Stylized Dialogue Response Generation Using Stylized Unpaired Texts

Yinhe Zheng¹³, Zikai Chen¹, Rongsheng Zhang², Shilei Huang², Xiaoxi Mao², Minlie Hang¹∗

¹ Department of Computer Science and Technology, Institute for Artificial Intelligence, State Key Lab of Intelligent Technology and Systems, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, China.
² Fuxi AI Lab, NetEase Inc., Hangzhou, China
³ Samsung Research China - Beijing (SRC-B), Beijing, 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, and 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, and 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.

1 Introduction

Building a conversational agent that can produce stylized and coherent responses has been one of the major challenges in dialogue systems (Huang et al., 2020). Such an agent can not only yield more vivacious dialogues but also 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.

* Corresponding Author: aihuang@tsinghua.edu.cn

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 ˜y using a text style transfer model.

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 that requires to use 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).

Few studies for dialogue modeling have been proposed to capture the stylistic features embedded in unpaired texts. Specifically, Niu and Bansal (2018) 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
Automatic and human evaluations show that our method outperforms competitive baselines with a large margin in producing stylized and coherent dialogue responses.

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.

2 Related Work

Stylized dialogue generation has attracted numerous attentions 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) or emotional (Zhou et al., 2018) dialogues. However, the training process of these dialogue model usually require 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 expressive ability of the dialogue model 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 contents when enforcing a negative sentiment. Unlike most previous works, our study investigates to model the writing styles that are “orthogonal” to the text semantic, so that 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 to 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 et al., 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-labelled 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 et al., 2016), dialogue data augmentation (Su et al., 2020), and text style transfer (Zhang et al., 2018; Lample et al., 2019; Dai et al., 2019). This approach is similar to the inverse dialogue model introduced in our study. However, different from previous approaches that focus on modeling the one-to-one mapping between the source and target languages, our inverse dialogue model tries to capture the one-to-many mappings between the responses and posts with the help of the proposed joint training process. In our study, the diversity of the generated pseudo posts are enhanced using a sampling approach.

### 3 Method

#### 3.1 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 $D_s = \{t_1, ..., t_M\}$ in the writing style $S_1$; 2) $N$ dialogue pairs $D_p = \{\langle x_1, y_1 \rangle, ..., \langle x_N, y_N \rangle\}$ 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 $S_i$ ($i = 0, 1$):

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

#### 3.2 Model Overview

![Figure 2: Overall framework.](image)

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 $D_p$ 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 sub-module 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 $D_s$, and all these posts are utilized in the training process. Further, a joint training process is also introduced to train these two sub-modules in an iterative fashion to enhance the coherency of the response.
3.3 Style Routing

There exist various approaches to condition the decoder to 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 input layer of the decoder, whereas the higher layers are not affected.

In this study, a style routing approach is devised to enhance existing approaches to stylize the stylized dialogue generation task. Specifically, in each decoder block, we first fuse the representation of the post (D) and previously decoded token sequence (yp) using the attention routing mechanism (Zheng et al., 2020), i.e., two sequences of representations, Rprev, Rpost ∈ Rl×h, are first calculated:

\[ R_{\text{prev}} = \text{MMHA}[e_w(y_p), e_w(y_p), e_w(y_p)], \]
\[ R_{\text{post}} = \text{MHA}[e_w(y_p), e(x), e(x)], \]

where \( e_w(y_p) \in R^{l \times h} \) 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. \( e_w(x) \) is the output of the encoder. A sequence of fused representations \( R_{\text{avg}} \) is obtained as:

\[ R_{\text{avg}} = (R_{\text{prev}} + R_{\text{post}})/2. \]

Then for a given style \( S_i \), a style embedding \( e_{\text{avg}}(S_i) \in R^{l \times h} \) is allocated and \( e_{\text{avg}}(S_i) \) is routed into \( R_{\text{avg}} \) by adding it to each time step of the sequence:

\[ R_{\text{merge}} = R_{\text{avg}} + e_{\text{avg}}(S_i). \]

Algorithm 1 Joint training process

**Input:** \( M \) unpaired texts: \( D_p=\{(x_i,y_i)\}_{i=1}^M \) in style \( S_p \), \( N \) dialogue pairs \( D_p=\{(d_i, y_i)\}_{i=1}^N \) in style \( S_o \).

**Output:** A stylized dialogue model

1: Init the stylized and inverse dialogue model \( e, d, \tilde{e}, \tilde{d} \)
2: while not converge do
3: Sample \( n_d \) dialogue pairs \( D^p_n = \{(x_i,y_i)\}_{i=1}^{n_d} \subset D_p \)
4: Train \( e \) and \( d \) by optimizing \( L_{2p} \) (Eq. 6) on \( D^p_n \)
5: Train \( \tilde{e} \) and \( \tilde{d} \) by optimizing \( L_{2p} \) (Eq. 7) on \( D^p_n \)
6: if Current Step > \( N_f \) then
7: \( D_{\text{inv}} \leftarrow \emptyset \)
8: Sample \( n_s \) stylized texts \( D^s_n = \{t_i\}_{i=1}^{n_s} \subset D_s \)
9: for each \( t_i \in D^s_n \) do
10: Decode \( \tilde{m} \) \( \{x_i^{\prime}\}_{i=1}^{m} \) from \( p_{\tilde{d}}(x|\tilde{e}(t_i)) \)
11: \( D_{\text{pp}} \leftarrow D_{\text{pp}} \cup \{\{x_i^{\prime}, t_i\}_{i=1}^{m} \}
12: end for
13: Train \( e \) and \( d \) by optimizing \( L_{inv} \) (Eq. 8) on \( D_{\text{pp}} \)
14: end if
15: end while

The loss \( L_{2p} \) and \( L_{2p} \) is used to train the stylized dialogue model and inverse dialogue model, respectively; 2) an inverse dialogue loss evaluated on texts from \( D_s \):

\[ L_{\text{inv}} = \mathbb{E}_{\tilde{x} \sim D_s, x^\prime \sim P_{\tilde{d}}(x^\prime|x)} - \log p_{\tilde{d}}(t|x^\prime), S_1, \]

in which \( x^\prime \) is the pseudo post sampled from the inverse dialogue model.

Note that the gradient back-propagation through the loss \( L_{\text{inv}} \) is intractable due to the in-differentiable sampling process in Eq. 8. In this study, we approximate the ideal back-propagation process through \( L_{\text{inv}} \) by truncating the gradients associated with the sampling operation. Specifically, when optimizing \( L_{\text{inv}} \), the parameters of the inverse dialogue model are fixed, and the stylized dialogue model is trained with pseudo posts \( x^\prime \) 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 \( x^\prime \) 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 literately. Specifically, in each training iteration, we first update the stylized and inverse dialogue model by optimizing the losses $L_{p2r}$ and $L_{r2p}$ using a batch of dialogue pairs sampled in $D_p$. Further, a batch of stylized sentences $D^b_s$ are sampled from $D_s$. For each sentence $t_i \in D^b_s$, $m$ pseudo posts $x_{i1}^{\prime},...,x_{im}^{\prime}$ are sampled from the inverse dialogue model, and $m$ pseudo dialogue pairs $\langle x_{ij}^{\prime},t_i \rangle$, $(j = 1,...,m)$ in the style $S_1$ are constructed. These pseudo pairs are used to train the stylized dialogue model with the loss $L_{inv}$. Moreover, to avoid corrupted pseudo posts at the beginning of the training process, we pre-train the inverse dialogue model on $L_{r2p}$ for $N_f$ steps before using it to decode pseudo posts. The detailed training process is summarized in Algorithm 1.

### 4 Experiment

#### 4.1 Dataset

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

1) **WDJN**: We collect 300K Weibo Dialogues (style $S_0$) as $D_s$ and sampled 95.1K stylized texts from Jinyong’s Novels (style $S_1$) as $D_s$. Moreover, we also extracted 2K dialogue pairs from Jinyong’s novels with hand-designed rules. These dialogues are used as the test set $D_t$ together with 2K additional Weibo dialogues. Note that all the Weibo dialogues in our WDJN dataset (both training and testing) are manually inspected and filtered by annotators.

Also note that to prevent the model from copying stylistic phrases in the post when producing Jinyong style responses in the testing phase, we erase the stylistic features related to Jinyong’s writing from the posts in these 2K Jinyong style dialogues in $D_s$ using the back translation approach (Zhang et al., 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 find the Appendix A. The WDJN dataset will be released for public use.

2) **TCFC** (Wu et al., 2020): This dataset focuses on the formality in English writing. We sampled 217.2K informal dialogue pairs (style $S_0$) as $D_p$ and 500.0K formal texts (style $S_1$) as $D_s$ from the original dataset, and used the test data in the original dataset as our test set $D_t$, which contains 1,956 manually-crafted dialogue pairs (978 informal pairs and 978 formal pairs).

#### 4.2 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 $D_p$ (WDJN) or after 8,000 steps of updates (TCFC). See Appendix B for more details of the reproduction guidance.

#### 4.3 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 $D_p$. 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 $D_s$, 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) **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 $D_s$ and $D_p$, and responses produced by the S2S baseline (in style $S_0$) are transferred to exhibit the target
Table 2: Automatic and manual evaluation results for responses with 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. |
| 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 40.4 89.6 85.8 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 3: Automatic and manual evaluation results for responses with style $S_0$. All differences between our model and baselines are significant with $p$-value < 0.01 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 92.3 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 |

4.4 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 the 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, the texts from $D_p$ and $D_s$ were used to train the classifiers for the WDJN experiments, and the formal/informal texts from the GY AFC dataset (Rao and Tetreault, 2018) were used to train the 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 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 to 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. 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 that 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 that the train data of these classifiers 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.

### 4.5 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 $D_1$ for each of these two datasets. Fleiss’s kappa $\kappa$ (Randolph, 2005) was used to measure the annotation agreement between annotators. Specifically, for Flu., Coh., and Style, the $\kappa$ 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 shows that our model achieved competitive performance when generating responses in style $S_0$.

Another interesting observation from the results in Table 2 and 3 is the trade-off between the coherency and style intensity when generating stylized dialogue responses, i.e., the high style intensity usually comes at the cost of a low coherency. For example, the model S2S+CT achieves the best style intensity score on the WDJN dataset (1.50) when producing responses in style $S_1$, but it obtains the
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 explicitly incorporated in each decoder block as in Eq.5. The decoder d is stylized by employing a style token as the 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 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 $L_{inv}$ 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.

| Post | WDJN dataset | TCF dataset |
|------|--------------|-------------|
| S2S+PTO Ours | I haven’t eaten hot pot in a long time too (我也好久没吃火锅了) | You’re not falling asleep yet. lol dude same here, my friend has a reason at night it’s almost 9 am here and i just got up... |
| S2S+BT Ours | I haven’t eaten hot pot in a long time too (我也好久没吃火锅了) | I have a headache and I can not stop drinking. isn’t it 5:30 in the morning? |
| S2S+CT Ours | 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, 老板娘去打火锅可惜他们没) | Same here but I think it’s gna say hello! She is not falling asleep yet. |
| S2S SRL | I’m almost done (我已经快好了) | That is not falling asleep then Maguties out for riddle. |
| S2S SLM | With that said, I want to eat too (这么一说, 你也想吃了吧) | I’m not falling asleep yet |
| S2S SFusion | I haven’t eaten in a long time. I really want to eat (好久没吃了) | I have a headache and I can not stop drinking. |

Table 5: Example responses produced by our model and the baselines on the TCFC and WDJN datasets.
because that it hard to build a smooth latent space for discrete texts. Moreover, 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 (Table 6). It can be seen that these pseudo pairs are generally of high quality both in fluency and coherency.

5 Conclusion

In this paper, we present 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 formality in English writing. 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.

References

Yu Cheng, Zhe Gan, Yizhe Zhang, Oussama Elachgar, Dianqi Li, and Jingjing Liu. 2020. Contextual text style transfer. ArXiv, abs/2005.00136.

Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. 2019. Style transformer: Unpaired text style transfer without disentangled latent representation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5997–6007, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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), pages 4171–4186.

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

Xiang Gao, Yizhe Zhang, Sungjin Lee, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 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), pages 1814–1823.

Sergey Golovanov, Rauf Kurbanov, Sergey Nikolenko, Kyryl Truskovskyi, Alexander Tselousov, and Thomas Wolf. 2019. Large-scale transfer learning for natural language generation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 6053–6058.

Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A probabilistic formulation of unsupervised text style transfer. In International Conference on Learning Representations.

Zhitong Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 1587–1596. JMLR. org.

Ming逛街 Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. ACM Transactions on Information Systems.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.
Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI Blog.

Justus J Randolph. 2005. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss’ fixed-marginal multirater kappa. Online submission.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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), pages 86–96, Berlin, Germany. Association for Computational Linguistics.

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

Xiaoyu Shen, Hui Su, Yanran Li, Wenjie Li, Shuzi Niu, Yang Zhao, Akiko Aizawa, and Guoping Long. 2017b. A conditional variational framework for dialogue generation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 504–509, Vancouver, Canada. Association for Computational Linguistics.

Hui Su, Xiaoyu Shen, Sanqiang Zhao, Zhou Xiao, Pengwei Hu, Randy Zhong, Cheng Niu, and Jie Zhou. 2020. Diversifying dialogue generation with non-conversational text. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7087–7097, Online. Association for Computational Linguistics.

Alexey Tikhonov, Viacheslav Shibaev, Aleksander Nagaev, Aigul Nugmanova, and Ivan P Yamshchikov. 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), pages 3927–3936.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems 30, pages 5998–6008. Curran Associates, Inc.
Ke Wang, Hang Hua, and Xiaojun Wan. 2019. Controllable unsupervised text attribute transfer via editing entangled latent representation. In Advances in Neural Information Processing Systems, pages 11034–11044.

Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyuan Zhu, and Minlie Huang. 2020. A large-scale chinese short-text conversation dataset. In NLPCC.

Chen Wu, Xuancheng Ren, Fuli Luo, and Xu Sun. 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, pages 4873–4883.

Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han, and Songlin Hu. 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, pages 5271–5277. International Joint Conferences on Artificial Intelligence Organization.

Yu Wu, Yunli Wang, and Shujie Liu. 2020. A dataset for low-resource stylized sequence-to-sequence generation. AAAI 2020.

Yi Zhang, Tao Ge, and Xu Sun. 2020. Parallel data augmentation for formality style transfer. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation.

Zhirui Zhang, Shuo Ren, Shujie Liu, Jianyong Wang, Peng Chen, Mu Li, Ming Zhou, and Enhong Chen. 2018. Style transfer as unsupervised machine translation. arXiv preprint arXiv:1805.08794.

Yinhe Zheng, Rongsheng Zhang, Xiaoxi Mao, and Minlie Huang. 2020. A pre-training based personalized dialogue generation model with persona-sparse data. AAAI.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyuan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In AAAI.