Content preserving text generation with attribute controls

Lajanugen Logeswaran, Honglak Lee, Samy Bengio
University of Michigan, Google Brain

March 19, 2019
Outline

1. Introduction

2. Formulation
   - Formulation
   - Content compatibility
   - Attribute compatibility

3. Experiments
   - Metrics
   - Sentiment Experiments
   - Monolingual Translation
   - Ablative study
   - Simultaneous control of multiple attributes

4. Conclusion
Outline

1. Introduction

2. Formulation
   - Formulation
   - Content compatibility
   - Attribute compatibility

3. Experiments
   - Metrics
   - Sentiment Experiments
   - Monolingual Translation
   - Ablative study
   - Simultaneous control of multiple attributes

4. Conclusion
The style transfer problem which aims to change more abstract properties of an image has seen significant advances. The discrete sequential natural of language makes it difficult to approach language problems in a similar manner. The focus of this work is on the problem of modifying textual attributes in sentences. Create a model that can control multiple attributes of generated text at the same time.
Outline

1. Introduction

2. Formulation
   - Formulation
   - Content compatibility
   - Attribute compatibility

3. Experiments
   - Metrics
   - Sentiment Experiments
   - Monolingual Translation
   - Ablative study
   - Simultaneous control of multiple attributes

4. Conclusion
Formulation

- K attributes of interest \( \{a_1, \ldots, a_K\} \).
- A set of labeled sentences \( D = \{(x^n, l^n)\}_{n=1}^N \) (\( l^n \) is a set of labels for a subset of the attributes)
- Given a sentence \( x \) and attributes \( l' = (l_1, \ldots, l_K) \), the goal is to produce a sentence that shares the content of \( x \), but reflects the attribute values specified by \( l' \).
Formulation

I will go to the airport.

- **mood=indicative, tense=past**
  - I went to the airport.

- **mood=indicative, tense=present**
  - I am going to the airport.

- **mood=subjunctive, tense=conditional**
  - I would go to the airport.
Model Overview

- Generative Model $G = (G_{enc}, G_{dec})$
- $G$ should generate a sentence that is closely related in meaning to the input sentence and consistent with the attributes.
- $G_{enc}: z_x = G_{enc}(x)$
- $G_{dec}: y \sim P_G(\cdot | z_x, l')$
Model Overview
Content compatibility

Two types of reconstruction losses to encourage content compatibility.

- **Autoencoding loss**
  \[
  \mathcal{L}^{ae}(x, l) = -\log P_G(x|z_x, l)
  \]

- **Back-translation loss**
  \[
  \mathcal{L}^{bt}(x, l) = -\log P_G(x|z_y, l)
  \]
Content compatibility

- A common pitfall of the auto-encoding loss in auto-regressive models.
  - Simply copy the input sequence without capturing and informative features.
- A de-nosing formulation is often considered: deleting, swapping or re-arranging words.
  - However, the generated sample $y$ can be mismatched in content from $x$. 
Content compatibility

This paper addresses these issues by interpolating the latent representations of ground truth sentence $x$ and generated sentence $y$.

- merge the autoencoding and back-translation losses by fusing the two latent representations $z_x, z_y$

$$z_{xy} = g \odot z_x + (1 - g) \odot z_y$$

where $g$ is a binary random vector of values sampled from a Bernoulli distribution.

- Interpolated reconstruction loss

$$\mathcal{L}^{int} = E_{(x,l) \sim p_{data}, y \sim p_G(\cdot|z_x,l')}[-\log p_G(x|z_{xy},l)]$$
Attribute compatibility

- Adversarial loss

\[ \mathcal{L}^{adv} \min_G \max_D \mathbb{E}[\log D(h_x, l) + \log(1 - D(h_y, l'))] \]

It is possible that the discriminator ignores the attributes and makes the real/fake decision based on just the hidden states, or vice versa.

- To prevent this situation, a new objective is proposed

\[ \mathcal{L}^{adv} \min_G \max_D \mathbb{E}[2 \log D(h_x, l) + \log(1 - D(h_y, l')) + \log(1 - D(h_x, l'))] \]

- The overall loss function

\[ \mathcal{L}^{int} + \mathcal{L}^{adv} \]
## Outline

1. Introduction

2. Formulation
   - Formulation
   - Content compatibility
   - Attribute compatibility

3. Experiments
   - Metrics
   - Sentiment Experiments
   - Monolingual Translation
   - Ablative study
   - Simultaneous control of multiple attributes

4. Conclusion
Metrics

• Attribute accuracy
  A pre-trained sentiment classifier

• Content compatibility

\[ f_{\text{content}}(M, M') = 0.5 \left[ \mathbb{E}_{x \sim D_{\text{src}}} \text{BLEU}(x, M' \circ M(x)) + \mathbb{E}_{x \sim D_{\text{tgt}}} \text{BLEU}(x, M \circ M'(x)) \right] \]

where \( M \circ M'(x) \) represents translating \( x \in D_{\text{src}} \) to domain \( D_{\text{tgt}} \) and then back to \( D_{\text{src}} \).

• Fluency
  A pre-trained language model
Sentiment Experiments

- Data
  Restaurant reviews dataset (447k/128k) & IMDB review corpus (128k/36k)
## Sentiment Experiments

- **Quantitative evaluation**

<table>
<thead>
<tr>
<th>Model</th>
<th>Yelp Reviews</th>
<th>IMDB Reviews</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Attribute</td>
<td>Content</td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
<td>B-1</td>
</tr>
<tr>
<td>Ctrl-gen [18]</td>
<td>76.36%</td>
<td>11.5</td>
</tr>
<tr>
<td>Cross-align [22]</td>
<td>90.09%</td>
<td>41.9</td>
</tr>
<tr>
<td>Ours</td>
<td>90.50%</td>
<td>53.0</td>
</tr>
</tbody>
</table>
### Sentiment Experiments

- **Qualitative evaluation**

<table>
<thead>
<tr>
<th>Restaurant reviews</th>
<th>negative → positive</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Query</strong></td>
<td>the people behind the counter were not friendly whatsoever.</td>
</tr>
<tr>
<td>Ctrl gen [18]</td>
<td>the food didn’t taste as fresh as it could have been either.</td>
</tr>
<tr>
<td>Cross-align [22]</td>
<td>the owners are the staff is so friendly.</td>
</tr>
<tr>
<td>Ours</td>
<td>the people at the counter were very friendly and helpful.</td>
</tr>
<tr>
<td></td>
<td>positive → negative</td>
</tr>
<tr>
<td><strong>Query</strong></td>
<td>they do an exceptional job here, the entire staff is professional and accommodating!</td>
</tr>
<tr>
<td>Ctrl gen [18]</td>
<td>very little water just boring ruined!</td>
</tr>
<tr>
<td>Cross-align [22]</td>
<td>they do not be back here, the service is so rude and do n’t care!</td>
</tr>
<tr>
<td>Ours</td>
<td>they do not care about customer service, the staff is rude and unprofessional!</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Movie reviews</th>
<th>negative → positive</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Query</strong></td>
<td>once again, in this short, there isn’t much plot.</td>
</tr>
<tr>
<td>Ctrl gen [18]</td>
<td>it’s perfectly executed with some idiotically amazing directing.</td>
</tr>
<tr>
<td>Cross-align [22]</td>
<td>but &lt;unk&gt;, the film is so good, it is.</td>
</tr>
<tr>
<td>Ours</td>
<td>first off, in this film, there is nothing more interesting.</td>
</tr>
<tr>
<td></td>
<td>positive → negative</td>
</tr>
<tr>
<td><strong>Query</strong></td>
<td>that’s another interesting aspect about the film.</td>
</tr>
<tr>
<td>Ctrl gen [18]</td>
<td>peter was an ordinary guy and had problems we all could &lt;unk&gt; with</td>
</tr>
<tr>
<td>Cross-align [22]</td>
<td>it’s the &lt;unk&gt; and the plot.</td>
</tr>
<tr>
<td>Ours</td>
<td>there’s no redeeming qualities about the film.</td>
</tr>
</tbody>
</table>
Sentiment Experiments

- Human evaluation

<table>
<thead>
<tr>
<th>Model</th>
<th>Attribute</th>
<th>Content</th>
<th>Fluency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ctrl-gen [18]</td>
<td>66.0%</td>
<td>6.94%</td>
<td>2.51</td>
</tr>
<tr>
<td>Cross-align [22]</td>
<td>91.2%</td>
<td>22.04%</td>
<td>2.54</td>
</tr>
<tr>
<td>Ours</td>
<td>92.8%</td>
<td>55.10%</td>
<td>3.19</td>
</tr>
</tbody>
</table>
Monolingual Translation

- Use a dataset of Shakespeare plays
- 17k pairs for training and 2k,1k pairs respectively for development and test. All remaining 80k are considered unpaired.
- Train the model using supervised learning and fine-tune on the unpaired data using the proposed objective.
The results show that the model is capable of finding sentence alignments by exploiting the unlabeled data.

<table>
<thead>
<tr>
<th>Supervision</th>
<th>Model</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paired data</td>
<td>Seq2seq</td>
<td>10.4</td>
</tr>
<tr>
<td></td>
<td>Seq2seq-bi</td>
<td>11.15</td>
</tr>
<tr>
<td>Unpaired data</td>
<td>Ours</td>
<td>7.65</td>
</tr>
<tr>
<td>Paired + Unpaired data</td>
<td>Ours</td>
<td>13.89</td>
</tr>
</tbody>
</table>
Ablative study

Content preserving text generation with attribute controls
Simultaneous control of multiple attributes

<table>
<thead>
<tr>
<th>Mood</th>
<th>Tense</th>
<th>Voice</th>
<th>Neg.</th>
<th>john was born in the camp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indicative</td>
<td>Past</td>
<td>Passive</td>
<td>No</td>
<td>john was born in the camp</td>
</tr>
<tr>
<td>Indicative</td>
<td>Past</td>
<td>Passive</td>
<td>Yes</td>
<td>john wasn’t born in the camp</td>
</tr>
<tr>
<td>Indicative</td>
<td>Past</td>
<td>Active</td>
<td>No</td>
<td>john had lived in the camp</td>
</tr>
<tr>
<td>Indicative</td>
<td>Past</td>
<td>Active</td>
<td>Yes</td>
<td>john didn’t live in the camp</td>
</tr>
<tr>
<td>Indicative</td>
<td>Present</td>
<td>Passive</td>
<td>No</td>
<td>john is born in the camp</td>
</tr>
<tr>
<td>Indicative</td>
<td>Present</td>
<td>Passive</td>
<td>Yes</td>
<td>john isn’t born in the camp</td>
</tr>
<tr>
<td>Indicative</td>
<td>Present</td>
<td>Active</td>
<td>No</td>
<td>john has lived in the camp</td>
</tr>
<tr>
<td>Indicative</td>
<td>Present</td>
<td>Active</td>
<td>Yes</td>
<td>john doesn’t live in the camp</td>
</tr>
<tr>
<td>Indicative</td>
<td>Future</td>
<td>Passive</td>
<td>No</td>
<td>john will be born in the camp</td>
</tr>
<tr>
<td>Indicative</td>
<td>Future</td>
<td>Passive</td>
<td>Yes</td>
<td>john will not be born in the camp</td>
</tr>
<tr>
<td>Indicative</td>
<td>Future</td>
<td>Active</td>
<td>No</td>
<td>john will live in the camp</td>
</tr>
<tr>
<td>Indicative</td>
<td>Future</td>
<td>Active</td>
<td>Yes</td>
<td>john will not survive in the camp</td>
</tr>
<tr>
<td>Subjunctive</td>
<td>Cond</td>
<td>Passive</td>
<td>No</td>
<td>john could be born in the camp</td>
</tr>
<tr>
<td>Subjunctive</td>
<td>Cond</td>
<td>Passive</td>
<td>Yes</td>
<td>john couldn’t live in the camp</td>
</tr>
<tr>
<td>Subjunctive</td>
<td>Cond</td>
<td>Active</td>
<td>No</td>
<td>john could live in the camp</td>
</tr>
<tr>
<td>Subjunctive</td>
<td>Cond</td>
<td>Active</td>
<td>Yes</td>
<td>john couldn’t live in the camp</td>
</tr>
</tbody>
</table>
Outline

1. Introduction

2. Formulation
   - Formulation
   - Content compatibility
   - Attribute compatibility

3. Experiments
   - Metrics
   - Sentiment Experiments
   - Monolingual Translation
   - Ablative study
   - Simultaneous control of multiple attributes

4. Conclusion
Conclusion

- Back-translation is useful for attribute control of discrete data.
- The proposed model can easily extend to the multiple attribute scenario.
- It would be interesting future work to consider attribute with continues values and a much larger set of semantic and syntactic attributes.