Evaluating Style Transfer

Peter W. Kong

***Abstract*— This project formalizes the problem of evaluation of style transfer and explores improves on the state of the art metrics.**

I. Introduction

Style identification is the identification of style in a text passage by an automated process, often a language model. Generative models can generate text in one or more styles, given content and style parameters. Style identification has proven utility in applications like author identification (Houvardas). Yet, while much research has focused on semantic transfer from one human language to another, there has been little attention paid to intra-lingual style transfer. [Intra-lingual] style transfer affords many interesting applications: authorship obfuscation, text simplification, and style enhancement.

The existing literature on style transfer uses a combination of BLEU scores, perplexity [Ficler], black box neural classifiers, and manual labeling to evaluation the quality of a transfer. Each of these has serious problems (see SECTION)

This papers objectives are three-fold:

- Identify characteristics desirable in a style transfer evaluation metric
- Formalize the evaluation objective
- Explore a specific approach and discuss results

II. Related work

There is active research in style transfer. Ficler uses fixed, human- intelligable stylistic parameters like length and sentiment, which can be extracted from specific training data with heuristics. They use perplexity as an evaluation metric.

[Fu, et al.] addresses the problem of learning style transfer without the use of parallel corpora, using style embeddings (a previous work) and an auto-encoder model with multiple decoders. Attempts are made to improve evaluation metrics, but they do not appear to improve on human judgment and black-box classifiers. Example outputs from models are hard to visualize/express in a relatable way.

[Hu] seeks to generate realistic (English) sentences using a VAE in a wake-sleep configuration. Semantic and stylistic signals can be injected to coerce specific meaning and styles. Outperforms stated previous work: a semi-supervised VAE (Kingma 2014). Reaches about 83pct sentiment classificiation accuracy, although performance varies greatly with datasets used. Metric: label accuracy. Method: External sentiment clasifiers.

[Prabhumoye]'s proposed model creates a style-agnostic representation of an English input sentence using A-¿B then B-¿A neural machine translators, with the objective of style transfer with minimal semantic drift. A style-specific transformation step adds specific style to the style-agnostic representation. This paper confirms that BLEU and ROUGE metrics are not suitable for capturing meaning.

III. Criticism of state of the art metrics

[TO EXPAND ON]

- Terrible accuracy (perplexity
- Easily fooled (bleu)
- Low explainability (neural models)
- Intractible (manual)

IV. Formalizing the problem

A. problem formalization

We formalize the style transfer problem as: 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.

B. Desired Evaluation metric traits

TO EXPAND ON

- 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

C. discussion of available paths

neural: pros and cons
human engineered features: pros and cons

D. Chosen path

human engineered features: include reasoning

V. Experiment setup

A. Data

All aligned sentences from the Shakespeare dataset were used, from the following plays: 'othello', 'antony-and-cleopatra', 'asyoulikeit', 'errors', 'hamlet', 'henryv', 'julius-caesar', 'lear', 'macbeth', 'merchant', 'msnd', 'muchado', 'richardiii', 'romeojuliet', 'shrew', 'tempest', 'twelfthnight'

42,000 parallel sentences: original and sparknotes version.

The topics encompass a wide variety of content and sentiment: comedies, tragedies, histories, although it is important to note that they are all written in the screenplay format. While this is an amazing corpus to work with, there are some natural limitations.

Here are some examples of parallel aligned sentences. As you can see, differences in percieved style can vary significantly from passage to passage (original style on the left):

TABLE I
Aligned Text Examples

What's the matter, lieutenant?	What's the matter, lieutenant?
Tell me this, I pray: Where have you left the money that I gave you?	Answer me this, please: where's the money I gave you?
Lucius, who's that knocking?	Lucius, who's that knocking?
Gentlemen, forward to the bridal dinner.	Gentlemen, on to the bridal dinner.

The original sentences were automatically assigned a "class 1" label, and the sparknotes sentences a "class 2" label, this generating a classical supervised learning dataset with "gold" labels.

This dataset was randomly shuffled. 70pct of corpus was designated the train set, 20pct the validation set, and 10pct the holdout test set.

We experimented with a variety of models. However, the choice of experiments was guided by our evaluator design ideals [see prev section].

B. Models

Models employed: logistic regression
SVM

C. feature engineering

Features created:

We used grid search to find successful combinations of engineered features.

VI. Results

include examples of type 1 and type 2 errors
include accuracy of best model
include statement of features used.

VII. Analysis

A. Corpora

Parallel, sentence aligned Shakespeare dataset. original + cliffnotes. [Coco xu] Analysis of bleu score:

style1 vs style2 averaged bleu: 44
randomized:.0175

BLEU scores range from 0 to 1. We can see from the differences of intra-style vs inter-style blue scores that BLEU certainly contains some signal. Unfortunately, since our labels of styles are binary, it is difficult to investigate this further and determine whether or not inter-style BLEU scores correlate smoothly/monotonically with inter-style change.

B. Error analysis

C. Acknowledged shortcomings

assumption of evenly distributed class (style) priors One dataset, where part of the stated task of style B was shortening.

VIII. Future work

expand approach to handle passages larger than sentences
...

IX. Conclusions

APPENDIX

...

References

[1] example of bib item