Towards Actual (Not Operational) Textual Style Transfer Auto-Evaluation

Richard Yuanzhe Pang [§]
New York University, New York, NY 10011, USA
yzpang@nyu.edu

There are advances on developing methods that do not require parallel corpora, but issues remain with automatic evaluation metrics. Current works (Pang and Gimpel, 2018; Mir et al., 2019) agree on the following three evaluation aspects. (1) Style accuracy of transferred sentences (measured by a pretrained classifier). (2) Semantic similarity between the original and transferred sentences. (3) Naturalness or fluency: researchers use perplexity of transferred sentences, using the language model pretrained on the original corpora.

Problem 1: Style Transfer Tasks. If we think about the practical use cases of style transfer (writing assistance, dialogue, author obfuscation or anonymity, adjusting reading difficulty in education, artistic creations such as works involving literature), we would find that the two would-be-collected non-parallel corpora have different vocabularies, *and* it is hard to differentiate style-related words from content-related words. For example, when transferring Dickens' to modern style literature (Pang and Gimpel, 2018), the former may contain "English farm", "horses"; the latter may contain "vampire", "pop music." But these words should stay the same, as they are content-related but not style-related. On the other hand, Dickens' literature may contain "devil-may-care" and ”flummox”, but these words *are* style-related and should be changed. Recent works, however, mostly deal with the **operational** style where corpus-specific content words are changed. The operational style transfer models work well on Yelp sentiment transfer which almost all researches focus on, but it does not inspire systems in practical use cases.

Problem 2: Metrics. Consider: *Oliver deemed the gathering in York a great success.* The expected transfer from Dickens to modern literature style should be similar to "Oliver thought the gathering was successful" (**actual** style transfer). However, the most likely transfer (if we use most existing models) will be "Karl enjoyed the party in LA" (**operational** style transfer). In evaluating semantic similarity, Mir et al. (2019) masked style keywords determined by a classifier. In this case, all corpus-specific content words (as well as style words) will be masked, and evaluation will fail. However, we can create the list of style keywords with outside knowledge. We can also consider keeping the words as they are without masking. Similar problems exist for the other two metrics.

Problem 3: Trade-off and Aggregation. Aggregation of metrics is especially helpful as there are tradeoffs (Pang and Gimpel, 2018; Mir et al., 2019), and we need to tune and select models systematically. Use A, B, C to represent the three metrics. For sentence s, define $G_{t_1,t_2,t_3,t_4}(s) = \big([A-t_1]_+ \cdot [B-t_2]_+ \cdot \min\{[t_3-C]_+, [C-t_4]_+\}\big)^{\frac{1}{3}}$ where t_i's are the parameters to be learned.[1] (Small and large C's are both bad.) The current research strives for a universal metric. We can randomly sample a few hundred pairs of *transferred* sentences from a range of style transfer outputs (from different models—good ones and bad ones) from a range of style transfer tasks, and ask annotators which of the two transferred sentences (from the same original sentence) is better. We can then train the parameters based on pairwise comparison. To make G more convincing, we may design more complicated functions $G = f(A, B, C)$. If we do not need a universal evaluator, then we can repeat the above procedure by only sampling pairs of transferred sentences from the dataset of interest, which is more accurate for the particular task.

[§] Abstract written while the author was a student at the University of Chicago.

[1] Inspired by geometric mean.

Proceedings of the 2019 EMNLP Workshop W-NUT: The 5th Workshop on Noisy User-generated Text, pages 444–445
Hong Kong, Nov 4, 2019. ©2019 Association for Computational Linguistics

References

Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating style transfer for text. *arXiv preprint arXiv:1904.02295*.

Yuanzhe Pang and Kevin Gimpel. 2018. Learning criteria and evaluation metrics for textual transfer between non-parallel corpora. *arXiv preprint arXiv:1810.11878*.