Shuhui Zhu
sz649

Critique 5: “Evaluating Style Transfer for Text”

Remi Mir, Bjarke Felbo, Nick Obradovich, Iyad Rahwan

While the paper attempts to offer a standard evaluation practice to text style transfer tasks, there are limits in approaching style transfer intensity and content preservation. To benchmark against human evaluation, the paper first aims to provide a reliable method to guide human evaluators via relative scoring. In human evaluating style transfer intensity, the study has human raters score the difference in style between input (x) and output (x’). Though this relative comparison may yield to a more precise and consistent evaluation, the method ignores one aspect, the comparison between output and target. What if the output text is drastically different from the input text but also quite different from the target style? In the automated evaluation, Earth Mover’s Distance (EMD) takes this aspect into consideration. It “penalizes the score if SC(x) displays a relative change of style in the wrong direction, away from the target style”. For human evaluators, specific weights on similarity to target style and difference from input text are not stated clearly in the paper.

Measuring intensity is a subjective task for humans. Given a scale from 1 to 5 in intensity to rate the same sentence, raters may have different expectations in the degree. The study should provide one set of concrete standard outputs corresponding to each scale. The comparison should be between the standards, not to input texts. Further, the original input texts have a degree in negativity or positivity. The intensity of “I really like apples” is different from “I like apples”. Take either as input, to compare with an output “I don’t like apples”, though the inputs both have positive labels and output is the same, the difference between the input and output would be different.

Another limit lies in evaluating content preservation. To settle the difference between human raters’ definition of content words, the paper proposes an additional step of masking before evaluation. The masking process first constructs a lexicon of style-related words using logistic regression classifier. While difficulty in identifying all style-related words is regarded as a minor issue by the paper, this approach does have consequences in masking and evaluating. As there is a discrepancy of understanding context words among human raters as well as between human and automated classifier, the question is whose judgement is more accurate, logistic regression classifier or humans. If humans’ judgement is more accurate, the additional step of masking though sets a standard, may set a wrong standard.

Meanwhile since identifying attribute is part of the style transfer task, evaluation after the masking process ignores the mistakes during attribute identification phase, one component of the transfer model to be evaluated. Thus, evaluation after masking on both input and output text is not independent of transfer models. Important aspect of model’s capacity to detect attribute is overlooked. One solution is to keep the input as no modification, apply style masking on the output. For human raters, they could evaluate both the accuracy of detection and the transformation of content by comparing output (the girls up front are <customstyle>) and input (the girls up front incompetent). For auto evaluation, we could add one comparison. Generate an output using the original label as target, apply METEOR or BLEU to compare with the no modification input. This attempts to capture model’s capability of differentiating content and attribute.