Style Transfer in Text: Exploration and Evaluation

1 Idea

The authors cite lack of parallel corpora and reliable evaluation metrics as the roadblocks for style transfer in natural language processing.

They aim to learn separate content representations and style representations, as is the case with pretty much any work dealing with style transfer in computer vision or natural language processing.

They measure 2 aspects of style transfer, namely transfer strength and content preservation.

2 Method

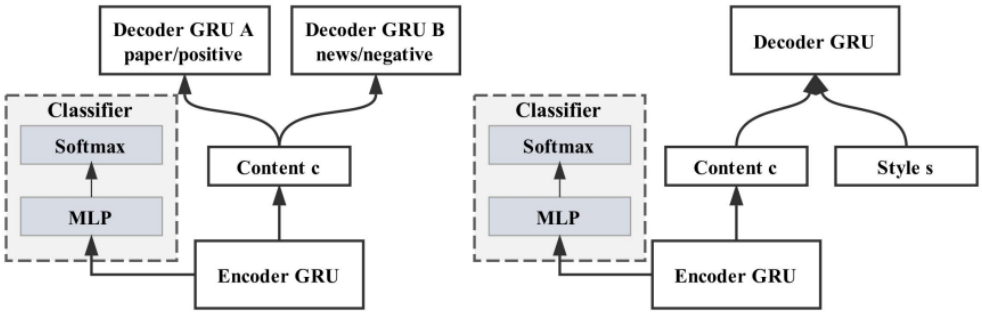
- The base method compared against is an autoencoder framework
- The authors employ 2 models:
  - Multi-decoder seq2seq model [1] that use the different decoders for different styles.
  - Style-embeddings augmented decoder (single decoder) to generate outputs in different styles.
- Adversarial objectives are applied to the content representation of both both. The objective is dissimilar to most adversarial objectives as it tries to maximize entropy of the predicted label from the content representation by minimizing

$$-\sum_{i=1}^{M}\sum_{j=1}^{N} H(P(j|Encoder(x_i;\theta_e);\theta_c))$$

where M is the size of the training data and N is the number of distinct styles.

- Similar to the persona-based neural conversation model [2], a style embedding is learned for each different style. The conditional generation is done using recurrent neural networks with the inputs being the recurrent networks current state, and the style embedding to apply.
- The style embeddings matrix is not directly parameterized by the encoder, but the learning algorithm propagates changes based on how well it combines with the content representation to reconstruct the original text.
- The methods are evaluated in the following manner:
  - Transfer strength is evaluated using a simple classifier
  - Content preservation is evaluated by computing the cosine distance between the original and the generated text embeddings.

3 Architecture



4 Observations

- The authors don't explain:
  - Why is a vanilla autoencoder the base model being compared with? It's objective does not optimize for transferring style.

  - What ratings qualify as a positive/negative review?
  - What kind of decoder strategy was used while predicting sequences? i.e. Greedy-search, Beam-search
  - Which dictionary was used to filter sentimentally polar words for the evaluation.

- The results indicate the the models proposed by the author in general perform better than the auto-encoder for the purposes of transfer strength.
- Although most of the generated sentences have some semblance of syntactic structure, the semantics are poor.
- The solution seems generalizable to multi-class problems, but the authors have conducted evaluations on only binary-class problems

References

[1] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[2] Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*, 2016.