1 Evaluation schema

Given:

- a style transfer model M_{AB} that attempts to transfer an input sentence in style A to style B
- a candidate output sentence from M_{AB}: x^i_{AB}
- a reference sentence and score pair $R^i_{AB}; y_i$ that is scored in a range of $[0, 1]$ by a trusted external source

The evaluation metric E(x^i_{AB}, R^i_{AB}, y_i) will output a score. A successful metric E will output a score $\hat{y}$, where $|y - \hat{y}|$ is minimized.

2 Desired attributes of an evaluation metric

- intuitive
- linear in the input to $E()$
- relies only on X, R, Y
- if it evaluates content in addition to style, it should do so independently