Revision in Continuous Space: Unsupervised Text Style Transfer without Adversarial Learning Dayiheng Liu¹, Jie Fu², Yidan Zhang¹, Chris Pal², Jiancheng Lv¹ ¹College of Computer Science, Sichuan University ²Mila, IVADO, Polytechnique Montreal

Introduction

Unsupervised Text Style Transfer :

- 1. Converting some attributes of a sentence (e.g., negative sentiment) to other attributes (e.g., positive sentiment)
- 2. Preserving attribute-independent content

Methodology

Core idea:

The proposed model consists of three components: (1) a variational auto-encoder (VAE), (2) attribute predictors, and (3) a content predictor. Predictors takes the continuous representation of a sentence as input and predicts its Bag-of-words content and other attributes. With the gradients obtained from these predictors, we can revise the continuous representation of the original sentence by gradient-based optimization to find a target sentence with the desired fine-grained attributes, and achieve the content-preserving text style transfer.

3. Accessing non-parallel, but style labeled sentences

Previous works: (1) seeking the explicit disentanglement of the content and the attributes. (2) troublesome adversarial learning

This paper:

- Easily Training. The method can be easily trained on the nonparallel dataset, avoiding the problem of training difficulties caused by adversarial learning and achieving higher performance
- Diverse, controllable, and interpretable. Our method revises the original sentence with gradient information for several steps during inference, which explicitly presents the process of the style transfer and can easily provide us multiple results with tuning the gradients. Therefore, the proposed method has higher interpretability and is more controllable
- **Control multiple fine-grained attributes**. Our approach is more generic in the sense that it naturally has the ability to control multiple fine-grained attributes, such as sentence



$$\mathcal{L}_{\text{VAE}}(\theta_{\text{enc}}, \theta_{\text{dec}}) = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{KL}}$$

= $-\mathbb{E}_{q_E(z|x)} \left[\log p_G(x|z)\right] + \mathcal{D}_{\text{KL}}(q_E(z|x)||p(z)),$

length and the existence of specific words

Experiments

🗖 Amaz	zon
Yelp	(sentiment)
Yelp	(Gender)

Metrics: Accuracy PPL Overlap Noun BLEU

Methods	Accuracy↑	PPL↓	Overlap↑	Noun%↑	BLEU ↑
Original	0.1	22.9	100.0	100.0	42.4
Human	91.8	76.9	47.2	78.5	100.0
Delete, Retrieve, & Generate (Li et al. 2018):					
TemplateBased	81.3	<u>183.6</u>	55.6	83.3	28.9
DeleteOnly	85.8	<u>81.4</u>	49.5	74.9	24.7
DeleteAndRetrieve	89.5	<u>96.1</u>	49.4	74.0	24.9
RetrievalOnly	98.4	25.7	<u>15.8</u>	<u>39.6</u>	<u>4.7</u>
StyleEmbedding (Fu et al. 2018)	7.2	93.9	75.4	74.2	31.9
MultiDecoder (Fu et al. 2018)	<u>48.8</u>	<u>166.5</u>	51.5	52.2	23.1
BTS (Prabhumoye et al. 2018)	94.8	32.8	<u>21.5</u>	<u>23.5</u>	<u>6.8</u>
CrossAligned (Shen et al. 2017)	<u>73.6</u>	<u>72.0</u>	41.1	<u>42.9</u>	18.4
Ours (content-strengthen)	88.2	26.5	46.6	77.4	21.8
Ours (style-content balance)	92.3	18.3	38.9	69.3	18.8
Ours (style-strengthen)	95.7	20.6	39.7	61.5	17.9
Methods	Accuracy↑	PPL↓	Overlap↑	Noun%↑	BLEU↑
Methods Original	$\frac{\text{Accuracy}}{23.4}$	PPL↓ 24.4	Overlap↑ 100.0	Noun%↑ 100.0	BLEU↑ 57.2
Methods Original Human	Accuracy↑ <u>23.4</u> 88.1	PPL↓ 24.4 62.9	Overlap↑ 100.0 60.5	Noun%↑ 100.0 85.0	BLEU↑ 57.2 100.0
Methods Original Human Delete, Retrieve, & Generate (Li et al. 2018):	Accuracy↑ <u>23.4</u> 88.1	PPL↓ 24.4 62.9	Overlap↑ 100.0 60.5	Noun%↑ 100.0 85.0	BLEU↑ 57.2 100.0
Methods Original Human Delete, Retrieve, & Generate (Li et al. 2018): TemplateBased	Accuracy↑ <u>23.4</u> 88.1 <u>69.6</u>	PPL↓ 24.4 62.9 <u>108.9</u>	Overlap↑ 100.0 60.5 73.3	Noun%↑ 100.0 85.0 87.9	BLEU↑ 57.2 100.0 42.8
Methods Original Human Delete, Retrieve, & Generate (Li et al. 2018): TemplateBased DeleteOnly	Accuracy \uparrow 23.4 88.1 69.6 51.6	PPL 24.4 62.9 108.9 49.3	Overlap↑ 100.0 60.5 73.3 74.4	Noun%↑ 100.0 85.0 87.9 95.1	BLEU↑ 57.2 100.0 42.8 44.7
MethodsOriginalHumanDelete, Retrieve, & Generate (Li et al. 2018):TemplateBasedDeleteOnlyDeleteAndRetrieve	Accuracy \uparrow 23.4 88.1 69.6 51.6 55.2	$ PPL 24.4 62.9 \frac{108.9}{49.3} 48.2 $	Overlap↑ 100.0 60.5 73.3 74.4 69.1	Noun%↑ 100.0 85.0 87.9 95.1 92.6	BLEU↑ 57.2 100.0 42.8 44.7 41.8
MethodsOriginalHumanDelete, Retrieve, & Generate (Li et al. 2018):TemplateBasedDeleteOnlyDeleteAndRetrieveRetrievalOnly	Accuracy \uparrow 23.4 88.1 69.6 51.6 55.2 87.2	$\begin{array}{r} PPL\downarrow \\ 24.4 \\ 62.9 \\ \hline \\ 108.9 \\ 49.3 \\ 48.2 \\ 28.7 \\ \end{array}$	Overlap↑ 100.0 60.5 73.3 74.4 69.1 21.0	Noun%↑ 100.0 85.0 87.9 95.1 92.6 44.5	BLEU↑ 57.2 100.0 42.8 44.7 41.8 <u>6.7</u>
MethodsOriginalHumanDelete, Retrieve, & Generate (Li et al. 2018):TemplateBasedDeleteOnlyDeleteAndRetrieveRetrievalOnlyStyleEmbedding (Fu et al. 2018)	Accuracy \uparrow 23.4 88.1 69.6 51.6 55.2 87.2 40.5	$\begin{array}{r} PPL \downarrow \\ 24.4 \\ 62.9 \\ \hline \\ 49.3 \\ 48.2 \\ 28.7 \\ \hline \\ 87.7 \\ \hline \end{array}$	Overlap↑ 100.0 60.5 73.3 74.4 69.1 <u>21.0</u> 42.2	Noun%↑ 100.0 85.0 87.9 95.1 92.6 44.5 41.8	BLEU↑ 57.2 100.0 42.8 44.7 41.8 <u>6.7</u> 22.1
MethodsOriginalHumanDelete, Retrieve, & Generate (Li et al. 2018):TemplateBasedDeleteOnlyDeleteAndRetrieveRetrievalOnlyStyleEmbedding (Fu et al. 2018)MultiDecoder (Fu et al. 2018)	Accuracy↑ 23.4 88.1 69.6 51.6 55.2 87.2 40.5 66.5	$\begin{array}{r} PPL \downarrow \\ 24.4 \\ 62.9 \\ \hline \\ 49.3 \\ 48.2 \\ 28.7 \\ \hline \\ 87.7 \\ \hline \\ 80.8 \\ \end{array}$	Overlap↑ 100.0 60.5 73.3 74.4 69.1 21.0 42.2 30.6	Noun%↑ 100.0 85.0 87.9 95.1 92.6 44.5 41.8 30.4	BLEU↑ 57.2 100.0 42.8 44.7 41.8 <u>6.7</u> 22.1 14.3
MethodsOriginalHumanDelete, Retrieve, & Generate (Li et al. 2018):TemplateBasedDeleteOnlyDeleteAndRetrieveRetrievalOnlyStyleEmbedding (Fu et al. 2018)MultiDecoder (Fu et al. 2018)BTS (Prabhumoye et al. 2018)	Accuracy↑ 23.4 88.1 69.6 51.6 55.2 87.2 40.5 66.5 82.6	$\begin{array}{r} PPL \downarrow \\ 24.4 \\ 62.9 \\ \hline \\ 49.3 \\ 48.2 \\ 28.7 \\ \hline \\ 87.7 \\ \underline{80.8} \\ 25.3 \\ \end{array}$	Overlap \uparrow 100.0 60.5 73.3 74.4 69.1 21.0 42.2 30.6 24.7	Noun%↑ 100.0 85.0 87.9 95.1 92.6 44.5 41.8 30.4 22.5	BLEU↑ 57.2 100.0 42.8 44.7 41.8 <u>6.7</u> 22.1 14.3 <u>9.2</u>
MethodsOriginalHumanDelete, Retrieve, & Generate (Li et al. 2018):TemplateBasedDeleteOnlyDeleteAndRetrieveRetrievalOnlyStyleEmbedding (Fu et al. 2018)MultiDecoder (Fu et al. 2018)BTS (Prabhumoye et al. 2018)CrossAligned (Shen et al. 2017)	Accuracy↑ 23.4 88.1 69.6 51.6 55.2 87.2 40.5 66.5 69.6 69.6	$\begin{array}{r} PPL\downarrow\\ 24.4\\ 62.9\\ \hline \\ 49.3\\ 48.2\\ 28.7\\ \hline \\ 87.7\\ \hline \\ 80.8\\ 25.3\\ 18.3\\ \end{array}$	Overlap \uparrow 100.0 60.5 73.3 74.4 69.1 21.0 42.2 30.6 24.7 19.3	Noun%↑ 100.0 85.0 87.9 95.1 92.6 44.5 41.8 30.4 22.5 20.4	$\begin{array}{r} \textbf{BLEU} \\ \hline 57.2 \\ 100.0 \\ \hline 42.8 \\ \textbf{44.7} \\ 41.8 \\ \underline{6.7} \\ 22.1 \\ 14.3 \\ \underline{9.2} \\ \underline{5.0} \\ \end{array}$
MethodsOriginalHumanDelete, Retrieve, & Generate (Li et al. 2018):TemplateBasedDeleteOnlyDeleteAndRetrieveRetrievalOnlyStyleEmbedding (Fu et al. 2018)MultiDecoder (Fu et al. 2018)BTS (Prabhumoye et al. 2018)CrossAligned (Shen et al. 2017)Ours (content-strengthen)	Accuracy↑ 23.4 88.1 69.6 51.6 55.2 87.2 40.5 66.5 82.6 69.6 81.9	$\begin{array}{r} PPL\downarrow\\ 24.4\\ 62.9\\ \hline \\ 49.3\\ 49.3\\ 48.2\\ 28.7\\ \hline \\ 87.7\\ \hline \\ 80.8\\ 25.3\\ \hline \\ 18.3\\ 35.0\\ \hline \end{array}$	Overlap \uparrow 100.0 60.5 73.3 74.4 69.1 21.0 42.2 30.6 24.7 19.3 37.7	Noun%↑ 100.0 85.0 87.9 95.1 92.6 44.5 41.8 30.4 22.5 20.4 76.0	$\begin{array}{r} \textbf{BLEU} \\ \hline 57.2 \\ 100.0 \\ \hline 42.8 \\ \textbf{44.7} \\ 41.8 \\ \underline{6.7} \\ 22.1 \\ 14.3 \\ \underline{9.2} \\ \underline{5.0} \\ \underline{11.5} \end{array}$
MethodsOriginalHumanDelete, Retrieve, & Generate (Li et al. 2018):TemplateBasedDeleteOnlyDeleteAndRetrieveRetrievalOnlyStyleEmbedding (Fu et al. 2018)MultiDecoder (Fu et al. 2018)BTS (Prabhumoye et al. 2018)CrossAligned (Shen et al. 2017)Ours (content-strengthen)Ours (style-content balance)	Accuracy↑ 23.4 88.1 69.6 51.6 55.2 87.2 40.5 66.5 82.6 69.6 81.9 85.1	$\begin{array}{r} PPL \downarrow \\ 24.4 \\ 62.9 \\ \hline \\ 49.3 \\ 48.2 \\ 28.7 \\ \hline \\ 87.7 \\ 80.8 \\ 25.3 \\ 18.3 \\ 35.0 \\ 21.8 \end{array}$	Overlap \uparrow 100.0 60.5 73.3 74.4 69.1 21.0 42.2 30.6 24.7 19.3 37.7 49.3	Noun%↑ 100.0 85.0 87.9 95.1 92.6 44.5 41.8 30.4 22.5 20.4 76.0 49.8	$\begin{array}{r} \textbf{BLEU} \\ \hline 57.2 \\ 100.0 \\ \hline 42.8 \\ \textbf{44.7} \\ 41.8 \\ \underline{6.7} \\ 22.1 \\ 14.3 \\ \underline{9.2} \\ \underline{5.0} \\ \underline{11.5} \\ 21.5 \\ \end{array}$

Content predictor:

$$f_{\text{bow}}(z) = \text{MLP}_{\text{bow}}(z) = p(x_{\text{bow}}|z).$$
$$\log p(x_{\text{bow}}|z) = \log \prod_{t=1}^{|x|} \frac{e^{f_{\text{bow}}^{(x_t)}}}{\sum_{j}^{\mathcal{V}} e^{f_{\text{bow}}^{(x_j)}}}$$
$$\mathcal{L}_{\text{BOW}}(\theta_{\text{bow}}, \theta_{\text{enc}}) = -\mathbb{E}_{q_E(z|x)} \log \left[p(x_{\text{bow}}|z) \right]$$

Attribute predictors:

$$\mathcal{L}_{\text{Attr},s_{j}}(\theta_{s_{j}},\theta_{\text{enc}}) = -\mathbb{E}_{q_{E}(z|x)}\log\left[f_{j}(z)\right]$$
$$\mathcal{L}_{\text{Attr},s_{j}}(\theta_{s_{j}},\theta_{\text{enc}}) = -\mathbb{E}_{q_{E}(z|x)}\log\left[f_{j}(z)\right]$$
$$\mathcal{L}'_{\text{Attr},s_{j}}(\theta_{s_{j}}) = -\mathbb{E}_{p(z)p_{G}(\hat{x}|z)}\log\left[p(\text{CNN}(\hat{x})|z)\right],$$
$$\mathcal{L}'_{\text{Attr},s_{j}}(\theta_{s_{j}}) = \mathbb{E}_{p(z)p_{G}(\hat{x}|z)}\left[(\hat{s}_{j} - f_{j}(z))^{2}\right].$$

Total loss:

$$\mathcal{L} = \mathcal{L}_{\text{VAE}} + \lambda_b \mathcal{L}_{\text{BOW}} + \lambda_s \sum_{i=1}^k \mathcal{L}_{\text{Attr},s_j}$$

Inference:

Our code and data are available at https://github.com/dayihengliu/Fine-Grained-Style-Transfer



