Stylized Dialogue Response Generation Using Stylized Unpaired Texts

Yinhe Zheng, Zikai Chen, Rongsheng Zhang, Shilei Huang, Xiaoxi Mao, Minlie Huang
Tsinghua University
Samsung Research
Fuxi AI Lab, NetEase Inc.

Equal Contribution
Background

- Text style is an interesting phenomenon to model
- Stylized dialogue systems are attractive to users
Motivation

• Most existing stylized dialog models need to train with stylized dialog pairs

\[ \text{Post, Stylized Response} \]

• However, most textual features are embedded in unpaired texts, (e.g. Novels)

“…忽见东首火把照耀，有七八人手执兵器，快步奔来。张杨二人忙在草丛中躲起。那干人奔到邻近…”

Stylized Unpaired Texts
Task Setting

- **Input**

  Dialogue pairs in style $S_0$: $\mathcal{D}_p = \{x_1, y_1\| x_2, y_2\| \ldots x_N, y_N\}$

  Unpaired texts in style $S_1$: $\mathcal{U}_t = \{t_1, t_2, \ldots, t_M\}$

- **Resulting model**

  A dialogue mode that can produce both $S_0$ and $S_1$ Responses
More on Task Setting

• Note that: This task can be tackled using existing unsupervised text style transfer model

Post $\rightarrow$ Dialogue model $\rightarrow$ Response in style $S_0$ $\rightarrow$ Unsupervised text style transfer $\rightarrow$ Response in style $S_1$

• But this may lead to in-consistent responses
Contribution

- A novel method is proposed to build a stylized dialogue model that can capture stylistic features embedded in unpaired texts. Specifically:
  - An inverse dialogue model is introduced to generate stylized pseudo dialogue pairs, which are further utilized in a joint training process.
  - An effective style routing approach is devised to intensify the stylistic features in the decoder.
- Automatic and human evaluations on two datasets show that our method outperforms competitive baselines with a large margin in producing stylized and coherent dialogue responses.
Inverse dialogue model

- Takes in a response, produce a post

Text in Style $S_1$
“大哥，有人来啦”

$\Rightarrow$

Inverse Enc-Dec

Pseudo Post in Style $S_0$
“今天有人来看我么？”

- Use produced pseudo dialogue pairs to train the stylized dialogue model

$D_1$ in style $S_1$
Response $y$

$\Rightarrow$

Inverse Encoder $e$

Inverse Decoder $d$

Decoder $d$

Encoder $e$

$\Rightarrow$

Post $x$

Pseudo posts

Style $S_i$
Style routing approach

- Add style embedding at the end of each attention block

\[
R_{prev} = \text{MMHA}[e_w(y_p), e_w(y_p), e_w(y_p)],
\]

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

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

\[
R_{merge} = R_{avg} + e_s(S_i).
\]
Joint Training

- Train the inverse dialogue model and stylized dialogue model interactively.

Post to response loss

\[ \mathcal{L}_{p2r} = \mathbb{E}_{(x,y) \sim \mathcal{D}_p} - \log p_{\hat{d}}(y | e(x), S_0), \]

Response to post loss

\[ \mathcal{L}_{r2p} = \mathbb{E}_{(x,y) \sim \mathcal{D}_p} - \log p_{\hat{d}}(x | \hat{e}(y)). \]

Inverse dialogue loss

\[ \mathcal{L}_{inv} = \mathbb{E}_{t \sim \mathcal{D}_a, \atop x' \sim p_{\hat{d}}(x | \hat{e}(t))} - \log p_{\hat{d}}(t | e(x'), S_1), \]
Experiments

- Chinese data: Weibo dialogue (S₀), Jinyong novel (S₁)
- English data: Reddits informal dialogue (S₀), Formal English writing (S₁)

<table>
<thead>
<tr>
<th>Model</th>
<th>WDJN Dataset</th>
<th></th>
<th></th>
<th>TCFC Dataset</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU-1,2</td>
<td>Dist. BERT SVM</td>
<td>Flu. Coh. Style</td>
<td>HAvg.</td>
<td>BLEU-1,2</td>
<td>Dist. BERT SVM</td>
</tr>
<tr>
<td>SLM</td>
<td>2.90</td>
<td>0.37</td>
<td>26.6 36.7</td>
<td>0.37</td>
<td>0.37</td>
<td>12.6</td>
</tr>
<tr>
<td>SRL</td>
<td>2.53</td>
<td>0.33</td>
<td>40.4 32.2</td>
<td>3.36</td>
<td>0.39</td>
<td>7.83</td>
</tr>
<tr>
<td>SFusion</td>
<td>3.84</td>
<td>0.20</td>
<td>33.1 8.24</td>
<td>1.63</td>
<td>0.40</td>
<td>5.51</td>
</tr>
<tr>
<td>S2S+BT</td>
<td>6.22</td>
<td>0.68</td>
<td>60.7 66.0</td>
<td>1.39</td>
<td>0.63</td>
<td>12.1</td>
</tr>
<tr>
<td>S2S+CT</td>
<td>11.3</td>
<td>0.62</td>
<td>32.4 7.23</td>
<td>0.45</td>
<td>0.35</td>
<td>8.05</td>
</tr>
<tr>
<td>S2S+PTO</td>
<td>3.57</td>
<td>0.44</td>
<td>32.9 35.1</td>
<td>1.54*</td>
<td>0.35</td>
<td>9.55</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>13.6</strong></td>
<td><strong>1.53</strong></td>
<td><strong>42.8 78.3</strong></td>
<td><strong>1.96</strong></td>
<td><strong>1.00</strong></td>
<td><strong>15.1</strong></td>
</tr>
<tr>
<td>Human</td>
<td>N/A</td>
<td>49.3</td>
<td>80.1 85.4</td>
<td>1.90</td>
<td>1.00</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Respons in Style S₁

<table>
<thead>
<tr>
<th>Model</th>
<th>WDJN Dataset</th>
<th></th>
<th></th>
<th>TCFC Dataset</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU-1,2</td>
<td>Dist. BERT SVM</td>
<td>Flu. Coh. Style</td>
<td>HAvg.</td>
<td>BLEU-1,2</td>
<td>Dist. BERT SVM</td>
</tr>
<tr>
<td>S2S</td>
<td>8.50</td>
<td>2.42</td>
<td>35.1 97.0</td>
<td>1.96*</td>
<td>1.73</td>
<td>6.92</td>
</tr>
<tr>
<td>SFusion</td>
<td>8.65</td>
<td>0.62</td>
<td>35.3 39.9</td>
<td>1.64*</td>
<td>0.74</td>
<td>4.61</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>11.6</strong></td>
<td><strong>2.93</strong></td>
<td><strong>39.0 93.5</strong></td>
<td><strong>1.97</strong></td>
<td><strong>1.85</strong></td>
<td><strong>6.96</strong></td>
</tr>
<tr>
<td>Human</td>
<td>N/A</td>
<td>56.4</td>
<td>97.9 94.4</td>
<td>1.89</td>
<td>1.86</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Respons in Style S₀
Conclusion

• An inverse dialogue model is introduced in our method to produce stylized pseudo dialogue pairs

• Automatic and manual evaluation shows that our method outperforms competitive base-lines in producing coherent and style-intensive responses.

Codes and data are coming soon:
https://github.com/silverriver/Stylized_Dialog
THANKS

2020.12.19