Structuring Latent Spaces for Stylized Response Generation

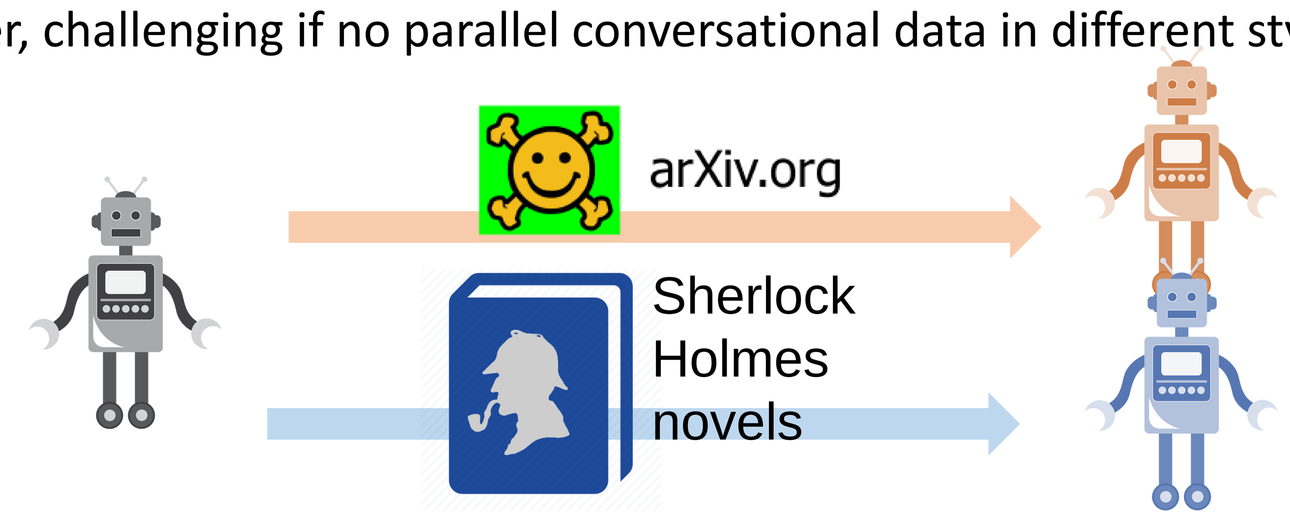
Xiang Gao, Yizhe Zhang, Sungjin Lee* , Michel Galley, Chris Brockett, Jianfeng Gao, Bill Dolan

Microsoft | Research

* Now at Amazon Alexa AI

Motivation & Task

- The master of response style is an important step towards humanlike chatbots.
- However, challenging if no parallel conversational data in different styles

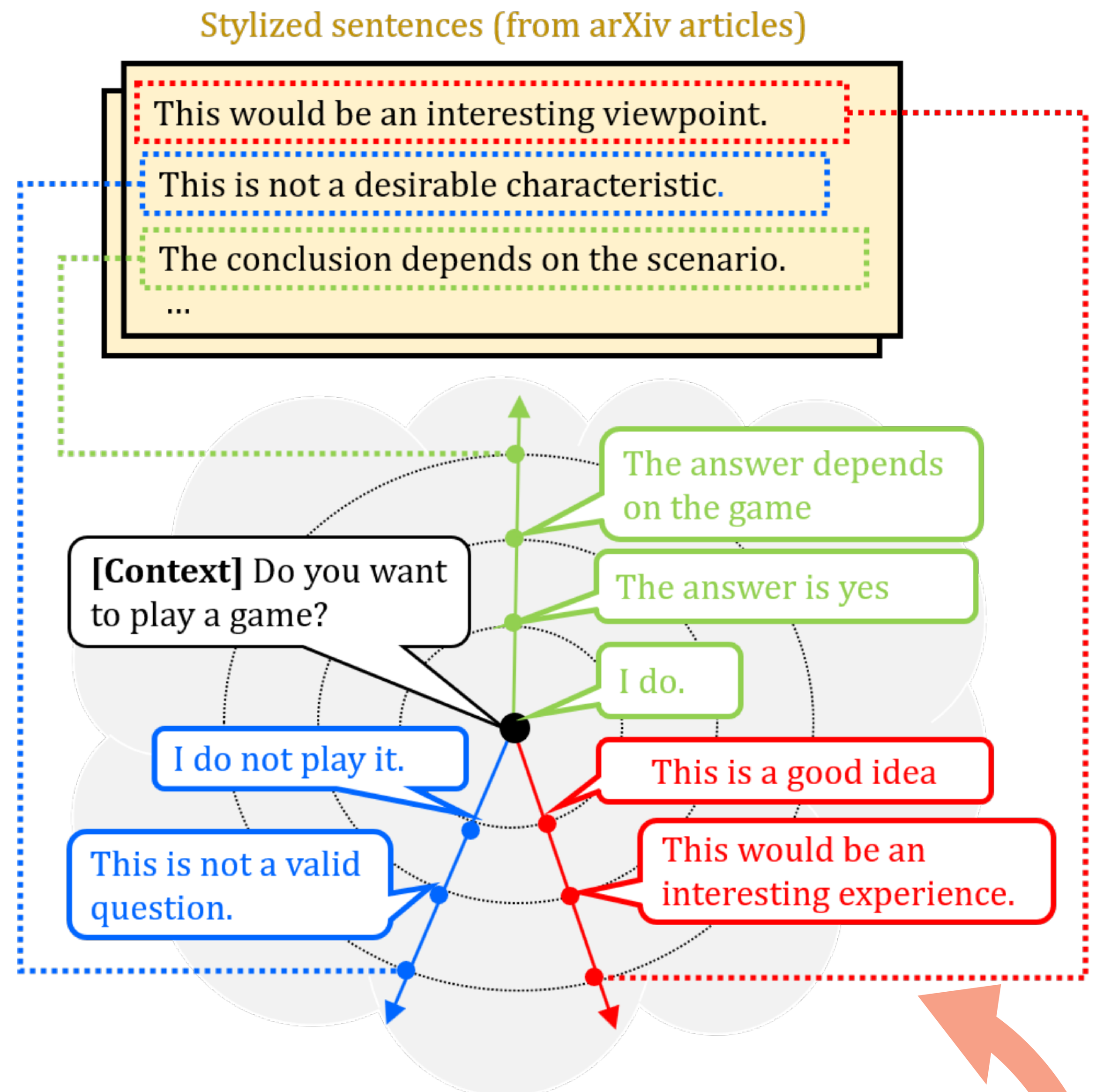


Our task

- Train an agent on conversational dataset *D*conv
- which learns style from non-conversational non-parallel text dataset of target style *D*style
- to generate response in target style and appropriate to context

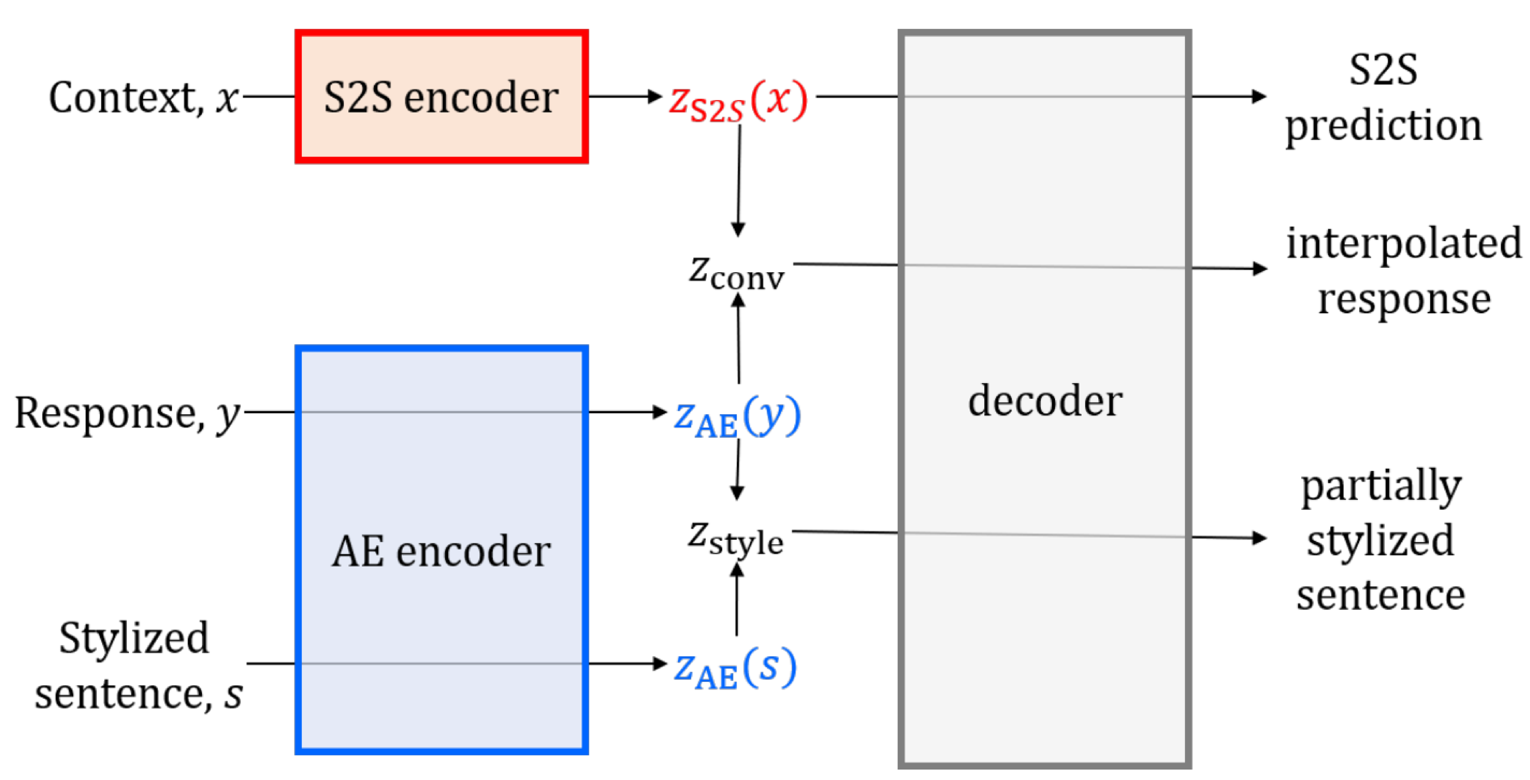
Intuition

Structure a latent space such that the stylized texts are mapped around the related conversation data



Approach

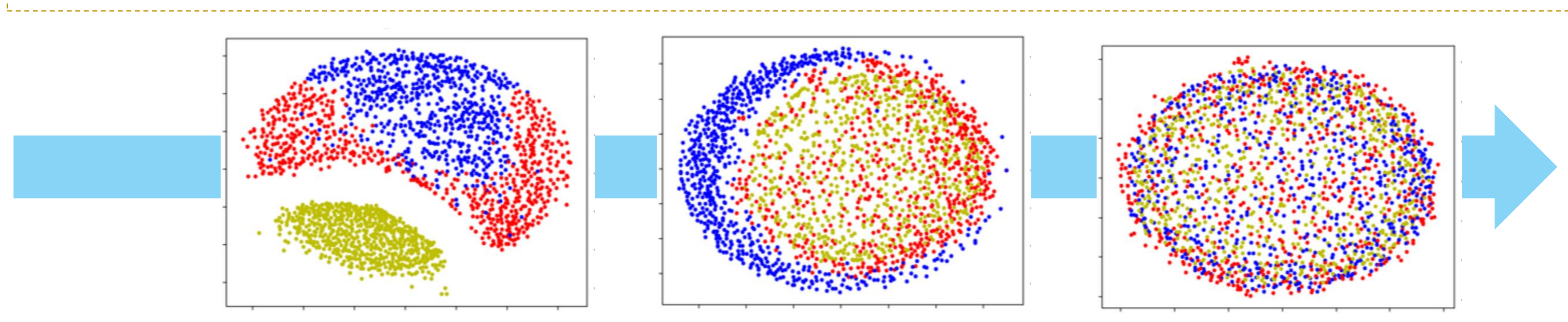
- We propose a regularized multi-task framework to align latent spaces of conversational and stylized datasets



- The loss function considers the generation probability as well as the structure of the latent space, as an extension of our previous SpaceFusion work (Gao et al., NAACL'19)

$$\mathcal{L} = \underbrace{\underbrace{\underbrace{-\frac{1}{|y|}\log p(y|z_{\mathrm{S2S}})}_{\textit{Vanilla S2S}} - \frac{1}{|y|}\log p(y|z_{\mathrm{AE}})}_{\textit{Vanilla MTask (Luan et al., IJCNLP'17)}} + \mathcal{L}_{\text{smooth,conv}} + \mathcal{L}_{\text{fuse,conv}}}_{\textit{SpaceFusion (Gao et al., NAACL'19)}} + \mathcal{L}_{\text{smooth,style}} + \mathcal{L}_{\text{fuse,style}}$$

"StyleFusion"



MDS visualization of the three learned latent spaces

- $z_{S2S}(x)$: prediction from context *x*, $z_{AE}(y)$: representation of the response *y*, $z_{AE}(s)$: representation of the stylized sentences *s*

- *L*smooth encourages smooth semantic interpolation on latent space

$$z_{\text{conv}} = (1-u)z_{\mathrm{AE}}(y) + u z_{\mathrm{S2S}}(x) + \epsilon \qquad z_{\text{style}} = (1-u)z_{\mathrm{AE}}(x) + u z_{\mathrm{AE}}(s) + \epsilon$$

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

- *L*fuse encourages different latent spaces to overlap with each other

$$\mathcal{L}_{\text{fuse,conv}} = d_{\text{conv}} - d_{\text{spread-out}} \qquad \mathcal{L}_{\text{fuse,style}} = d_{\text{style}} - d_{\text{spread-out}}$$

- d_{conv} measures the distance between a pair of $z_{S2S}(x)$ and $z_{AE}(y)$
- d_{style} measures the distance between a $z_{S2S}(x)$ point and its nearest neighbor from $z_{AE}(s)$
- $d_{spread\text{-}out}$ measures the average distance between a point and its nearest neighbor from the same latent space

Inference

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

$$z = z_{\mathrm{S2S}}(x) + r \qquad \rho = |r|/(\sigma\sqrt{l})$$

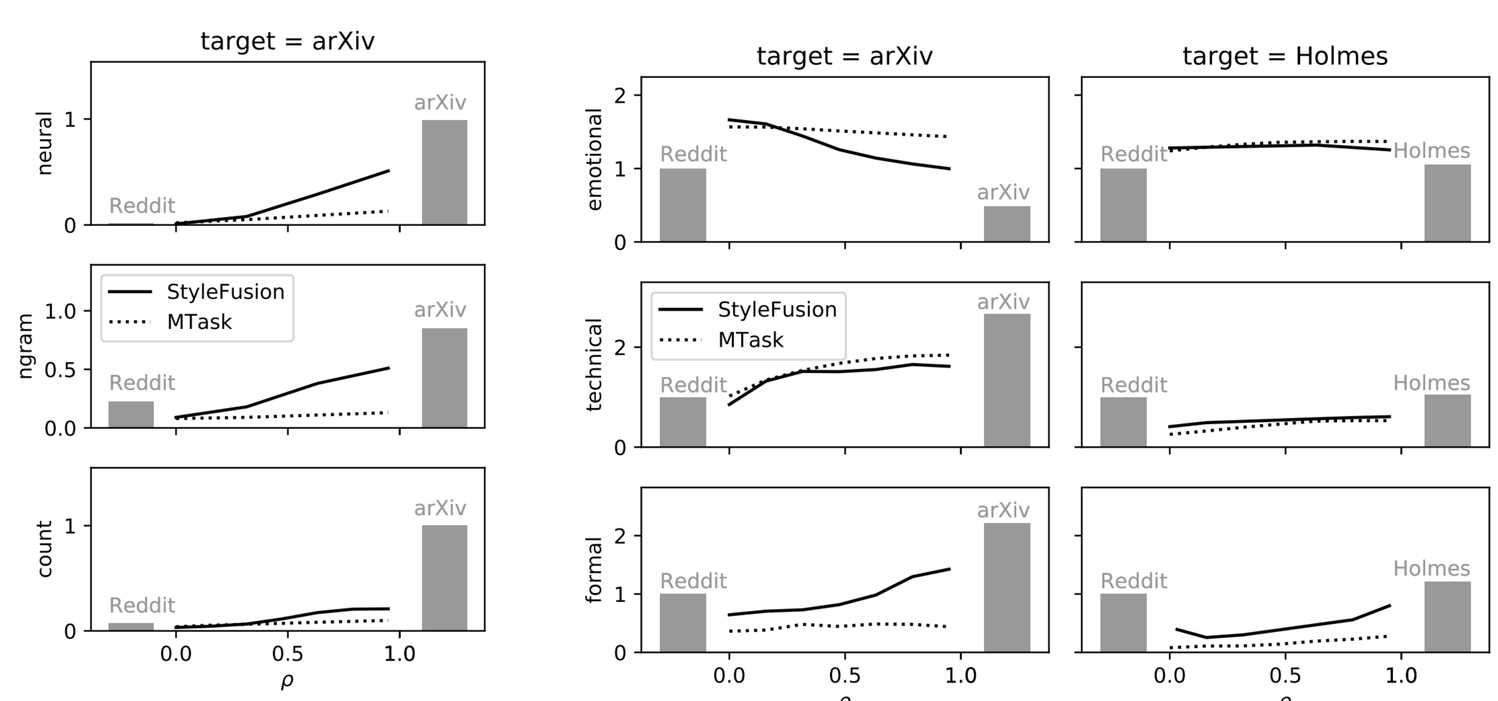
$$\text{score}(h_i) = (1-\lambda)P(h_i|z_{\mathrm{S2S}}(x)) + \lambda P_{\text{style}}(h_i)$$

Results

- Direction controls the content, radius ρ controls style intensity

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.

- As radius ρ increases, fine-granularity styles (emotional, formal, technical) also become similar to the target style



- Human evaluation shows that, our proposed approach, StyleFusion, jointly improve response appropriateness and style intensity, Compared with competitive baselines,

target	model	appropriateness	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

Notes

- S2S+LM refers to the method proposed by Niu and Bansal (2018), which uses the weighted average of a S2S model, trained on conversational dataset, and a LM model, trained on style dataset, as the token probability distribution at inference time.
- MTask refers to the vanilla multi-task learning model proposed in (Luan et al., 2017) trainedon both Dconv and Dstyle.

EMNLP-IJCNLP 2019

Paper: arxiv.org/abs/1909.05361

Code/Data: github.com/golsun/StyleFusion