Evaluating Style Transfer

Peter W. Kong

***Abstract*— This project formalizes the problem of evaluation of monolingual sequence to sequence style transfer models and explores improvements 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 paper's 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

[TODO: expand on each bullet below]

- Terrible accuracy (perplexity)
- Easily fooled (bleu)
- low explanatory power (neural models). [TODO: mention work in developing neural model explanatory power.]
- 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. [TODO: remove y_i ?] A successful metric E will output a score $\hat{y}$, where $|y - \hat{y}|$ is minimized. Some variations on this loss function are detailed later (TODO: reference Brier score).

B. Desired Evaluation metric traits

After reviewing the pros and cons of various evaluation metrics in the current literature, we distilled four desirable traits in an evaluation metric.

- The ideal metric $E()$ should be accurate. More precisely, it should track ground truth values closely. When the ground truth score for x^i_{AB} is low, E(x^i_{AB}, R^i_{AB}, y_i) should also be low. Likewise for high ground truth scores.
- $E()$ should be intuitive. It is common for research tasks to use [TODO: add reference] LSTM classifiers to evaluate style transfer (when, of course, human scores are not sufficiently available). While some of these neural classifiers achieve reasonable accuracy, their explanatory power is quite low. Humans cannot

easily draw intuitions from the weight matrices of neural layers.
- $E()$ speed should be reasonable. We define this having a run time that is linear in the input data.
- $E()$ relies only on X, R, Y. This allows $E()$ to be portable between experiments and corpora.
- If $E()$ evaluates content in addition to style, it should do so independently We quickly realized that an improvement on the existing metrics required an ensemble of multiple submetrics.

C. Discussion of exploratory paths

With these traits in mind, we avoided including neural network, or other 'black box' classifiers in our evaluator. Instead, we focused on simple linear classifiers with intuitive engineered features.

V. Experiment setup

A. Data

Experimental data was draw from Coco Xu's Shakespeare dataset. This dataset is drawn from two distinct sources: Shakespeare's original play scripts, and Sparknote's modernized versions. Sentences are aligned, so that the same sentence may be accessed simultaneously in both styles. All aligned sentences from the Shakespeare dataset were used, from the following plays:

'othello', 'antony-and-cleopatra', 'asyoulikeit', 'errors', 'hamlet', 'henryv', 'juliuscaesar', 'lear', 'macbeth', 'merchant', 'msnd', 'muchado', 'richardiii', 'romeojuliet', 'shrew', 'tempest', 'twelfthnight'

42,000 parallel sentences in total.
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 by sentence pair, preserving style-to-style alignments. 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].

We used 28 [TODO: update number] features, including: sentence length, bleu score, counts of all major parts of speech, counts of all major punctuations, and adv-verb and adj-noun ratios.

B. NER and BLEU score exploration

We attempted to use NER scores as well, but found that they provided almost no predictive power [TODO: expound?] even between content-aligned inter-style sentence pairs. And of course, using NER in a non-aligned context reduces portability. Furthermore, NER, even if configured to yield predictive power, adds a distracting content signal into the evaluator, violating one of our ideal traits.

We knew that BLEU scores, while incomplete on their own, provided a useful style (and not merely content) signal, because of this exploratory experiment:
We took a single play ("Othello"), and summed the difference the differences in BLEU scores of randomly paired intra-style sentence pairs. This yielded a normalized score of .44 [TODO: update score]

Then we performed the same procedure, except that summed BLEU score differences between randomly paired inter-style sentence pairs. this yielded a normalized score of .17 [TODO: update score]

POS features were generated by the Spacy POS tagger, trained on a standard web corpus. We found the tagger to be quite accurate, even on the relatively ancient shakespearean corpus. tagging example:

"Go bid the priests do present sacrifice And bring me their opinions of success"

Go VERB
bid VERB
the DET
priests NOUN
do VERB
present ADJ
sacrifice
And CCONJ
bring VERB
me PRON
their ADJ
opinions NOUN
of ADP
success NOUN

C. Models

We reviewed linear models that had good classification ability as well as the ability to output probability distributions of class predictions, rather than just the binary predictions.

Models employed: SVM [TODO: expound]

logistic regression [TODO: expound]

Multinomial Naive Bayes: has the ability to predict probabilities, which, while less helpful given our binary labeled data set, would be quite helpful in further verification of the evaluator for a future, ranged labeled dataset. With this in mind, we made the assumption that our binary training data labels were actually 100pct-weighted probability values. We then used Brier, a normalizad form of SSE, as our loss function during evaluation (see Results).

D. grid search

[TODO: expound on grid search methodology]

VI. Results

TABLE II
MODEL PERFORMANCE

Model	Validation Brier Score	Test-set Accuracy
MNB w/ Bleu, POS features	.2924	51.5%

TODO: insert table with final scores

VII. Analysis

Since this is an evaluation metric, it's important to strive for a non-binary output that closely tracks gold scoring along the entire range of possible values (i.e. 0 to 1)..
Our best model, Multinomial Naive bayes with (XXX features) underperformed our intuition-driven expectations.

A. Corpora

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

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.

[Include explanation of why Multinomial naive bayes performed best. Probably should include both Brier loss and zero-one loss]

B. Error analysis

TODO: add discussion of validation-set vs. test-set error difference (overfitting)
TODO: add and analyze example of both AB and BA misprediction

F1 score not sensible for multiclass, dynamic class problem like this, so did not use..

Additional parallel corpora (from authors besides Shakespeare) with scalar rather than binary gold labels would undoubtedly guide use towards a better performing evaluator. Unfortunately, these parallel style corpora are virtually nonexistent.

Cherry-picking intuition-guided features to engineer adds a subjective element to our evaluator, which is responsible for loss in accuracy, to a degree that is difficult to measure precisely. Doubtless, additional features exist that would improve accuracy.

C. Acknowledged shortcomings

Assumption of evenly distributed class (style) priors. This is not a portable assumption.

The way forward will be to either prove evaluator success with heuristically seeded priors, or require the injection of class priors from each project's dataset.

We only used one dataset, the Shakespeare dataset. It is simplistic to assume that Sparknotes data resembles a single human author's. It probably resembles other business objectives, like truncation.

VIII. Future work

[TODO: expand] Expand approach to handle passages larger than single sentences
What is the class prior for a style? Can we pick a meaningful value?

How do we deal with fluidity and overlapping between styles, or shifting style boundaries?

We assumed Style-author hypothesis: That an author always writes in the same style. This is probably not always the case. Do we tackle this issue by including author identification as a sub-metric in a future evaluator?

...

IX. Conclusions

APPENDIX

...

REFERENCES

[1] example of bib item
[2] example of bib item
[3] example of bib item
[4] example of bib item
[5] example of bib item
[6] example of bib item
[7] example of bib item
[8] example of bib item