Style Transfer Through Back-Translation

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, Alan W Black
What is Style Transfer

- Rephrasing the text to contain specific stylistic properties without changing the intent or affect within the context.
What is Style Transfer

- Rephrasing the text to contain specific stylistic properties without changing the intent or affect within the context.

  “Shut up! the video is starting!”

  “Please be quiet, the video will begin shortly.”
Motivation

I have an exam today.

May the Force be with you!

Best of Luck!
User Adaptation

User: I am frustrated with work. My models are not working!

Siri: No! Try not! Do or do not, there is no try!
I am frustrated with work. My models are not working!

Have the courage to follow your heart and intuition. They somehow know what you truly want to become.
Applications
Anonymization

- To preserve anonymity of users online, for personal security concerns (Jardine, 2016), or to reduce stereotype threat (Spencer et al., 1999).
Balanced Data

- Demographically-balanced training data for downstream applications.
Our Goal

To create a representation that is devoid of style but holds the meaning of the input sentence.
Challenges

Content

Style
Challenges

- No Parallel Data!
  - “The movie was very long.”
  - “I entered the theatre in the bloom of youth and emerged with a family of field mice living in my long, white mustache.”
- Hard to detect style
Our Solution

- Back-Translation
  - Translating an English sentence to a pivot language and then back to English.
- Reduces the stylistic properties
- Helps in grounding meaning
Overview

How it works?

How to train?

Evaluation
Architecture

MT e → f

encoder decoder
I thank you, Rep. Visclosky

MT $e \rightarrow f$

encoder decoder

je vous remercie, Rep. Visclosky
I thank you, Rep. Visclosky

MT e → f
encoder decoder

je vous remercie, Rep. Visclosky

MT f → e
encoder decoder
I thank you, Rep. Visclosky

MT $e \rightarrow f$

encoder decoder

je vous remercie, Rep. Visclosky

MT $f \rightarrow e$

coder

Architecture
I thank you, Rep. Visclosky.

MT $e \rightarrow f$

encoder  decoder

je vous remercie, Rep. Visclosky

MT $f \rightarrow e$

encoder

$z$

Style 1

decoder

I thank you, senator Visclosky

Style 2

decoder

I'm praying for you sir.

Carnegie Mellon University
Language Technologies Institute
Overview

How it works?

How to train?

Evaluation
Train Pipeline

- Style 1 decoder
  - $\hat{x}_{style1}$
- Style 2 decoder
  - $\hat{x}_{style2}$
Train Pipeline

- Style 1 decoder
- Style 2 decoder
- \( \hat{x}_{style1} \)
- \( \hat{x}_{style2} \)
- classifier
Experimental Settings

- Encoder-Decoders follow sequence-to-sequence framework (Sutskever et al., 2014; Bahdanau et al., 2015)

\[
\min_{\theta_{gen}} \mathcal{L}_{gen} = \mathcal{L}_{recon} + \lambda_c \mathcal{L}_{class}
\]
Figure 2: Cross-aligning between \( x_1 \) and transferred \( x_2 \). For \( x_1 \), \( G \) is teacher-forced by its words \( w_1 \)\( w_2 \)\( \cdots \)\( w_t \). For transferred \( x_2 \), \( G \) is self-fed by previous output logits. The sequence of hidden states \( h^0, \cdots, h^t \) and \( \tilde{h}^0, \cdots, \tilde{h}^t \) are passed to discriminator \( D_1 \) to be aligned. Note that our first variant aligned auto-encoder is a special case of this, where only \( h^0 \) and \( \tilde{h}^0 \), i.e. \( z_1 \) and \( z_2 \), are aligned.
Neural Machine Translation

- WMT 15 data
  - News, Europarl and Common Crawl
  - ~5M parallel English - French sentences

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>English - French</td>
<td>32.52</td>
</tr>
<tr>
<td>French - English</td>
<td>31.11</td>
</tr>
</tbody>
</table>
## Style Tasks

<table>
<thead>
<tr>
<th>Task</th>
<th>Labels</th>
<th>Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Male, Female</td>
<td>Yelp</td>
</tr>
<tr>
<td>Political Slant</td>
<td>Republican, Democratic</td>
<td>Facebook Comments</td>
</tr>
<tr>
<td>Sentiment Modification</td>
<td>Negative, Positive</td>
<td>Yelp</td>
</tr>
</tbody>
</table>
Overview

How it works?

How to train?

Evaluation
Evaluation

- Style Transfer Accuracy
- Meaning Preservation
- Fluency
Style Transfer Accuracy

- generated sentences are evaluated using a pre-trained style classifier
- Transfer the style of test sentences and test the classification accuracy of the generated sentences for the desired label.

<table>
<thead>
<tr>
<th>Classifier Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>82%</td>
</tr>
<tr>
<td>Political Slant</td>
<td>92%</td>
</tr>
<tr>
<td>Sentiment Modification</td>
<td>93.23%</td>
</tr>
</tbody>
</table>
Style Transfer Accuracy

Accuracy

<table>
<thead>
<tr>
<th>Category</th>
<th>Baseline</th>
<th>BST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>60.4</td>
<td>57.04</td>
</tr>
<tr>
<td>Political Slant</td>
<td>75.82</td>
<td>88.01</td>
</tr>
<tr>
<td>Sentiment Modification</td>
<td>80.43</td>
<td>87.22</td>
</tr>
</tbody>
</table>

Carnegie Mellon University
Language Technologies Institute
Preservation of Meaning

- Human Annotation: A/B Testing
- The annotators are given instructions.
- Annotators are presented with the *original* sentence.
Instructions

- “Which transferred sentence maintains the same sentiment of the source sentence in the same semantic context (i.e. you can ignore if food items are changed)”
- “Which transferred sentence maintains the same semantic intent of the source sentence while changing the political position”
- “Which transferred sentence is semantically equivalent to the source sentence with an opposite sentiment”
Preservation of Meaning

![Bar Chart]

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>BST</th>
<th>No Preference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>15.23</td>
<td>14.55</td>
<td>43.41</td>
</tr>
<tr>
<td>Political Slant</td>
<td>35.91</td>
<td>23.18</td>
<td>39.55</td>
</tr>
<tr>
<td>Sentiment</td>
<td>40.91</td>
<td>41.36</td>
<td>45.9</td>
</tr>
</tbody>
</table>
Fluency

- Human annotators were asked to annotate the generated sentences for fluency on a scale of 1-4.
- 1: Unreadable
- 4: Perfect
Discussion

- The loss function of the generators includes two competing terms, one to improve meaning preservation and the other to improve the style transfer accuracy.
- Sentiment modification task is not well-suited for evaluating style transfer.
- The style-transfer accuracy for gender is lower for BST model but the preservation of meaning is much better for the BST model, compared to CAE model and to “No preference” option.
Gender Examples

- Male -- Female

  my wife ordered country fried steak and eggs.

  My husband ordered the chicken salad and the fries.

- Female -- Male

  Save yourselves the huge headaches,

  You are going to be disappointed.
Political Slant Examples

- Republican -- Democratic

  *I will continue praying for you and the decisions made by our government!*

  *I will continue to fight for you and the rest of our democracy!*

- Democratic -- Republican

  *As a hoosier, I thank you, Rep. Vislosky.*

  *As a hoosier, I’m praying for you sir.*
Sentiment Modification Examples

- Negative -- Positive
  
  *This place is bad news!*

  *This place is amazing!*

- Positive -- Negative
  
  *The food is excellent and the service is exceptional!*

  *The food is horrible and the service is terrible.*
Future Directions

- Enhance back-translation by pivot through several languages
  - to learn a better grounded latent meaning representation.
- Use multiple target languages with single source language as described in (Johnson et al., 2016) to see whether pivoting via multiple languages captures better semantic representations.
Future Directions

- Deploy the system in a real world conversational agent to analyze the effect on user satisfaction
- Caring for more styles!
Thank You
References

References
