PRELIMINARY VERSION: DO NOT CITE
The AAAI Digital Library will contain the published
version some time after the conference

Stylized Dialogue Response Generation Using Stylized Unpaired Texts

Yinhe Zheng[1,2*] **, Zikai Chen**[1*]**, Rongsheng Zhang**[3]**, Shilei Huang**[3]**, Xiaoxi Mao**[3]**, Minlie Huang**[1†]

[1] Department of Computer Science and Technology, Institute for Artifical Intelligence, State Key Lab of Intelligent Technology and Systems, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, China.

[2] Samsung Research China - Beijing (SRC-B), Beijing, China

[3] Fuxi AI Lab, NetEase Inc., Hangzhou, China

yh.zheng@samsung.com, natnstart@gmail.com, {zhangrongsheng, huangshilei, maoxiaoxi}@corp.netease.com, aihuang@tsinghua.edu.cn

Abstract

Generating stylized responses is essential to build intelligent and engaging dialogue systems. However, this task is far from well-explored due to the difficulties of rendering a particular style in coherent responses, especially when the target style is embedded only in unpaired texts that cannot be directly used to train the dialogue model. This paper proposes a stylized dialogue generation method that can capture stylistic features embedded in unpaired texts. Specifically, our method can produce dialogue responses that are both coherent to the given context and conform to the target style. In this study, an inverse dialogue model is first introduced to predict possible posts for the input responses. Then this inverse model is used to generate stylized pseudo dialogue pairs based on these stylized unpaired texts. Further, these pseudo pairs are employed to train the stylized dialogue model with a joint training process. A style routing approach is proposed to intensify stylistic features in the decoder. Automatic and manual evaluations on two datasets demonstrate that our method outperforms competitive baselines in producing coherent and style-intensive dialogue responses.

Introduction

Building a dialogue agent that can produce stylized and coherent responses has been one of the major challenges in dialogue systems (Dinan et al. 2019). Such an agent can yield more vivacious dialogues and deliver more engaging conversations by taking advantage of the linguistic style matching phenomenon (Niederhoffer and Pennebaker 2002), which suggests that people tend to adjust their linguistic style during communication to pursue higher engagement.

Generating stylized dialogue responses has been investigated in various studies, where the definition of styles covers a variety of subtle concepts, such as sentiment (Shen et al. 2017b), emotion (Zhou et al. 2018), or persona (Li et al. 2016b). Despite the success, previous studies are generally conducted in a fully supervised setting requiring dialogue pairs in the target style. However, in most cases, the stylistic features we want to capture are embedded in unpaired texts that can not be directly utilized by these supervised models (Gao et al. 2019).

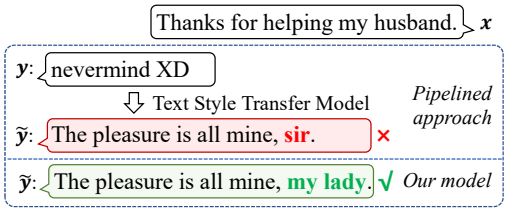

Figure 1: A pipelined approach to produce formal dialogue responses. For a post x, a response y is first produced using a dialogue model and then it is transferred to a formal response $\widetilde{y}$ using a text style transfer model. This approach may introduce incoherent contents in $\widetilde{y}$.

Few studies for dialogue modeling have been proposed to capture the stylistic features embedded in unpaired texts. Specifically, Niu and Bansal (2018) and Su et al. (2019) employs a style-aware reinforce loss, and Gao et al. (2019) resorts to a joint continuous latent space. However, despite the reported feasibility, we argue that due to the discrete nature of texts and subtle definition of text styles, it is hard to produce coherent and style-specific responses by relying on sparse reinforce signals or controlling continuous representations.

Note that we can also implement a straightforward stylized dialogue generation pipeline with the help of an unsupervised text style transfer model (Hu et al. 2017), which can be trained using stylized unpaired texts. Specifically, for a post x, a non-stylized dialogue response y is first generated using a regular dialogue model. Then y is transferred to a stylized response $\widetilde{y}$ using a text style transfer model. However, this approach may hurt the coherence between x and $\widetilde{y}$ since the style transferring process is unaware of x and may introduce inappropriate content. As shown in Figure 1, the style transfer model generates a strong stylistic word "sir" to emphasize the formality of $\widetilde{y}$. However, this makes $\widetilde{y}$ incoherent with x since x is most likely to be issued by a female.

In this paper, we propose to build a stylized dialogue generation model that can capture stylistic features embedded in a set of unpaired texts $\mathcal{D}_s$. Specifically, to tackle the problem of lacking stylized dialogue pairs, an inverse dialogue

*Equal contribution

†Corresponding Author: aihuang@tsinghua.edu.cn

model is built to predict posts based on the responses, and a set of stylized pseudo dialogue pairs are constructed by producing pseudo posts for texts in $\mathcal{D}_s$. A stylized dialogue model is then trained using these pseudo pairs, and a joint training process is introduced to enhance the coherency between the post and the resulting responses. Moreover, our dialogue models are parameterized using the Transformer-based encoder-decoder framework and initialized with the pre-trained GPT weights (Radford et al. 2018). A style routing approach is devised to fuse a style embedding in each decoder block of the stylized dialogue model to intensify the stylistic features in the decoding process.

We evaluate our method on two datasets with two distinct writing styles: 1) Jinyong novels[1] in Chinese, and 2) formality in English writing. Automatic and human evaluations show that our method significantly outperforms competitive baselines with a large margin in generating coherent dialogue responses while rendering stronger stylistic features.

Our contributions can be summarized as:

1) A novel method is proposed to build a stylized dialogue model that can capture stylistic features embedded in unpaired texts. Specifically, an inverse dialogue model is introduced to generate stylized pseudo dialogue pairs, which are further utilized in a joint training process. An effective style routing approach is devised to intensify the stylistic features in the decoder.

2) Automatic and human evaluations on two datasets show that our method outperforms competitive baselines with a large margin in producing stylized and coherent dialogue responses.

Related Work

Stylized dialogue generation has attracted numerous attention in recent years (Gao et al. 2019; Niu and Bansal 2018). With a rather wide definition of styles, various studies that focus on controllable dialogue generation have been categorized as "stylized" dialogue generation, such as generating personalized (Li et al. 2016b; Luan et al. 2017; Su et al. 2019) or emotional (Zhou et al. 2018) dialogues. However, these dialogue models' training process usually requires dialogue pairs in the target style, whereas our study aims to capture stylistic features embedded in unpaired texts.

Moreover, the styles defined in most previous studies are deeply fused with the text contents (Tikhonov et al. 2019). Enforcing these styles may limit the dialogue model's expressive ability because there are contradictions between certain semantic contents and style categories. For example, it is hard, if not impossible, for a service agent to yield comforting content when enforcing a negative sentiment. Unlike most previous works, our study investigates to model the writing styles that are mostly "orthogonal" to the text semantic. The contents we want to deliver will not be constrained by the style we intend to render.

Text style transfer is a related but different task compared to our work. Specifically, these text style transfer models aim to preserve the style-agnostic contents of the input text (Fu et al. 2018). In contrast, our study aims to produce coherent responses rather than preserve the contents of the posts. Early works on this task focus to disentangle the representation of styles and contents (Hu et al. 2017; Shen et al. 2017a; Prabhumoye et al. 2018). However, recent studies argue the effectiveness of such disentanglement (Lample et al. 2019) and propose to revise the latent codes using classifiers (Liu et al. 2020; Wang, Hua, and Wan 2019). Some works are also proposed to render the target styles by replacing stylistic words (Wu et al. 2019a,b).

We have also noticed a recent work that considers a contextual constraint in the text style transferring process (Cheng et al. 2020). However, although being feasible, the training of this model requires style-labeled parallel data. This hinders us from directly employing this model in our study since these parallel data are usually unavailable.

Back translation is a popular approach that has been widely employed in various NLP tasks such as machine translation (Sennrich, Haddow, and Birch 2016), dialogue data augmentation (Su et al. 2020), text style transfer (Zhang et al. 2018; Lample et al. 2019; Dai et al. 2019), and even stylized dialogue generation (Wang et al. 2017). This approach is similar to the inverse dialogue model introduced in our study. However, different from previous approaches that only try to model the one-to-one mapping between the source and target inputs, our inverse dialogue model tries to capture the one-to-many mappings between the responses and posts with the help the proposed joint training process. In our study, the diversity of the generated pseudo posts are enhanced using a sampling approach.

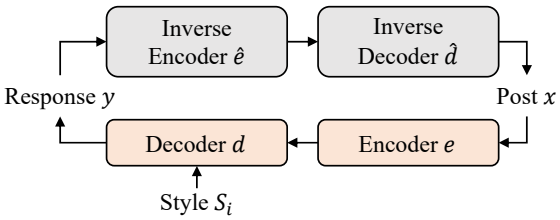

Figure 2: Overall framework.

Method

Task Definition

In this study, we propose to build a stylized dialogue model without utilizing dialogue pairs in the target style. Specifically, our method takes as input two sets of data in the training stage: 1) M unpaired texts $\mathcal{D}_s = \{t_1, ..., t_M\}$ in the writing style S_1; 2) N dialogue pairs $\mathcal{D}_p = \{\langle x_1, y_1\rangle, ..., \langle x_N, y_N\rangle\}$ with style S_0, where x_i and y_i is the post and response, respectively. Our stylized dialogue model aims to generate a response y that is coherent to a given post x while exhibiting a certain style S_i $(i = 0, 1)$:

$$y = \arg\max_{y'} p(y'|x, S_i). \quad (1)$$

[1] Jinyong is a famous Chinese writer who wrote many Kung Fu novels.

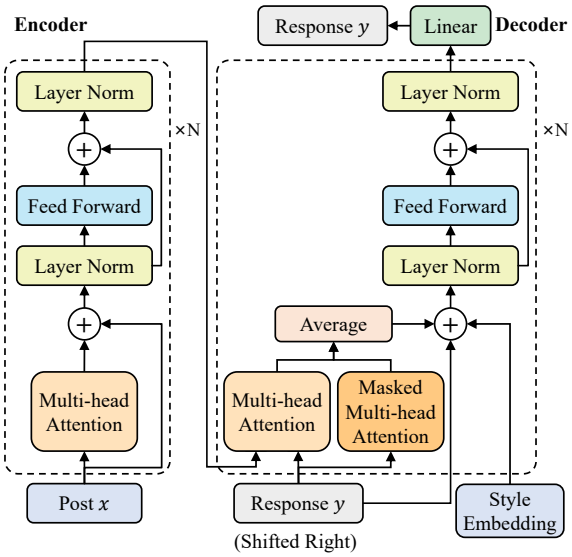

Figure 3: Architecture of the stylized dialogue model.

Model Overview

Our model consists of two mirrored sub-modules (Figure 2): (1). A stylized dialogue module (i.e., e and d in Figure 2) that can produce a stylized response y based on a given post x and a style label $S_i (i = 0, 1)$. A style routing approach is devised to incorporate stylistic features in d; (2). An inverse dialogue module (i.e., $\hat{e}$ and $\hat{d}$ in Figure 2) that aims to produce pseudo posts x based on an input response y. Note that the inverse dialogue model is introduced to tackle the problem of lacking dialogue pairs in style S_1, i.e., we can regard the texts in $\mathcal{D}_s$ as possible dialogue responses and use the predicted pseudo posts to construct pseudo dialogue pairs in style S_1. Therefore, we omit the style label in the inverse decoder $\hat{d}$ to encourage it to focus more on the semantic aspect of the dialogue.

The dialogue modules in our study are parameterized using the Transformer-based encoder and decoder architecture (Vaswani et al. 2017) and are initialized using pretrained GPT (Radford et al. 2019) weights. Further, we also follow previous works (Golovanov et al. 2019) to share the weights of the encoder and decoder from the same submodule to save memories. Particularly, the weights of e and d are shared, and the weights of $\hat{e}$ and $\hat{d}$ are shared.

Moreover, to better capture the one-to-many phenomenon and alleviate the problem of producing trivial posts in the inverse dialogue model, a top-k sampling scheme is employed to sample multiple pseudo posts for each stylized text in $\mathcal{D}_s$. All these sampled posts are utilized in the training process. Further, a joint training process is also introduced to train these two sub-modules iteratively to enhance the coherency of the response.

Style Routing

There exist various approaches to condition the decoder d on the style label. For example, employing a special style token as the start token (Lample et al. 2019), or adding a style embedding to each word embedding (Zheng et al. 2020). However, these approaches only incorporate the style representation in the decoder's input layer, whereas the higher layers are not explicitly affected.

In this study, a style routing approach is devised to enhance existing approaches to stylize d in the stylized dialogue model (see Figure 3). Specifically, in each decoder block, we first fuse the representation of the post x and previously decoded token sequence y_p using the attention routing mechanism (Zheng et al. 2020), i.e., two sequences of representations, $\boldsymbol{R}_{prev}, \boldsymbol{R}_{post} \in R^{l \times h}$, are first calculated:

$$\boldsymbol{R}_{prev} = \mathrm{MMHA}[\boldsymbol{e_w}(y_p), \boldsymbol{e_w}(y_p), \boldsymbol{e_w}(y_p)], \quad (2)$$

$$\boldsymbol{R}_{post} = \mathrm{MHA}[\boldsymbol{e_w}(y_p), \boldsymbol{e}(x), \boldsymbol{e}(x)], \quad (3)$$

where $\boldsymbol{e_w}(y_p) \in R^{l \times h}$ denotes the embedding of y_p and it is used as the query in MMHA and MHA, which represent the masked and un-masked multi-head attention operation, respectively. l is the length of y_p, and h is the hidden size. $\boldsymbol{e_w}(x)$ is the output of the encoder. A sequence of fused representations $\boldsymbol{R}_{avg}$ is obtained as:

$$\boldsymbol{R}_{avg} = (\boldsymbol{R}_{prev} + \boldsymbol{R}_{post})/2. \quad (4)$$

Then for a given style S_i, a style embedding $\boldsymbol{e_s}(S_i) \in R^{1 \times h}$ is allocated and $\boldsymbol{e_s}(S_i)$ is routed into $\boldsymbol{R}_{avg}$ by adding it to each time step of the sequence:

$$\boldsymbol{R}_{merge} = \boldsymbol{R}_{avg} + \boldsymbol{e_s}(S_i). \quad (5)$$

Also note that the fusion operation in Eq. 4 and 5 is similar to some previous studies that try to incorporate additional contexts in a transformer-based decoder (Golovanov et al. 2019). However, different from these approaches that focus to model sequential contexts, the styles modeled in our study are categorical, and more priority is allocated to the style representation in our model. Moreover, we are the first to use such a style routing approach in the stylized dialogue generation task.

Joint Training

The training of our model involves the following losses: 1) standard maximum log likelihood losses evaluated on dialogue pairs from $\mathcal{D}_p$:

$$\mathcal{L}_{p2r} = \mathop{\mathbb{E}}_{\langle x,y \rangle \sim \mathcal{D}_p} -\log p_d(y|\boldsymbol{e}(x), S_0), \quad (6)$$

$$\mathcal{L}_{r2p} = \mathop{\mathbb{E}}_{\langle x,y \rangle \sim \mathcal{D}_p} -\log p_{\hat{d}}(x|\hat{\boldsymbol{e}}(y)). \quad (7)$$

The loss $\mathcal{L}_{p2r}$ and $\mathcal{L}_{r2p}$ is used to train the stylized dialogue model and inverse dialogue model, respectively; 2) an inverse dialogue loss evaluated on texts from $\mathcal{D}_s$:

$$\mathcal{L}_{inv} = \mathop{\mathbb{E}}_{\substack{t \sim \mathcal{D}_s, \\ x' \sim p_{\hat{d}}(x|\hat{\boldsymbol{e}}(t))}} -\log p_d(t|\boldsymbol{e}(x'), S_1), \quad (8)$$

in which x' is the pseudo post sampled from the inverse dialogue model.

Note that the gradient back-propagation through the loss $\mathcal{L}_{inv}$ is intractable due to the in-differentiable sampling process in Eq. 8. In this study, we approximate the ideal back-propagation process through $\mathcal{L}_{inv}$ by truncating the gradients associated with the sampling operation. Specifically,

Algorithm 1 Joint training process

Input: M unpaired texts: $\mathcal{D}_s$=$\{t_i\}_{i=1}^{M}$ in style S_1,
N dialogue pairs $\mathcal{D}_p$=$\{\langle x_i, y_i\rangle\}_{i=1}^{N}$ in style S_0.
Output: A stylized dialogue model

Init the stylized and inverse dialogue model e, d, $\hat{e}$, $\hat{d}$
while not converge **do**
  Sample n_d dialogue pairs $\mathcal{D}_p^b = \{\langle x_i, y_i\rangle\}_{i=1}^{n_d} \subset \mathcal{D}_p$
  Train e and d by optimizing $\mathcal{L}_{p2r}$ (Eq. 6) on $\mathcal{D}_p^b$
  Train $\hat{e}$ and $\hat{d}$ by optimizing $\mathcal{L}_{r2p}$ (Eq. 7) on $\mathcal{D}_p^b$
  if Current Step $> N_f$ **then**
    $\mathcal{D}_{pp} \leftarrow$ empty set.
    Sample n_s stylized texts $\mathcal{D}_s^b = \{t_i\}_{i=1}^{n_s} \subset \mathcal{D}_s$
    for each $t_i \in \mathcal{D}_s^b$ **do**
      Decode m posts $\{x'_{ij}\}_{j=1}^{m}$ from $p_{\hat{d}}(x|\hat{\boldsymbol{e}}(t_i))$
      $\mathcal{D}_{pp} \leftarrow \mathcal{D}_{pp} \bigcup \{\langle x'_{ij}, t_i\rangle\}_{j=1}^{m}$
    end for
    Train e and d by optimizing $\mathcal{L}_{inv}$ (Eq. 8) on $\mathcal{D}_{pp}$
  end if
end while

Dataset		Train		Test	
		$\mathcal{D}_p$	$\mathcal{D}_s$	$\mathcal{D}_t$	
WDJN	Size	300.0K	95.13K	2.0K	2.0K
	Style	Weibo	Jinyong	Weibo	Jinyong
TCFC	Size	217.2K	500.0K	0.97K	0.97K
	Style	Informal	Formal	Informal	Formal

Table 1: Statistics of datasets

when optimizing $\mathcal{L}_{inv}$, the parameters of the inverse dialogue model are fixed, and the stylized dialogue model is trained with pseudo posts x' that are sampled from the inverse dialogue model. Similar approaches have been proven to be effective in other NLP tasks (Lample et al. 2018; He et al. 2020). However, unlike previous works that use the greedy decoding scheme, our study employs the top-k sampling scheme with beam search to produce x' since the mapping between dialogue responses and posts is not unique. The greedy decoding scheme may limit the diversity of the decoded pseudo posts and lead to sub-optimal performance.

To facilitate the learning with the above gradient approximation approach, a joint training process is introduced to train the model iteratively. Specifically, in each training iteration, we first update the stylized and inverse dialogue model by optimizing the losses $\mathcal{L}_{p2r}$ and $\mathcal{L}_{r2p}$ using a batch of dialogue pairs sampled in $\mathcal{D}_p$. Further, a batch of stylized sentences $\mathcal{D}_s^b$ are sampled from $\mathcal{D}_s$. For each sentence $t_i \in \mathcal{D}_s^b$, m pseudo posts $x'_{i1}, ..., x'_{im}$ are sampled from the inverse dialogue model, and m pseudo dialogue pairs $\langle x'_{ij}, t_i\rangle, (j = 1, ..., m)$ in the style S_1 are constructed. These pseudo pairs are used to train the stylized dialogue model with the loss $\mathcal{L}_{inv}$. To avoid corrupted pseudo posts at the beginning of the training process, we pre-train the inverse dialogue model on $\mathcal{L}_{r2p}$ for N_f steps before using it to decode pseudo posts. The detailed training process is summarized in Algorithm 1.

Experiment

Dataset

Our method is evaluated on two datasets with two distinct styles (see statistics in Table 1).

1) WDJN: We collected 300K Weibo Dialogues (style S_0) as $\mathcal{D}_p$ and sampled 95.1K stylized unpaired texts that are wrapped in quotation marks in Jinyong's Novels (style S_1) as $\mathcal{D}_s$. The texts in $\mathcal{D}_s$ are mostly "spoken utterances" that are issued by novel characters.

The testing data of the WDJN dataset $\mathcal{D}_t$ also involves two parts: The first part contains 2.0K additional dialogue pairs collected from Weibo; The second part contains 2.0K dialogue pairs extracted from Jinyong's novel. Specifically, we regard two consecutive spoken utterances that are wrapped in quotation marks as a dialogue pair. We filtered $\mathcal{D}_s$ and $\mathcal{D}_t$ to avoid overlaps.

Also note that to prevent the model from copying stylistic phrases in the post when producing Jinyong style responses in the testing phase, we erased the stylistic features related to Jinyong's writing from the *posts* in these 2K Jinyong style dialogues in $\mathcal{D}_t$ using the back translation approach (Zhang, Ge, and Sun 2020). Moreover, all the resulting posts are manually checked and revised to ensure the stylistic features related to style S_1 are erased. More details about the WDJN dataset can be found in Appendix A. The WDJN dataset will be released for public use.

2) TCFC (Wu, Wang, and Liu 2020): This dataset focuses on the formality in English writing. We sampled 217.2K informal dialogue pairs (style S_0) as $\mathcal{D}_p$ and 500.0K formal texts (style S_1) as $\mathcal{D}_s$ from the original dataset, and used the test data in the original dataset as our test set $\mathcal{D}_t$, which contains 1,956 manually-crafted dialogue pairs (978 informal pairs and 978 formal pairs).

Implementation Details

For experiments on the WDJN and TCFC dataset, we used the pre-trained CDial-GPT (Wang et al. 2020) and DialoGPT (size 345M) (Zhang et al. 2019) model to initialize the dialogue modules, respectively. The top-K sampling process in Algorithm 1 employs a $K = 20$ and beam size of 4 (WDJN) or 2 (TCFC). The value of N_f is set to 300. The training of our model stops after 10 iteration epochs on $\mathcal{D}_p$ (WDJN) or after 8,000 steps of updates (TCFC). See Appendix B for more details of the reproduction guidance.

Baselines

We choose two groups of baselines:

The first group contains dialogue models with different style modeling scheme: **1) *S2S*** (Golovanov et al. 2019): a strong Transformer-based dialogue model that is only trained on $\mathcal{D}_p$. This baseline can only produce responses in style S_0; **2) *SLM***: the "Fusion" model proposed by Niu and Bansal (2018), in which an independent stylized language model is trained on $\mathcal{D}_s$, and the distributions decoded from the *S2S* baseline and the stylized LM are fused when producing responses in style S_1; **3) *SRL***: the "RL" model proposed by Niu and Bansal (2018), in which a reinforce signal produced by a style classifier is used to enforce the style S_1; **4)**

Model	WDJN Dataset									TCFC Dataset							
	BLEU-1,2		Dist.	BERT	SVM	Flu.	Coh.	Style	HAvg.	BLEU-1,2		Dist.	BERT	SVM	Flu.	Coh.	Style	HAvg.
SLM	2.90	0.37	26.6	26.7	40.7	1.96*	1.52	0.37	0.79	12.6	0.99	42.5	85.6	87.2	1.90*	0.89	1.78	1.36
SRL	2.53	0.33	40.4	36.2	43.2	1.83	1.52	0.39	0.82	7.83	0.70	42.7*	47.6	53.5	1.76	0.72	1.25	1.09
SFusion	3.84	0.20	33.1	8.24	19.8	1.63	0.69	0.40	0.67	5.51	0.28	**61.0**	21.9	39.0	1.47	0.56	1.17	0.90
S2S+BT	6.22	0.68	30.7	66.0	83.6	1.89	1.53*	0.63	1.09	12.1	1.25	42.0	86.3	86.8	1.58	0.72	1.66	1.14
S2S+CT	11.3	0.62	32.4	72.3	76.4	0.45	0.19	**1.50**	0.38	8.05	0.64	60.9	67.7	67.8	0.37	0.12	0.64	0.24
S2S+PTO	3.57	0.44	32.9	35.1	43.3	1.82	1.54*	0.35	0.75	9.55	0.84	34.5	28.6	50.3	0.35	0.26	0.39	0.32
Ours	**13.6**	**1.53**	**42.8**	**78.3**	**89.3**	**1.96**	**1.60**	1.16	**1.48**	**15.1**	**1.71**	43.4	**97.3**	**96.1**	**1.90**	**1.01**	**1.89**	**1.46**
Human	N/A		49.3	80.1	85.4	1.93	1.60	1.53	1.67	N/A		62.7	89.6	85.8	1.91	1.18	1.83	1.56

Table 2: Automatic and manual evaluation results for responses in style S_1. All differences between our model and baselines are significant with p-value <0.05 except for the ones marked with *.

Model	WDJN Dataset									TCFC Dataset							
	BLEU-1,2		Dist.	BERT	SVM	Flu.	Coh.	Style	HAvg.	BLEU-1,2		Dist.	BERT	SVM	Flu.	Coh.	Style	HAvg.
S2S	8.50	2.42	35.1	97.0	93.0	1.96*	1.73	1.86	1.85*	6.92*	0.61*	54.8	70.1*	60.9	1.82*	1.16*	1.68*	1.50*
SFusion	8.65	0.82	35.3	**99.9**	**99.2**	1.41	0.74	1.92*	1.16	4.61	0.22	**62.8**	**70.3**	**61.1**	1.57	0.76	**1.77***	1.19
Ours	**11.6**	**2.93**	**39.0**	93.5	89.2	**1.97**	**1.85**	**1.93**	**1.92**	**6.96**	**0.67**	49.4	69.4	59.2	**1.85**	**1.16**	1.70	**1.51**
Human	N/A		56.4	97.9	94.4	1.89	1.86	1.98	1.91	N/A		72.6	72.0	72.1	1.76	1.19	1.76	1.52

Table 3: Automatic and manual evaluation results for responses in style S_0. All differences between our model and baselines are significant with p-value <0.05 except for the ones marked with *.

SFusion (Gao et al. 2019): A fused latent space is built using a multi-task training scheme. Specifically, for each post, six responses are sampled, and two classifiers are used to rank these responses for the styles.

The second group of baselines are built using the pipelined approach, i.e., different unsupervised text style transfer models are trained on texts from $\mathcal{D}_s$ and $\mathcal{D}_p$, and responses produced by the *S2S* baseline (in style S_0) are transferred to exhibit the target style S_1 using these models: **5) *S2S+BT***: a back-translation-based text style transfer model (He et al. 2020); **6) *S2S+CT***: a model that tries to entangle the latent code for styles and contents (Wang, Hua, and Wan 2019); **7) *S2S+PTO***: a model that renders the target style by replacing stylistic words (Wu et al. 2019a).

Note that for baselines 2, 3 and 5-7, the responses generated by the S2S baseline are used as their responses for the S_0 style since they can only produce responses in S_1 once trained. Moreover, we implemented baselines 1-3 using the same architecture and hyper-parameters as our model for fair comparisons. For baselines 4-7, we used the official codes released by the authors. Note that it is non-trivial to utilize the pre-trained GPT model in the baseline *SFusion* since it handles fixed-length latent codes.

Automatic Evaluation

Metrics: We first used automatic metrics to evaluate the response quality of our model: 1). ***BLEU*** (Papineni et al. 2002) was used to measure n-gram (n=1, 2) overlap between the generated responses and the reference responses; 2). *Distinct* (**Dist.**) (Li et al. 2016a) measures the proportion of unique n-grams in the generated responses (n=2).

To evaluate the style intensity of each model, we first trained two text style classifiers (i.e., ***BERT*** (Devlin et al. 2019) and ***SVM***) and then calculated the style intensity score as the portion of generated responses that conform to the target style based on these classifiers. In our study, texts from $\mathcal{D}_p$ and $\mathcal{D}_s$ were used to train classifiers for the WDJN experiments, and the GYAFC dataset (Rao and Tetreault 2018) was used to train classifiers for the TCFC experiments. The accuracy of the BERT and SVM classifier on the holdout test set was 98.52% and 94.20% respectively for the WDJN experiments, and 93.98% and 89.57% respectively for the TCFC experiments (see Appendix C for more details).

Results: We separately evaluated the responses in style S_1 (Table 2) and S_0 (Table 3). Note that the baseline *S2S* is not included in Table 2 since it can not produce responses in style S_1. Similarly, only the baselines *S2S* and *SFusion*

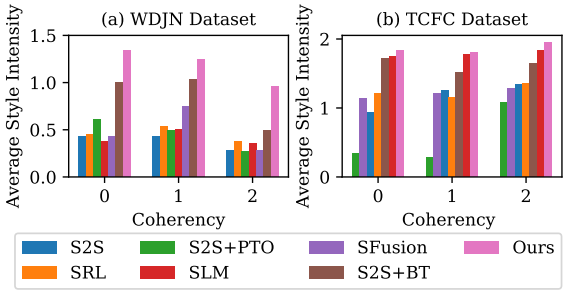

Figure 4: Averaged style intensity scores for responses with different coherency scores.

Model	WDJN Dataset									TCFC Dataset							
	BLEU-1,2		Dist.	BERT	SVM	Flu.	Coh.	Style	HAvg.	BLEU-1,2		Dist.	BERT	SVM	Flu.	Coh.	Style	HAvg.
Ours	**12.6**	**2.23**	**40.9**	85.9	89.3	**1.96**	**1.69**	1.54	**1.71**	**11.0**	**1.19**	46.4	83.3	77.7	**1.87**	**1.08**	1.79	1.49
w/o Rout.	12.5*	2.19*	40.3*	84.0	87.1	1.51	1.25	1.48	1.41	**11.0***	1.13*	**46.8***	82.8*	77.0	1.86*	1.03	1.78*	1.45
w/o JointT	7.57	1.71	30.1	**91.3**	**94.5**	1.51	1.23	1.54*	1.41	9.84	0.97	46.2*	83.9*	78.7	**1.87***	0.95	**1.81***	1.40
w/o Samp.	9.79	1.62	39.7	90.2	92.8	1.26	1.07	1.58*	1.27	10.7*	1.18*	46.7*	83.4*	78.3	**1.87***	0.99	**1.81***	1.43
w/o PreT	10.9	1.43	16.9	91.2	91.9	1.46	0.90	**1.60***	1.24	10.1*	0.74	32.5	**84.8**	**78.8**	1.86*	0.65	1.80*	1.15

Table 4: Automatic and manual evaluation results of ablation models for responses in style S_0 and S_1. All differences between our model and ablation models are significant with p-value <0.05 except for the ones marked with *.

are contained in Table 3. Significance tests are performed between the results of our model and all the baselines using the t-test with bootstrap resampling (Koehn 2004).

As can be seen from the automatic results, our method outperforms all the baselines with large margins when generating dialogue responses in style S_1 (Table 2), and achieves competitive performance when producing responses in style S_0 (Table 3). This indicates that our model can produce high-quality responses that are both coherent with the given context and consistent to the target style. We can further observe that:

1). The pipelined approaches achieve lower BLEU scores comparing to our method. This verifies our claim that the response coherency is affected by the style transferring process. Similar results are also observed in manual evaluation. Also note that some non-pipelined baselines achieve lower BLEU scores comparing to the pipelined baselines, e.g. SRL and SFusion on the TCFC dataset. This indicates that it is hard to capture the stylistic features by modifying the latent spaces.

2). The high diversity (i.e., *Dist.* scores) of the baselines on the TCFC dataset come along with a dramatic decrease on the BLEU scores. This is because these baselines overfit to the diverse colloquial phrases in the informal responses and fail to render responses in style S_1, which are more formal and less diverse.

Also note that the style intensity scores for human-generated responses (last row in Table 2 and 3) do not match the accuracy of our style classifiers. This is because these classifiers' train data involve non-conversational texts, which leads to mismatches when testing using conversational responses. To alleviate this mismatch, we performed manual evaluations to concrete our analysis.

Manual Evaluation

Metrics: For a given post, dialogue responses with different styles were generated using our model and all the baselines. Three annotators were recruited from the crowd-sourcing platform to evaluate these responses from three aspects: 1) *Fluency* (**Flu.**): whether the response is fluent and free from grammar errors; 2) *Coherency* (**Coh.**): whether the response is coherent with the dialogue context; 3) *Style Intensity* (**Style**): whether the response conforms to the given style. Each metric is rated among {0, 1, 2}, in which 0 means worst and 2 best. Moreover, the *Harmonic Average* (i.e, **HAvg.**) of above measures is also reported.

Results: We sampled 300 posts from $\mathcal{D}_t$ for each of these two datasets. Fleiss's kappa κ (Randolph 2005) was used to measure the annotation agreement between annotators. Specifically, for *Flu.*, *Coh.*, and *Style*, the κ value was 0.69, 0.50, 0.86, respectively on the WDJN dataset (indicating substantial, moderate, and substantial agreement), and 0.44, 0.31, 0.42, respectively on the TCFC dataset (indicating moderate, fair, and moderate agreement).

As shown in Table 2, our model surpasses all the baselines significantly on style intensity (except for S2S+CT on the WDJN dataset, which comes with dramatic decreases on the fluency and coherency scores) when producing responses in style S_1, and it achieves competitive or higher fluency and coherency scores. This verifies the superiority of our method in producing coherent and style intensified dialogue responses. Moreover, results in table 3 also show that our model achieved competitive performance when generating responses in style S_0.

We can also observe from Table 2 and 3 that:

1). There are trade-offs between the coherency and style intensity when generating stylized dialogue responses on the WDJN dataset, i.e., the high style intensity usually comes at the cost of a low coherency. For example, the S2S+CT achieves the best style intensity score on WDJN (1.50) when producing responses in style S_1, but it obtains the worst coherency (0.19) score. This phenomenon is also observed in various previous studies (Niu and Bansal 2018; Zheng et al. 2020). Nevertheless, our model achieves competitive coherency while producing style-intensive responses.

2). The distribution mismatch between texts in $\mathcal{D}_s$ and $\mathcal{D}_p$ may bring an adverse impact on the response quality in style S_1. For example, on the TCFC dataset, the coherency score of our model drops from 1.18 (human performance) to 1.01 in Table 2. This is because that the data in $\mathcal{D}_s$ of TCFC originate from the QA texts that are usually not in a conversation form. In contrast, our model obtains the same coherency score with human responses on WDJN since the data in $\mathcal{D}_s$ of WDJN are mostly unpaired conversational texts.

3). The baselines SFusion, SRL, and S2S+CT generally yield low *HAvg.* scores on both datasets. This verifies our claim that it is hard to generate stylized and coherent responses relying on the sparse reinforce signals (i.e., SRL) or continuous latent codes (i.e., SFusion and S2S+CT).

Our method's superiority to generate stylized dialogue responses is further demonstrated by analyzing the style intensity scores of responses with different coherency levels.

		WDJN dataset	TCFC dataset
Post		Haven't eaten hot pot in a long time (好久没吃火锅了)	It's only 9:57 pm and I'm already falling asleep.
S_0	S2S	I haven't eaten hot pot in a long time too (我也好久没吃火锅)	You're not falling asleep yet, lol
	SFusion	I also want to eat, just started (我也想吃,刚刚开始)	dude same here, my friend has a reason at night
	Ours	I also want to eat (我也想吃了)	it's almost 9 am here and i just got up...
S_1	SLM	With that said, I want to eat too (这么一说,我也想吃了)	I have a headache and I can not stop drinking.
	SRL	I haven't eaten in a long time. I really want to eat (好久没吃了,好想吃啊)	isn't it 5:30 in the morning?
	SFusion	I'm almost done (我已经快好了)	Same here but I think it's gna say hello!
	S2S+BT	We haven't eaten hot pot in a long time (我们好久没吃火锅)	She is not falling asleep yet
	S2S+CT	I have no problem for a long time too. I went to the hot pot but unfortunately they didn't (我也好久没问题,老衲去打了火锅可惜他们没)	That is not falling asleep then Maguties out for riddle
	S2S+PTO	I haven't eaten hot pot in a long time too (我也好久没吃火锅)	/ ' re not falling asleep yet
	Ours	Pretty good, but hero, you are hungry for a whole day. Let's eat first! (不错,大侠饿了一天,现下先吃饭吧！)	Yes, it is 9:06 pm here, and I am still on the couch

Figure 5: Example responses produced by our model and the baselines on the TCFC and WDJN datasets.

TCFC dataset	
Pseudo Post:	Are you enjoying the new album?
Text in S_1:	Yes, I am. I am loving her last cd.
Pseudo Post:	I'm so tired of golf.
Text in S_1:	What is the point of golf?
Pseudo Post:	Hey, are you going to the game tonight?
Text in S_1:	Hardly, I live up north. Maybe next time.
WDJN dataset	
Pseudo Post:	I am very, very sad today (今天的我,伤心的不得了)
Text in S_1:	Are your hurt? (你受伤了么？)
Pseudo Post:	Did anyone come to see me today? (今天有人来看我么?)
Text in S_1:	Brother, someone is coming. (大哥，有人来啦。)
Pseudo Post:	I'm going to kill you today (今天我要杀了你)
Text in S_1:	Dude, you could have killed me, but you didn't. (老兄，刚才你本可杀我，没有下手。)

Figure 6: Example pseudo pairs generated by the inverse dialogue model in the training process.

Specifically, all the annotated responses in style S_1 were collected and categorized into three groups based on the coherency scores (i.e., 0, 1, or 2) they received. The averaged style intensity score for each group was calculated and shown in Figure 4. It can be seen that our model achieves the highest style intensity scores in all coherency groups. This further demonstrates that the responses produced by our method are more style-intensive than those by the baselines. Note that the baseline S2S+CT is not included in Figure 4 because its fluency score is extremely low. There is no point in comparing its coherency to other baselines.

Also note that the results in Figure 4 very much by the dataset. It is easier to capture the stylistic features in the TCFC dataset. This interesting phenomenon may dues to the fact that there are larger gaps between texts in style S_0 and S_1 in the dataset WDJN. Specifically, style S_0 texts in WDJN originate from the web corpus (i.e., Weibo) and style S_1 texts originate from Kung-fu novels written in 1960s. Such gaps in TCFC is much smaller since style S_0 and S_1 texts in TCFC all originate from the web corpus. This also explains why the the trade-offs between the coherency and style intensity in TCFC dataset is not that significant. For example, it is hard, if not impossible, to describe some modern events (e.g. ipod or 5G networks) using phases in 1960s' Kung-fu novels.

Ablation Study

Ablation studies were performed to verify the effect of each component in our method. Specifically, the following variants were tested: 1) without the style routing approach (**w/o Rout.**), i.e., the style embedding is not incorporated in each decoder block as in Eq.5. The decoder d is stylized by employing stylized start token and adding a style embedding to each word embedding; 2) Without the joint training process (**w/o JointT**), i.e., an inverse dialogue model is first trained and fixed, and then a fixed set of pseudo pairs are generated and used to train the stylized dialogue model. Note that the same amount of pseudo pairs were used to optimize the loss $\mathcal{L}_{inv}$ in this variant as it is used in Algorithm 1; 3) Without using the top-K sampling scheme when producing pseudo posts (**w/o Samp.**), i.e., pseudo pairs are decoded greedily; 4) Without using the pre-trained GPT weights (**w/o PreT**).

As shown in Table 4, our model achieves the highest *BLEU* and *Coh.* scores among all the ablation models. We can further observe that: 1) Almost all our variants surpass the baselines with a large margin on the style intensity score. This verifies the feasibility of our framework in capturing stylistic features; 2) Removing the joint training process (**w/o JointT**) or the top-K sampling scheme (**w/o Samp.**) makes the dialogue models over-fit to render more stylistic features while failing to achieve high *BLEU* and *Coh.* scores. However, we argue that since our stylized decoder is already strong in capturing stylistic features, it is critical to utilize the proposed joint training and top-K sampling scheme to improve the response coherency; 3) The pre-training approach significantly improves the diversity and coherency of the generated responses.

Case Study

Figure 5 shows some dialogue responses generated by our model and the baselines on the two datasets. We can observe that the models that directly manipulate the continuous latent space (i.e., SFusion and S2S+CT) yield non-fluent responses. This is because it hard to build a smooth latent space for discrete texts. Moreover, on the TCFC dataset, the baseline SFusion in style S0, and baselines SLM, SFusion, S2S+BT, S2S+CT, S2S+PTO in style S1 fail to produce coherent responses to the post. Further, pipelined approaches either fail to convert the inputs to the target style (i.e., S2S+PTO on the WDJN dataset) or hurt the coherency between the response and the post (i.e., S2S+BT, S2S+CT, and S2S+PTO on the TCFC dataset).

In addition, we sampled some of these pseudo pairs generated by the inverse dialogue model in the training phase (Figure 6). It can be seen that these pseudo pairs are generally of high quality both in fluency and coherency.

Conclusion

This paper presents a stylized dialogue generation method that can produce coherent and style-intensive responses by utilizing stylized unpaired texts. An inverse dialogue model is introduced in our method to produce stylized pseudo dialogue pairs, which are used in a joint training process to train the stylized dialogue model. Further, a style routing approach is introduced to intensify stylistic features in the decoding process. We demonstrate our method on two datasets with two different styles: Chinese Jinyong novels and English formality writing. The automatic and manual evaluation shows that our method outperforms competitive baselines in producing coherent and style-intensive responses. As future works, we will extend this method to other stylized text generation tasks.

Acknowledgement

This work was jointly supported by the Major Project of the New Generation of Artificial Intelligence(No. 2018AAA0102900), and NSFC projects (Key project with No. 61936010 and regular project with No. 61876096). This work was also supported by the Guoqiang Institute of Tsinghua University, with Grant No. 2019GQG1. We thank THUNUS NExT Joint-Lab for the support.

References

Cheng, Y.; Gan, Z.; Zhang, Y.; Elachqar, O.; Li, D.; and Liu, J. 2020. Contextual Text Style Transfer. *ArXiv* abs/2005.00136.

Dai, N.; Liang, J.; Qiu, X.; and Huang, X. 2019. Style Transformer: Unpaired Text Style Transfer without Disentangled Latent Representation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5997–6007. Florence, Italy: Association for Computational Linguistics. doi:10.18653/v1/P19-1601. URL https://www.aclweb.org/anthology/P19-1601.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

Dinan, E.; Logacheva, V.; Malykh, V.; Miller, A. H.; Shuster, K.; Urbanek, J.; Kiela, D.; Szlam, A.; Serban, I.; Lowe, R.; Prabhumoye, S.; Black, A. W.; Rudnicky, A. I.; Williams, J.; Pineau, J.; Burtsev, M.; and Weston, J. 2019. The Second Conversational Intelligence Challenge (ConvAI2). *CoRR* abs/1902.00098. URL http://arxiv.org/abs/1902.00098.

Fu, Z.; Tan, X.; Peng, N.; Zhao, D.; and Yan, R. 2018. Style transfer in text: Exploration and evaluation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Gao, X.; Zhang, Y.; Lee, S.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2019. Structuring Latent Spaces for Stylized Response Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 1814–1823.

Golovanov, S.; Kurbanov, R.; Nikolenko, S.; Truskovskyi, K.; Tselousov, A.; and Wolf, T. 2019. Large-scale transfer learning for natural language generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6053–6058.

He, J.; Wang, X.; Neubig, G.; and Berg-Kirkpatrick, T. 2020. A Probabilistic Formulation of Unsupervised Text Style Transfer. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=HJlA0C4tPS.

Hu, Z.; Yang, Z.; Liang, X.; Salakhutdinov, R.; and Xing, E. P. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 1587–1596. JMLR. org.

Koehn, P. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 388–395. Barcelona, Spain: Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W04-3250.

Lample, G.; Conneau, A.; Denoyer, L.; and Ranzato, M. 2018. Unsupervised Machine Translation Using Monolingual Corpora Only. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=rkYTTf-AZ.

Lample, G.; Subramanian, S.; Smith, E.; Denoyer, L.; Ranzato, M.; and Boureau, Y.-L. 2019. Multiple-Attribute Text Rewriting. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=H1g2NhC5KQ.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016a. A Diversity-Promoting Objective Function for Neural Conversation Models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 110–119. San Diego, California: Association for Com-

putational Linguistics. doi:10.18653/v1/N16-1014. URL https://www.aclweb.org/anthology/N16-1014.

Li, J.; Galley, M.; Brockett, C.; Spithourakis, G.; Gao, J.; and Dolan, B. 2016b. A Persona-Based Neural Conversation Model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 994–1003. Berlin, Germany: Association for Computational Linguistics. doi:10.18653/v1/P16-1094. URL https://www.aclweb.org/anthology/P16-1094.

Liu, D.; Fu, J.; Zhang, Y.; Pal, C.; and Lv, J. 2020. Revision in Continuous Space: Unsupervised Text Style Transfer without Adversarial Learning. *national conference on artificial intelligence* .

Luan, Y.; Brockett, C.; Dolan, B.; Gao, J.; and Galley, M. 2017. Multi-Task Learning for Speaker-Role Adaptation in Neural Conversation Models. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 605–614. Taipei, Taiwan: Asian Federation of Natural Language Processing. URL https://www.aclweb.org/anthology/I17-1061.

Niederhoffer, K. G.; and Pennebaker, J. W. 2002. Linguistic style matching in social interaction. *Journal of Language and Social Psychology* 21(4): 337–360.

Niu, T.; and Bansal, M. 2018. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics* 6: 373–389.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics. doi:10.3115/1073083.1073135. URL https://www.aclweb.org/anthology/P02-1040.

Prabhumoye, S.; Tsvetkov, Y.; Salakhutdinov, R.; and Black, A. W. 2018. Style Transfer Through Back-Translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 866–876. Melbourne, Australia: Association for Computational Linguistics. doi:10.18653/v1/P18-1080. URL https://www.aclweb.org/anthology/P18-1080.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving language understanding by generative pre-training. *OpenAI Blog* .

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* .

Randolph, J. J. 2005. Free-Marginal Multirater Kappa (multirater K [free]): An Alternative to Fleiss' Fixed-Marginal Multirater Kappa. *Online submission* .

Rao, S.; and Tetreault, J. 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)*, 129–140. New Orleans, Louisiana: Association for Computational Linguistics. doi:10.18653/v1/N18-1012. URL https://www.aclweb.org/anthology/N18-1012.

Sennrich, R.; Haddow, B.; and Birch, A. 2016. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 86–96. Berlin, Germany: Association for Computational Linguistics. doi:10.18653/v1/P16-1009. URL https://www.aclweb.org/anthology/P16-1009.

Shen, T.; Lei, T.; Barzilay, R.; and Jaakkola, T. 2017a. Style transfer from non-parallel text by cross-alignment. In *Advances in neural information processing systems*, 6830–6841.

Shen, X.; Su, H.; Li, Y.; Li, W.; Niu, S.; Zhao, Y.; Aizawa, A.; and Long, G. 2017b. A Conditional Variational Framework for Dialog Generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 504–509. Vancouver, Canada: Association for Computational Linguistics. doi:10.18653/v1/P17-2080. URL https://www.aclweb.org/anthology/P17-2080.

Su, F.-G.; Hsu, A. R.; Tuan, Y.-L.; and Lee, H.-Y. 2019. Personalized Dialogue Response Generation Learned from Monologues. In *INTERSPEECH*, 4160–4164.

Su, H.; Shen, X.; Zhao, S.; Xiao, Z.; Hu, P.; Zhong, R.; Niu, C.; and Zhou, J. 2020. Diversifying Dialogue Generation with Non-Conversational Text. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7087–7097. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.634. URL https://www.aclweb.org/anthology/2020.acl-main.634.

Tikhonov, A.; Shibaev, V.; Nagaev, A.; Nugmanova, A.; and Yamshchikov, I. P. 2019. Style Transfer for Texts: Retrain, Report Errors, Compare with Rewrites. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3927–3936.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30*, 5998–6008. Curran Associates, Inc. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Wang, D.; Jojic, N.; Brockett, C.; and Nyberg, E. 2017. Steering Output Style and Topic in Neural Response Generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2140–2150. Copenhagen, Denmark: Association for Computational Linguistics. doi:10.18653/v1/D17-1228. URL https://www.aclweb.org/anthology/D17-1228.

Wang, K.; Hua, H.; and Wan, X. 2019. Controllable Unsupervised Text Attribute Transfer via Editing Entangled La-

tent Representation. In *Advances in Neural Information Processing Systems*, 11034–11044.

Wang, Y.; Ke, P.; Zheng, Y.; Huang, K.; Jiang, Y.; Zhu, X.; and Huang, M. 2020. A Large-Scale Chinese Short-Text Conversation Dataset. In *NLPCC*. URL https://arxiv.org/abs/2008.03946.

Wu, C.; Ren, X.; Luo, F.; and Sun, X. 2019a. A Hierarchical Reinforced Sequence Operation Method for Unsupervised Text Style Transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4873–4883.

Wu, X.; Zhang, T.; Zang, L.; Han, J.; and Hu, S. 2019b. Mask and Infill: Applying Masked Language Model for Sentiment Transfer. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 5271–5277. International Joint Conferences on Artificial Intelligence Organization. doi:10.24963/ijcai.2019/732. URL https://doi.org/10.24963/ijcai.2019/732.

Wu, Y.; Wang, Y.; and Liu, S. 2020. A Dataset for Low-Resource Stylized Sequence-to-Sequence Generation. *AAAI 2020* .

Zhang, Y.; Ge, T.; and Sun, X. 2020. Parallel Data Augmentation for Formality Style Transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.-C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, B. 2019. DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation.

Zhang, Z.; Ren, S.; Liu, S.; Wang, J.; Chen, P.; Li, M.; Zhou, M.; and Chen, E. 2018. Style transfer as unsupervised machine translation. *arXiv preprint arXiv:1808.07894* .

Zheng, Y.; Zhang, R.; Mao, X.; and Huang, M. 2020. A Pre-training Based Personalized Dialogue Generation Model with Persona-sparse Data. *AAAI* .

Zhou, H.; Huang, M.; Zhang, T.; Zhu, X.; and Liu, B. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *AAAI*.