Style Transfer Through Back-Translation

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, Alan W Black

Carnegie Mellon University
Language Technologies Institute

What is Style Transfer

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

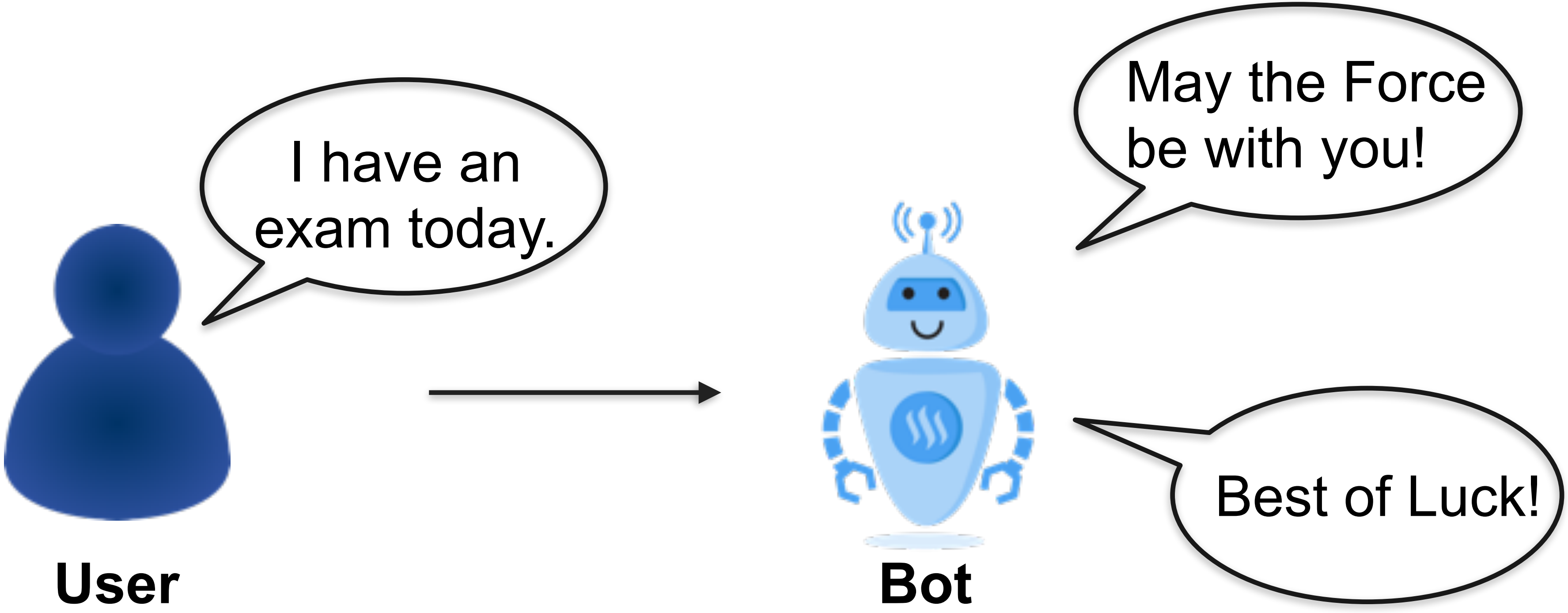
What is Style Transfer

- 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
Bot

Applications

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

- 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.

Prior Work

- (Hu et al., 2017) - VAE with classifier feedback
- (Shen et al., 2017) - Cross aligned auto encoder with two discriminators
- (Li et al., 2018) - delete, retrieve and generate
- (Fu et al., 2018) - multiple decoders and style embeddings

Toward Controlled Generation of Text

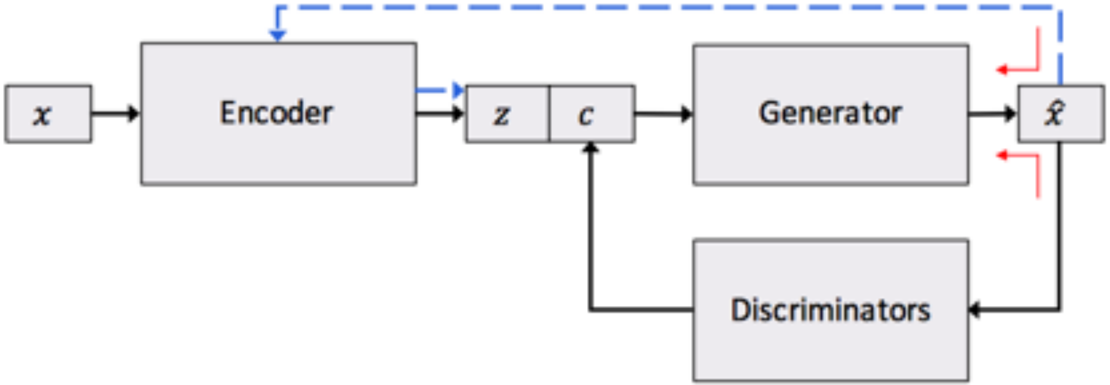


Figure 1. The generative model, where $\boldsymbol{z}$ is unstructured latent code and $\boldsymbol{c}$ is structured code targeting sentence attributes to control. Blue dashed arrows denote the proposed independency constraint (section 3.2 for details), and red arrows denote gradient propagation enabled by the differentiable approximation.

Hu et. al. ICML, 2017

Style Transfer from Non-Parallel Text by Cross-Alignment

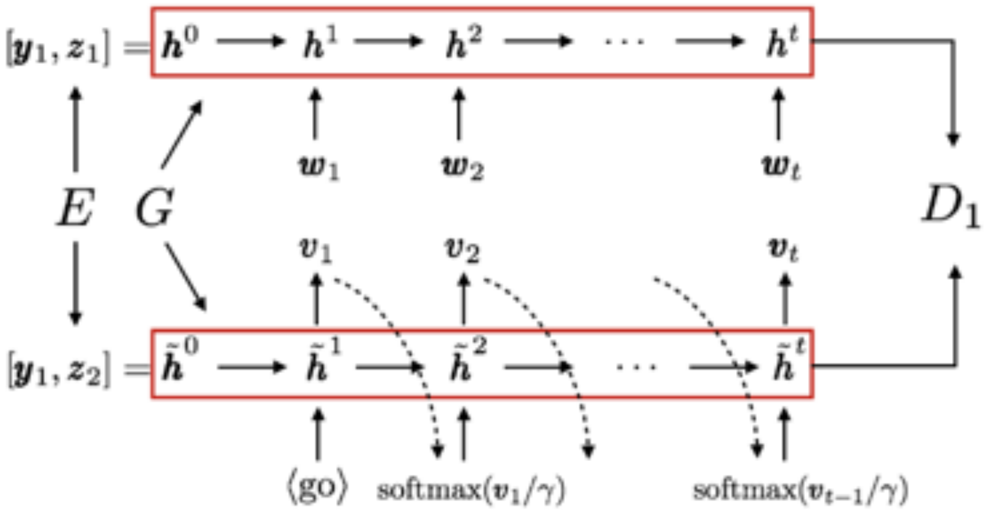


Figure 2: Cross-aligning between x_1 and transferred x_2. For x_1, G is teacher-forced by its words $w_1 w_2 \cdots w_t$. For transfered x_2, G is self-fed by previous output logits. The sequence of hidden states $h^0, \cdots, h^t$ and $\tilde{h}^0, \cdots, \tilde{h}^t$ are passed to discriminator D_1 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.

Shen et. al. NIPS, 2017

Delete, Retrieve, Generate: A Simple Approach to Sentiment and Style Transfer

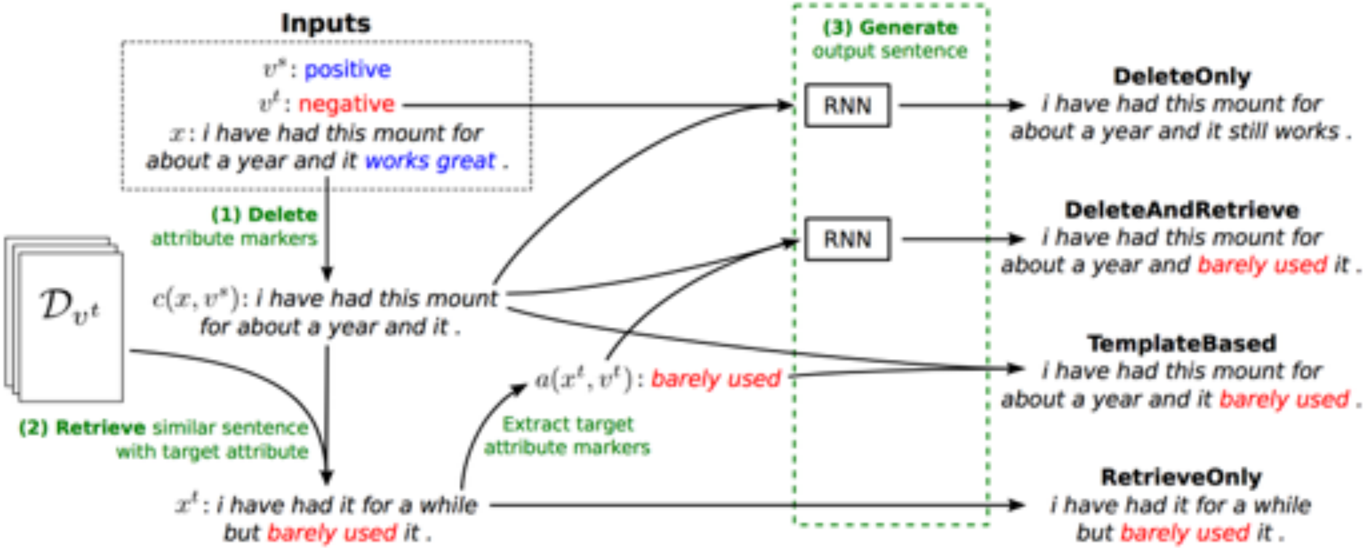
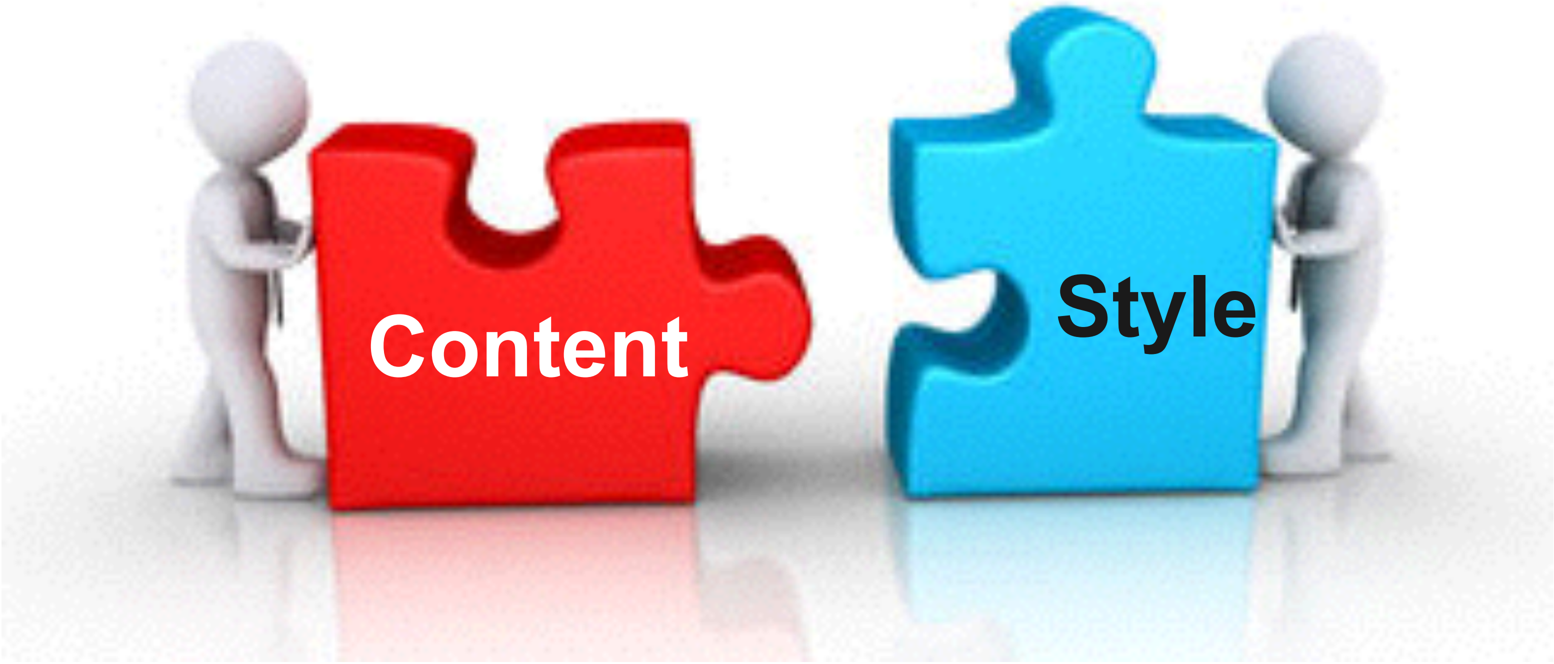


Figure 2: Our four proposed methods on the same sentence, taken from the AMAZON dataset. Every method uses the same procedure (1) to separate attribute and content by deleting attribute markers; they differ in the construction of the target sentence. RETRIEVEONLY directly returns the sentence retrieved in (2). TEMPLATEBASED combines the content with the target attribute markers in the retrieved sentence by slot filling. DELETEANDRETRIEVE generates the output from the content and the retrieved target attribute markers with an RNN. DELETEONLY generates the output from the content and the target attribute with an RNN.

Li et. al NAACL, 2018

Challenges



Challenges

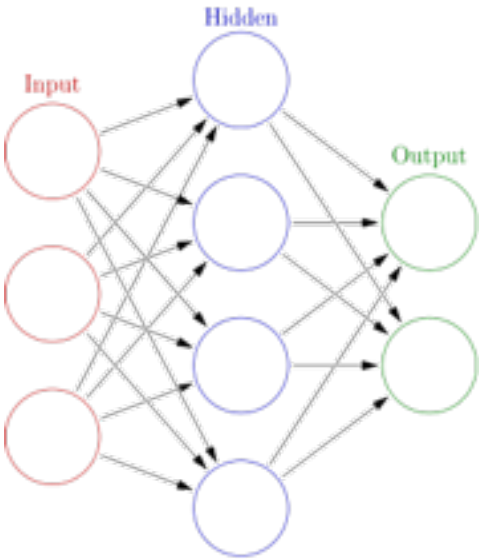
- No Parallel Data!
  - “The movie was very long.”
  - “I entered the theatre in the bloom of youth and emerged with a family of field mice living in my long, white mustache.”
- Style is subtle

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
- Creates a representation independent of the generative model
- Representation is agnostic to the style task

Overview

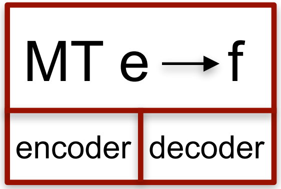
How it works?



How to train?

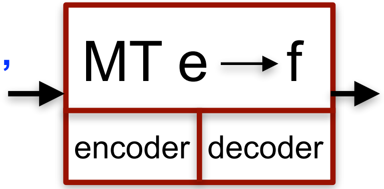
Evaluation

Architecture



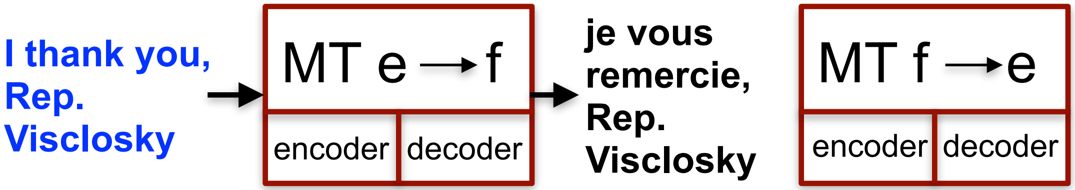
Architecture

I thank you, Rep. Visclosky

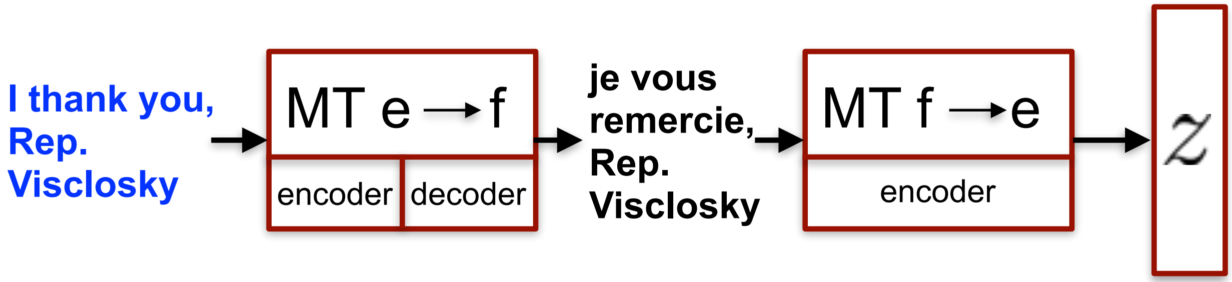


je vous remercie, Rep. Visclosky

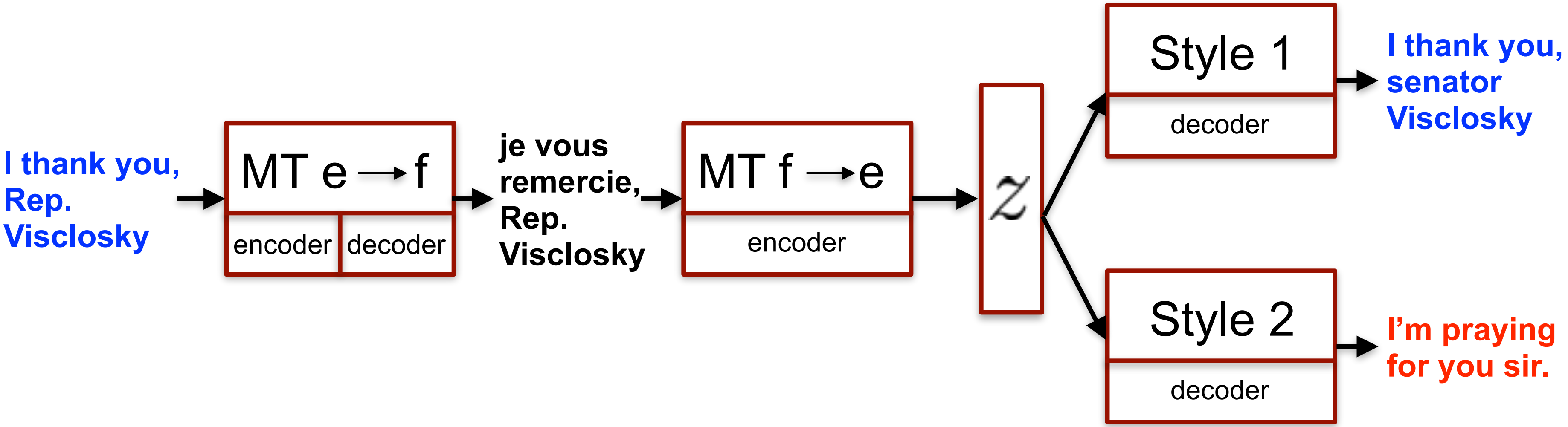
Architecture



Architecture

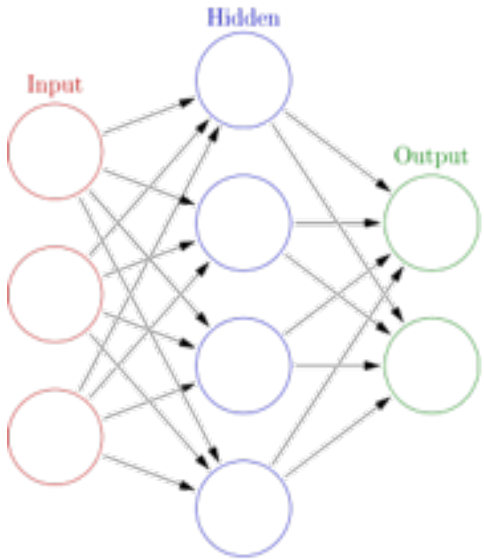


Architecture



Overview

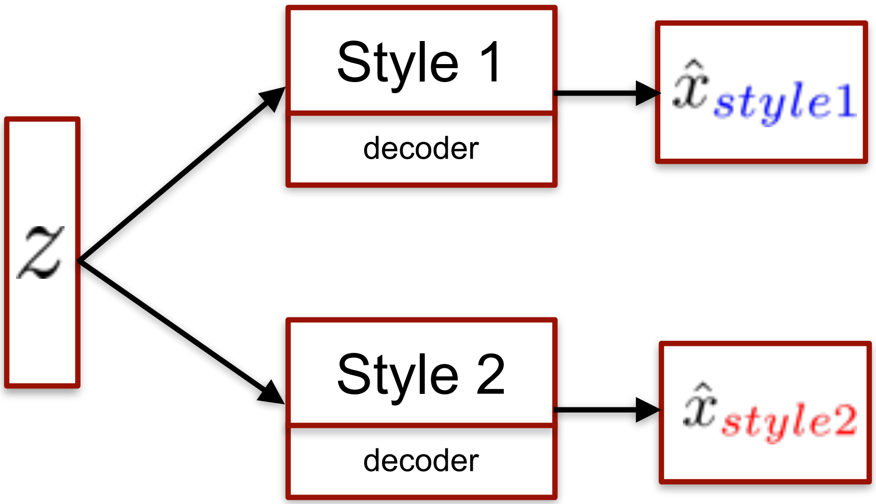
How it works?



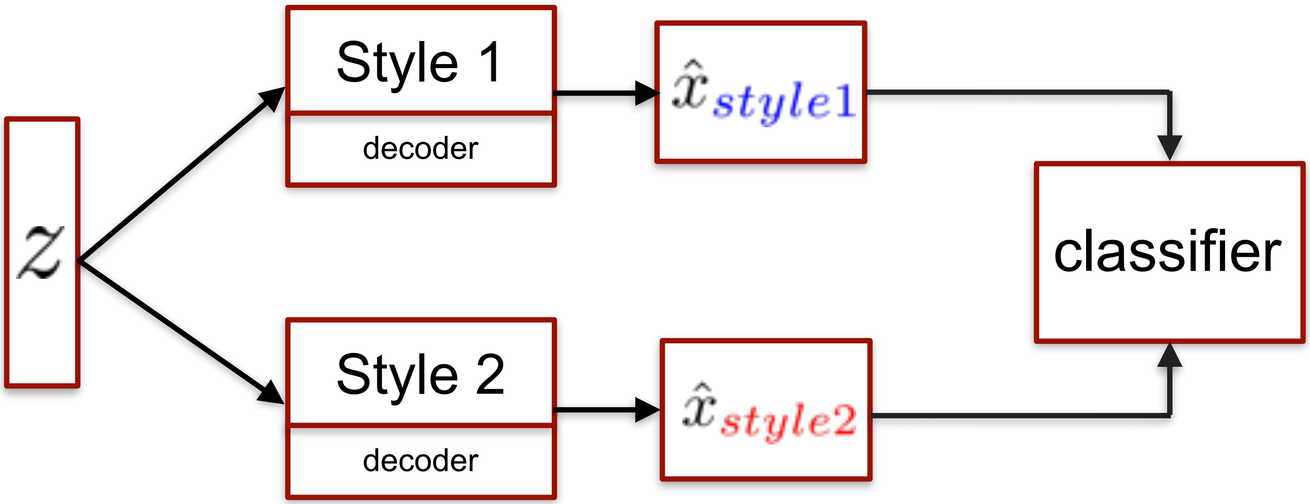
How to train?

Evaluation

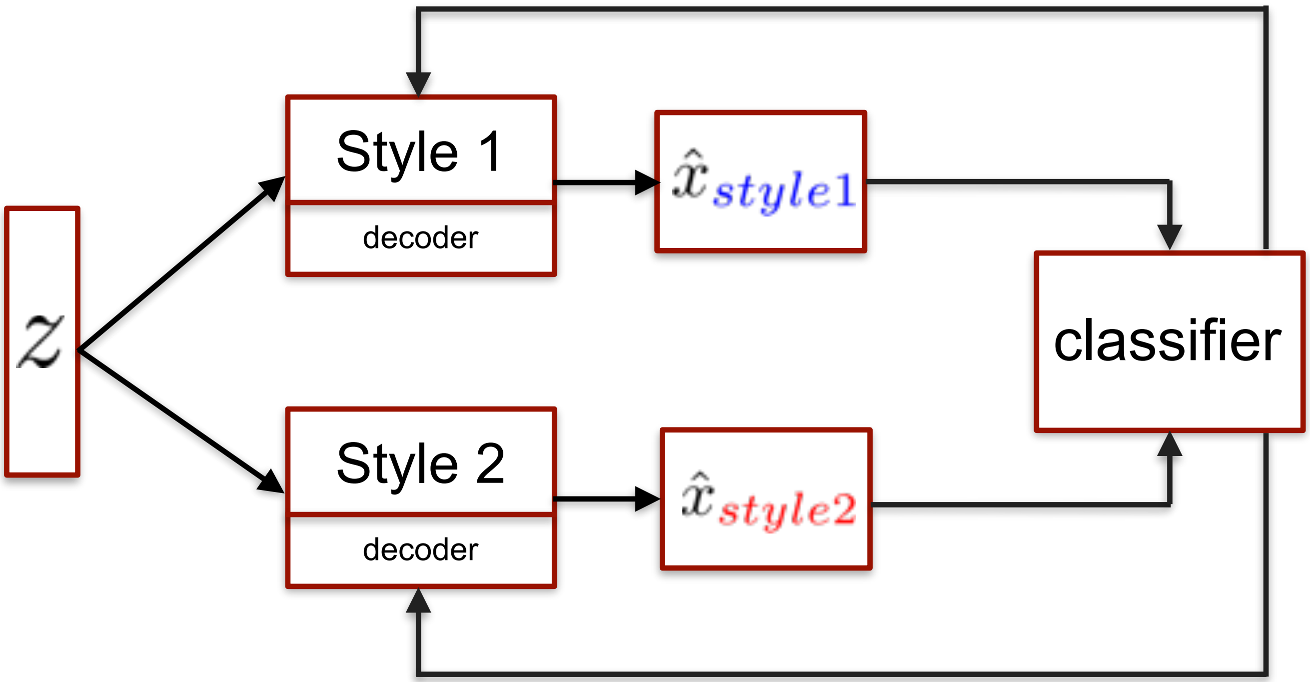
Train Pipeline



Train Pipeline



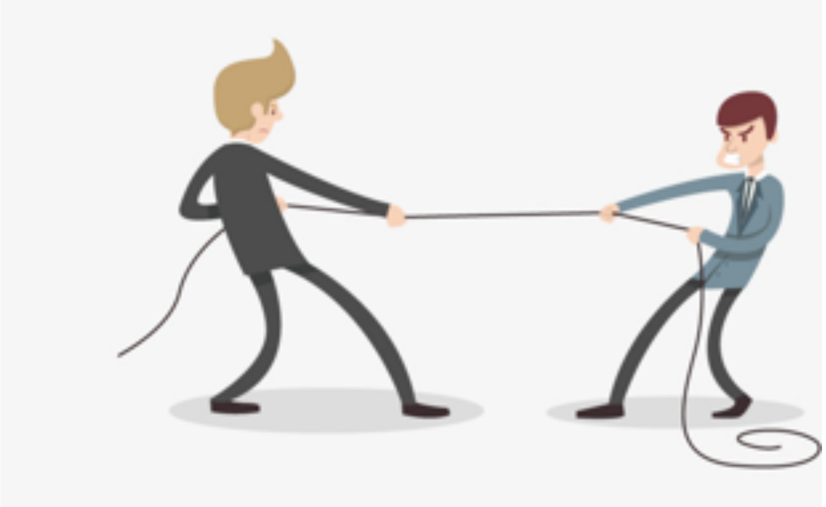
Train Pipeline



Experimental Settings

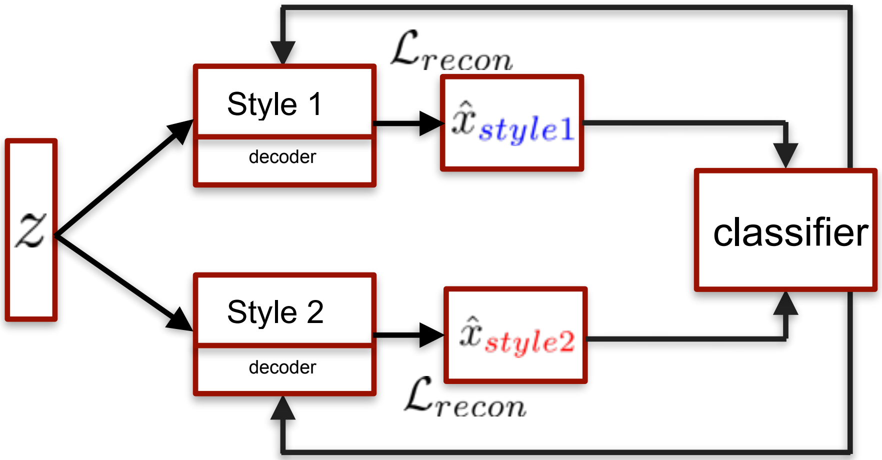
- Encoder-Decoders follow sequence-to- sequence framework (Sutskever et al., 2014; Bahdanau et al., 2015)

$$\min_{\theta_{gen}} \mathcal{L}_{gen} = \mathcal{L}_{recon} + \lambda_c \mathcal{L}_{class}$$



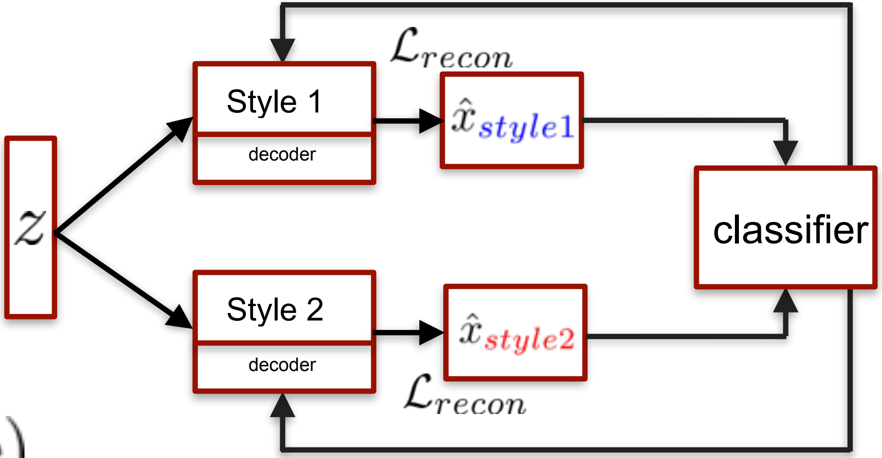
Loss Functions

- Reconstruction loss $\mathcal{L}_{recon}$ is Cross Entropy Loss



Loss Functions

- Reconstruction loss $\mathcal{L}_{recon}$ is Cross Entropy Loss



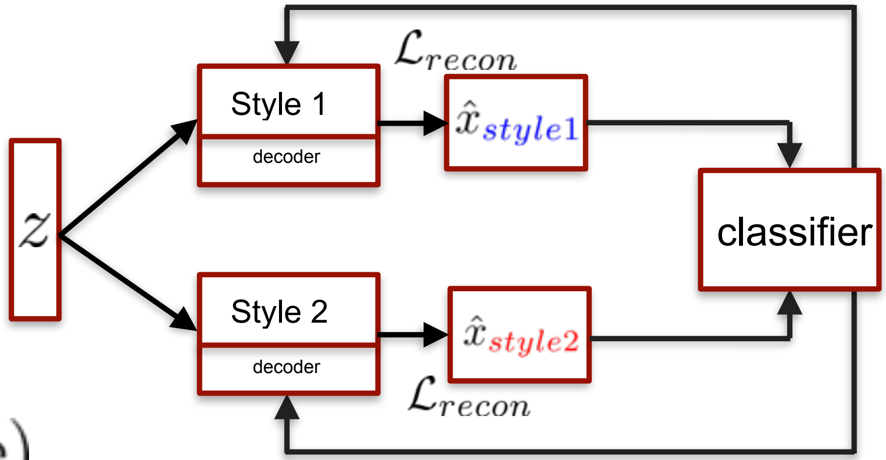
- Input to the classifier:

$$\text{Linear}(o_t, \text{vocab\_size})$$

output of the decoder

Loss Functions

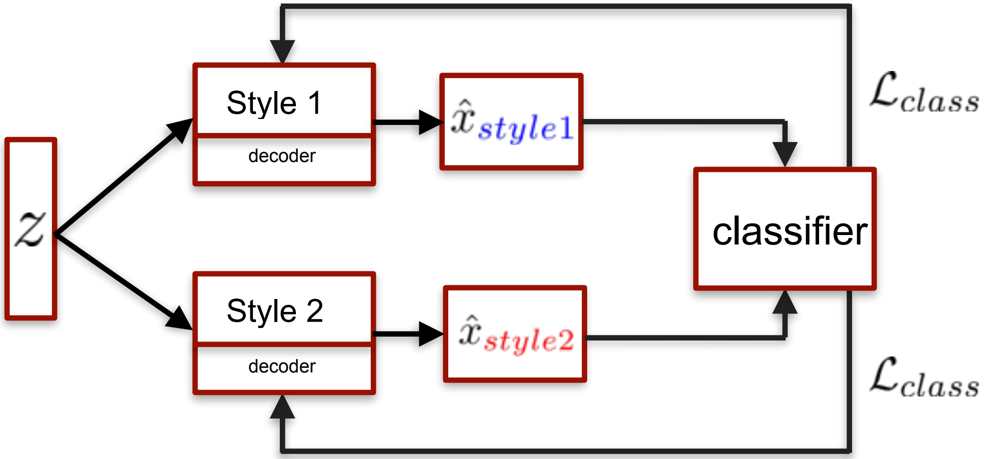
- Reconstruction loss $\mathcal{L}_{recon}$ is Cross Entropy Loss



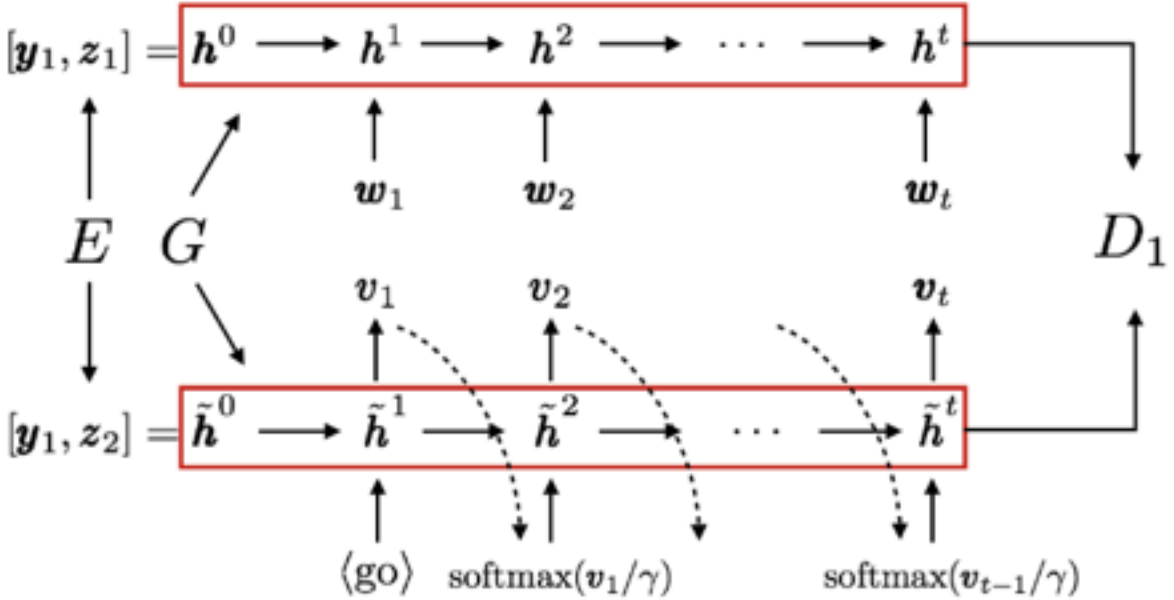
- Input to the classifier:
  - $\mathrm{Linear}(o_t, \mathrm{vocab\_size})$
  - Softmax

Classifier

- Convolutional Neural Network Classifier
- Filter Size: 5 and 100 filters.
- Maximum sentence length of 50.
- Loss $\mathcal{L}_{class}$ is Binary Cross Entropy Loss



Baseline (Shen et al., 2017)



Neural Machine Translation

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

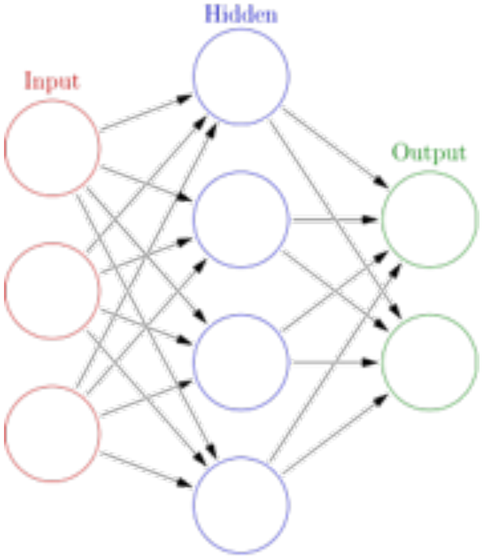
Model	BLEU	WMT 15 Best System
English - French	32.52	34.00
French - English	31.11	33.00

Style Tasks

Task	Labels	Corpus
Gender	Male, Female	Yelp (Reddy and Knight's, 2016)
Political Slant	Republican, Democratic	Facebook Comments (Voigt et al., 2018)
Sentiment Modification	Negative, Positive	Yelp (Shen et al., 2017)

Overview

How it works?



How to train?

Evaluation

Evaluation

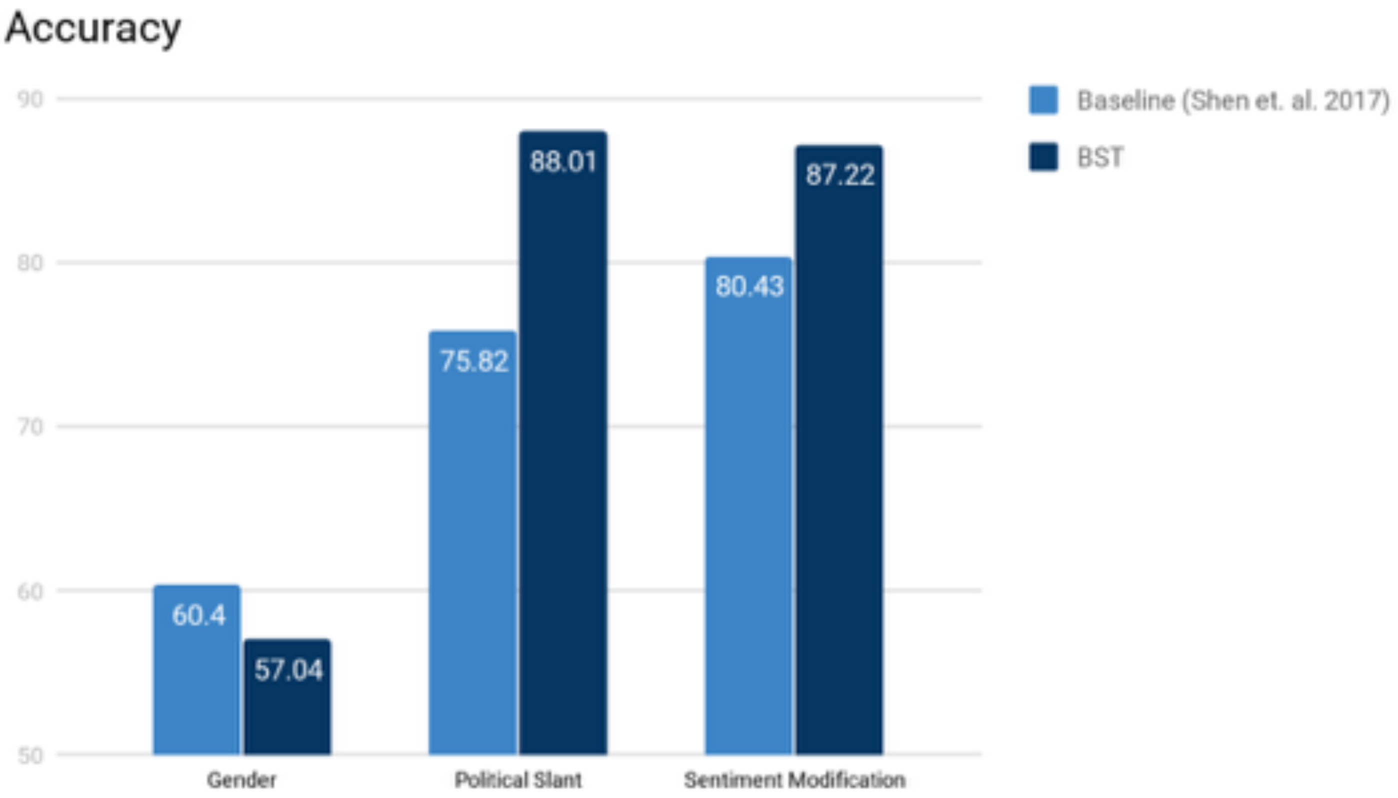
- Style Transfer Accuracy
- Meaning Preservation
- Fluency

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.

Classifier Model	Accuracy
Gender	82%
Political Slant	92%
Sentiment Modification	93.23%

Style Transfer Accuracy



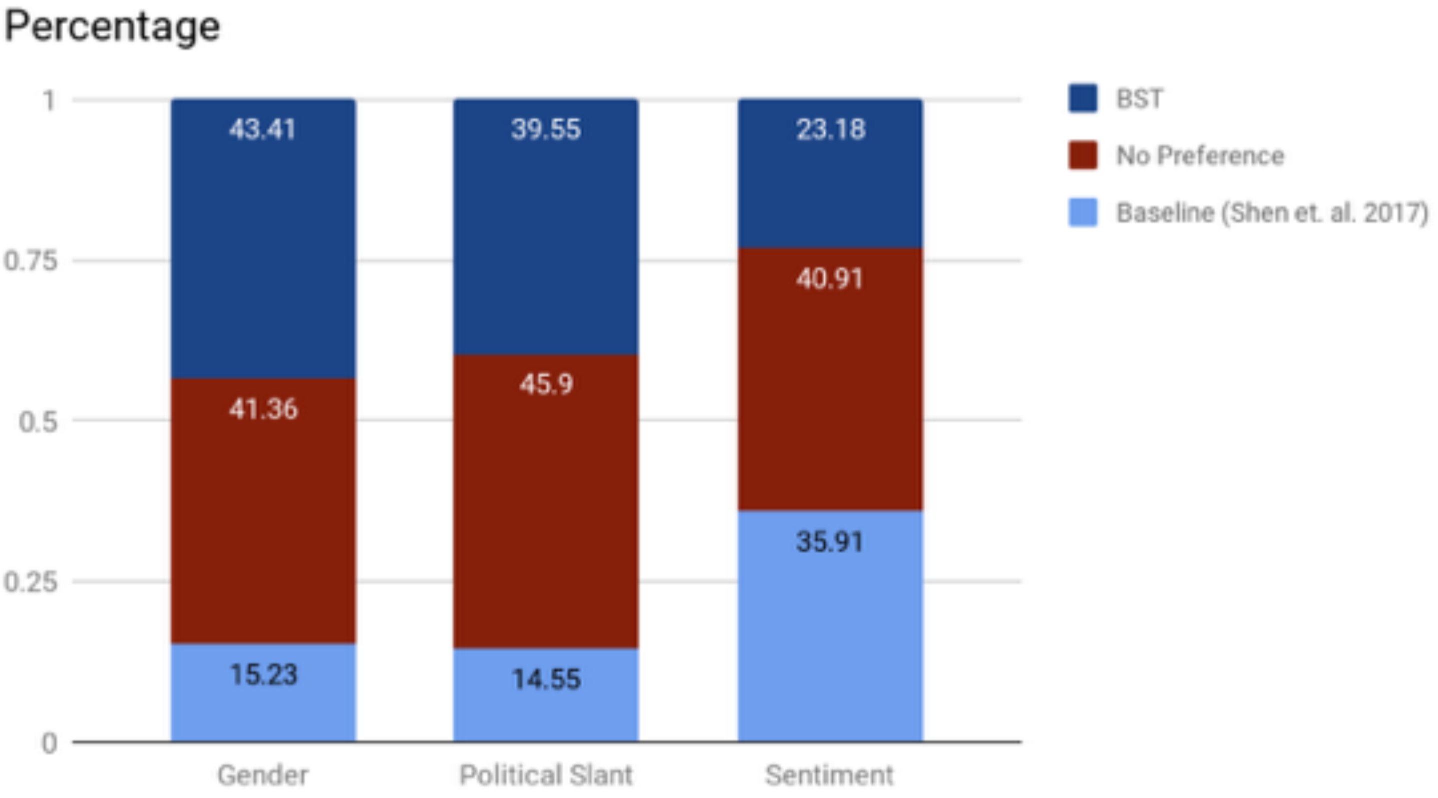
Preservation of Meaning

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

Instructions

- Gender Instruction:
  - “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)”
- Political Slant Instruction:
  - “Which transferred sentence **maintains the same semantic intent** of the source sentence while **changing the political position**”
- Sentiment Instruction:
  - “Which transferred sentence is **semantically equivalent** to the source sentence with an **opposite sentiment**”

Preservation of Meaning



Discussion

Generator loss function

Improve meaning preservation

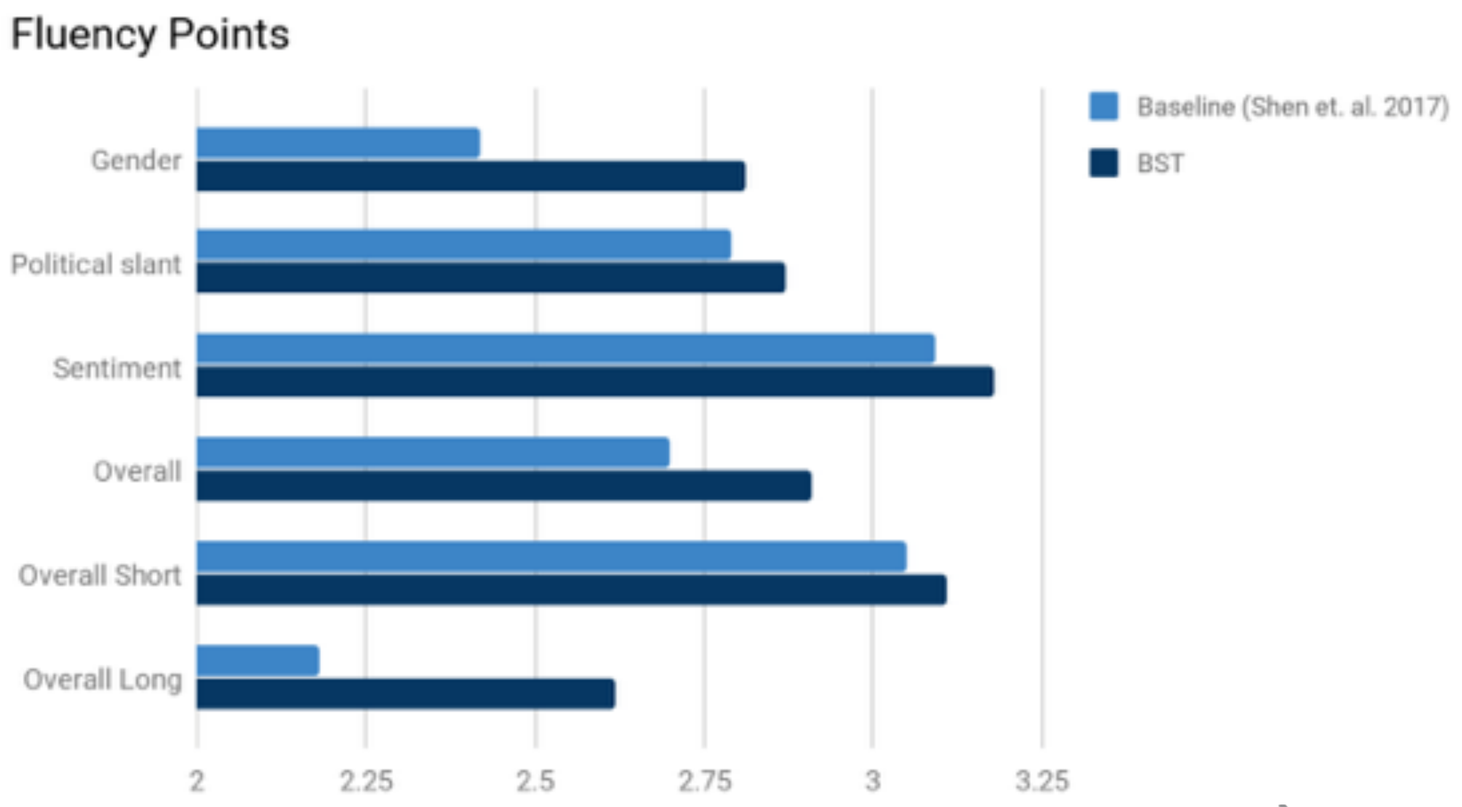
Improve style transfer

- Sentiment modification: not well-suited, evaluating transfer
- Gender style-transfer accuracy $\longrightarrow$ lower BST model but preservation of meaning $\longrightarrow$ much better BST model

Fluency

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

Fluency


Fluency Points
Baseline (Shen et. al. 2017)
BST
Gender
Political slant
Sentiment
Overall
Overall Short
Overall Long
2
2.25
2.5
2.75
3
3.25

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: pivot multiple languages
  - to learn a better grounded latent meaning representation.
- Use multiple target languages with single source language

Future Directions

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

Thank You

Code and data could be found at https://github.com/shrimai/Style-Transfer-Through-Back-Translation

References

- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural net- works. In Proc. NIPS, pages 3104–3112.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Ben- gio. 2015. Neural machine translation by jointly learning to align and translate. In Proc. ICLR.
- Eric Jardine. 2016. Tor, what is it good for? political repression and the use of online anonymity-granting technologies. New Media & Society.
- Steven J. Spencer, Claude M. Steele, and Diane M. Quinn. 1999. Stereotype Threat and Women’s Math Performance. Journal of Experimental Social Psy- chology, 35:4–28.

References

- Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Vie´gas, Martin Wattenberg, Greg Corrado, et al. 2016. Google's multilingual neural machine translation system: enabling zero-shot translation. arXiv preprint arXiv:1611.04558.
- Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In Proc. NIPS.
- Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward con- trolled generation of text. In Proc. ICML, pages 1587–1596.
- J. Li, R. Jia, H. He, and P. Liang. 2018. Delete, Re- trieve, Generate: A Simple Approach to Sentiment and Style Transfer. ArXiv e-prints.