Evaluating Style Transfer

Peter W. Kong

***Abstract*— This project formalizes the problem of evaluation of intra-lingual sequence to sequence style transfer models and explores improvements on the state of the art evaluation metrics.**

I. Introduction

Style transfer is the transfer 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, a companion research problem, has proven utility in applications like author identification [Houvardas]. 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, to name a few.

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

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] 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]. Hu reaches about 83% sentiment classification accuracy, although performance varies greatly with datasets used. Hu's metric is label accuracy using external (i.e. blackbox) sentiment classifiers.

[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 style-agnostic precursor representation. This paper confirms that BLEU and ROUGE metrics are not suitable for capturing meaning.

III. Criticism of existing style transfer metrics

In this section we comment on existing metrics used to evaluate style transfer (or similar tasks).

- Perplexity and BLEU
  Perplexity, according to [Li, et al.] is the the normalized inverse probability of a test set, formalized as:

  $$Perplexity(W) = \frac{1}{P(w_1 ... w_n)}^{\frac{1}{N}}$$

  This metric is quite intuitive, simple, and portable. However, it suffers significant failures when evaluating across styles that use different words or word orderings. Consider this example inter-style sentence alignment:

  (Reference R) Mary ate her cheeseburger.
  (Candidate A) A meal was eaten by Mary.
  (Candidate B) Mary said that Bob ate his cheeseburger but I think she's lying.

  Assume, reasonably, that (R) and (A) are both drawn from the same style corpus, and (B) is not: (R) and (A) are both concise and factual, while B uses indirect voice and a narrative effect.

  Candidate (A) will be heavily penalized because P(meal), P(was), P(A), and P(by) are all equal to zero. The perplexity of (A) with respect to (R) will be high, since the metric accounts for specific words, not semantics or style. In contrast, perplexity will

value (B)'s word overlap with (R), even though the overlapped words "ate" and "cheeseburger" are attached to different subjects.

For the same reasons, the BLEU score incorrectly penalizes candidates that express the same style and concepts, but with different words.

- Neural Models
  As mentioned in Section II, neural net classifiers are heavily used in sequence to sequence evaluation (including in style transfer), but they suffer from poor explanatory power due to their composition of usually immense weight vectors that are not individually attached to human intelligable concepts. [Li, et al.] and others have attacked this problem with some success. However, achievements are still modest and tend to explain predictions row-by-row, rather than explaining the classifier as a whole.

- Manual evaluation
  Manual evaluation in theory provides great results, and indeed the majority of research in style transfer employs it so some degree. However, manual evaluation is obviously intractible for large amounts of training data. It is also subjective with respect to the human evaluators.

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}) will output a score.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.

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 LSTM classifiers [Fu, et al., Hu, et al., Prabhumoye] 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 drawn from Coco Xu's Shakespeare dataset [Xu]. 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'

There are 42,192 parallel sentences in total. The topics encompass a wide variety of content and sentiment: comedies, tragedies, and histories. 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 perceived 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, thus generating a classical supervised learning dataset with ”gold” labels.

This dataset was randomly shuffled by sentence pair, preserving style-to-style alignments. 80% of corpus was designated the train set, 10% the validation set, and 10% the holdout test set.

We experimented with a variety of models. However, the choice of experiments was guided by our evaluator design ideals (see previous section).

B. Early Exploration - NER and BLEU

We attempted to use Named Entity Recognition (NER) match scores as a subm-etric, but found that they provided almost no predictive power even between content-aligned inter-style sentence pairs. And of course, using NER in a non-aligned context reduces portability. Furthermore, even if NER could be configured to yield predictive power, it adds a distracting content signal into the evaluator, violating one of our ideal traits.

We wanted to employ BLEU scoring as a sub-metric. BLEU scores similar documents closer to ”1”, and dissimilar documents closer to ”0”. To crudely abstract content from style in the Shakespeare dataset, we aggregated BLEU scores on randomized sentence pairs in three experiments: pairs drawn from Style A, pairs drawn from Style B, and pairs drawn from Styles A and B.

We hypothesized that if BLEU scores computed from same-style pairs were better than inter-style, BLEU might differentiate between styles agnostic to content. Disappointingly, BLEU was not up to the task:

TABLE II
CONTENT-AGNOSTIC BLEU SCORING

Same-style, A	.016
Same-style, B	.021
Inter-style	.017

C. Feature Engineering

We used a total of of 1645 features in our final model, all human-intelligible.

Drawing on insight from Strunk and White’s ”The Elements of Style”, we tracked the counts of the following punctuation marks in each sentence of each style:

: ; ’ (... !

We constructed the remainder of the features from POS tagging and dependency parsing artifacts. 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. Example:

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

yields the mapping shown in Table III.

TABLE III
POS TAGGING EXAMPLE

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

We also used adjective-noun ratios and adverb-verb ratios as separate features.

The last grouping of features, representing the majority, were one-hot POS sequence features. For example, this sentence:

”A sleep that ends all the heartache.”

would result in the feature ”NOUN_VERB_NOUN” acquiring the value 1, and the feature ”VERB_NOUN” acquiring the value 0. Note that we ignored all parts of speech except nouns, verbs, and conjunctions. We did this in order to capture as much signal from the ”core” tag-shape of a sentence while avoiding rare or singleton features that result from accounting for ”non-core” parts of speech. We also truncated sequence features to a cardinality of 7 to prevent feature explosion. We found no accuracy improvement past this cardinality.

D. Models

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

- Multinomial Naive Bayes
  MNB 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 100%-weighted probability values.
- SVM
  SVM also has the ability to predict probabilities and

handles reasonably large numbers of one-hot features well.

VI. Results

We show the two best performing models in the Model Performance table.

TABLE IV
MODEL PERFORMANCE

Model	Validation Accuracy	Test (holdout) Accuracy
SVM, all features, kernel=linear, reg. penalty=.1, all plays	.643	.618
Naive Bayes, all features, alpha=.1, all plays	.607	.607

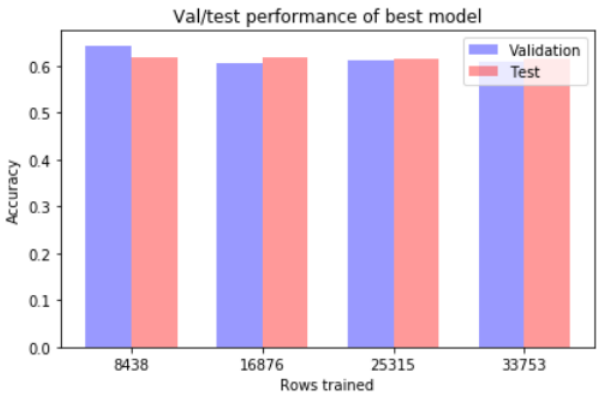
VII. Analysis

We found that SVM, with all final features present, yielded the best classification accuracy across both models attempted. We used grid search to tune hyperparameters: regularization penalties and feature selection on both models, and RBF/linear/poly kernels on SVM in particular.

For SVM, test set accuracy hardly diverged from validation set accuracy. Thus, we are confident that overfitting is not a serious issue. We attribute the minimal differential to a large dataset and use of regularization within the SVM model.

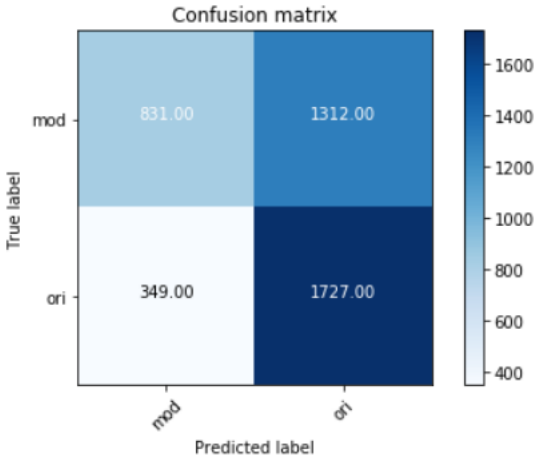
A. Error Analysis

While we initially trained and tuned both models on the entire dataset, we found that both test and validation accuracy peaked when using about 25% of the dataset. Test (holdout) accuracy tracked validation accuracy very closely, again indicating that our ample training data minimized overfitting.



Evaluators should be quite accurate if they are to judge other models. Unfortunately, at 64.3% accuracy we consider this evaluator a failure.

The confusion matrix below shows that the evaluator made most of it's mistakes in misidentifying 'modern' sentences as 'original'.



Without additional parallel style data, it is difficult to anticipate how the model will perform on non-Shakespearean datasets.

We estimate that a significant amount of accuracy loss is attibutable to the binary nature of the dataset gold labels. 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.

B. Hyperparameter tuning

After experimenting with using RBF and polynomial kernels, we found that a linear kernel yielded the best result.

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, SVM with linear kernel, modest regularization, and full 1645 feature set underperformed our intuition-driven expectations (Table IV).

C. Acknowledged shortcomings

We assumed evenly distributed class (style) priors for this experiment. This assumption fits the Shakespearean dataset, but is not expected to be portable to other datasets. 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, or avoid using Bayesian models that depend on class priors.

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

VIII. Future work

There are opportunities to extend this work in many different directions. We enumerate below the opportunities that we feel will be most fruitful.

- We formulated the evaluation problem with the assumption that the evaluator can accept any style transfer model M_{AB} between any two styles. Scope and dataset contraints limited us to only two styles. We set up our experiment with a multi-class extension in mind. All models and programs can be used for multiclass prediction. However, until we actually run the evaluator on multiple classes of styles, we risk overfitting our evaluator on only two styles.

- We assumed the "style-author" hypothesis: That any given author always writes in the same style. This is certainly not always the case - counterexamples are legion. An extension to the experimental setup would be to use unsupervised clustering techniques to identify stylistic drift within an author's oeuvre before using his/her data as training data. If stylistic drift is found to be common and significant, it may be necessary to programmatically differentiate multile styles within an oeuvre.

- Related to the above extension, using a black box author intentification system as a submetric for the evaluator may boost evaluator performance.

- Gains in evaluator perfomance may be found when using larger-than-sentence sequences. Most known work in style transfer occurs at the sentence level. Intuitively, stylistic signal should be attainable from larger sequences. For example, irony may only be detected in a sentence because of context from an earlier sentence.

IX. Conclusions

We hope that we have motivated the need for a better style transfer metric, and that our problem formalization will be helpful in future work. This project also revealed that style evades rigorous rules and definition. Future work may benefit from defining "style" more technically and narrowly than the broad, colloquial definition assumed in this paper.

Unfortunately, our strategy of maximizing explanatory power through human-intelligible engineering features resulted in disappointing accuracy results in our best performing model, a Support Vector Machine classifier.

References

[1] [Li, et al.] Understanding Neural Networks through Representation Erasure. Jiwei Li, Will Monroe and Dan Jurafsky. https://arxiv.org/pdf/1612.08220.pdf

[2] [Ficler] Controlling Linguistic Style Aspects in Neural Language Generation. Jessica Ficler and Yoav Goldberg. https://arxiv.org/pdf/1707.02633.pdf

[3] [Fu] Style Transfer in Text: Exploration and Evalution. Fu, et al. https://arxiv.org/abs/1711.06861

[4] [Hu] Toward Controlled Generation of Text. Hu, et al. https://arxiv.org/abs/1703.00955

[5] [Potthast] Overview of the Author Obfuscation Task at PAN 2018: A New Approach to Measuring Safety. Potthast, et al. http://ceur-ws.org/Vol-2125/invited_paper_16.pdf

[6] [Prabhumoye] Style Transfer Through Back-Translation. Prabhumoye, et al. https://arxiv.org/abs/1804.09000

[7] [Ruseti] Authorship Identification Using a Reduced Set of Linguistic Features. Ruseti, et al. https://www.uni-weimar.de/medien/webis/events/pan-12/pan12-papers-final/pan12-author-identification/ruseti12-notebook.pdf

[8] [Kingma 2014] Auto-Encoding Variational Bayes. Kingma, et al. https://arxiv.org/abs/1312.6114

[9] [Houvardas] N-Gram Feature Selection for Authorship Identification. Houvardas. http://www.icsd.aegean.gr/website_files/diplomatikes/msc/561623167.pdf

[10] [Xu] Paraphrasing for Style. Xu, Wei and Ritter, Alan and Dolan, Bill and Grishman, Ralph and Cherry, Colin. http://www.aclweb.org/anthology/C12-1177

[11] The Elements of Style. William Strunk, Jr., and E.B. White. Allyn and Bacon (1999): Boston, United States.