Neural Stylistic Response Generation with Disentangled Latent Variables

Qingfu Zhu[♯], **Weinan Zhang**[♯]*, **Ting Liu**[♯], **William Yang Wang**[♭]
[♯]Harbin Institute of Technology, Harbin, China
[♭]University of California, Santa Barbara, USA
{qfzhu, wnzhang, tliu}@ir.hit.edu.cn
william@cs.ucsb.edu

Abstract

Generating open-domain conversational responses in the desired style usually suffers from the lack of parallel data in the style. Meanwhile, using monolingual stylistic data to increase style intensity often leads to the expense of decreasing content relevance. In this paper, we propose to disentangle the content and style in latent space by diluting sentence-level information in style representations. Combining the desired style representation and a response content representation will then obtain a stylistic response. Our approach achieves a higher BERT-based style intensity score and comparable BLEU scores, compared with baselines. Human evaluation results show that our approach significantly improves style intensity and maintains content relevance.

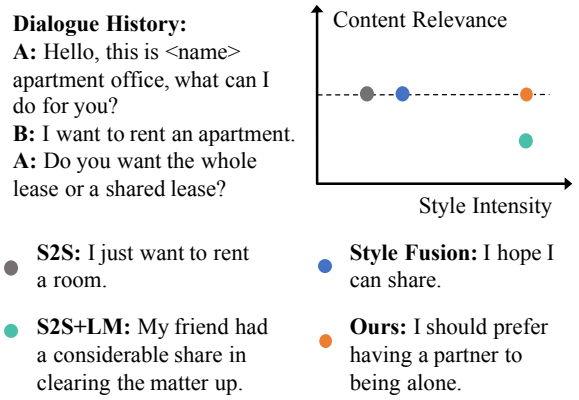


Figure 1: An example of responses generated by S2S, S2S+LM (Niu and Bansal, 2018), Style Fusion (Gao et al., 2019b), and our approach, targeting the Holmes style, which is quite formal and polite.

1 Introduction

Linguistic style is an essential aspect of natural language interaction and provides particular ways of using language to engage with the audiences (Kabbara and Cheung, 2016). In human-bot conversations, it is crucial to generate stylistic responses for increasing user engagement to conversational systems (Gan et al., 2017). Currently, most of the existing parallel datasets are not stylistically consistent. Samples in these datasets are usually contributed by a variety of users, resulting in an averaging effect across style characteristics (Zhang et al., 2018a). Meanwhile, constructing a parallel stylistic dataset for training the open-domain conversational agents is both labor-intensive and time-consuming.

Recent studies show the effect of stylizing responses using a monolingual dataset in the desired style and a conventional conversational dataset (Niu and Bansal, 2018; Gao et al., 2019b). However, increasing style intensity often leads to the expense of decreasing content relevance between dialogue history and response. As an example in Figure 1 shows, Niu and Bansal (2018) independently train a response generation model and a stylistic language model and subsequently interpolates them in the inference phase. Lacking the interaction between the stylistic language model and response generation encoder, it usually yields a trade-off between style intensity and content relevance. Gao et al. (2019a,b) fuse a structured latent space where the direction denotes the diversity, and the distance denotes style intensity and content relevance. The main issue is that style intensity and content relevance are contradictory in measurement but are coupling to the same "distance" metric of the latent space. To sum up, the key issue of the above studies is the improper entanglement of style and content.

To address the issue, we propose to disentangle the style and content of a response. The disentanglement is conducted on the structured latent space, where each sentence (dialogue history, response,

*Corresponding author.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 4391–4401
August 1–6, 2021.

and stylistic sentence) is projected into a vector representation. We further split the representation into two components: style and content representations. The former is a corpus-level feature since sentences within a dataset have the same style. In contrast, the content representation is a sentence-level feature decided by a sentence itself. We thus disentangle the content and style by diluting sentence-level information in the style representation. This encourages the encoding of content information into the content representation. Otherwise, the content information will be corrupted in the style representation, making it hard to reconstruct the original content in the subsequent decoding process. We conduct experiments on DailyDialogue conversational dataset (Li et al., 2017) and Holmes monolingual stylistic dataset (Gao et al., 2019b). Experimental results show that our proposed approach improves style intensity and maintains content relevance. Our contributions are listed below:

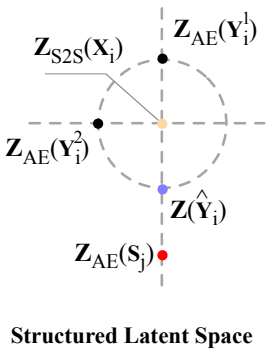
- We propose a unified framework to simultaneously improve style intensity and maintain content relevance for neural stylistic response generation.
- We introduce a scheme of learning latent variables by a diluting strategy to disentangle the style and content.
- Experimental results show that our approach achieves higher performance in style intensity without decreasing content relevance, compared with previous approaches.

2 Method

2.1 Task Definition

The task of stylistic response generation is defined as follows: given a monolingual stylistic dataset $\boldsymbol{S} = \{\boldsymbol{S}_1, ..., \boldsymbol{S}_N\}$[1] and a conversational dataset $\boldsymbol{C} = \{(\boldsymbol{X}_1, \boldsymbol{Y}_1), ..., (\boldsymbol{X}_M, \boldsymbol{Y}_M)\}$, where $\boldsymbol{S}_i$, $\boldsymbol{X}_i$, and $\boldsymbol{Y}_i$ denote a stylistic sentence, dialogue history, and a response respectively, the goal is to learn a generation model $P(\hat{\boldsymbol{Y}}|\boldsymbol{X})$, where $\hat{\boldsymbol{Y}}$ is a generated response expected to be in the style of $\boldsymbol{S}$ (called the desired style in the following sections). We will first briefly review the concept of structured latent space and then introduce our disentanglement approach.

[1]Throughout the paper, we use bold letters to denote vectors, i.e., $\boldsymbol{V} = \{V_1, V_2, ..., V_N\}$.



$\mathbf{X}_\text{i}$: We are going hiking this weekend. Do you want to join us?

$\mathbf{Y}_\text{i}^1$: Yes, of course.

$\mathbf{Y}_\text{i}^2$: I don't like hiking.

$\hat{\mathbf{Y}}_\text{i}$: I would like to join.

$\mathbf{S}_\text{j}$: Thanks for your help. I would like to book.

Figure 2: An example of a dialogue in the structured latent space. The center point corresponds to the dialogue history representation $\boldsymbol{Z}_{\text{S2S}}(\boldsymbol{X}_i)$. The k-th response representation $\boldsymbol{Z}_{\text{AE}}(\boldsymbol{Y}_i^k)$ (denoted by a black point) is optimized to be distributed around $\boldsymbol{Z}_{\text{S2S}}(\boldsymbol{X}_i)$. The red point $\boldsymbol{Z}_{\text{AE}}(\boldsymbol{S}_j)$ and the purple point $\boldsymbol{Z}(\hat{\boldsymbol{Y}}_i)$ are representations of a monolingual stylistic sentence and a stylistic response, respectively.

2.2 Background: Structured Latent Space

Overview The structured latent space is constructed by two main mechanisms: (i) sharing a decoder between a sequence-to-sequence (S2S) model and an auto-encoder (AE), and (ii) fusion and smoothness objectives. As an example in Figure 2 shows, a response representation $\boldsymbol{Z}_{\text{AE}}(\boldsymbol{Y}_i)$ is regularized by the two mechanisms to be distributed around its dialogue history representation $\boldsymbol{Z}_{\text{S2S}}(\boldsymbol{X}_i)$. The notations $\boldsymbol{Z}_{\text{AE}}(\cdot)$ and $\boldsymbol{Z}_{\text{S2S}}(\cdot)$ denote the representations computed by AE encoder and S2S encoder, respectively. Such a latent space makes it possible to predict a response $\hat{\boldsymbol{Y}}$ by sampling nearby the dialogue history representation. Based on that, Gao et al. (2019b) further align stylistic sentence representations into the latent space, which improves the style intensity of generated responses. In summary, the construction of the structred latent space is a process of aligning the three spaces ($\boldsymbol{Z}_{\text{S2S}}(\boldsymbol{X}_i)$, $\boldsymbol{Z}_{\text{AE}}(\boldsymbol{Y}_i)$, and $\boldsymbol{Z}_{\text{AE}}(\boldsymbol{S}_j)$) by two mechanisms (sharing the decoder, and fusion and smoothness objectives).

Fusion Objective cross-aligns sentences of different spaces. Since $\boldsymbol{X}_i$ and $\boldsymbol{Y}_i$ are paired, we align them by minimizing their pair-wise dissimilarity:

$$d_{\text{conv}} = \sum_{i \in batch} \frac{d_E(\boldsymbol{Z}_{\text{S2S}}(\boldsymbol{X}_i), \boldsymbol{Z}_{\text{AE}}(\boldsymbol{Y}_i))}{n\sqrt{l}}, \quad (1)$$

where d_E denotes the Euclidean distance, n is the batch size, and l is the dimensionality of the latent space. In contrast, the pair-wise dissimilarity can-

not be applied to stylistic sentences since they are not paired with conversational data. To this end, the fusion objective instead optimizes the nearest neighbor distance between the two datasets:

$$d_{\text{style}} = \frac{1}{2} d_{\text{NN}}^{\text{cross}}(\{\boldsymbol{Z}_{\text{S2S}}(\boldsymbol{X}_i)\}, \{\boldsymbol{Z}_{\text{AE}}(\boldsymbol{S}_j)\}) + \frac{1}{2} d_{\text{NN}}^{\text{cross}}(\{\boldsymbol{Z}_{\text{AE}}(\boldsymbol{S}_j)\}, \{\boldsymbol{Z}_{\text{S2S}}(\boldsymbol{X}_i)\}), \quad (2)$$

where $d_{\text{NN}}^{\text{cross}}(\{a_i\}, \{b_j\})$ denotes the batch average distance between a_i and its nearest neighbor in the set $\{b_j\}$. To further encourage the representations spread-out the latent space, a inner-distance loss is introduced:

$$d_{\text{spread-out}} = \min\{d_{\text{NN}}^{\text{inner}}(\boldsymbol{Z}_{\text{S2S}}(\boldsymbol{X}_i)), d_{\text{NN}}^{\text{inner}}(\boldsymbol{Z}_{\text{AE}}(\boldsymbol{Y}_i)), d_{\text{NN}}^{\text{inner}}(\boldsymbol{Z}_{\text{AE}}(\boldsymbol{S}_j))\}, \quad (3)$$

where $d_{\text{NN}}^{\text{inner}}(\{a_i\})$ denotes the batch average distance between a_i and its nearest neighbor in the set $\{a_i\}$. The final fusion objective is defined as:

$$L_{\text{fuse}} = d_{\text{conv}} + d_{\text{style}} - d_{\text{spread-out}}. \quad (4)$$

Smoothness Objective aims to make the structured latent space a continuous space, where each point can decode a natural sentence. Given three discrete points $\boldsymbol{Z}_{\text{S2S}}(\boldsymbol{X}_i)$, $\boldsymbol{Z}_{\text{AE}}(\boldsymbol{Y}_i)$, and $\boldsymbol{Z}_{\text{AE}}(\boldsymbol{S}_j)$, the objective encourages points in the area between $\boldsymbol{Z}_{\text{S2S}}(\boldsymbol{X}_i)$ and $\boldsymbol{Z}_{\text{AE}}(\boldsymbol{Y}_i)$ to generate $\boldsymbol{Y}_i$:

$$\boldsymbol{Z}_{\text{conv}} = U\boldsymbol{Z}_{\text{S2S}}(\boldsymbol{X}_i) + (1-U)\boldsymbol{Z}_{\text{AE}}(\boldsymbol{Y}_i) + \epsilon,$$
$$L_{\text{smooth,conv}} = -\log P(\boldsymbol{Y}_i|\boldsymbol{Z}_{\text{conv}}), \quad (5)$$

where $\epsilon \sim N(0, \sigma^2 I)$, and $U \sim U(0,1)$. Meanwhile, as a point moves from $\boldsymbol{Z}_{\text{AE}}(\boldsymbol{Y}_i)$ to $\boldsymbol{Z}_{\text{AE}}(\boldsymbol{S}_j)$, the corresponding generation is expected to gradually move from $\boldsymbol{Y}_i$ to $\boldsymbol{S}_j$:

$$\boldsymbol{Z}_{\text{style}} = U\boldsymbol{Z}_{\text{AE}}(\boldsymbol{Y}_i) + (1-U)\boldsymbol{Z}_{\text{AE}}(\boldsymbol{S}_j) + \epsilon$$
$$L_{\text{smooth,style}} = -U\log P(\boldsymbol{Y}_i|\boldsymbol{Z}_{\text{style}}) - (1-U)\log P(\boldsymbol{S}_j|\boldsymbol{Z}_{\text{style}}). \quad (6)$$

The smoothness objective L_{smooth} is the sum of $L_{\text{smooth,conv}}$ and $L_{\text{smooth,style}}$, and is added to the final loss function along with the fusion objective and response generation loss of S2S.

2.3 Our Method

Despite aligning monolingual stylistic sentences into the structured latent space helps stylize generated responses, their style intensity is still limited. We conjecture this is due to the coupling of the style and the content in sentence representations. To this end, we propose to disentangle the two aspects in the structured latent space.

In our proposed approach, a sentence representation $\boldsymbol{Z} \in \mathbb{R}^l$ in the latent space consists of two components: content representation $\boldsymbol{Z}^c \in \mathbb{R}^{l_c}$ and style representation $\boldsymbol{Z}^s \in \mathbb{R}^{l_s}$, where l is the dimensionality of latent space and $l_c + l_s = l$. $\boldsymbol{Z}^s$ encodes all the style information of a sentence. It is a corpus-level feature because $\boldsymbol{Z}^s$ for different sentences in the same corpus should be similar. In contrast, $\boldsymbol{Z}^c$ can be seen as a sentence-level feature which only decided by the content of its corresponding sentence.

Figure 3 shows an example of our approach, where $\boldsymbol{Z}^c$ and $\boldsymbol{Z}^s$ can be seen as two "containers". Colored squares represent the content and style information. We encourage the disentanglement of the two types of information by diluting sentence-level content information in $\boldsymbol{Z}^s$. As an example in Figure 3 (a) shows, the content and style information may be mixed in both $\boldsymbol{Z}^c$ and $\boldsymbol{Z}^s$. During the decoding process of a sentence, i.e., $\boldsymbol{Y}_i$, we replace its style representation $\boldsymbol{Z}^s_{\text{AE}}(\boldsymbol{Y}_i)$ with its batch average style representation $\bar{\boldsymbol{Z}}^s_{\text{AE}}(\boldsymbol{Y}_i) = \frac{1}{n}\sum_{j \in batch} \boldsymbol{Z}^s_{\text{AE}}(\boldsymbol{Y}_j)$. In this way, its sentence-level content information will be diluted since it greatly varies from other sentences' content information, which introduces extra noise. In contrast, its corpus-level style information, which is similar to that of other sentences within the batch, will remain unaffected. As the training processes, the content information will be encouraged to be encoded into $\boldsymbol{Z}^c$ where it can remain unchanged, as an example in Figure 3 (b) shows. Otherwise, the content information will be corrupted in $\boldsymbol{Z}^s$, making it hard to recover the content of $\boldsymbol{Y}_i$. As a result, the encoding process will be punished by the response generation loss of S2S and the reconstruction loss of AE, as shown in Figure 3 (a).

Based on that, we update the response generation process by replacing its style representation $\boldsymbol{Z}^s$ with the corresponding batch average style representation $\bar{\boldsymbol{Z}}^s$:

$$L_{\text{S2S}} = -\log P(\boldsymbol{Y}_i|[\boldsymbol{Z}^c_{\text{S2S}}(\boldsymbol{X}_i) : \bar{\boldsymbol{Z}}^s_{\text{S2S}}(\boldsymbol{X}_i)]), \quad (7)$$

where the bracket [:] denotes concatenation. The decoding process in the smoothness objective is updated similarly. Note that when we move from

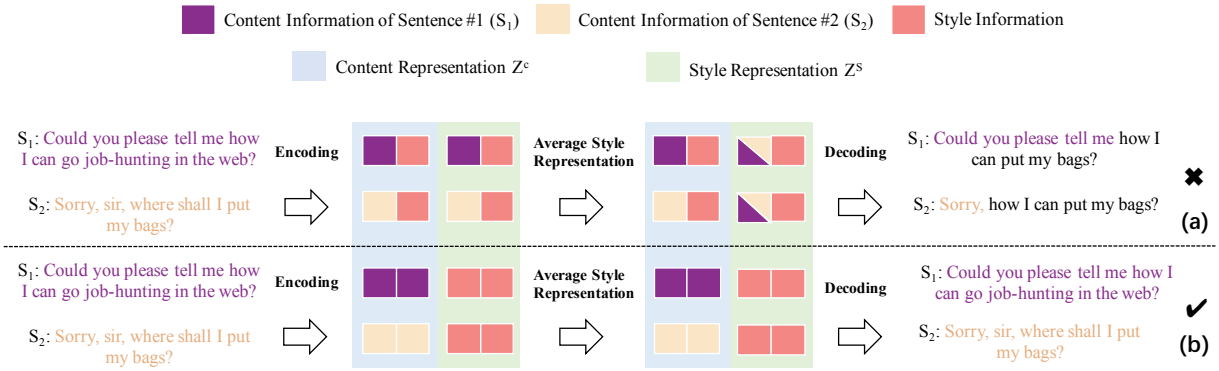


Figure 3: An example of disentangling content and style. The purple block is the content information of the first sentence. The yellow block is the content information of the second sentence. Style information in both two sentences is denoted by red blocks as it is a corpus-level feature shared among samples within the corpus. (a): A negative example whose content and style information is mixed in $\boldsymbol{Z}^c$ and $\boldsymbol{Z}^s$. Its content information is corrupted after averaging $\boldsymbol{Z}^s$ within the batch and fails to recover the input content. (b): A positive example. Content information in $\boldsymbol{Z}^c$ and style information in $\boldsymbol{Z}^s$ will not be affected after averaging $\boldsymbol{Z}^s$.

$\boldsymbol{Y}_i$ to $\boldsymbol{S}_j$, and from $\boldsymbol{X}_i$ to $\boldsymbol{Y}_i$, we only interpolate their content representations $\boldsymbol{Z}^c$ in the latent space:

$$\begin{aligned}\boldsymbol{Z}^c_{\text{conv}} &= U\boldsymbol{Z}^c_{\text{S2S}}(\boldsymbol{X}_i) + (1-U)\boldsymbol{Z}^c_{\text{AE}}(\boldsymbol{Y}_i) + \epsilon,\\ \boldsymbol{Z}^c_{\text{style}} &= U\boldsymbol{Z}^c_{\text{AE}}(\boldsymbol{Y}_i) + (1-U)\boldsymbol{Z}^c_{\text{AE}}(\boldsymbol{S}_j) + \epsilon.\end{aligned} \tag{8}$$

The batch average style representation $\bar{\boldsymbol{Z}}^s$ remains consistent with the target, i.e., being $\bar{\boldsymbol{Z}}^s_{\text{AE}}(\boldsymbol{S}_j)$ when the target is $\boldsymbol{S}_j$. The updated smoothness objective is as follows:

$$\begin{aligned}L_{\text{smooth,conv}} &= -\log P(\boldsymbol{Y}_i|[\boldsymbol{Z}^c_{\text{conv}} : \bar{\boldsymbol{Z}}^s_{\text{AE}}(\boldsymbol{Y}_i)]),\\ L_{\text{smooth,style}} &= -U\log P(\boldsymbol{Y}_i|[\boldsymbol{Z}^c_{\text{style}} : \bar{\boldsymbol{Z}}^s_{\text{AE}}(\boldsymbol{Y}_i)])\\ &\quad -(1-U)\log P(\boldsymbol{S}_j|[\boldsymbol{Z}^c_{\text{style}} : \bar{\boldsymbol{Z}}^s_{\text{AE}}(\boldsymbol{S}_j)]).\end{aligned} \tag{9}$$

The final training loss is the sum of the response generation loss, fusion objective, and smoothness objective:

$$L = L_{\text{S2S}} + L_{\text{fuse}} + L_{\text{smooth}}. \tag{10}$$

Here, we do not employ pre-training models, i.e., DialoGPT (Zhang et al., 2020b) and OpenAI GPT2 (Radford et al., 2019). This is because the disentanglement is usually conducted on a sentence representation. While most of the pre-training models depend on the attention mechanism, and there is no static global sentence representation during the decoding process.

2.4 Inference

To generate a stylistic response $\hat{\boldsymbol{Y}}_i$ given dialogue history $\boldsymbol{X}_i$ during the inference process, we first obtain $\boldsymbol{Z}^c_{\text{S2S}}(\boldsymbol{X}_i)$ by S2S encoder and subsequently sample $\boldsymbol{Z}^c(\hat{\boldsymbol{Y}}_i)$ from the hypersphere of $\boldsymbol{Z}^c_{\text{S2S}}(\boldsymbol{X}_i)$ with a mannually tuned radius r. After that, we generate $\hat{\boldsymbol{Y}}_i$ by concatenating $\boldsymbol{Z}^c(\hat{\boldsymbol{Y}}_i)$ and $\bar{\boldsymbol{Z}}^s_{\text{AE}}(\boldsymbol{S}_j)$, which is the batch average style representation of randomly sampled stylistic sentences.

Considering the discrepancy between training and inference that content and style representations in different corpora have never been concatenated for generation, we propose a soft combination approach to introduce the desired style by interpolating $\boldsymbol{Z}^s_{\text{S2S}}(\boldsymbol{X}_i)$ and $\bar{\boldsymbol{Z}}^s_{\text{AE}}(\boldsymbol{S}_j)$:

$$\boldsymbol{Z}^s_{\text{soft}} = \boldsymbol{Z}^s_{\text{S2S}}(\boldsymbol{X}_i) + \alpha * \bar{\boldsymbol{Z}}^s_{\text{AE}}(\boldsymbol{S}_j), \tag{11}$$

where α is the weight of the desired style. After that, $\hat{\boldsymbol{Y}}_i$ is generated by the decoder whose hidden state is set to $[\boldsymbol{Z}^c(\hat{\boldsymbol{Y}}_i) : \boldsymbol{Z}^s_{\text{soft}}]$.

To further balance style intensity and content relevance, we also employ the re-ranking strategy following Gao et al. (2019b). It samples N_y candidate responses and re-ranks them by:

$$s_r = \gamma * P_{\text{S2S}}(\hat{\boldsymbol{Y}}_i|\boldsymbol{X}_i) + (1-\gamma) * P_{\text{style}}(\hat{\boldsymbol{Y}}_i), \tag{12}$$

where $P_{\text{S2S}}(\hat{\boldsymbol{Y}}_i|\boldsymbol{X}_i)$ is the generation probability under a S2S model measuring the relevance. $P_{\text{style}}(\hat{\boldsymbol{Y}}_i)$ is the probability that $\hat{\boldsymbol{Y}}_i$ has the desired style. It is a interpolation between the probabilities of a neural-based classifier and a n-gram classifier:

$$\begin{aligned}P_{\text{style}}(\hat{\boldsymbol{Y}}_i) =& \eta * P_{\text{neural}}(\hat{\boldsymbol{Y}}_i) + (1-\eta)\\ &* \sum_{n=1}^{N} w_n * P_{\text{n-gram}}(\hat{\boldsymbol{Y}}_i),\end{aligned} \tag{13}$$

Training Dialogues	11,118
Validation Dialogues	1,000
Test Dialogues	1,000
Average Tokens Per Dialogue	114.7
Average Tokens Per Utterance	14.6

Table 1: Statistics of the DailyDialog dataset.

Model	Time (s)	# of parameters
S2S	4.55	63M
Style Fusion	4.60	75M
Ours	4.60	75M

Table 2: The average running time (in seconds per batch) and the number of parameters.

where w_n is a weight which is set to the accuracy of the corresponding classifier.

3 Experiments

3.1 Data

Conversational Dataset We employ *DailyDialog*[2] (Li et al., 2017) as our conversational dataset $\boldsymbol{C}$. It is a human-written multi-turn dataset covering various topics of daily life. Table 1 shows some statistics of its training, validation, and test set. We split dialogue of K utterances into K-1 samples. Each sample consists of at most three continuous utterances. The last utterance of a sample is regarded as the response. The previous utterances of the response are concatenated as its dialogue history. Here, *Reddit* dataset is not employed as Gao et al. (2019b) because the post-reply format data collected from social networks is noisy and different from real conversations (Li et al., 2017).

Monolingual Stylistic Dataset Following Gao et al. (2019b), we use *Holmes*[3] as the stylistic dataset $\boldsymbol{S}$. It is collected from the Sherlock Holmes novel series and consists of roughly 38k sentences. We do not use the *arXiv* dataset as it contains too many special tokens, i.e., equations, and incomplete sentences, such as "is concerned" and "exactly identical restrictions".

3.2 Baselines

We compare the proposed approach with the following baselines:

- **S2S**, the sequence-to-sequence response generation model (Shang et al., 2015).
- **S2S+LM**, a S2S trained on $\boldsymbol{C}$ and a stylistic language model trained on $\boldsymbol{S}$ (Niu and Bansal, 2018). During the inference process, it generates a stylistic response by interpolating outputs of the two models.
- **Style Fusion**, a multi-task learning based model whose latent space fuses dialogue history, responses, and stylistic sentences with a specific structure (Gao et al., 2019b).

Note that we do not consider the Label-Fine-Tuning model and Polite Reinforcement Learning model (Niu and Bansal, 2018), because they require some training samples in the conversational dataset to have the desired style (Gao et al., 2019b).

3.3 Experiment Settings

We implement the proposed approach based on the released code of Style Fusion model[4]. The vocabulary table consists of the most frequent 20,000 words. S2S encoder, AE encoder, and the shared decoder are two-layer LSTMs. The number of their hidden units is 1000, which is also the size of the structured latent space. The dimension of $\boldsymbol{Z}^c$ and $\boldsymbol{Z}^s$ is 950 and 50, respectively. The maximum length is set to 90 for the dialogue history and 30 for the response.

During the training process, we use the ADAM optimizer, whose learning rate is 0.0003. σ^2 for sampling ϵ in Equation 8 is 0.1^2. Table 2 shows the average running time on a single TITAN X (Pascal) GPU. During the inference process, the weights γ and η for re-ranking are set to 0.5. The weight (accuracy) of n-gram classifier is 0.93, 0.87, 0.77, and 0.65 for n from 1 to 4. The number of candidate responses, N_y, is set to 10. The radius r is set to 3.

4 Results

4.1 Evaluation Metrics

Automatic Evaluation Considering that it is unfair to evaluate a response by the classifiers that are used for selecting the response (Song et al., 2020), we fine-tune a BERT (Devlin et al., 2019) to measure style intensity. Concretely, positive samples are the stylistic sentences. Negative samples are

[2]http://yanran.li/dailydialog

[3]https://github.com/golsun/StyleFusion

[4]https://github.com/golsun/StyleFusion

Model	SI(%)	Dist-1	Dist-2	BLEU-3	BLEU-4	Mean
S2S (Shang et al., 2015)	6.32	0.035	0.227	0.70	0.20	0.10
S2S+LM (Niu and Bansal, 2018)	32.79	0.015	0.086	0.55	0.08	0.13
Style Fusion (Gao et al., 2019b)	10.58	**0.043**	0.280	**0.82**	0.22	0.14
Ours (α=0.25)	11.91	0.041	0.275	0.79	**0.23**	0.16
Ours (α=0.50)	20.67	0.040	0.275	0.64	0.17	**0.19**
Ours (α=0.75)	**34.85**	0.038	**0.285**	0.47	0.10	0.16

Table 3: Automatic evaluation results of SI, Dist-1, Dist-2, and BLEU. The last column is the harmonic mean of SI and BLEU-4 measuring the overall performance of style intensity and content relevance.

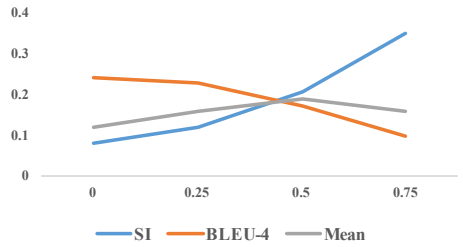


Figure 4: The trade-off between style intensity measured by SI and content relevance measured by BLEU-4. The x-axis corresponds to α. The harmonic mean achieves the maximum around α=0.5.

	Content Relevance		Style Intensity	
	Win	Lose	Win	Lose
vs. S2S	40.21	39.79	49.37	36.84
vs. S2S+LM	65.00	20.00	53.30	32.50
vs. Style Fusion	43.32	42.67	48.77	36.68

Table 4: Pair-wise human evaluation results of content relevance and style intensity.

randomly selected from DailyDialog's responses, which are of the same amount of sentences as the positive samples. Given the fine-tuned BERT classifier (whose accuracy achieves 0.96 on the validation set), we report the average probability of responses being positive as a measurement of the style intensity. For brevity, we denote this metric as *SI*. The content relevance is evaluated by BLEU. Since it may correlate weakly with human judgments of quality in a single reference setting (Liu et al., 2016), we employ the expanded responses in multi-reference DailyDialog test set (Gupta et al., 2019) as references to alleviate the problem. Meanwhile, we evaluate the diversity by *Dist-k* (Li et al., 2016), which is the number of distinct k-grams normalized by the total number of words of responses.

Human Evaluation We randomly sample 200 messages from the test set of C to conduct the human evaluation from two aspects: style intensity and content relevance. Each aspect is independently evaluated by five Amazon Mechanical Turk (AMT)[5] workers whose approval rate is greater than 95%, and the number of approved is greater than 500. Given dialogue history and two responses generated by a baseline and our approach, the workers are asked to give a preference of which one is better (ties are also permitted).

[5]https://www.mturk.com

4.2 Results

Figure 4 shows the trade-off between style intensity and content relevance in our approach. There is an improvement in SI and a decrease in BLEU associated with the increase of α in Equation 11. To assess the overall performance, we also compute their harmonic mean, whose maximum lies around $\alpha = 0.5$. We thus conduct the human evaluation and analysis in this parameter setting.

We report the human evaluation results in Table 4. Our approach is clearly preferred in style intensity because the percentage of Win is significantly higher than that of Lose (p <0.001, T-test). In terms of content relevance, the ratios of Win in "vs. S2S" and "vs. Style Fusion" are similar to those of Lose. This suggests that our approach can significantly improve the style intensity without decreasing the content relevance. In contrast, S2S+LM loses in most of the cases in the content relevance. Following Zhou et al. (2018) and Ke et al. (2018), we evaluate the agreement of annotators via inter-rater consistency. The percentage of samples that at least three annotators have the same preference (3/5 agreement) is 81.80%. And the percentage for 4/5 agreement is 32.15%.

Table 3 shows the results of the automatic evaluation. Our approach has the highest mean score, which indicates that it achieves the best overall performance. S2S+LM has a high SI score, but its BLEU scores are not as good as others, i.e., S2S.

	SI	BLEU-3	BLEU-4	Mean
Full Model	11.71	0.67	0.17	0.14
-Disentangle	7.52	0.68	0.17	0.11
-L_{fuse}	6.46	0.59	0.15	0.09
-L_{smooth}	6.02	0.63	0.17	0.09

Table 5: Results of the ablation study.

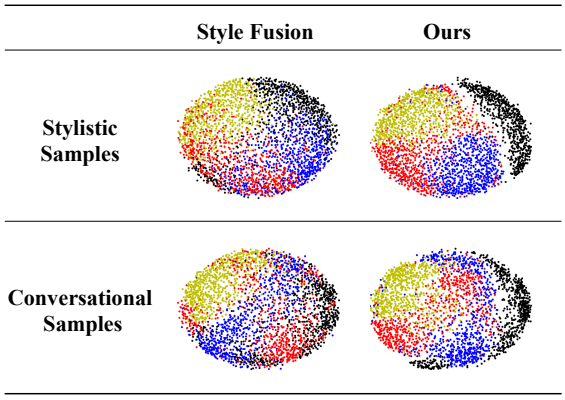


Figure 5: MDS visualization of $\boldsymbol{Z}^s$ (black) and three continuous sub-sequences extracted from the head (yellow), middle (red), and tail (blue) of $\boldsymbol{Z}^c$.

This is in line with our human evaluation results and Niu and Bansal (2018)'s observation that biasing a decoder with a stylistic language model may harm the content relevance. In contrast, our approach ($\alpha = 0.25$) significantly outperforms S2S and is comparable to Style Fusion. By increasing α to 0.5, the BLEU score drops slightly but is comparable to baselines (evidenced by the human evaluation results). Meanwhile, there is a significant improvement (up to 95.37%) in SI comparing with Style Fusion. This verifies the effectiveness of our disentanglement approach in improving the style intensity and maintaining the content relevance. Besides, the Dist-k results in Table 3 also indicate that the diversity of our approach is comparable to the best-performed Style Fusion.

4.3 Ablation Study

We conduct ablation studies to investigate the contributions of the fusion objective, smoothness objective, and our disentanglement approach. To focus on their effects on the generation process, in this section, we sample a single response without using the re-ranking strategy (Equation 12).

Table 5 shows the results of the ablation study. There is a significant decline in SI and a slight change in BLEU-3 and BLEU-4 after removing each component. This indicates that a multi-task learning architecture without the three components can achieve a good content relevance performance but fails to stylize a response. By removing the disentanglement component, our approach degenerates into Style Fusion. In this case, the SI score decreases significantly while BLEU scores are nearly unchanged, which demonstrates the disentanglement could improve the style intensity and maintain the relevance at the same time. The decreases in SI after removing the fusion objective and smoothness objective are more significant than that after removing the disentanglement. This is because the two objectives are bottom components for constructing the structured latent space, where our approach and Style Fusion are built upon.

	$[\boldsymbol{Z}^c : \boldsymbol{Z}^s]$	$\boldsymbol{Z}^s$
Style Fusion	0.83	0.72 (-13.02%)
Ours	0.88	0.86 (-1.71%)

Table 6: Style classification accuracy of the full latent variable ($[\boldsymbol{Z}^c : \boldsymbol{Z}^s]$) and $\boldsymbol{Z}^s$.

4.4 Analysis

In this section, we analyze whether style information is disentangled into $\boldsymbol{Z}^s$. To achieve this goal, we train style classifiers taking as input a latent variable and use the validation accuracy as an indicator. Taking our approach as an instance, we first freeze the parameters of our well-trained model. Then we independently learn two style classifiers whose inputs are the full latent variable ($[\boldsymbol{Z}^c : \boldsymbol{Z}^s]$) and $\boldsymbol{Z}^s$ respectively. Note that $\boldsymbol{Z}^c$ and $\boldsymbol{Z}^s$ in Style Fusion are a simple partition of its latent variable. There are not any disentanglement approaches applied to obtain the two representations. As shown in Table 6, Style Fusion achieves 0.83 validation accuracy training on its full latent variable. And the accuracy decreases by 13.02% when the classification is only based on $\boldsymbol{Z}^s$. In contrast, the decrease of our approach is only 1.71%, indicating that most of the style information is disentangled into $\boldsymbol{Z}^s$.

We show a visualization of the disentanglement of the latent variable by MDS (Borg and Groenen, 2005) in Figure 5. Each figure consists of $\boldsymbol{Z}^s$ (black) and three continuous sub-sequences extracted from the head (yellow), middle (red), and tail (blue) of $\boldsymbol{Z}^c$. The sub-sequences are of the same length with $\boldsymbol{Z}^s$. For both stylistic and conversational samples, all the sub-sequences and $\boldsymbol{Z}^s$ are mixed in Style Fusion. In contrast, there is a clear separation between $\boldsymbol{Z}^s$ and the sub-sequences

Dialogue History	– Yes , after my graduation, I worked in a trade company in Macao for one year. – What section did you work in? – The export business section.
S2S	– Have you received any other honors?
S2S+LM	– Did you ever find a job that you did not?
Style Fusion	– What was your major in?
Ours	– I 'm glad that you have been doing great.
Dialogue History	– I will always have you come in for a discussion if there is a major problem. – Will I be able to read the results on the website and understand them? – Yes, if you go there, you can see what each test is about.
S2S	– OK, I will ask the manager to take care of that.
S2S+LM	– I'm not sure I'll have you to do that,
Style Fusion	– What are my parents for me?
Ours	– I am looking forward to getting the information about that.

Table 7: An example of responses generated by baselines and our approach.

in our approach. This is because most of the style information is disentangled into $\boldsymbol{z}^s$ in our approach, making its distribution different from sub-sequences of $\boldsymbol{z}^c$.

4.5 Case Study

Table 7 shows some examples of generated responses. There is no significant Holmes style in responses of S2S. Similarly, the style intensity of responses in Style Fusion is also limited. The semantics of S2S+LM's response in the first example is not very clear, making it less relevant to the dialogue history than other responses. We believe this is also due to the lack of interaction between the response generation encoder and the stylistic language model. In contrast, our approach not only achieves a good content relevance performance but also has a significant Holmes style, which is quite polite and formal.

5 Related Work

5.1 Text Style Transfer without Parallel Data

The task of text style transfer aims at transferring the style of a sentence while preserving its meaning. One way is to disentangle the content and style, and subsequently combine the content with the desired style. The disentanglement can be achieved by adversarial learning (Shen et al., 2017; Hu et al., 2017; Fu et al., 2018; Yang et al., 2018; Logeswaran et al., 2018), reinforcement learning (Jain et al., 2019), back-translation (Prabhumoye et al., 2018; Nogueira dos Santos et al., 2018), multi-task learning (John et al., 2019), and removing stylistic phrases (Li et al., 2018; Xu et al., 2018; Zhang et al., 2018b). The other way transfers the style without disentangled representations, for example using generator-evaluator architecture (Gong et al., 2019), cycle reconstruction (Dai et al., 2019), parameter sharing (Wang et al., 2020), and data augmentation (Zhang et al., 2020a).

The main difference between our task and text style transfer lies in two aspects. First, all the content to be generated is available in the input in text style transfer, while our task needs to create new (response) content. And the key is content relevance to the dialogue history, rather than content preservation of the input. Second, the data for text style transfer is isomorphic. Data in different styles are in the same free-text format. However, our conversational data are context-response pairs while the stylistic data are free-texts, which is heterogeneous and requires more sophisticated structures, i.e., the structured latent space (Gao et al., 2019b).

5.2 Stylistic Response Generation without Parallel Stylistic Data

Niu and Bansal(2018) propose three weak-supervised models based on reinforcement learning, conditional text generation, and language model. Gao et al. (2019b) fuses the latent spaces of a response generation model and a stylistic auto-encoder to improve the style intensity of sampled responses. Yang et al. (2020) inject the style information by introducing a word-level KL loss and a sentence-level style classifier to the fine-turning process of DialoGPT (Zhang et al., 2020b). Distinct from previous work, we explicitly disentangle the style and content in the latent space and employ a unified architecture to jointly optimize the style intensity and content relevance.

6 Conclusion

We propose a uniform framework to simultaneously improve the style intensity and maintain the content relevance for neural stylistic response generation. In contrast to existing approaches, our approach

disentangles the style and the content in the latent space by a diluting strategy. Experiments show that our approach improves the style intensity of generated responses and maintains the content relevance at the same time, which demonstrates the effectiveness of this approach.

Acknowledgments

The authors would like to thank all the anonymous reviewers for their insightful comments. The authors from HIT are supported by the National Natural Science Foundation of China (No. 62076081, No. 61772153, and No. 61936010) and Science and Technology Innovation 2030 Major Project of China (No. 2020AAA0108605). The author from UCSB is not supported by any of the projects above.

Ethical Statement

This paper honors the ACL Code of Ethics. Stylistic response generation intends to improve the engagement of a dialogue system in human-bot conversations. It responds to users with the desired style, i.e., being polite, humorous, or romantic, rather than imitating any specific person. Meanwhile, style is a linguistic aspect of natural language interaction. There is not any identity characteristic being used as a variable.

References

Ingwer Borg and Patrick JF Groenen. 2005. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media.

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.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

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*.

Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. Stylenet: Generating attractive visual captions with styles. In *Proceedings of CVPR*, pages 955–964. IEEE.

Xiang Gao, Sungjin Lee, Yizhe Zhang, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2019a. Jointly optimizing diversity and relevance in neural response generation. 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 1229–1238, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiang Gao, Yizhe Zhang, Sungjin Lee, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2019b. 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, Hong Kong, China. Association for Computational Linguistics.

Hongyu Gong, Suma Bhat, Lingfei Wu, JinJun Xiong, and Wen-mei Hwu. 2019. Reinforcement learning based text style transfer without parallel training corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 3168–3180.

Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskenazi, and Jeffrey Bigham. 2019. Investigating evaluation of open-domain dialogue systems with human generated multiple references. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 379–391, Stockholm, Sweden. Association for Computational Linguistics.

Zhiting 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.

Parag Jain, Abhijit Mishra, Amar Prakash Azad, and Karthik Sankaranarayanan. 2019. Unsupervised controllable text formalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6554–6561.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434.

Jad Kabbara and Jackie Chi Kit Cheung. 2016. Stylistic transfer in natural language generation systems using recurrent neural networks. In *Proceedings of the Workshop on Uphill Battles in Language Processing: Scaling Early Achievements to Robust Methods*, pages 43–47.

Pei Ke, Jian Guan, Minlie Huang, and Xiaoyan Zhu. 2018. Generating informative responses with controlled sentence function. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1)*, pages 1499–1508.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. 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*, pages 110–119.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics,Volume 1*, pages 1865–1874.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Lajanugen Logeswaran, Honglak Lee, and Samy Bengio. 2018. Content preserving text generation with attribute controls. In *Advances in Neural Information Processing Systems*, pages 5103–5113.

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

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1)*, pages 866–876.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Cicero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. 2018. Fighting offensive language on social media with unsupervised text style transfer. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2)*, pages 189–194, Melbourne, Australia.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1)*, pages 1577–1586.

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

Haoyu Song, Wei-Nan Zhang, Jingwen Hu, and Ting Liu. 2020. Generating persona consistent dialogues by exploiting natural language inference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8878–8885.

Yunli Wang, Yu Wu, Lili Mou, Zhoujun Li, and Wenhan Chao. 2020. Formality style transfer with shared latent space. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2236–2249.

Jingjing Xu, Xu Sun, Qi Zeng, Xiaodong Zhang, Xuancheng Ren, Houfeng Wang, and Wenjie Li. 2018. Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Volume 1*, pages 979–988.

Ze Yang, Wei Wu, Can Xu, Xinnian Liang, Jiaqi Bai, Liran Wang, Wei Wang, and Zhoujun Li. 2020. StyleDGPT: Stylized response generation with pretrained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1548–1559, Online. Association for Computational Linguistics.

Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. 2018. Unsupervised text style transfer using language models as discriminators. In *Advances in Neural Information Processing Systems*, pages 7287–7298.

Ye Zhang, Nan Ding, and Radu Soricut. 2018a. SHAPED: Shared-private encoder-decoder for text style adaptation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 1528–1538.

Yi Zhang, Tao Ge, and Xu Sun. 2020a. Parallel data augmentation for formality style transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3221–3228, Online. Association for Computational Linguistics.

Yi Zhang, Jingjing Xu, Pengcheng Yang, and Xu Sun. 2018b. Learning sentiment memories for sentiment modification without parallel data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1108.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing

Liu, and Bill Dolan. 2020b. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In *the 27th International Joint Conference on Artificial Intelligence and the 23rd European Conference on Artificial Intelligence*, pages 4623–4629.