Mask & Infill

Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han and Songlin Hu

IJCAI 2019

Presented by: Egill Ian Gudmundsson
Outline

- Problem definition
- Previous efforts
- Unsolved difficulties
- Mask and Infill method
- Results
- Discussion
Problem Definition

- Style transfer and sentiment transfer usually grouped together
- Take a sentence, change the sentiment (the attribute words) while preserving content
- Sentiment can be, e.g. positive and negative

**Positive:** Parasite was an incredible film, showing societal issues in a riveting and entertaining manner.

**Negative:** Parasite was an uninspiring film, showing societal issues in a cliché and trite manner.
Problem Definition

- Corpora $D$ only consists of a sentence and its sentiment label $(x, s)$
- We do not have the same sentence with a different sentiment $(x, s')$
- Need to implement unsupervised learning since there’s no parallel data
- Reconstruction loss is our friend
Previous work

- Delete, Retrieve, Generate (2018), J.Li et al
  - Used LSTMs to achieve better results
  - Picks out the attribute words using saliency scores and deletes them
  - Gets a similar sentence from the corpus (using TF-IDF)
  - Generates the output sentence by feeding the incomplete sentence and the similar sentence into a neural network
- The saliency method is also called the frequency-ratio method
Previous work

- Delete, Retrieve, Generate (2018), J.Li et al

Parasite was an incredible film, showing societal issues in a riveting and entertaining manner.

It was an uninspiring film, both cliché and trite

Parasite was an film, showing societal issues in a and manner.

Parasite was an uninspiring film, showing societal issues in a cliché and trite manner.
Previous work

- **Delete, Retrieve, Generate with Transformers (2019), Akhilesh Sudhakar et al**
  - Swapped the generative network with a Guided Generative Style Transformer and a Blind Generative Style Transformer
  - Delete Transformer added in stead of pure saliency score
  - Still uses the 3 steps in DRG
Unsolved difficulties

- Longer sentences still tricky
- May be hard to adapt for multi-sentiment transfer
- Trade-off between content and attributes
Mask & Infill

- Find attribute markers in sentence and mask them
- Generate new attribute markers in the sentence where masked
Masking

- **Attribute** words identified with a mix of frequency-ratio and attention methods
- Frequency-ratio counts how often an n-gram appears in a sentence with a sentiment

\[ s_c(u, a) = \frac{\text{count}(u, D_a) + \lambda}{\left( \sum_{a' \in A, a' \neq a} \text{count}(u, D_{a'}) \right) + \lambda} \]  

- If the score is above threshold, marked as attribute marker
- Relies on good corpus, fails if corpus quality is subpar
Masking

- Attribute words identified with a mix of frequency-ratio and attention methods
- Frequency-ratio counts how often an n-gram appears in a sentence with a sentiment

\[ s_c(u, a) = \frac{\text{count}(u, D_a) + \lambda}{\left( \sum_{a' \in A, a' \neq a} \text{count}(u, D_{a'}) \right) + \lambda} \]  

- If the score is above threshold, marked as attribute marker
- Relies on good corpus, fails if corpus quality is subpar
Masking

- Attention method uses bidirectional LSTM trained to extract the extent to which words contribute to sentiment
- If the attention on a word is more than the average in a sentence, it is marked as attribute

\[
\begin{align*}
H &= (h_1, h_2, \cdots, h_N) \quad (2) \\
\mathbf{a} &= \text{softmax}(\mathbf{w} \cdot \tanh(\mathbf{W}H^T)) \quad (3) \\
\mathbf{c} &= \mathbf{a} \cdot \mathbf{H} \quad (4) \\
y &= \text{softmax}(\mathbf{W}' \cdot \mathbf{c}) \quad (5)
\end{align*}
\]

- Can fail if too many attribute words vying for attention
Masking

- Using a combination of both attention and frequency make up for each other’s faults
- Use frequency for saliency score and attention for likelihood of attribute word being selected
  \[ s(u, a) = s_c(u, a) \times p \]  
  (6)
- If the score is above a certain threshold, mark as attribute
- If frequency method fails (too few or too many words) => default to attention method
Infill

- Don’t have parallel sentences => need to use reconstruction loss
  \[
  \mathcal{L}_{rec} = - \sum_{a \in A, t_i \in M} \log p(t_i | \bar{S}, a) \tag{7}
  \]

- Discriminator is used to classify sentence attribute
  \[
  \mathcal{L}_{acc} = -\log p(\hat{a} | \hat{S}) \tag{9}
  \]

- Combination of the two used for loss
  \[
  \min_{\theta} \mathcal{L} = \mathcal{L}_{rec} + \eta \mathcal{L}_{acc} \tag{10}
  \]
Infill

- Uses Attribute Conditional Masked Language Model (AC-MLM) for language modelling
- Uses soft-sampling to back-propagate gradients by using approximations of the sampled word vector

\[ \hat{t}_i \sim \text{softmax}(o_t/\tau) \] (11)
Algorithm 1 Implementation of “Mask and Infill” approach.

1: Pre-train attention-based classifier $C_l$s (Eq.2-5)
2: Construct attribute marker vocabulary $\mathcal{V}$ (Eq.1.6)
3: for every sentence $S$ in $\mathcal{D}$ do
4: Mask attribute markers within $S$ by looking up $\mathcal{V}$, getting $\overline{S}$
5: if $\overline{S}$ is too short or $\overline{S}$ is the same as $S$ then
6: Re-mask with attention weights calculated by $C_l$s (Eq.3)
7: end if
8: end for
9: for each iteration $i=1,2,...,M$ do
10: Sample a masked sentence $\overline{S}$ with attribute $a$
11: Reconstruct $S$ with $\overline{S}$ and $a$, calculating $\mathcal{L}_{rec}$ based (Eq.7)
12: $\hat{a}$ = the target attribute
13: Construct $\hat{S}$ (Eq.8)
14: calculating $\mathcal{L}_{acc}$ (Eq.9)
15: Update model parameters $\theta$
16: end for
The Boring Stuff

- Model uses BERT\textsubscript{base} as a language model with original parameters. 12 layers, 12 attention heads, 110 million parameters, 512 token max input
- Segment embedding layer replaced with attribute embedding layer
- Pre-trained discriminator is CNN-based classifier using Word-Piece embeddings
- Balancing parameter and temperature in (10) and (11) respectively are selected via grid-search
- BERT is fine-tuned as AC-MLM for 10 epochs and then 6 epochs are trained with discriminator constraint applied
## Results

<table>
<thead>
<tr>
<th></th>
<th>YELP</th>
<th></th>
<th></th>
<th>AMAZON</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Gra</td>
<td>Con</td>
<td>Att</td>
<td>Gra</td>
<td>Con</td>
<td>Att</td>
</tr>
<tr>
<td>DeleteAndRetrieval</td>
<td>3.4</td>
<td>3.5</td>
<td>3.6</td>
<td>3.5</td>
<td>3.2</td>
<td>3.3</td>
</tr>
<tr>
<td>w/frequency-ratio</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AC-MLM-SS</td>
<td>3.9</td>
<td>3.2</td>
<td>4.2</td>
<td>3.8</td>
<td>3.6</td>
<td>3.7</td>
</tr>
<tr>
<td>w/attention-based</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AC-MLM-SS</td>
<td>4.0</td>
<td>3.8</td>
<td><strong>4.4</strong></td>
<td>3.9</td>
<td>3.7</td>
<td>3.7</td>
</tr>
<tr>
<td>w/fusion-method</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AC-MLM-SS</td>
<td><strong>4.2</strong></td>
<td><strong>4.0</strong></td>
<td><strong>4.4</strong></td>
<td><strong>4.1</strong></td>
<td><strong>4.0</strong></td>
<td><strong>4.0</strong></td>
</tr>
</tbody>
</table>

Table 5: Human evaluation results on two datasets. We show average human ratings for grammaticality (Gra), content preservation (Con), target attribute match (Att).

<table>
<thead>
<tr>
<th></th>
<th>YELP</th>
<th></th>
<th></th>
<th>AMAZON</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ACC (%)</td>
<td>BLEU</td>
<td></td>
<td>ACC (%)</td>
<td>BLEU</td>
<td></td>
</tr>
<tr>
<td>CrossAligned</td>
<td>73.7</td>
<td>3.1</td>
<td></td>
<td>74.1</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>StyleEmbedding</td>
<td>8.7</td>
<td>11.8</td>
<td></td>
<td>43.3</td>
<td>10.0</td>
<td></td>
</tr>
<tr>
<td>MultiDecoder</td>
<td>47.6</td>
<td>7.1</td>
<td></td>
<td>68.3</td>
<td>5.0</td>
<td></td>
</tr>
<tr>
<td>CycledReinforce</td>
<td>85.2</td>
<td>9.9</td>
<td></td>
<td>77.3</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>TemplateBased</td>
<td>81.7</td>
<td>11.8</td>
<td></td>
<td>68.7</td>
<td>27.1</td>
<td></td>
</tr>
<tr>
<td>RetrievalOnly</td>
<td>95.4</td>
<td>0.4</td>
<td></td>
<td>70.3</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>DeleteOnly</td>
<td>85.7</td>
<td>7.5</td>
<td></td>
<td>45.6</td>
<td>24.6</td>
<td></td>
</tr>
<tr>
<td>DeleteAndRetrieval</td>
<td>88.7</td>
<td>8.4</td>
<td></td>
<td>48.0</td>
<td>22.8</td>
<td></td>
</tr>
<tr>
<td>w/frequency-ratio</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AC-MLM</td>
<td>55.0</td>
<td>12.7</td>
<td></td>
<td>28.7</td>
<td>31.0</td>
<td></td>
</tr>
<tr>
<td>AC-MLM-SS</td>
<td>90.5</td>
<td>11.6</td>
<td></td>
<td>75.7</td>
<td>26.0</td>
<td></td>
</tr>
<tr>
<td>w/attention-based</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AC-MLM</td>
<td>41.5</td>
<td></td>
<td><strong>15.9</strong></td>
<td>31.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AC-MLM-SS</td>
<td>97.3</td>
<td>14.1</td>
<td></td>
<td>75.9</td>
<td>28.5</td>
<td></td>
</tr>
<tr>
<td>w/fusion-method</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AC-MLM</td>
<td>43.5</td>
<td>15.3</td>
<td></td>
<td>42.9</td>
<td>30.7</td>
<td></td>
</tr>
<tr>
<td>AC-MLM-SS</td>
<td><strong>97.3</strong></td>
<td>14.4</td>
<td>84.5</td>
<td>28.5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Automatic evaluation performed by tools from [Li et al., 2018]³. “ACC” indicates accuracy, “BLEU” measures content similarity between the outputs and the human references. “AC-MLM”, represents attribute conditional masked language model. “w/” represents “with”.”-SS” represents with soft-sampling.
Discussion

- Hasn’t been tested with multi-sentiment transfer
- Still hasn’t fully tackled the content-attribute trade-off

\[ \min_\theta \mathcal{L} = \mathcal{L}_{rec} + \eta \mathcal{L}_{acc} \]  \hspace{1cm} (10)

Figure 3: The trend of BLEU with the increase of accuracy.
Discussion

- Can it be used for *style* transfer?

Shakespeare

There are more things in heaven and earth, Horatio, than are dreamt of in your philosophy

? 

Lt. Com. Data

Horatio, there are many items that are left unaccounted for in your philosophy
References


