Style Transfer in Text
Exploration and Evaluation

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, Rui Yan

Association for Advancement of Artificial Intelligence, 2018

Problems Addressed

1. Progress in Natural Language Generation Tasks has lagged behind other fields because of lack of parallel data and other evaluation metrics.
2. The authors provide an architecture to achieve style transfer in non-parallel data. This architecture is also capable of separating style from content.
3. The authors also provide two test metrics - one to judge content preservation, and another to judge style transfer strength. The evaluation metric is highly correlated to human judgment.
4. Finally, the authors compose a dataset of paper-news titles to facilitate the research in NLP.

Proposed Solution

Model

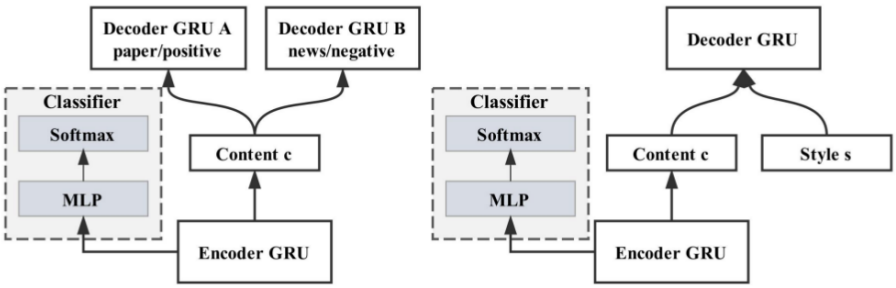

Figure 1: Two models in this paper, multi-decoder (left) and style-embedding (right). Content c represents output of the encoder. Multi-layer Perceptron (MLP) and Softmax constitute the classifier. This classifier aims at distinguishing the style of input X. An adversarial network is used to make sure content c does not have style representation. In style-embedding, content c and style embedding s are concatenated and $[c, e]$ is fed into decoder GRU.

1. The provided architectures solve the above-mentioned problems by Multitask Learning and Adversarial Training.

   The paper suggests an Encoder model and two Decoder models, the common ground of the two models being to learn a representation for the input sentence that only contains the content information.

   #### Encoder

   - The encoder utilises an adversarial framework to separate the content information c from the style information of the given text x.
   - The adversarial network is composed of two parts:
     - The first part aims at classifying the style of x given the representation learnt by the encoder.

– The loss function minimizes the negative log probability of the style labels in the training data:

$$L_{adv1}(\Theta_c) = -\sum_{i=1}^{M} \log p(l_i|Encoder(x_i;\Theta_e);\Theta_c)$$

Where Θ_c represent the parameters of the Multi Layer Perceptron (MLP) that acts as the classifier, and Θ_e stands for the parameters of the RNN encoder.

– The second part of the adversarial network comprises of the encoder, which aims at making the classifier unable to identify the style of x by maximising the entropy of predicted style labels.

$$L_{adv2}(\Theta_e) = -\sum_{i=1}^{M}\sum_{j=1}^{N} H(p(j|Encoder(x_i;\Theta_e);\Theta_c))$$

Where N stands for the number of styles, which is pre-decided.

– Both loss functions update different set of parameters, together ensuring that the output of the encoder does not contain any style information.

First model

- The first model implements a multi-decoder seq2seq model.
- The encoder captures the content c of input X, and the multi-decoder contains $n(n \geq 2)$ decoders to generate outputs in different styles.
- The loss function of each decoder is (similar to)

$$L_{seq2seq}(\Theta_e, \Theta_d) = -\sum_{i=1}^{M} \log P(y_i|x_i;\Theta_e,\Theta_d)$$

(Note: In auto-encoder, output sequence y is set same as x)

- The total generation loss is the sum of the generation loss of each decoder

$$L_{gen1}(\Theta_e, \Theta_d) = \sum_{i=1}^{L} L^i_{seq2seq}(\Theta_e, \Theta^i_d)$$

- The final loss is simply an unweighted sum of losses discussed so far:

$$L_{total}(\Theta_e, \Theta_d, \Theta_c) = L_{gen1}(\Theta_e, \Theta_d) + L_{adv1}(\Theta_c) + L_{adv2}(\Theta_e)$$

Second model

- This model uses the same encoding strategy, but introduces style embeddings $E \in R^{N\times d_s}$, where N represents number of styles and d_s is the dimensionality of style vector, that are jointly trained with the model.
- A single decoder is trained here, that augments the content c by picking an embedding e and concatenating it to c, thus ensuring that learning one decoder is sufficient to generate outputs in different styles.
- The loss function of this model is:

$$L_{total2}(\Theta_e, \Theta_d, \Theta_c, E) = L_{gen1}(\Theta_e, \Theta_d, E) + L_{adv1}(\Theta_c) + L_{adv2}(\Theta_e)$$

Parameter Estimation

- Adadelta with an initial learning rate of 0.0001 and batch size of 128 was used to learn all the parameters.
- The best parameters are decided on the basis of perplexity on the validation of data with a maximum of 50 training epochs for paper-news task and 10 training epochs for positive-negative task.
- For multi-decoder model, the decoders are trained alternately, using the data in the corresponding style.
- For the style embeddings model, the data was randomly shuffled during training, and the style embeddings were jointly learnt with encoder-decoder part.

Evaluation

Transfer Strength

1. The metric for transfer strength is implemented using the LSTM-sigmoid classifier based upon Keras examples.
2. Transfer strength is defined as:

$$TransferStrength = \frac{N_{right}}{N_{total}}$$

Where N_{total} is the total number of test data, and N_{right} is the number of correct case which is transferred to correct style.

Content Preservation

1. Content preservation rate is defined as cosine distance between source sentence embedding v_s and target sentence embedding v_t.
2. Sentence embedding is the combination of max, min, mean pooling of word embeddings. (For word embeddings, pre-trained GloVe is used.)

$$v_{min}[i] = minw_1[i], ..., w_n[i] \tag{1}$$
$$v_{mean}[i] = meanw_1[i], ..., w_n[i] \tag{2}$$
$$v_{max}[i] = maxw_1[i], ..., w_n[i] \tag{3}$$
$$v = [v_{min}, v_{mean}, v_{max}] \tag{4}$$
$$score = \frac{v_s^T v_t}{\|v_s\| \cdot \|v_t\|} \tag{5}$$
$$score_{total} = \sum_{i}^{M_{test}} score_i \tag{6}$$

Single metrics that combine transfer strength and content preservation such as $F1$ score are avoid as sometimes style transfer may be more important, while in other cases content preservation may be more important, thus making a weighted integration of separate metrics ideal.

Experiments

Datasets

1. Two datasets are used : paper-news title dataset (composed by the authors), and positive-negative review dataset, both of which are non-parallel.

2. Both datasets are divided into 3 parts: training, validation, and test data, with the sizes of validation and test data being 2000 sentences each.

3. Sentences with less that 20 words were ignored, and all characters were converted to lower cases. Also, all numbers were converted to a special string "<NUM>".

Model Settings

1. For paper-news title transfer, word embedding of size 64, encoder hidden vector size among 32, 64, 128, and style embedding size among 32, 64, 128 were explored.

2. For paper-news title transfer, word embedding of size 64 for multi-decoder and 64, 128 for style embedding model, encoder hidden vector size among 16, 32, 64, and style embedding size among 16, 32, 64 were explored.

Evaluation Settings

1. LSTM-sigmoid classifier is used to measure transfer strength, which is trained with input word embedding dimension and hidden state dimensions of 128.

2. For content preservation metric, pre-trained 100 dimensional word embeddings are used to compute sentence similarities.

3. For positive and negative review transfer task, sentiment words are filtered out to make sure that the content preservation metrics indeed measure the content similarity. For this, a positive and negative word dictionary is used.

Results

Comparison with Human Judgements

1. The human judgements are obtained by randomly sampling 200 paper-news title transferred pairs from test data, and ask 3 different people to assign a score of 0, 1, or 2 to each, with 2 denoting that the sentences are very similar, and 0 denoting that the sentences are not similar at all.

2. Spearman's coefficient between human judgements and content preservation metric is found to be 0.5656 with p-value < 0.0001, indicating a high correlation between human judgement scores and content preservation metric.

Model Performances

1. Transfer strength and content preservation are found to be negatively correlated across all tasks, with the trade-off slopes being less steep in the models proposed by the authors than the traditional autoencoder models.

Analysis in Multitask Learning View

1. Autoencoder, Style-Embedding and Multi-Decoder can be seen as an implementation of multitask learning.

    - For auto-encoder, two tasks share all the parameters, so it does not have the ability to generate different style sequence.
    - For style-embedding, two tasks share encoder and decoder with separate style embedding, so it has weak ability to generate different style sequence.
    - For multi-decoder, two tasks share encoder with two separate decoders, so it shows high ability to generate different style sequence.
    - For content preservation, more parameters are shared, less distinction between two tasks and more content is preserved.
    - Since the style-embedding model shares more parameters among tasks, less training data is needed to train the model, but the style embeddings have heavier burden to encode the style information.

Lower Bound for Content Preservation

1. The lower bound of the content preservation metric is estimated to gauge how well the model performed in preserving the content.
    - The lower bound is estimated by randomly sampling 2,000 sentence pairs from the two datasets, respectively.
    - Results show that the estimated lower bound of content preservation on the paper-news title dataset is 0.609 and 0.863 on the positive-negative review dataset.
2. For both datasets, the proposed models achieved much higher content preservation scores than the lower bound. This indicates that the proposed model learned to preserve the content of the source sentence well.

Qualitative Study

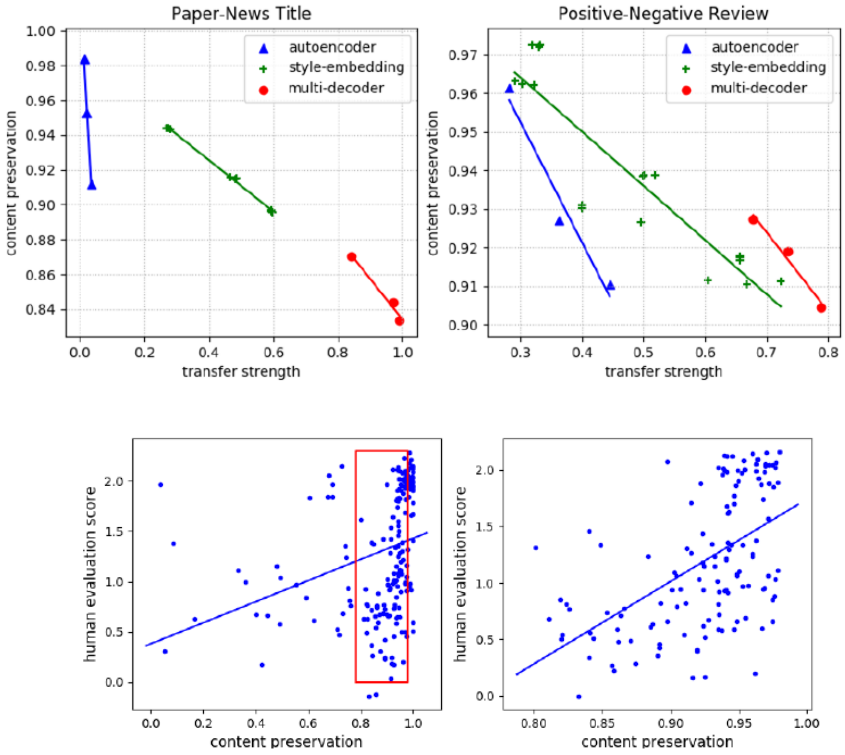

Figure 3: Score correlation of content preservation and human evaluation. Gaussian noise is added to human evaluation for better visualization. The partial enlarged graph is shown on the right.

Dimension Influence

1. Results of auto-encoder in paper-news task : With the increment of encoder (decoder) dimension, ability to recover source sequence increases. So content preservation is larger and transfer strength is smaller (more indeterminacy decreasing).
2. Results of multi-decoder in paper-news task : Similar to auto-encoder, with the dimension increasing, it shows higher ability to recover sentence.
3. Results of style-embedding in paper-news task : Encoder dimension usually has little influence on results, but larger style embedding dimension (also decoder dimension) allows preserving more content. The performance of decoder to recover sentence also increases with dimension.

Bibliography

1. (2011) Banchs, R. E., and Li, H. 2011. Am-fm: a semantic framework for translation quality assessment. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2, 153–158. Association for Computational Linguistics.

2. (2005) Banerjee, S., and Lavie, A. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, volume 29, 65–72.

3. (2016) Bousmalis, K.; Trigeorgis, G.; Silberman, N.; Krishnan, D.; and Erhan, D. 2016. Domain separation networks. In Advances in Neural Information Processing Systems, 343–351.

4. (2017) Braud, C., and Søgaard, A. 2017. Is writing style predictive of scientific fraud? arXiv preprint arXiv:1707.04095.

5. (1998) Caruana, R. 1998. Multitask learning. In Learning to learn. Springer. 95–133.

6. (2017) Chen, X.; Shi, Z.; Qiu, X.; and Huang, X. 2017. Adversarial multi-criteria learning for chinese word segmentation. In ACL.

7. (2000) Doddington, G. R.; Przybocki, M. A.; Martin, A. F.; and Reynolds, D. A. 2000. The nist speaker recognition evaluation–overview, methodology, systems, results, perspective. Speech Communication 31(2):225–254.

8. (2017) Ficler, J., and Goldberg, Y. 2017. Controlling linguistic style aspects in neural language generation. arXiv preprint arXiv:1707.02633.

9. (2015) Ganin, Y., and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In International Conference on Machine Learning, 1180–1189.

10. (2016) Gatys, L. A.; Bethge, M.; Hertzmann, A.; and Shechtman, E. 2016. Preserving color in neural artistic style transfer. arXiv preprint arXiv:1606.05897.

11. (2016) Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2016. Image style transfer using convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2414–2423.

12. (2016) Ghazvininejad, M.; Shi, X.; Choi, Y.; and Knight, K. 2016. Generating topical poetry. In EMNLP, 1183–1191.

13. (2014) Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In Advances in neural information processing systems, 2672–2680.

14. (2016) He, R., and McAuley, J. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In Proceedings of the 25th International Conference on World Wide Web, 507–517. International World Wide Web Conferences Steering Committee.

15. (2017) Jhamtani, H.; Gangal, V.; Hovy, E.; and Nyberg, E. 2017. Shakespearizing modern language using copyenriched sequence-to-sequence models. arXiv preprint arXiv:1707.01161.

16. (2016) Li, J.; Galley, M.; Brockett, C.; Spithourakis, G. P.; Gao, J.; and Dolan, B. 2016. A persona-based neural conversation model. arXiv preprint arXiv:1603.06155.

17. (2017) Li, Y.; Wang, N.; Liu, J.; and Hou, X. 2017. Demystifying neural style transfer. In IJCAI.

18. (2013) Lichman, M. 2013. UCI machine learning repository. (2004) Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out: Proceedings of the ACL-04 workshop, volume 8. Barcelona, Spain.

19. (2017) Long, M.; Wang, J.; and Jordan, M. I. 2017. Deep transfer learning with joint adaptation networks. In ICML.

20. (2017) Mueller, J.; Gifford, D.; and Jaakkola, T. 2017. Sequence to better sequence: continuous revision of combinatorial structures. In International Conference on Machine Learning, 2536–2544.

21. (2002) Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics, 311–318. Association for Computational Linguistics.

22. (2014) Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In EMNLP, volume 14, 1532–1543.

23. (1985) Rumelhart, D. E.; Hinton, G. E.; and Williams, R. J. 1985. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science.

24. (2017) Shen, T.; Lei, T.; Barzilay, R.; and Jaakkola, T. 2017. Style transfer from non-parallel text by cross-alignment. arXiv preprint arXiv:1705.09655.

25. (2014) Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In Advances in neural information processing systems, 3104– 3112.

26. (2017) Tao, C.;Mou, L.; Zhao, D.; and Yan, R. 2017. Ruber: An unsupervised method for automatic evaluation of opendomain dialog systems. arXiv preprint arXiv:1701.03079.

27. (2015) Vinyals, O.; Fortunato, M.; and Jaitly, N. 2015. Pointer networks. In Advances in Neural Information Processing Systems, 2692–2700.

28. (2013) Yan, R.; Jiang, H.; Lapata, M.; Lin, S.-D.; Lv, X.; and Li, X. 2013. i, poet: Automatic chinese poetry composition through a generative summarization framework under constrained optimization. In IJCAI, 2197–2203.

29. (2016) Yan, R. 2016. i, poet: Automatic poetry composition through recurrent neural networks with iterative polishing schema. In IJCAI, 2238–2244.

30. (2012) Zeiler, M. D. 2012. Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701.

31. (2017) Zhou, H.; Huang, M.; Zhang, T.; Zhu, X.; and Liu, B. 2017. Emotional chatting machine: Emotional conversation generation with internal and external memory. arXiv preprint arXiv:1704.01074.

32. (2017) Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. arXiv preprint arXiv:1703.10593.