A Dual Reinforcement Learning Framework for Unsupervised Text Style Transfer

Fuli Luo\textsuperscript{1}, Peng Li\textsuperscript{2}, Jie Zhou\textsuperscript{2}, Pengcheng Yang\textsuperscript{1,3}, Baobao Chang\textsuperscript{1,4}, Xu Sun\textsuperscript{1}, Zhifang Sui\textsuperscript{1,4}

\textsuperscript{1}Key Lab of Computational Linguistics, School of EECS, Peking University
\textsuperscript{2}Pattern Recognition Center, WeChat AI, Tencent Inc, China
\textsuperscript{3}Deep Learning Lab, Beijing Institute of Big Data Research, Peking University
\textsuperscript{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.\textsuperscript{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.

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 $f$ transfers an informal sentence $x$ into a formal sentence $y'$, while the backward one-step mapping model $g$.

![Figure 1: The proposed DualRL framework for unsupervised text style transfer with an informal-to-formal text example, where both $f_θ$ and $g_ϕ$ are a sequence-to-sequence mapping model.](image-url)
Get Style Classifier Reward for Changing Style
Gotta see both sides of the story.
You have to consider both... Reward
x y′ Start
\( f_\theta : X \rightarrow Y \)
\( g_\phi : Y \rightarrow X \)
\( f_\theta : X \rightarrow Y \)
\( g_\phi : Y \rightarrow X \)

Figure 2: Training process of DualRL. We alternately train the two mapping models \( f_\theta \) and \( g_\phi \).

g transfers a formal sentence \( y \) into an informal sentence \( x' \).
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. Specifically, the two signals are combined as a reward for the reinforcement learning (RL) method to alternately train the model \( f \) and \( g \) (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 \( D_X = \{x^{(i)}\}_{i=1}^m \) and \( D_Y = \{y^{(i)}\}_{i=1}^n \) 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 \( (x^{(i)}, y^{(j)}) \) 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 \( f_\theta : X \rightarrow Y \) transfers the sequence \( x \) with style \( s_x \) into a sequence \( y' \) with style \( s_y \), while the backward model \( g_\phi : Y \rightarrow X \) transfers the sequence \( y \) with style \( s_y \) into a sequence \( x' \) 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 \( y' \) matches the target style. For the latter, the probability of feeding \( y' \) to the backward model \( g \) to reconstruct \( x \) can reflect how much content information preserved in the source sentence \( x \). Because of the discrete connection \( y' \) 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.

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 \( x \) from corpus \( D_X \), model \( f \) will be trained based on two rewards provided by the style classifier \( cls_\phi \) and model \( g \). Meanwhile, starting from sampling a sequence \( y \) from \( D_Y \), model \( g \) can be trained based on the two rewards provided by the style classifier \( cls_\phi \) and model \( f \). The overview of DualRL is shown in Algorithm 1. The definitions of the two rewards and the gradients for model \( f \) are introduced as follows, and those for model \( g \) is computed in a similar way.

**Reward**

Since the gold transferred result of input \( x \) is unavailable, the quality of the generated sentence \( y' \) 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 \( y' \) matches the target style. Formally, the style classifier reward is formulated as

\[ R_s = P(s_y | y'; \varphi) \]  

where \( \varphi \) is the parameter of the classifier and is fixed during the training process.

---

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

1: Pre-train text style transfer models \( f_\theta \) and \( g_\phi \) using pseudo-parallel sentence pairs from corpora \( D_X \) and \( D_Y \)
2: Pre-train a binary style classifier \( cls_\phi \)
3: for each iteration \( i = 1, 2, \ldots, M \) do
4:   \( \triangleright \) Start to train model \( f_\theta \)
5:   Sample sentence \( x \) from \( D_X \)
6:   Generate sentence \( y' \) of opposite style via model \( f_\theta \)
7:   Compute style reward \( R_s \) based on Eq. 1
8:   Compute content reward \( R_c \) based on Eq. 1
9:   Compute total reward \( R \) based on Eq. 3
10: Update \( \theta \) using reward \( R \) based on Eq. 4
11: Update \( \theta \) using annealing teacher-forcing via MLE
12: \( \triangleright \) Start to train model \( g_\phi \)
13: Sample sentence \( y \) from \( D_Y \)
14: Generate sentence \( x' \) of opposite style via model \( g_\phi \)
15: Compute style reward \( R_s \) similar to Eq. 1
16: Compute content reward \( R_c \) similar to Eq. 2
17: Compute total reward \( R \) based on Eq. 3
18: Update \( \phi \) using reward \( R \) similar to Eq. 4
19: Update \( \phi \) using annealing teacher-forcing via MLE
20: end for
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 $y'$ by means of the probability that the model $g$ reconstructs $x$ when taking $y''$ as input. Formally, the corresponding reconstruction reward is formulated as
\[ R_c = \log(P(x|y'; \phi)) \] (2)
where $\phi$ is the parameter of model $g$. Another intuitive way to measure the content preservation is to calculate the BLEU score [Papineni et al., 2002] of $x''$ with the input $x$ as reference, where $x''$ is the output of the backward model $g$ when taking $y'$ 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
\[ R = \frac{1 + \beta^2}{\beta^2 \cdot R_c + R_s} \] (3)
where $\beta$ 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 $y'$, whose gradient w.r.t. the parameter $\theta$ of the forward model $f$ is estimated by sampling as
\[
\nabla_\theta \mathbb{E}[R] = \nabla_\theta \sum_k P(y'_k|x; \theta) R_k
\]
\[
= \sum_k P(y'_k|x; \theta) R_k \nabla_\theta \log(P(y'_k|x; \theta))
\]
\[
\simeq \frac{1}{K} \sum_{k=1}^K R_k \nabla_\theta \log(P(y'_k|x; \theta))
\] (4)
where $R_k$ is the reward of the $k_{th}$ sampled sequence $y'_k$ from model $f$, and $K$ is the sample size.

2.2 DualRL for Unsupervised Task
When applied to the field of text generation, RL faces two inained 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.

1: Initialize the iteration interval $p$
2: for each iteration $i = 1, 2, ..., M$ do
3: \( \triangleright \) Start to train model $f_\theta$
4: Update parameter $\theta$ via RL based on Eq. 4
5: if $i \% p = 0$ then
6: \( \triangleright \) Pseudo Teacher-Forcing
7: Generate a pair of data $(x'_i, y_i)$, where $x'_i = g(y_i)$
8: Update $\theta$ using data $(x'_i, y_i)$ via MLE
9: end if
10: \( \triangleright \) Start to train model $g_\phi$
11: Update parameter $\phi$ via RL similar to Eq. 4
12: if $i \% p = 0$ then
13: \( \triangleright \) Pseudo Teacher-Forcing
14: Generate a pair of data $(y'_i, x_i)$, where $y'_i = f(x_i)$
15: Update $\phi$ using data $(y'_i, x_i)$ via MLE
16: end if
17: Exponential increase in $p$ based on Eq. 5
18: end for

Annealing Pseudo Teacher-Forcing
Teacher-forcing is the strategy that feeds the parallel data $(x, y)$ 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 $f_\theta/g_\phi$ at previous iteration $i - 1$ to generate a higher quality of pseudo-parallel data $(x'_i, y_i)/(y'_i, x_i)$ than those used during pre-training on-the-fly to update model $g_\phi/f_\theta$ 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 $p$
\[
p = \min(p_0 \times r^i, p_{max})
\] (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.

$^2y_i$ and $y'_i$, denote the original data and $y'_i$ and $x'_i$, denote the corresponding generated pseudo-parallel data at $i$-th iteration.
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 crowd-workers 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 \(f\) and \(g\) are pre-trained 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 \(\beta\) 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 \(f\) and \(g\) 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] \(^*\) 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.

\(^*\)The BLEU score is computed using multi-bleu.perl.
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.
Table 5: Example outputs on the YELP and GYAFC datasets. Improperly generated words and grammar errors are colored.

<table>
<thead>
<tr>
<th>Source</th>
<th>From informal to formal (GYAFC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CrossAlign</td>
<td>Everyone on the fish, they were fresh and filling.</td>
</tr>
<tr>
<td>Template</td>
<td>Moving past the shape, they are a wonderful truly.</td>
</tr>
<tr>
<td>Del-Retri</td>
<td>Moving past the shape is awesome, and they will definitely be back!</td>
</tr>
<tr>
<td>UnsuperMT</td>
<td>Moving moving the shape, they were juicy and truly delicious.</td>
</tr>
<tr>
<td>DualRL</td>
<td>Moving past the shape, they were tasty and truly delicious.</td>
</tr>
</tbody>
</table>

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 relieve 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.

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 relieve 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


