Evaluating Style Modification in Text

by

Remi Mir

B.S., Massachusetts Institute of Technology (2017)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2018

Author .....
Department of Electrical Engineering and Computer Science
May 25, 2018

Certified by .....
Associate Professor Iyad Rahwan, Thesis Supervisor
May 25, 2018

Accepted by .....
Katrina LaCurts, Chair, Masters of Engineering Thesis Committee

Evaluating Style Modification in Text

by

Remi Mir

Submitted to the Department of Electrical Engineering and Computer Science
on May 25, 2018, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

In this thesis, we identify best practices for evaluating style modification, or style transfer, for text. Research of style transfer is bottlenecked by a lack of standard evaluation practices. We define three key aspects of interest (style transfer intensity, content preservation, and naturalness) and show how to obtain more reliable measures of them from human evaluation than in previous work. We also demonstrate stronger correlation between human judgment and a new set of automated metrics: the Wasserstein distance, word mover's distance on texts with style masked out, and adversarial classification for the respective aspects. Lastly, we illustrate aspect tradeoff curves for three state-of-the-art style transfer models to highlight the importance of evaluating style transfer models at specific points on the curves. This can enable direct comparison of the models, facilitating future research in style transfer.

Thesis Supervisor: Associate Professor Iyad Rahwan

Acknowledgements

I would like to express my utmost gratitude to the following people:

- My supervisors, Bjarke Felbo and Iyad Rahwan, for the chance to join the lab and for guidance and support of this work. I felt incredibly encouraged to go beyond and submit my first conference paper, with helpful feedback and suggestions.

- Juncen Li, Tianxiao Shen, and Junbo (Jake) Zhao for help in using their text modification models, which are markers of major progress in the area of style transfer research and without which this work would not have been possible.

- Nick Obradovich for conversations that opened up new dimensions of exploration for this thesis, and for helping me approach problems with a different lens. I will also never look at the sentence "I love beanbags" in the same way again.

- All members of the Scalable Cooperation group for creating a vibrant, fun environment in which to work and collaborate. There is probably no other lab like this one, and I am lucky to have been a part of it.

- My family, especially my mum, for giving all the love and support one could possibly give and more, and for lending an ear, whether at 3AM or 3PM. She is my gem.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Style transfer in text is a growing area of research in natural language processing. Style transfer is the task of changing a single attribute (style) of an input, while retaining non-attribute related content of the original input. In this chapter, we define style transfer in more detail. We then define three key aspects by which to evaluate the outputs of a style transfer model: style transfer intensity, content preservation, and naturalness.

Because various papers evaluate transfer models with different techniques of evaluation, there is a need for a set of standard evaluation metrics and tools that improve upon what currently exist. In Chapter 2, we review what is used in practice, and in Chapter 3, we discuss the approaches for developing this collection of improved metrics, for both human evaluation and automated evaluation of the aspects of interest. In Chapter 4, we analyze the effectiveness of this collection, with respect to currently used metrics. We also illustrate aspect tradeoff plots that result from using the best correlated automated metrics to evaluate different style transfer models, and discuss how we can directly compare the models using specific points on the curves. Finally, Chapter 5 assesses our work with respect to previous approaches and proposes ideas for future work.

1.1 Definition of Style Transfer

Exploration of style transfer in general began in the field of computer vision in 2015, with the application of artistic styles, such as Salvador Dalí's surrealism, to images using convolutional neural networks [11]. These neural networks were trained to take in a "content" image and a "style" image (e.g. a picture of a cat and *The Persistence of Memory*), separate and extract

style and content features, and apply the style to the content image. Style itself can comprise of multiple elements to varying degrees, including brush stroke, color palettes, and use of shapes. Style transfer in images and video has largely been assessed with model-dependent approaches (content and style losses associated with the same model used for training), as well as subjective or qualitative approaches [12, 14]. While research with respect to images has seen the emergence of several transfer models, introducing new capabilities such as multi-style transfer to the same content image [40], the problem of separation and the definitions of content and style remain open-ended.

Style transfer in text poses similar challenges in term of defining content, style, and the degree of separation between the two. While content can be understood as the semantic meaning of a text, style does not lend itself to being as easily defined, but has been interpreted as the social meaning of a text [8]. It can be embodied in several forms, including but not limited to formality [21, 26], sarcasm [17], and binary or multinary sentiment [7]. In particular, previous work has modified text to make it more negative in sentiment [32], related to a different topic [42], and humorous [23]. Like style in images and video, style in text can comprise of multiple, possibly subjective factors, such as varied use of punctuation, repetition of certain words for emphasis, and diction (slang, abbreviations, etc.).

Unlike style transfer for vision-related tasks, style transfer for text is still in the early stages. Text is more challenging as it is a discrete form of data, precluding the extraction of style and content features in the same manner as in the computer vision domain. Some transfer models modify text by manipulating their latent representations [10, 32, 42], whereas others identify and replace style-related words directly [23]. Regardless of approach, they are difficult to compare as there is currently no standard set of evaluation practices, nor a clear definition of which particular aspects to evaluate.

1.2 Aspects of Evaluation

We consider three aspects of interest on which to evaluate the output text x' of a style transfer model, potentially with respect to an input text x:

1. *style transfer intensity* $STI(SC(x), SC(x'))$ quantifies the difference in style, where $SC(\cdot)$ maps inputs to a style distribution

2. *content preservation* $CP(x, x')$ quantifies the relative difference in content between the input and the output

3. *naturalness* $NT(x')$ quantifies the degree to which the output appears as if it could have been written by humans

Style transfer models should be compared across all three aspects to properly and comprehensively characterize their differences. For example, given a style of binary sentiment, it is insufficient for a style transfer model to simply change the sentiment of a text from positive to negative (and vice versa). If, for example, it also alters objective words like the names of places, it does not preserve content well. If it repeats certain words such as "the" needlessly, it does not appear to be natural, by our aforementioned definition.

Conversely, a model that overemphasizes text reconstruction would yield high content preservation and possibly high naturalness, but little to no style transfer, especially if it leaves the inputs largely unchanged. Consequently, all three aspects are critical in a style transfer evaluation system. Current approaches for evaluating these and related aspects are reviewed in the next chapter.

Chapter 2

Related Work

We focus on three models: the cross-aligned autoencoder (CAAE) [32], the adversarially regularized autoencoder (ARAE) [42], and the delete-and-retrieve (DAR) model [23]. They have been evaluated with combinations of human and automatic evaluation (summarized in Table 2.1) that employ different instructions and scales. Reproducing and comparing evaluation scores are thus difficult, time-consuming tasks. Additionally, without knowing which metrics have the strongest correlations with human judgments in the style transfer setting, one cannot conclude that the current automated metrics, described below, are the most appropriate to use. Note that some of them rely on training models on the corpus of input sentences, X, and/or the corpus of output sentences, X'.

2.1 Style Transfer Intensity

The current approach for evaluating this aspect is training a classifier on X and corresponding style labels, and measuring the ratio of X' that is classified as having the target style [23, 32, 42]. Because the same classifier model is not used across evaluations, resulting ratios using this *target style probability* approach are not directly comparable.

2.2 Content Preservation

BLEU, used widely in machine translation [25], has also been used in the style transfer setting [23, 42]. Measured on a scale of 0 to 1, higher scores indicate more similarity between candidates (output texts) and references (input texts). BLEU is based on the number of

n-gram matches between candidates and references (clipped by the number of reference n-grams), divided by the total number of candidate n-grams, according to the following modified precision formula [25]:

$$p_n = \frac{\sum_{x' \in X'} \sum_{ngram \in x'} Count_{clip}(ngram)}{\sum_{x' \in X'} \sum_{ngram \in x'} Count(ngram)} \tag{2.1}$$

In the machine translation setting, BLEU is often used to compare candidates with multiple reference translations, as it has been shown to have high correlation with human judgment, particularly with an n-gram where $n = 4$ [25]. The same cannot be said in the style transfer setting, as there is only one "reference" (x) per candidate (x'), and its correlation in the style transfer setting is unknown.

Another issue is that using BLEU directly to evaluate content preservation is inconsistent with the aim of style transfer, which is not to entirely reconstruct text, but to alter style words while preserving the remaining semantics. BLEU penalizes the necessary difference between X and X'.

Current evaluations also do not employ a standard implementation of BLEU. Multiple implementations may have different evaluation settings, e.g. the value of n for the considered n-grams or sentence-level smoothing for any value of n that yields no n-gram overlap between candidates and references [13, 19]. Ambiguity in these parameter settings prevents direct comparison of BLEU scores reported by different papers.

2.3 Naturalness

Current evaluations of naturalness rely on human ratings on a variety of scales under different names: fluency/readability [32, 42], grammaticality [23], and naturalness itself [42].

An issue with measuring grammaticality is that text with proper syntax can still be semantically nonsensical, e.g. "Colorless green ideas sleep furiously" [6]. Furthermore, input texts such as the Yahoo! Q&A dataset [41] may not demonstrate correct grammaticality, despite being written by humans and thus being natural by our definition in Section 1.2. This undermines the effectiveness of grammaticality measures for transfer texts.

While existing evaluations of naturalness have largely depended on human workers, Zhao et al. use perplexity to automate evaluation on a corpus level. To obtain perplexity PPL,

	Style transfer intensity			Content Preservation			Naturalness	
	HRC(x')	HRD(x')	SC(x')	HRC(x,x')	HRR(x,{x'})	BLEU(x,x')	HRC(x')	PPL(x')
CAAE		x	x		x		x^F	
ARAE		x	x	x		x	x	x^F
DAR	x		x	x		x	x^G	

Table 2.1: Summary of current evaluation methods. HRC is human rating on a continuous scale (e.g. 1 to 5), HRD is human rating on a discrete scale (e.g. positive/negative), and HRR is human ranking. $\{x'\}$ is the set of x' from models trained on different parameters. SC is a style classifier. PPL is perplexity. Superscripts denote that the evaluation is done for fluency (F) or grammar (G), which we count as a subset of naturalness.

a language model LM is first trained on X. Perplexity is based on the log likelihood of texts, normalized by m, the number of words in the vocabulary of X (Eq. 2.2 and 2.3). The higher the probability of a text t, according to LM, the lower the perplexity (Eq. 2.3).

$$p(t) = p(t_1)p(t_2|t_1)p(t_3|t_1,t_2)\dots p(t_m|t_1,\dots t_{m-1}) \tag{2.2}$$

$$PPL(t) = 2^{-\frac{1}{m}log_2(p(t))} \tag{2.3}$$

Low perplexity signifies less uncertainty of which words can be used to continue a sequence, pointing to a language model's ability to predict gold texts [5, 39]. However, because transfer outputs are not gold standard, perplexity is not necessarily a valid proxy for quantifying naturalness. Furthermore, the correlation between sentence-level perplexity and human judgments of transfer texts is unknown in the style transfer setting.

Chapter 3

Methods

In this chapter, we describe the methods we use to obtain human evaluations and calculate automated metrics. Experiments confirming the effectiveness of these methods are detailed in the next chapter.

3.1 Human Evaluation

Existing work from fields outside NLP has demonstrated that human raters provide more accurate scores when asked to evaluate items in relation to each other, compared to when they are asked to score a single item on an absolute scale [2, 35]. We use these findings to devise potentially more reliable forms of human evaluations with *relative scoring* instead of *absolute scoring*, particularly for the aspects of style transfer intensity and naturalness.

3.1.1 Style Transfer Intensity

The ratio of outputs with a target style does not provide information about the degree of style transfer for a given sentence. For example, if style is binary sentiment, "I love you" could be turned into "I like you", "I dislike you", or "I hate you" by a style transfer model, each possibility exhibiting more negative sentiment than the next. To capture this more nuanced behavior, we ask workers to perform a relative scoring task: rate the difference in style between x and x', on a scale of 1 (identical styles) to 5 (completely different styles). This differs from the absolute scoring of existing work, which asks raters about the degree to which x' displays the target style (Table 2.1).

3.1.2 Content Preservation

For content preservation, we consider the difficulty of asking raters to ignore style-related words as done in [32]. Because not all workers may identify the same words as stylistic, they may potentially vary substantially in their evaluations. We run experiments without this variable by asking workers to evaluate content preservation on the same texts, but where we have masked style words using a lexicon of style words (see A.1 for details). With this new *masking* task, workers only rate text similarity, without also having to take style into account. We use the same scale as is used by existing evaluations: 1 (not similar at all) to 5 (very similar).

3.1.3 Naturalness

Similar to style transfer intensity, human evaluations of naturalness only consider x'. In contrast, we ask workers to choose which of x and x' is more natural. Assuming that inputs, even if not completely grammatical or coherent, are produced by humans and are thus natural by definition (1.2), an x' that is marked as more natural by a worker indicates some success on the part of the style transfer model because it is able to fool the worker.

3.2 Automated Evaluation

In this section, to address the issue of costly and time-consuming human scoring, we propose a set of automated metrics for evaluating the aspects of interest.

3.2.1 Style Transfer Intensity

To capture potentially nuanced changes in the style distributions of x and x' (3.1.1), we use the Wasserstein distance (also known as earth mover's distance) [27, 28, 30]: $WD(SC(x), SC(x'))$. For the style classifier SC, we train text-CNN [18, 22] and fastText [15]. WD itself can be interpreted as the minimum cost to turn one distribution into the other. We use an identity matrix of size N, where N is the number of style classes, as the distance matrix used in determining the total cost. Distances of 1 on the diagonal represent the measurement of changes only with respect to the same style class in $SC(x)$ and $SC(x')$, for each style class.

Because WD is able to handle any number of classes in the distributions, this method is scalable and can easily be applied to datasets with more than two style classes.

Modification	Texts	Modified Bigram Score (p_2)
None	x: "I love *ya ,* tomorrow !" x': "I hate *ya ,* today !"	$\frac{1}{5}$
Style Removal	x: "*I ya ,* tomorrow !" x': "*I ya ,* today !"	$\frac{2}{4}$
Style Masking	x: "*I <masked> ya ,* tomorrow !" x': "*I <masked> ya ,* today !"	$\frac{3}{5}$

Table 3.1: Effects of text modification on p_2 values (Eq. 2.1) that contribute to BLEU scores. The style is binary sentiment and the style words include "love" and "hate". Italicized words represent overlap between x and x'.

3.2.2 Content Preservation

Similar to introducing masking to obtain potentially improved forms of human evaluation (3.1.2), we subject texts to style removal and style masking, which are different types of modification prior to calculating content preservation (see examples in Table 3.1).

For style removal, we delete style words from both x and x' using a style lexicon (see A.1 for details on lexicon construction). For style masking, we instead replace the style words with a *<masked>* placeholder token.[1] Style masking acknowledges the notion that style is not entirely disjoint from content and can contribute to the semantic meaning of a sentence. For example, if the style is binary sentiment, "love" acts as both style and content because it is a verb and contributes to sentence structure. If a style transfer model outputs "I you" for the input text "I love you," the change in sentiment comes at the expense of content, which must be accounted for.

With text modification prior to calculating content preservation metrics, we address the undesired penalization of those metrics on texts that are expected to demonstrate changes due to style transfer (Table 3.1). The metrics we test are BLEU-4 and those not previously used in style transfer evaluation: METEOR [1] and word embedding metrics, including Word Mover's Distance (WMD) [20]. Details about these metrics can be found in A.2.

[1]Another approach to style masking is to use more fine-grained placeholders, such as *<verb>* and *<adverb>*. However, labeling the texts with part-of-speech tags would require them to have proper syntax, which is not always feasible. Many transfer texts are noisy, often containing repeating words, e.g. "I hate hate this this restaurant" that do not allow for effective part-of-speech parsing.

3.2.3 Naturalness

From the perspective of automation, a method evaluating naturalness should be exposed to human-sourced texts in X in order to have a baseline understanding of what is considered natural. For each transfer model, we train adversarial unigram classifiers and neural classifiers [4] on X and X'. An adversarial classifier AC essentially performs a Turing test [37], in that it must distinguish natural texts (seemingly written by humans) from texts generated by models. The more natural x' is, the more likely it is to fool AC [16].

Chapter 4

Experiments

To compare our evaluation practices with existing ones, we use the CAAE, ARAE, and DAR models. We train each model across a range of parameters to construct the aspect tradeoff plots and to examine the performance of metrics across different settings. We use the same Yelp restaurant review dataset as in [32], where the original reviews are split into individual sentences. Those given ratings of 3 or higher are marked as positive in sentiment, while those below are marked as negative.

4.1 Style Transfer Models

For each of the CAAE, ARAE, and DAR models, we choose a wide range of training parameters to allow for variation of content preservation, and indirectly, of style transfer intensity, in X.

CAAE uses autoencoders [38] that are cross-aligned, under the assumption that texts already share a hidden content distribution [32]. It transfers styles of non-parallel texts, using hidden states of the RNN and multiple discriminators to align distributions of the x' in X' exhibiting one style with the distributions of the x in X that exhibit another. It makes use of adversarial components to separate style information from the latent space where inputs are represented. For our experiments, we train the CAAE model on different values (0.01, 0.1, 0.5, 1, 5) of ρ, a weight on the adversarial loss.

CAAE is used as a baseline for later style transfer models. One such model is the ARAE, which has a separate decoder per style class [42]. We evaluate outputs of ARAE models trained on different values (1, 5, 10) of γ. This parameter has a role similar to that of ρ in

Model	Text Modification	
	None	Style Masking
CAAE	0.158	**0.289**
ARAE	0.201	**0.321**
DAR	0.161	**0.281**
Average	0.173	**0.297**

Table 4.1: Fleiss' kappa scores for human ratings of content preservation of unmasked and style masked Yelp outputs.

CAAE because ARAE also takes an adversarial approach to separate style and content.

The third model that we evaluate, which also uses CAAE as a baseline, avoids adversarial methods altogether and employs an entirely different approach, called Delete-and-Retrieve (DAR) [23]. It identifies and removes style words from texts, searches for related words pertaining to a new target style, and combines the de-stylized text with the search results using a neural model. We train all DAR models on a single γ value of 15, a threshold parameter that influences the maximum number of style words that can be removed from texts, with respect to the size of the corpus vocabulary. For this single training value, we experiment with a range of γ values (0.1, 1, 15, 500) on test time because the model does not need to be retrained [23].

4.2 Results

4.2.1 Human Evaluation

We use Fleiss' kappa κ of inter-rater reliability [9] to identify the more effective human scoring task for different aspects of interest. We observe that kappa scores for ratings on style-masked texts are greater than those on unmasked texts, demonstrating that tests with masking enable more reliable human evaluation of content preservation (Table 4.1).

Similarly, for the aspect of naturalness, we find that all kappa scores for the relative scoring tasks exceed those of the absolute scoring tasks (Table 4.2). Despite the two types of tasks having different numbers of categories (2 vs 5), we are able to compare them by using a threshold to bin the absolute score for each x into a "natural" group or an "unnatural" group. These correspond to the two groups in relative scoring, where x' is considered to be more natural than x, and where x is considered to be more natural than x'. For example,

Model	Absolute		Relative
	Threshold = 3	Threshold = 2	
CAAE	0.193	0.321	**0.579**
ARAE	0.215	0.415	**0.741**
DAR	0.103	0.201	**0.259**
Average	0.170	0.312	**0.526**

Table 4.2: Fleiss' kappa scores for human ratings of Yelp outputs in relative and absolute naturalness tasks.

Model	fastText		text-cnn	
	Target Style Probability	Wasserstein	Target Style Probability	Wasserstein
CAAE	0.566 ± 0.038	**0.587 ± 0.037**	0.587 ± 0.037	**0.602 ± 0.036**
ARAE	0.513 ± 0.053	**0.519 ± 0.053**	0.515 ± 0.053	**0.522 ± 0.053**
DAR	0.470 ± 0.049	**0.536 ± 0.045**	0.508 ± 0.047	**0.565 ± 0.043**
Average	0.516 ± 0.047	**0.547 ± 0.045**	0.537 ± 0.046	**0.563 ± 0.044**

Table 4.3: Correlations of automated metrics for style transfer intensities with human ratings of differences in style between Yelp input and output texts.

given the absolute scale of 1 to 5, a threshold of 2 places texts for which absolute scores are greater than or equal to 2 into the natural group. In examining multiple thresholds (2 and 3), we again find that the kappa scores for relative tasks consistently exceed those of absolute tasks. Therefore, if researchers prefer to execute future human evaluations of naturalness, the relative scoring paradigm is preferable as it demonstrates greater inter-rater reliability.

4.2.2 Automated Evaluation

For each aspect of interest, we compute Pearson correlations between scores from the existing metric and scores from human raters. We do the same for our proposed metrics, in order to identify which metric is more reliable for automated evaluation of the given aspect. To ensure statistically significant differences between correlation coefficients, we use Chi square testing. We calculate standard error using the standard deviation of correlation coefficients, as formulated in [3].

Style Transfer Intensity

For both the fastText and the text-CNN style classifiers, the Wasserstein distance between style distributions of x and x' has higher correlations with human scores than target style

Model	BLEU	METEOR	Embed Avg	Greedy Match	Vector Extrema	WMD
CAAE	0.458 ± 0.044	**0.498 ± 0.042**	0.370 ± 0.048	0.488 ± 0.043	0.496 ± 0.042	0.496 ± 0.042
ARAE	0.337 ± 0.064	0.387 ± 0.062	0.313 ± 0.065	0.419 ± 0.060	0.423 ± 0.060	**0.445 ± 0.058**
DAR	0.440 ± 0.051	0.455 ± 0.050	0.379 ± 0.054	0.472 ± 0.049	0.472 ± 0.049	**0.484 ± 0.048**
Average	0.412 ± 0.053	0.447 ± 0.051	0.354 ± 0.056	0.460 ± 0.051	0.464 ± 0.050	**0.475 ± 0.049**

Table 4.4: Absolute correlations of content preservation metrics with human scores on Yelp texts with style removal.

Model	BLEU	METEOR	Embed Avg	Greedy Match	Vector Extrema	WMD
CAAE	0.488 ± 0.043	0.517 ± 0.041	0.356 ± 0.049	0.490 ± 0.043	0.496 ± 0.042	**0.517 ± 0.041**
ARAE	0.356 ± 0.063	0.374 ± 0.062	0.302 ± 0.066	0.405 ± 0.061	0.422 ± 0.060	**0.457 ± 0.057**
DAR	0.444 ± 0.050	0.454 ± 0.050	0.370 ± 0.054	0.450 ± 0.050	0.473 ± 0.049	**0.475 ± 0.049**
Average	0.429 ± 0.052	0.448 ± 0.051	0.343 ± 0.056	0.448 ± 0.051	0.464 ± 0.050	**0.483 ± 0.049**

Table 4.5: Absolute correlations of content preservation metrics with human scores on Yelp texts with style masking.

probabilities. In terms of classifiers, text-CNN exhibits stronger correlation across all experiments (Table 4.3).

Content Preservation

METEOR, which has been shown to have higher correlation with human judgment than BLEU[1] in the machine translation setting [1], exhibits the same relationship in the style transfer setting (Tables 4.4 and 4.5). However, of all the content preservation metrics, WMD generally shows the strongest absolute correlations with human scores in both text modification settings (style removal and style masking) (Tables 4.4 and 4.5). Because WMD is lower when texts are more similar, making it anti-correlated with human scores, we take absolute correlations to facilitate comparison with correlations from the other content preservation metrics.

With respect to text modification, style masking may be more suitable than style removal because, on average for the WMD scores, masking yields higher correlations with human ratings.

[1]One reason BLEU may not perform as well as in the machine translation setting is that there is only one reference text per candidate text for style transfer. Generally, several reference texts, as is usual in machine translation tasks, increase the chance of n-gram overlap (such as $n = 3$ or higher) with the candidate. Having a single reference text reduces this likelihood and the overall effectiveness of the metric.

Model	Unigram Adv. Clf.		Neural Adv. Clf.	
	Absolute	Relative	Absolute	Relative
CAAE	72.5	**77.0**	73.0	**78.7**
ARAE	49.5	**66.7**	50.7	**67.9**
DAR	65.2	**65.6**	61.1	**62.3**
Average	62.4	**69.8**	61.6	**69.6**

Table 4.6: Percent agreement between adversarial classifiers and human raters on which of the input and output texts is more natural, for both relative and absolute naturalness scoring tasks.

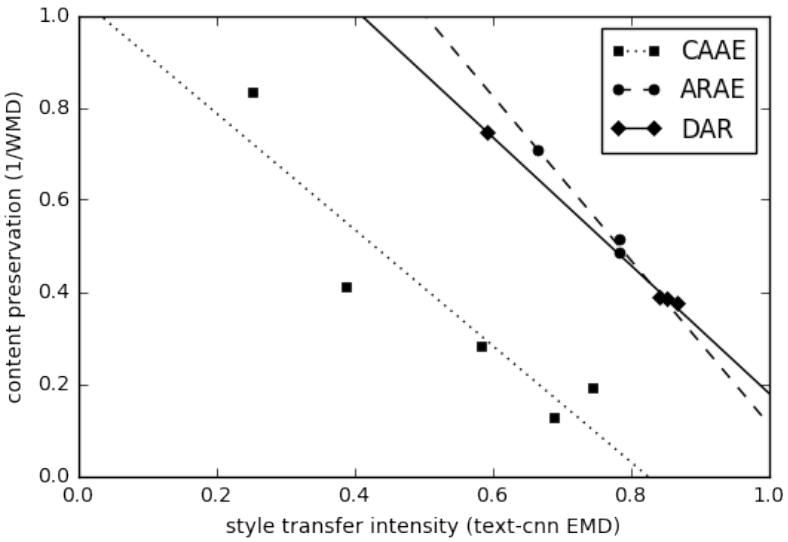

Figure 4-1: Tradeoff plots between style transfer intensity and content preservation.

Naturalness

For the adversarial classifiers, neither feature set (unigram or neural) yields particularly stronger agreements with human scores than the other. However, on average, there is greater agreement on which texts are more natural between the adversarial classifier and relative scorers than between the classifier and absolute scorers (Table 4.6).

We also trained language models on X and obtained sentence-level perplexities for each text in X', but found no significant correlation with human scores ($p = 0.05$). These results indicate that adversarial classifiers can be a more useful tool for automating the measurement of naturalness.

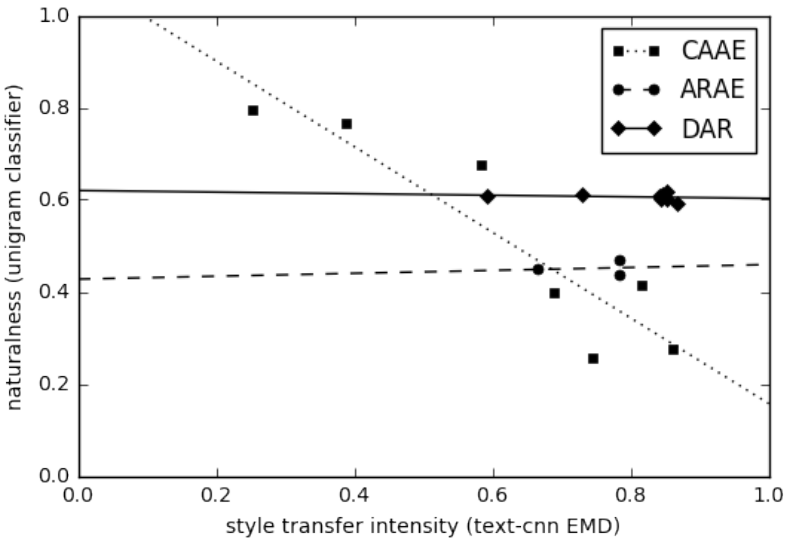

Figure 4-2: Tradeoff plots between style transfer intensity and naturalness.

4.2.3 Aspect Tradeoff Plots

We construct a tradeoff plot[2] for each style transfer model, using scores from the metrics that best correlated with human judgment: WD for style transfer intensity, WMD for content preservation, and adversarial classification for naturalness. Because scores for WMD are lower when texts are more similar, we take the normalized inverse scores to measure the degree of content preservation on a 0 to 1 scale.

Given that each point on the plots represents a different parameter setting on which a style transfer model was trained (Section 4.1), the plots allow us to visualize model behavior across parameters. We observe a clear trend of reduction in content preservation as style transfer intensity increases (Figure 4-1). We also observe that the CAAE model has a substantial decrease in naturalness as style transfer intensity increases ($p<0.01$), whereas the other models do not (Figure 4-2).

The tradeoff plots are also useful for directly comparing one model with others at specific points. If multiple models are able to achieve similar degrees of two different aspects, e.g. style transfer intensity and content preservation, this may be represented on the curves as points of intersection. Experiments that aim to evaluate multiple models can make use of the training parameters that yielded those points. Note that these types of comparisons can

[2]Concurrent work on the DAR model also examines tradeoffs, but only for style versus content. Specifically, it only focuses on producing curves for variants of its own model, not for other models, like CAAE and ARAE [23].

only be made with respect to the same dataset, as it is expected that the tradeoff plots may vary with dataset.

Chapter 5

Conclusion

Previous work on style transfer models used a variety of evaluation methods (Table 2.1), making it difficult to compare results across papers. It is also not clear from existing research how to define particular aspects of interest, and what methods (both human and automated) are best suited for evaluating and comparing different style transfer models.

To address these issues, we defined key aspects of interest (style transfer intensity, content preservation, and naturalness) and showed how to obtain more reliable measures of them from human evaluation than in previous work. We also showed that our set of proposed metrics for automated evaluation (the Wasserstein distance, word mover's distance on style-masked texts, and adversarial classification) exhibit stronger correlations than existing automated metrics do with human scores. While human evaluation may still be useful, such metrics may facilitate future evaluation of style transfer models where it is infeasible to collect many human ratings, which are expensive and time-consuming to process. Finally, with aspect tradeoff plots that are based on the scores from the better automated metrics, we can visualize style transfer model behavior and directly compare multiple models. We can also make more informed decisions about hyperparameter selection when training, designing, or revising models for future experiments.

With regards to future work, we can examine the behavior of style transfer models on other datasets, such as those that have different types of styles or a different number of style classes. Performing inter-rater reliability and correlation testing on scores associated with these datasets may enable us to approach the evaluation of style transfer models in a broader setting.

Appendix A

Supplementary Material

A.1 Constructing a Style Lexicon

For the style of sentiment, there are several open-source lexica of style words, including *WordNet-Affect* [36], which groups words under affective labels, and *DepecheMood*, which assigns emotion scores to individual words for the following categories: AFRAID, AMUSED, ANGRY, ANNOYED, DONT_CARE, HAPPY, INSPIRED [34].

While building a new lexicon or extending existing ones may be feasible with sentiment as the style, it is not scalable to manually do so for other types of styles, such as formality, topic, and sarcasm. Static lexica also might not take context into account, which is an issue for sentences with words that are ambiguous in terms of style weight, e.g. "dog" in "That is a man with a dog" vs. "That man is a dog." This *lexical affinity* is also an issue when stylistic phrasing is incorporated throughout a sentence, e.g. "I can't tell you what this means to me!" [33]. Thus, it is more appropriate to automatically construct a lexicon for each dataset of interest.

To construct a style lexicon for Yelp, we train logistic regression classifiers of varying regularization strengths and types (lasso, ridge, and no regression) on X and corresponding style labels.[1] We interpret the most strongly weighted features (words with the highest absolute weights) as words that have the most impact on the outcome of the style label, which we can use to construct the style lexicon.

We select 1800 of the top occurring words, as well as 1800 random words, from the

[1] As long as style labels are available, this approach can be used for any type of style. We use the one-vs-rest strategy to train a separate classifier per style class, allowing for generalizability to datasets where there are more than two style classes.

Negative Sentiment	Positive Sentiment
'ruined': -11.5	'mouthwatering': 9.4
'worst': -11.4	'delightfully': 9.5
'failure': -11.2	'wonderfully': 10.5
'lackluster': -10.5	'marvelous': 10.6
'horrible': -10.4	'refreshing': 11.5

Table A.1: Sample words and weights for the Yelp style lexicon.

vocabulary of X. From these, we construct gold sets of style and content words, based on words that human workers mark as containing style, e.g. negative or positive sentiment for the Yelp dataset. In order to identify which model best separates the style and non-style groups, we visualize the spread of the gold words over the distribution of weights for each model. We quantify the point of separation as the number of standard deviations D from the mean of all feature weights. All words whose weights are at or beyond D standard deviations away from the mean are added to the lexicon.

We choose this heuristic because weights further from the mean signify more impact of the corresponding words on the outcome of the style labels. The larger D is, the "cleaner" the style lexicon because it has fewer, but more strongly weighted, words. There is inherently a tradeoff of capturing more style words and reducing noise in the construction of the lexicon.

After training the classifiers, we observe that those with lasso regression resulted in a high number of zero (or close to zero) weighted features, several of which were words marked as having stylistic meaning by workers. This can also be problematic for style words that may not appear frequently in a dataset. The issue was less severe for ridge regression classifiers, which overall gave better separation between the gold style and non-style words.

We restrict the size of the lexicon with the separation point $D = 2$. At the expense of not capturing some style words, we opt for a smaller D to reduce noise and minimize the risk of removing content words, which are critical to evaluations of content preservation. Sample style words for the Yelp dataset are shown in Table A.1.

A.2 Content Preservation Metrics

In addition to BLEU, we measure content preservation with METEOR and word embedding metrics. METEOR is a word-overlap based metric like BLEU, but handles sentence-level

scoring more robustly, allowing it to be both a sentence-level and corpus-level metric [1]. It attempts to map words between x and x' using matches in single tokens, synonyms, stems, and paraphrases. An alignment with minimal crosses between words of x and words of x' signifies greater similarity of word order between the two texts. Optimal alignments like these are then used to calculate a weighted harmonic mean of unigram precision and recall between x and x'. While BLEU penalizes the final score if the total length of all texts in X' is less than that of all texts in X, METEOR levies a penalty that increases with the number of non-adjacent mappings in x and x'. In other words, METEOR penalizes texts that do not have longer n-gram matches [1].

For the word embedding metrics, embeddings can be obtained with widely used methods like Word2Vec [24] and GloVe [29]. Cosine similarities can also be calculated between the sentence-level embeddings of x and of x'. Sentence-level embeddings can be comprised of the most extreme values of word embeddings over each embedding dimension (*vector extrema*), or the averages over individual word embeddings (*embedding average*) [10].

Word Mover's Distance (WMD), based on the Wasserstein distance, calculates the minimum "distance" between word embeddings of x and of x', where a smaller distance signifies greater similarity [20]. *Greedy matching* greedily matches words in x and x' based on cosine similarities taken over word embeddings, and then averages those similarities. It repeats the process in the reverse direction and takes the mean of those two scores [31].

We examine all of these content preservation metrics to identify the one most strongly correlated with human judgment in the style transfer setting (Section 4.2.2).

A.3 Tradeoff Space

To better understand the space in which the tradeoff plots operate, we can define extreme cases of relationships between aspects, such as style transfer intensity and content preservation. In all cases, we assume the measurement of content preservation ignores stylistic content.

The worst models produce outputs with a wide range of style transfer intensity, but consistently poor content preservation (Figure A-1a) or a wide range of content preservation, but consistently poor style transfer intensity (Figure A-1b). Figure A-1c, however, demonstrates the behavior of an ideal transfer model, which produces outputs with a wide range

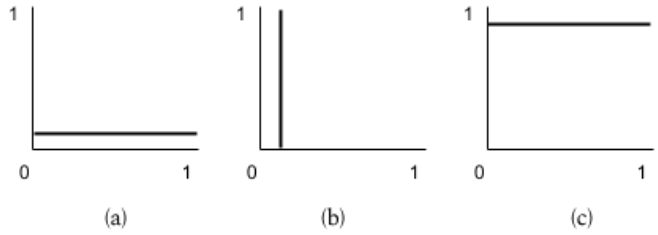

Figure A-1: Cases of extreme aspect tradeoff plots.

of style transfer intensity and consistently high content preservation. Given these cases, we can interpret the models with better performance to be the ones whose tradeoff plots are closer to that of the ideal model, and farther from those of the worst case models.

References

[1] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[2] Tammo HA Bijmolt and Michel Wedel. The effects of alternative methods of collecting similarity data for multidimensional scaling. *International Journal of Research in Marketing*, 12(4):363–371, 1995.

[3] AL Bowley. The standard deviation of the correlation coefficient. *Journal of the American Statistical Association*, 23(161):31–34, 1928.

[4] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. Generating Sentences from a Continuous Space. *CoRR*, abs/1511.06349, 2015.

[5] Peter F. Brown, Vincent J. Della Pietra, Robert L. Mercer, Stephen A. Della Pietra, and Jennifer C. Lai. An Estimate of an Upper Bound for the Entropy of English. *Comput. Linguist.*, 18(1):31–40, March 1992.

[6] Noam Chomsky. *Syntactic Structures*. Mouton and Co., The Hague, 1957.

[7] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

[8] Jessica Ficler and Yoav Goldberg. Controlling Linguistic Style Aspects in Neural Language Generation. *CoRR*, abs/1707.02633, 2017.

[9] Joseph L. Fleiss and Jacob Cohen. The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability. *Educational and Psychological Measurement*, 33(3):613–619, 1973.

[10] Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. Style Transfer in Text: Exploration and Evaluation. *arXiv preprint arXiv:1711.06861*, 2017.

[11] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A Neural Algorithm of Artistic Style. *CoRR*, abs/1508.06576, 2015.

[12] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 2414–2423. IEEE, 2016.

[13] Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. Sockeye: A Toolkit for Neural Machine Translation. *arXiv preprint arXiv:1712.05690*, December 2017.

[14] Haozhi Huang, Hao Wang, Wenhan Luo, Lin Ma, Wenhao Jiang, Xiaolong Zhu, Zhifeng Li, and Wei Liu. Real-time neural style transfer for videos. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 7044–7052. IEEE, 2017.

[15] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of Tricks for Efficient Text Classification. *CoRR*, abs/1607.01759, 2016.

[16] Dan Jurafsky and James H Martin. *Speech and language processing*, volume 3. Pearson London:, 2014.

[17] Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. A Large Self-Annotated Corpus for Sarcasm. *CoRR*, abs/1704.05579, 2017.

[18] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[19] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints*.

[20] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966, 2015.

[21] Shibamouli Lahiri. SQUINKY! A Corpus of Sentence-level Formality, Informativeness, and Implicature. *CoRR*, abs/1506.02306, 2015.

[22] Dongjun Lee. text-cnn. `https://github.com/DongjunLee/text-cnn-tensorflow`, 2018.

[23] J. Li, R. Jia, H. He, and P. Liang. Delete, Retrieve, Generate: A Simple Approach to Sentiment and Style Transfer. *ArXiv e-prints*, April 2018.

[24] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[25] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[26] Ellie Pavlick and Joel Tetreault. An Empirical Analysis of Formality in Online Communication. *Transactions of the Association for Computational Linguistics*, 4:61–74, 2016.

[27] Ofir Pele and Michael Werman. A linear time histogram metric for improved sift matching. In *European conference on computer vision*, pages 495–508. Springer, 2008.

[28] Ofir Pele and Michael Werman. Fast and robust earth mover's distances. In *Computer vision, 2009 IEEE 12th international conference on*, pages 460–467. IEEE, 2009.

[29] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[30] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. A metric for distributions with applications to image databases. In *Computer Vision, 1998. Sixth International Conference on*, pages 59–66. IEEE, 1998.

[31] Vasile Rus and Mihai Lintean. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 157–162. Association for Computational Linguistics, 2012.

[32] Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. Style Transfer from Non-Parallel Text by Cross-Alignment. *CoRR*, abs/1705.09655, 2017.

[33] Shiv Naresh Shivhare and Saritha Khethawat. Emotion detection from text. *arXiv preprint arXiv:1205.4944*, 2012.

[34] Jacopo Staiano and Marco Guerini. DepecheMood: a Lexicon for Emotion Analysis from Crowd-Annotated News. *CoRR*, abs/1405.1605, 2014.

[35] Neil Stewart, Gordon DA Brown, and Nick Chater. Absolute identification by relative judgment. *Psychological review*, 112(4):881, 2005.

[36] Carlo Strapparava, Alessandro Valitutti, et al. Wordnet affect: an affective extension of wordnet. In *Lrec*, volume 4, pages 1083–1086. Citeseer, 2004.

[37] A. M. Turing. Computers &Amp; Thought. chapter Computing Machinery and Intelligence, pages 11–35. MIT Press, Cambridge, MA, USA, 1995.

[38] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *ICML*, 2008.

[39] Steve J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. *The HTK Book Version 3.4*. Cambridge University Press, 2006.

[40] Hang Zhang and Kristin J. Dana. Multi-style Generative Network for Real-time Transfer. *CoRR*, abs/1703.06953, 2017.

[41] Xiang Zhang and Yann LeCun. Text understanding from scratch. *arXiv preprint arXiv:1502.01710*, 2015.

[42] Junbo Jake Zhao, Yoon Kim, Kelly Zhang, Alexander M. Rush, and Yann LeCun. Adversarially Regularized Autoencoders for Generating Discrete Structures. *CoRR*, abs/1706.04223, 2017.