, Ruslan Salakhutdinov, and Shrimai Prabhumoye. Style transfer through back-translation. In *Proceedings of ACL*, 2018.

[Williams, 1992] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8, 1992.

[Xu *et al.*, 2018] Jingjing Xu, Xu Sun, Qi Zeng, Xiaodong Zhang, Xuancheng Ren, Houfeng Wang, and Wenjie Li. Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. In *Proceedings of ACL*, 2018.

[Yang *et al.*, 2018a] Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. SGM: sequence generation model for multi-label classification. In *Proceedings of COLING 2018*, pages 3915–3926, 2018.

[Yang *et al.*, 2018b] Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. Unsupervised text style transfer using language models as discriminators. In *Proceedings of NeurIPS*. 2018.

[Zhang *et al.*, 2018a] Yi Zhang, Jingjing Xu, Pengcheng Yang, and Xu Sun. Learning sentiment memories for sentiment modification without parallel data. In *Proceedings of the EMNLP 2018*, pages 1103–1108, 2018.

[Zhang *et al.*, 2018b] Zhirui Zhang, Shuo Ren, Shujie Liu, Jianyong Wang, Peng Chen, Mu Li, Ming Zhou, and Enhong Chen. Style transfer as unsupervised machine translation. *CoRR*, 2018.