Structuring Latent Spaces for Stylized Response Generation

Xiang Gao Yizhe Zhang Sungjin Lee [*]
Michel Galley Chris Brockett Jianfeng Gao Bill Dolan
Microsoft Research, Redmond, WA, USA
{xiag,billdol}@microsoft.com

Abstract

Generating responses in a targeted style is a useful yet challenging task, especially in the absence of parallel data. With limited data, existing methods tend to generate responses that are either less stylized or less context-relevant. We propose STYLEFUSION, which bridges conversation modeling and non-parallel style transfer by sharing a structured latent space. This structure allows the system to generate stylized relevant responses by sampling in the neighborhood of the conversation model prediction, and continuously control the style level. We demonstrate this method using dialogues from Reddit data and two sets of sentences with distinct styles (arXiv and Sherlock Holmes novels). Automatic and human evaluation show that, without sacrificing appropriateness, the system generates responses of the targeted style and outperforms competitive baselines. [1]

1 Introduction

A social chatbot designed to establish long-term emotional connections with users must generate responses that not only match the *content* of user input and context, but also do so in a desired target *style* (Zhou et al., 2018; Li et al., 2016b; Luan et al., 2016; Gao et al., 2019a). A conversational agent that speaks in a polite, professional tone is likely to facilitate service in customer relationship scenarios; likewise, an agent that sounds like an cartoon character or a superhero can be more engaging in a theme park. The master of response style is also an important step towards human-like chatbots. As highlighted in social psychology studies (Niederhoffer and Pennebaker, 2002a,b), when two people are talking, they tend to match linguistic style of each other, sometime even regardless of their intentions. Achieving this level of performance, however, is challenging. Lacking parallel data in different conversational styles, researchers often resort to what we will term *style datasets* that are in non-conversational format (e.g. news, novels, blogs). Since the contents and formats of these are quite different from conversation data, existing approaches tend to generate responses that are either less style-specific (Luan et al., 2017) or less context-relevant (Niu and Bansal, 2018).

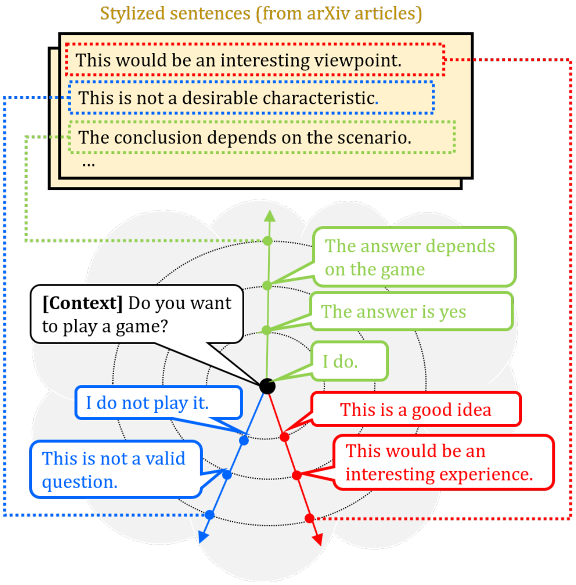
Stylized sentences (from arXiv articles)
This would be an interesting viewpoint.
This is not a desirable characteristic.
The conclusion depends on the scenario.
...
[Context] Do you want to play a game?
The answer depends on the game
The answer is yes
I do.
I do not play it.
This is not a valid question.
This is a good idea
This would be an interesting experience.

Figure 1: STYLEFUSION helps conversational model to distill style from non-conversational, non-parallel sentences by mapping them to points surrounding the related conversations in the structured latent space. Direction and distance from the model prediction (center black dot) roughly correspond to contents and style intensity, respectively, illustrated by examples taken from Table 2.

We suggest that this trade-off between appropriateness and style stems from profound differences

[*] Now at Alexa AI, Amazon.

[1] An implementation of our model and the scripts to generate the datasets are available at https://github.com/golsun/StyleFusion.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 1814–1823,
Hong Kong, China, November 3–7, 2019. ©2019 Association for Computational Linguistics

between conversation and style datasets in format, style and contents that impede joint learning. One approach has been to combine these only during decoding: Niu and Bansal (2018) trained two models separately, a Sequence-to-Sequence (S2S) (Sutskever et al., 2014) on a conversation dataset and a language model (LM) on a style dataset. At inference time, they take a weighted average of the token probability distribution of the two models to predict the next token. This forced bias, however, degrades output relevance. An alternative approach attempts to map the two datasets into the same latent space: Luan et al. (2017) use multi-task learning to train a S2S model on a conversation dataset and an autoencoder (AE) on a style dataset. Gao et al. (2019b) point out that the two datasets still form separate clusters in the latent space; below we observe that this leads to a low style intensity in generated responses (Section 5).

We propose to bridge conversation modeling and non-parallel style transfer by structuring a shared latent space using novel regularization techniques, that we dub STYLEFUSION. In contrast to Luan et al. (2017), the two datasets are well aligned in the latent space by generalizing SPACEFUSION [2] (Gao et al., 2019b), an approach that aligns latent spaces for paired samples, to non-parallel datasets. In the structured shared latent space, stylized sentences are nudged adjacent to semantically related conversations, thereby allowing the system to generate style-specific relevant responses by sampling in the neighborhood of the model prediction. Distance and direction from the model prediction roughly match the style intensity and content diversity of generated responses, respectively, as illustrated in Fig. 1

We demonstrate this method using dialogues from Reddit data and two sets of sentences with distinct styles (arXiv and Sherlock Holmes novels). Automatic and human evaluation show that, without sacrificing appropriateness, our system can generate responses in a targeted style and outperforms competitive baselines.

Our contribution can be summarized thus: 1) We introduce an end-to-end approach that generates style-specific responses from conversational data and non-parallel non-conversation style data. 2) We generalize the SPACEFUSION model of (Gao et al., 2019b) to non-parallel data by a new regularization method. 3) We present a visualization analysis that provides intuitive insights into the drawbacks of alternative approaches.

2 Related Work

Text style transfer is a related but distinct task. It usually preserves the content (Yang et al., 2018; Hu et al., 2017; Fu et al., 2018; Shen et al., 2017; Gong et al., 2019). In contrast, content of conversational responses in a given context can be semantically diverse. Various approaches have been proposed for non-parallel data setup. Fu et al. (2018) proposed to use separate decoders for different styles and a classifier to measure style strength. Shen et al. (2017) proposed to map texts of two different styles into a shared latent space where the ”content” information is preserved and ”style” information is discarded. An adversarial discriminator is used to align the latent spaces of two different styles. However, Yang et al. (2018) point out the difficulty of training an adversarial discriminator and proposed instead the use of language models as discriminator. Like Shen et al. (2017); Yang et al. (2018), we align latent spaces for different styles. However we also align latent spaces encoded by different models (S2S and AE).

Stylized response generation is a relatively new task. Akama et al. (2017) use a stylized conversation corpus to fine-tune a conversation model pretrained on a background conversation dataset. However, stylized texts are usually in non-conversational format, as in the present setting. Niu and Bansal (2018) proposed a method that takes the weighted average of the token probability distribution predicted by a S2S trained on background conversational dataset and that predicted by a LM trained on style dataset as the token probability. They observed reduced relevance and attributed this to the fact that the LM was not trained to attend to conversation context and S2S was not trained to learn style during training. In contrast, we jointly learn from conversation and style datasets during training. Niu and Bansal (2018) have proposed label-fine-tuning, but this is limited to scenarios where a reasonable portion of the conversational dataset is in the target style, which is not always the case.

Persona-grounded conversation modeling Li et al. (2016b); Luan et al. (2017) aim to generate responses mimicking a speaker. It is closely re-

[2] Integrated into Microsoft Icecaps toolkit (Shiv et al., 2019) https://github.com/microsoft/icecaps.

lated to the present task, since persona is, broadly speaking, the manifestation of a type of style. Li et al. (2016b) feeds a speaker ID to the decoder to promote generation of response for that target speaker. However non-conversational data cannot be used. Luan et al. (2017) applied a multi-task learning approach to utilize non-conversational data. A S2S model, taking in conversational data, and an autoencoder (AE), taking in non-conversational data, share the decoder and are trained alternately. However, Gao et al. (2019b) observed that sharing the decoder may not truly allow S2S and AE to share the latent space, and thus S2S may not fully utilize what is learned by AE. Unlike Li et al. (2016b) using labelled persona IDs, Zhang et al. (2019) have proposed using a self-supervised method to extract persona features from conversation history. This allows modeling persona dynamically, which agrees with the fact that even the same person can speak in different style in different scenarios.

Multi-task learning McCann et al. (2018); Liu et al. (2019); Luan et al. (2017); Gao et al. (2019b); Zhang et al. (2017) aggregates the strengths of each specific task, and induces regularization effects (Liu et al., 2019) as the model is trained to learn a more universal representation. However a simple multi-task approach (Luan et al., 2017) may learn separate representations for each dataset (Gao et al., 2019b). To address this, in previous work (Gao et al., 2019b), we proposed the SPACEFUSION model featuring a regularization technique that explicitly encourages alignment of latent spaces for a universal representation. SPACEFUSION, however, is only designed for paired samples. We generalize SPACEFUSION to non-parallel datasets in this paper.

3 The STYLEFUSION Model

3.1 Problem statement

Let $\mathcal{D}_{conv} = [(x_0, y_0), (x_1, y_1), \cdots, (x_n, y_n)]$ denote a conversation dataset, where x_i and y_i are context sentences and a corresponding response, respectively. x_i consists of one or more utterances and y_i is only one utterance. $\mathcal{D}_{style} = [s_0, s_1, \cdots, s_m]$ denotes a non-conversational style dataset, where s_i is a sentence sampled from a corpus of the targeted style. Samples from $\mathcal{D}_{style}$ do not have a labelled corresponding relation with samples from $\mathcal{D}_{conv}$ (thus ”non-parallel”). Our aim is to train a model jointly on $\mathcal{D}_{style}$ and $\mathcal{D}_{conv}$ to generate appropriate responses in the style similar to sentences from $\mathcal{D}_{style}$, to a given context. The iven context may or may not be in the target style.

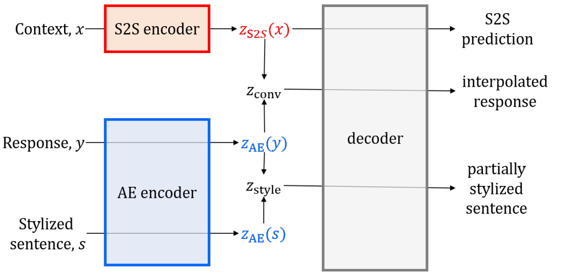

Figure 2: STYLEFUSION model architecture.

3.2 Training

In contrast to SPACEFUSION(Gao et al., 2019b), which only fuses context-response pairs, our goal is to additionally map related stylized sentences to points surrounding the context in the shared latent representation space. The system can then generate relevant stylized responses by sampling in the neighborhood of the prediction based on the context.

Model overview. As illustrated in figure 2, the model consists of a sequence-to-sequence (S2S) module and an autoencoder (AE) module that shares a decoder. We use S2S to produce the prediction representation $z_{S2S}(x_i)$, and AE to obtain the representation of the corresponding responses $z_{AE}(y_i)$ and stylized sentences $z_{AE}(s_j)$. We use generalized regularization terms, fusion and smoothness, to align the three latent spaces $z_{S2S}(x_i)$, $z_{AE}(y_i)$, and $z_{AE}(s_j)$.

Fusion objectives encourage different latent spaces to be close to each other. Accordingly, we define the cross-latent-space distances to be minimized. For response appropriateness, as x_i and y_i are paired as context and response, we use their pair-wise dissimilarity, following Gao et al. (2019b)

$$d_{\text{conv}} = \sum_{i \in \text{batch}} \frac{d_E(z_{\text{S2S}}(x_i), z_{\text{AE}}(y_i))}{n\sqrt{l}} \tag{1}$$

where n is the batch size, l is the dimension of latent space , and we use d_E, the Euclidean distance in latent space, as the dissimilarity.

For style transfer, however, x_i and s_j is not paired. Thus, we instead minimize the distance between a point and its nearest neighbor from another dataset to pull these two datasets close to each other in the shared latent space.

$$d_{\text{style}} = \frac{1}{2} d_{\text{NN}}^{\text{cross}}(\{z_{\text{S2S}}(x_i)\}, \{z_{\text{AE}}(s_i)\}) + \frac{1}{2} d_{\text{NN}}^{\text{cross}}(\{z_{\text{AE}}(s_i)\}, \{z_{\text{S2S}}(x_i)\}) \quad (2)$$

where $d_{NN}^{\text{cross}}(\{a_i\}, \{b_i\})$ is the batch average of the distance between a_i and $b_{\text{NN of } a_i}$ – the nearest neighbor (NN) of a_i from set $\{b_i\}$

$$d_{\text{NN}}^{\text{cross}}(\{a_i\}, \{b_i\}) = \sum_{i \in \text{batch}} \frac{d_E(a_i, b_{\text{NN of } a_i})}{n\sqrt{l}} \quad (3)$$

While minimizing the cross-latent-space distances, d_{conv} and d_{style}, we want the samples from the same latent space spread out, following Gao et al. (2019b). For this purpose, Gao et al. (2019b) maximized the average of capped distance between points from the same latent space. However, we found that the results are sensitive to the cap value. Instead, we define the following nearest-neighbor-based characteristic distance

$$d_{\text{spread-out}} = \min[d_{\text{NN}}^{\text{same}}(\{z_{\text{AE}}(y_i)\}), d_{\text{NN}}^{\text{same}}(\{z_{\text{AE}}(s_i)\}), d_{\text{NN}}^{\text{same}}(\{z_{\text{S2S}}(x_i)\})] \quad (4)$$

$$d_{\text{NN}}^{\text{same}}(\{a_i\}) = \sum_{i \in \text{batch}} \frac{d_E(a_i, a_{\text{NN of } a_i})}{n\sqrt{l}} \quad (5)$$

Combining these loss terms we have the following two objectives:

$$\mathcal{L}_{\text{fuse,conv}} = d_{\text{conv}} - d_{\text{spread-out}} \quad (6)$$

$$\mathcal{L}_{\text{fuse,style}} = d_{\text{style}} - d_{\text{spread-out}} \quad (7)$$

Smoothness objective encourages smooth semantic transition in the shared latent space. For response appropriateness, following Gao et al. (2019b), we encourage the interpolation between the prediction $z_{S2S}(x_i)$ and the target response $z_{AE}(y_i)$ to generate the target response y_i.

$$z_{\text{conv}} = (1 - u) z_{\text{AE}}(y) + u z_{\text{S2S}}(x) + \epsilon \quad (8)$$

$$\mathcal{L}_{\text{smooth,conv}} = -\frac{1}{|y|} \log p(y | z_{\text{conv}}) \quad (9)$$

where $u \sim U(0, 1)$ is a uniformly distributed random variable, and ϵ is a Gaussian noise with zero mean and covariance matrix of $\sigma^2 I$.

For style transfer, as we move from a non-stylized sentence $z_{\text{AE}}(x)$ to a random stylized sentence $z_{\text{AE}}(s)$, we expect to generate a partially stylized sentence and encourage the generated sentence to gradually change from x to s.

$$z_{\text{style}} = (1 - u) z_{\text{AE}}(x) + u z_{\text{AE}}(s) + \epsilon \quad (10)$$

$$\mathcal{L}_{\text{smooth,style}} = -(1 - u) \frac{1}{|x|} \log p(x | z_{\text{style}}) - u \frac{1}{|s|} \log p(s | z_{\text{style}}) \quad (11)$$

Training objective to be minimized is a combination of a vanilla S2S and the above regularization terms [3]. $\mathcal{L}_{\text{smooth,style}}$ and $\mathcal{L}_{\text{fuse,style}}$ are new terms not existing in (Gao et al., 2019b). A more compact definition $\mathcal{L}_{\text{conv}} = \mathcal{L}_{\text{smooth,conv}} + \mathcal{L}_{\text{fuse,conv}}$ and $\mathcal{L}_{\text{style}} = \mathcal{L}_{\text{smooth,style}} + \mathcal{L}_{\text{fuse,style}}$ yields

$$\mathcal{L} = -\frac{1}{|y|} \log p(y | z_{\text{S2S}}) + \mathcal{L}_{\text{conv}} + \mathcal{L}_{\text{style}} \quad (12)$$

For the case $\mathcal{D}_{style}$ is much smaller than $\mathcal{D}_{conv}$, as in the present work, the model may overfit on $\mathcal{D}_{style}$. We propose to firstly pretrain the model on $\mathcal{D}_{conv}$ only [4], then continue training on both $\mathcal{D}_{conv}$ and $\mathcal{D}_{style}$. Furthermore, to reduce overfitting, we applied a data augmentation technique by randomly mask tokens in s_i by a special out-of-vocab token. The masking probability of a token is inversely proportional to its frequency in training data.

3.3 Inference

Following (Gao et al., 2019b), we sample in the neighborhood of $z_{S2S}(x)$ by adding a noise r of a given length $|r|$ towards a direction randomly drawn from the uniform distribution.

$$z = z_{\text{S2S}}(x) + r \quad (13)$$

As r depends on l, the dimension of z, We define a normalized value

$$\rho = |r| / (\sigma \sqrt{l}) \quad (14)$$

As the stylized texts are usually sparse, it is possible to generate non-stylized hypothesis as we vary ρ along some direction. Thus we rank the hypotheses considering both relevance and style intensity.

$$\text{score}(h_i) = (1 - \lambda) P(h_i | z_{\text{S2S}}(x)) + \lambda P_{\text{style}}(h_i) \quad (15)$$

[3] More generally, one may use a weighted sum of these terms instead. We set them equally weighted for simplicity

[4] by setting terms $\mathcal{L}_{\text{smooth,style}}$ and $\mathcal{L}_{\text{fuse,style}}$ as zero

where $\lambda = 0.5$ unless otherwise specified, $P(h_i|z_{\text{S2S}}(x))$ estimates the relevancy, and $P_{style}(h_i)$ is the probability of hypothesis h_i being targeted style predicted by pretrained classifiers.

We considered two style classifiers: a "**neural**" based on two stacked GRU (Cho et al., 2014) cells, and a "**ngram**" classifier which is a logistic regressor using ngram (n=1,2,3,4) multi-hot features. Both classifiers are trained using $\{y_i\}$ as negative samples and $\{s_i\}$ as positive samples. $P_{style}(h_i)$ is computed by taking average of the prediction of these two classifiers.

4 Experiment Setup

4.1 Tasks and datasets

We experiments with two tasks: generating arXiv-like and Holmes-like responses, respectively, using the datasets summarized in Table 1

Task	Training		Testing
	$\mathcal{D}_{\text{conv}}$	$\mathcal{D}_{\text{style}}$	$\mathcal{D}_{\text{test}}$
arXiv-like	Reddit	arXiv	arXiv-like Reddit
Holmes-like	Reddit	Holmes	Holmes-like Reddit

Table 1: Summary of tasks and datasets

(i) *Reddit* is a conversation dataset constructed from posts and comments on Reddit.com [5] during 2011, consisting of 10M pairs of context and response . (ii) *arXiv* is a non-conversational dataset extracted from articles on arXiv.org [6] from 1998 to 2002, consisting of 1M sentences . (iii) *Holmes* refers to another non-conversational dataset extracted from Sherlock Holmes novel series[7] by Arthur Conan Doyle, with 38k sentences.

$\mathcal{D}_{\text{test}}$ is the test set with stylized reference responses, constructed by filtering the Reddit dataset from year 2013 using the trained neural and ngram classifiers. For each context, there are at least 4 reference responses approximately in the targeted style ($P_{\text{style}} > 0.3$). The style intensity of the context is not filtered.

4.2 Human evaluation

We designed the following two tasks.

Response appropriateness measurement task presents a context and a set of hypotheses (from the present and baseline systems), and for each hypothesis annotators choose from one of the following options that best fits the quality of the response: ok, marginal, bad (generic or irrelevant), and then map them to numerical score 1, 0.5, and 0, respectively.

Style classification task presents a hypothesis and two groups of example sentences, from Reddit and style corpus (Holmes or arXiv). Then crowd-sourced annotators judge whether the hypothesis is more similar to the Reddit group, not sure, or more similar to the style corpus group. We then map these to numerical scores 0, 0.5, and 1, respectively.

For all tasks, the hypotheses of different systems of the same set of 500 randomly selected x_i are presented in random order and the identity of the system is invisible to annotators. Each sample is judged by 5 annotators individually.

4.3 Automatic evaluation

We measure relevance using multi-reference BLEU Papineni et al. (2002), and diversity by entropy 4-gram (Zhang et al., 2018), and distinct 1,2-gram (Li et al., 2016a).

For style intensity evaluation, besides the **neural** and **ngram** classifier prediction (Section 3.3), we also use simple word-counting (hereafter **count** metric) to minimize model-specific effects. We first construct a training corpus with balanced positive (from $\mathcal{D}_{\text{style}}$) and negative (responses sampled from $\mathcal{D}_{\text{conv}}$) samples. Then, for each word that appears in more than 5 sentences in the training corpus, we compute the average style intensity of sentences that contain this word. The top k words of highest style intensity are chosen as the keywords in this style. For a test corpus, we compute the average ratio of words that are keywords of a style, as its "count" style metric.

Besides the overall style comparisons (Reddit vs. Holmes, and Reddit vs. arXiv), we also crowd-sourced three sets of sentences with human labeled levels in three finer styles: how formal, emotional, and technical each sentence is, and build the keyword list for the count metric.

4.4 Baselines

We compare the following baseline systems.

[5] using raw data collected by a third party http://files.pushshift.io/reddit/comments/

[6] from KDD 2003 http://www.cs.cornell.edu/projects/kddcup/datasets.html

[7] from https://gutenberg.org

The first category is generative models. (i) *MTask* refers to the vanilla multi-task learning model proposed in (Luan et al., 2017) trained on both $\mathcal{D}_{\text{conv}}$ and $\mathcal{D}_{\text{style}}$. (ii) *S2S+LM* refers to the method proposed by Niu and Bansal (2018)[8], which uses the weighted average of a S2S model, trained on $\mathcal{D}_{\text{conv}}$, and a LM model, trained on $\mathcal{D}_{\text{style}}$, as the token probability distribution at inference time.

The second category draws a training sample as hypothesis. (iii) *Retrieval* refers to a simple retrieval system which returns the sentence from $\mathcal{D}_{\text{style}}$ of highest generation probability by the MTask model. (iv) *Rand* is a system that randomly picks a sentence from $\mathcal{D}_{\text{style}}$. (v) *Human* refers to the system randomly picks one of the multiple reference responses in the given context from $\mathcal{D}_{\text{test}}$.

4.5 Model setup

StyleFusion and trainable baselines, MTask and S2S+LM, use two stacked GRU (Cho et al., 2014) cells for encoders and decoders with $l = 1000$. The word embedding is also 1000 dimension, trained from random initialization. The variance of the noise ϵ is set to $\sigma^2 = 0.1^2$. The state of the top layer of encoder GRU cell is used as z. z is used as the initial state of all layers of the decoder. All trainable models are trained with the ADAM method (Kingma and Ba, 2014) with a learning rate of 0.0003. For StyleFusion and MTask, we first train on $\mathcal{D}_{\text{conv}}$ for 2 epochs, and then continue the training on both $\mathcal{D}_{\text{conv}}$ and $\mathcal{D}_{\text{style}}$ until convergence [9]. For all systems except "Rand" and "Retrieval", we use the ranking method Eq 15 to select the top one hypothesis from 100 candidates.

5 Results and Analysis

5.1 Modulating the style

By leveraging the structure of the shared latent space, we can modulate the style intensity by ρ, as illustrated by examples in Table 2. For example, given the context *"Do you want to play a game"*, the hypothesis generated from $\rho = 0$ is *"I do"*, which is non-stylized. While moving towards z_{AE} of an arXiv-style sentence *"This would be an interesting viewpoint"*, the responses gradually change to "This would be an interesting experience" at $\rho = 1.0$, which remains relevant but is more similar to the target style. Similar trends can be observed when moving in the other direction *"The conclusion depends on the scenario"* and *"This is not a desirable characteristic"*. It also shows that the contents are affected by the direction, a desired property inherited from SpaceFusion models.

context	Do you want to play a game?
towards	The conclusion depends on the scenario .
$\rho = 0.0$	I do.
$\rho = 0.5$	The answer is yes.
$\rho = 1.0$	The answer depends on the game.
towards	This would be an interesting viewpoint.
$\rho = 0.4$	This is a good idea.
$\rho = 0.9$	This would be an interesting experience
towards	This is not a desirable characteristic.
$\rho = 0.5$	I don't play it.
$\rho = 1.0$	This is not a valid question.

Table 2: Example StyleFusion outputs for arXiv-like response generation task at different distance ρ and direction (towards z_{AE} of a given sentence)

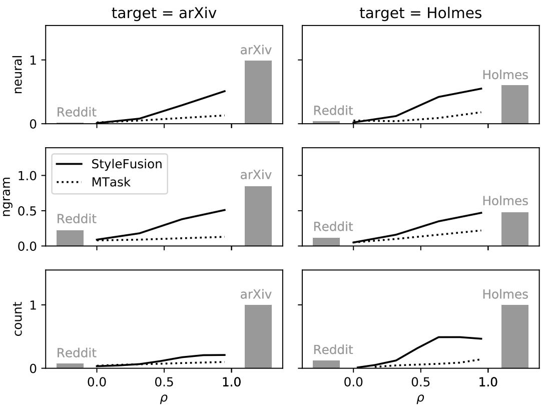

Figure 3: Change of the overall style intensity with ρ, as measured by two classifiers "neural" and "ngram" (Section 3.3) and the "count" metric. The "count" metric is normalized by the value of the target style corpus. The barplot shows the desired trend (from Reddit to arXiv or Holmes), and the lines the actual trends.

The relation between style intensity and ρ is further confirmed by automatic measurement. As illustrated in Fig. 3, as ρ increases, responses come to resemble the targeted style within the depicted range. In contrast, the style intensity of MTask outputs rises only slightly as ρ increases.

The increase of overall style intensity is coupled with change in the style's finer granularity, as illustrated in Fig. 4. Compared to Reddit, arXiv is less emotional, and more formal and technical. Consistent with this, StyleFusion outputs exhibit

[8] This method was referred as "Fusion" in (Niu and Bansal, 2018) but to avoid confusing readers with our StyleFusion method, we refer it as "S2S+LM"

[9] approximately one pass of arXiv and 5 passes of Holmes

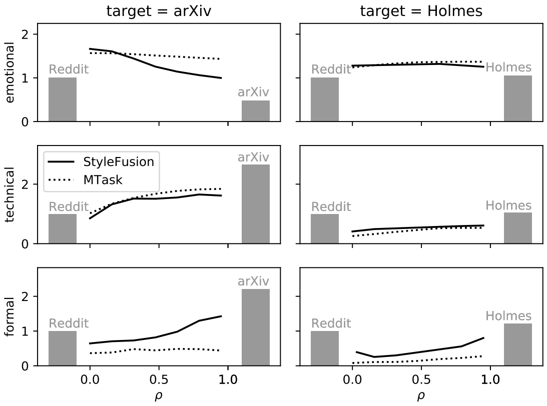

Figure 4: Change of the styles at finer granularity of ρ, measured by the ”count” metric normalized by the value of Reddit dataset. The barplot shows the desired trend (from Reddit to arXiv or Holmes), and the lines the actual trends.

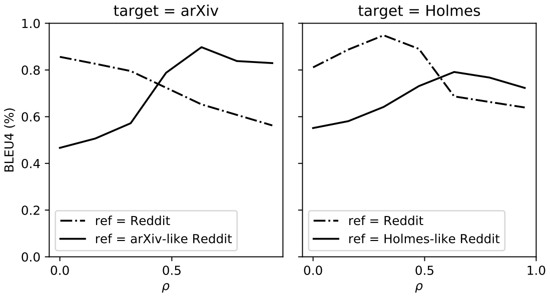

Figure 5: Relevancy of the StyleFusion outputs at different ρ as measured by BLEU4 with references of different styles

less emotion, but become much more technical and formal as ρ increases. MTask, however, tends only to show increased technical level, but fails to be less emotional and more formal, inconsistent with the target style. Where Holmes is the target, the emotional and technical levels do not significantly change compared to Reddit, but Holmes is stylistically more formal. StyleFusion captures these trends, whereas MTask outputs are insufficiently formal, shown in Fig. 4(lowest panel).

We also measured the BLEU4 score at different ρ as shown in Fig. 5. Besides $\mathcal{D}_{\text{test}}$ (stylized references), we also tested on held-out Reddit data (i.e. non-stylized references). We observed that at smaller ρ, the StyleFusion outputs are relevant to context and style is more similar to Reddit, indicated by the relatively high BLEU4 computed using non-stylized references. At greater radius, BLEU4 rises when calculated on arXiv-like and Holmes-like references, indicating that the responses generated remain relevant but are closer to the targeted style. Combined with the case at small radius, the implication is that, although the style gradually changes, the responses generated by StyleFusion remain relevant at a relatively wide range of radius.

context	Okay, but can we host it in the cloud?
StyleFusion	It would be an interesting experiment. It is a possibility.
S2S	I think it might be a bit of a stretch.
MTask	Yes, yes you can.
S2S+LM	The relevant question is , as to this , is you referring to the casimir energy program?
Retrieval	Even though this issue can be analyzed in more detail, we shall content ourselves with some simple statements.
rand	In particular, it involves the combination equation
human	To the cloud !
context	The universe has no edge.
StyleFusion	Does this mean the universe is expanding?
S2S	The universe is infinite.
MTask	The universe is finite.
S2S+LM	The universe can be on the infinite loop.
Retrieval	Is there really any difference in the low energy physics?
rand	In other words , the wave function in equation.
human	So... a sphere?

Table 3: Example system outputs on arXiv-like response generation task at $\rho = 1$

5.2 Fused latent space

As illustrated by the MDS (Borg and Groenen, 2003) visualization in Fig. 6, MTask cannot align different latent spaces, not only those from different model ($z_{\text{AE}}(y)$ and $z_{\text{S2S}}(x)$), but also for those from same model that have different styles ($z_{\text{AE}}(y)$ and $z_{\text{AE}}(s)$). SpaceFusion (Gao et al., 2019b) can align $z_{\text{AE}}(y)$ and $z_{\text{S2S}}(x)$ better, but $z_{\text{AE}}(s)$ forms a separate cluster, indicating that the conversation dataset and style dataset remain unaligned in the latent space. This is because SpaceFusion was not designed to align non-parallel samples. The separation between the conversation dataset and style dataset in latent space, as is the case for MTask and SpaceFusion, makes it difficult for the conversation model to use style knowledge. In contrast, StyleFusion aligns all three latent spaces well as evidenced by Fig. 6.

5.3 Human evaluation

Human evaluation results are presented in Table 5. As in the automatic evaluation results, StyleFusion and MTask show the highest appropriate-

context	First step: go check behind the cabinet and see what's back there.
StyleFusion	I would advise against this.
S2S	I'm going to try this next time.
MTask	I'll be back in a few minutes.
S2S+LM	I've got a little time that i had not been engaged yet.
Retrieval	i was thinking the same thing.
Rand	Holmes shrugged his shoulders.
human	Then there would be nothing , and that would be the answer.
context	that 's a surprisingly large amount of police coming to see a *** teenager.
StyleFusion	What is this crime you speak of?
S2S	I'm pretty sure that 's the only thing that 's going to happen.
MTask	I'm not sure what you mean by that.
S2S+LM	I should be on the same business.
Retrieval	well, yes.
Rand	I shall be back in an hour or two.
human	Must have feared what he was packin'

Table 4: Example system outputs on Holmes-like response generation task at $\rho = 1$

target	model	appropr-iateness	style intensity	harmonic mean
arXiv	StyleFusion	**0.17**	0.26	**0.20**
	MTask	**0.17**	0.11	0.14
	S2S+LM	0.09	0.51	0.15
	Retrieval	0.07	0.71	0.14
	Rand	0.04	**0.96**	0.07
	Human	0.51	0.28	0.36
Holmes	StyleFusion	**0.22**	0.28	**0.25**
	MTask	**0.23**	0.15	0.18
	S2S+LM	0.03	0.55	0.05
	Retrieval	0.14	0.30	0.19
	Rand	0.08	**0.69**	0.14
	Human	0.63	0.26	0.37

Table 5: Human evaluation results. The top models (and those models that are not statistically different, except "human") are in **bold**.

ness (not statistically different) apart from the Human system. However StyleFusion outputs are much more stylized. Rand, Retrieval and S2S+LM tend to generate stylized but irrelevant responses. To make the overall trends sharper, following (Gao et al., 2019b), we compute the harmonic mean of appropriateness and style intensity, in terms of which StyleFusion outperforms all baselines except the Human system. Additional examples of the system outputs and human responses are provided in Table 3 and Table 4

5.4 Ablation study and automatic evalution

The automatic evaluation results for arXiv-like and Holmes-like response generation tasks are presented in Table 6. In both instances, StyleFusion achieved relatively high BLEU, and showed high style intensity. The Rand baseline has the highest style intensity but lowest relevance. S2S+LM has the comparable style intensity to StyleFusion but BLEU is much lower, consistent with the observation made by (Niu and Bansal, 2018). MTask shows significantly less style intensity than StyleFusion. Moreover, MTask's diversity, as measured by entropy4 and distinct1,2, is much lower, indicating that outputs of this model tend to be bland. Adding $\mathcal{L}_{\text{conv}}$ regularization, which is SpaceFusion, increases diversity, relevance and style intensity slightly, consistent with the finding in (Gao et al., 2019b). Style intensity is further boosted by the addition of term $\mathcal{L}_{\text{style}}$. relevancy and diversity are not significantly affected by the addition of $\mathcal{L}_{\text{style}}$.

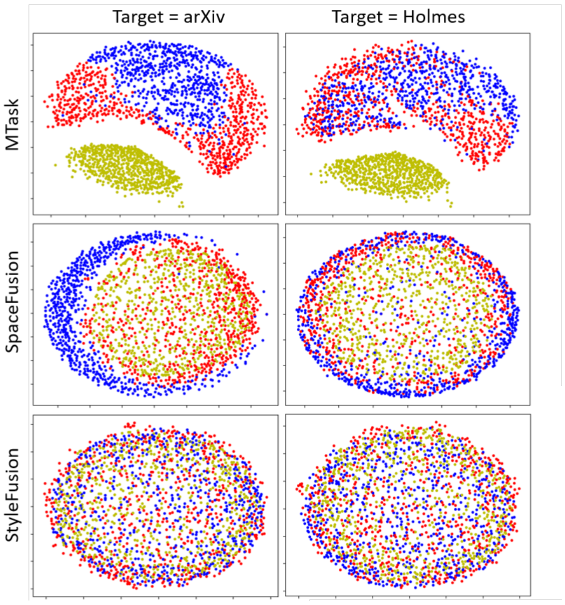

Figure 6: MDS visualization of learned latent spaces. yellow dots for $z_{\text{S2S}}(x)$, **blue** dots for $z_{\text{AE}}(s)$ and **red** dots for $z_{\text{AE}}(y)$

6 Conclusion

We propose a regularized multi-task learning approach, StyleFusion, that bridges conversation models and non-parallel style transfer by structuring a shared latent space. This structure allows the system to generate stylized relevant responses by sampling in the neighborhood of the model prediction, and to continuously control style intensity by modulating the sampling radius. We demonstrate this method in two tasks: generating arXiv-like and Holmes-like conversational responses. Automatic and human evaluation show that, with-

system	style intensity			relevancy				diversity		
sampled (∈) or generated	neural	ngram	count	BLEU1	BLEU2	BLEU3	BLEU4	entropy4	distinct1	distinc2
target = arXiv										
Rand ($\in \mathcal{D}_{\text{style}}$)	0.99	0.85	1.00	12.1	1.7	0.6	0.34	9.4	0.13	0.56
Retrieval ($\in \mathcal{D}_{\text{style}}$)	0.84	0.77	0.59	15.5	2.3	0.8	0.49	7.6	0.06	0.19
Human ($\in \mathcal{D}_{\text{test}}$)	0.43	0.47	0.35	29.0	16.3	10.6	7.44	8.6	0.31	0.81
S2S+LM	0.36	**0.48**	**0.34**	14.2	2.5	0.7	0.41	**9.4**	**0.11**	**0.54**
MTask	0.13	0.13	0.10	16.5	2.9	1.2	0.66	5.7	0.04	0.13
+$\mathcal{L}_{\text{conv}}$ (SpaceFusion)	0.27	0.41	0.17	**18.1**	3.9	1.4	0.75	6.9	0.04	0.13
+$\mathcal{L}_{\text{style}}$ (StyleFusion)	**0.40**	0.46	0.21	17.9	**4.4**	**1.6**	**0.83**	7.9	0.05	0.20
target = Holmes										
Rand ($\in \mathcal{D}_{\text{style}}$)	0.60	0.48	1.00	13.1	1.9	0.6	0.37	9.0	0.15	0.62
Retrieval ($\in \mathcal{D}_{\text{style}}$)	0.20	0.21	0.10	10.7	1.7	0.7	0.45	6.5	0.04	0.15
Human ($\in \mathcal{D}_{\text{test}}$)	0.46	0.43	0.67	26.5	13.7	9.2	6.65	9.3	0.16	0.60
S2S+LM	0.50	0.44	**0.59**	16.3	3.0	0.8	0.44	**8.7**	**0.07**	**0.38**
MTask	0.17	0.22	0.14	19.5	4.5	1.5	0.73	6.9	0.03	0.12
+$\mathcal{L}_{\text{conv}}$ (SpaceFusion)	0.39	0.33	0.22	18.9	4.6	1.5	**0.76**	7.7	0.04	0.17
+$\mathcal{L}_{\text{style}}$ (StyleFusion)	**0.55**	**0.48**	0.47	**20.6**	**5.1**	**1.6**	0.73	7.8	0.04	0.17

Table 6: Performance of each model on automatic measures. The highest score for generative models in each column is in **bold** for each target. the ”count” metric is normalized by the value of the targeted style dataset. Note that the unit for BLEU is percentage.

out sacrificing relevance, the system generates responses of the targeted style and outperforms competitive baselines. In future work, we will use this technique to distill information from other non-parallel datasets, such as external informative text (Qin et al., 2019; Galley et al., 2019).

References

Reina Akama, Kazuaki Inada, Naoya Inoue, Sosuke Kobayashi, and Kentaro Inui. 2017. Generating stylistically consistent dialog responses with transfer learning. In *IJCNLP*, pages 408–412.

Ingwer Borg and P Groenen. 2003. Modern multidimensional scaling: theory and applications. *Journal of Educational Measurement*, 40(3):277–280.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *SSST-8*, pages 103–111.

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*.

Michel Galley, Chris Brockett, Xiang Gao, Jianfeng Gao, and Bill Dolan. 2019. Grounded response generation task at dstc7. In *AAAI Dialog System Technology Challenges Workshop*.

Jianfeng Gao, Michel Galley, and Lihong Li. 2019a. Neural approaches to conversational ai. *Foundations and Trends in Information Retrieval*, 13(2-3):127–298.

Xiang Gao, Sungjin Lee, Yizhe Zhang, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2019b. Jointly optimizing diversity and relevance in neural response generation. *NAACL-HLT 2019*.

Hongyu Gong, Suma Bhat, Lingfei Wu, Jinjun Xiong, and Wen-mei Hwu. 2019. Reinforcement learning based text style transfer without parallel training corpus. *arXiv preprint arXiv:1903.10671*.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *ICML*, pages 1587–1596. JMLR. org.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *NAACL*, pages 110–119.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *ACL*, volume 1, pages 994–1003.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*.

Yi Luan, Chris Brockett, Bill Dolan, Jianfeng Gao, and Michel Galley. 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)*, volume 1, pages 605–614.

Yi Luan, Yangfeng Ji, Hannaneh Hajishirzi, and Boyang Li. 2016. Multiplicative representations for unsupervised semantic role induction. In *ACL*.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

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

Kate G Niederhoffer and James W Pennebaker. 2002b. Sharing one's story: On the benefits of writing or talking about emotional experience. *Handbook of positive psychology*.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Lianhui Qin, Michel Galley, Chris Brockett, Xiaodong Liu, Xiang Gao, Bill Dolan, Yejin Choi, and Jianfeng Gao. 2019. Conversing by reading: Contentful neural conversation with on-demand machine reading. *arXiv preprint arXiv:1906.02738*.

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

Vighnesh Leonardo Shiv, Chris Quirk, Anshuman Suri, Xiang Gao, Khuram Shahid, Nithya Govindarajan, Yizhe Zhang, Jianfeng Gao, Michel Galley, Chris Brockett, et al. 2019. Microsoft icecaps: An open-source toolkit for conversation modeling. In *Proceedings of the 57th Conference of the Association for Computational Linguistics: System Demonstrations*, pages 123–128.

I. Sutskever, O. Vinyals, and Q. Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*.

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.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating informative and diverse conversational responses via adversarial information maximization. In *Advances in Neural Information Processing Systems*, pages 1813–1823.

Yizhe Zhang, Xiang Gao, Sungjin Lee, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2019. Consistent dialogue generation with self-supervised feature learning. *arXiv preprint arXiv:1903.05759*.

Yizhe Zhang, Dinghan Shen, Guoyin Wang, Zhe Gan, Ricardo Henao, and Lawrence Carin. 2017. Deconvolutional paragraph representation learning. In *Advances in Neural Information Processing Systems*, pages 4169–4179.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2018. The design and implementation of xiaoice, an empathetic social chatbot. *arXiv preprint arXiv:1812.08989*.