Evaluating Style Transfer for Text

Remi Mir[1], **Bjarke Felbo**[2], **Nick Obradovich**[2], **Iyad Rahwan**[2]
[1]Department of EECS, Massachusetts Institute of Technology
[2]Media Lab, Massachusetts Institute of Technology
{rmir, bfelbo, nobradov, irahwan}@mit.edu

Abstract

Research in the area of style transfer for text is currently bottlenecked by a lack of standard evaluation practices. This paper aims to alleviate this issue by experimentally identifying best practices with a Yelp sentiment dataset. We specify three 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 propose a set of metrics for automated evaluation and demonstrate that they are more strongly correlated and in agreement with human judgment: direction-corrected Earth Mover's Distance, Word Mover's Distance on style-masked texts, and adversarial classification for the respective aspects. We also show that the three examined models exhibit tradeoffs between aspects of interest, demonstrating the importance of evaluating style transfer models at specific points of their tradeoff plots. We release software with our evaluation metrics to facilitate research.

1 Introduction

Style transfer in text is the task of changing an attribute (style) of an input, while retaining non-attribute related content (referred to simply as *content* for brevity in this paper).[1] For instance, previous work has modified text to make it more positive (Shen et al., 2017), romantic (Li et al., 2018), or politically slanted (Prabhumoye et al., 2018).

Some style transfer models enable modifications by manipulating latent representations of the text (Shen et al., 2017; Zhao et al., 2018; Fu et al., 2018), while others identify and replace style-related words directly (Li et al., 2018). Regardless of approach, they are hard to compare as there is currently neither a standard set of evaluation practices, nor a clear definition of which exact aspects to evaluate. In Section 2, we define three key aspects to consider. In Section 3, we summarize issues with previously used metrics. Many rely on human ratings, which can be expensive and time-consuming to obtain.

To address these issues, in Section 4, we consider how to obtain more reliable measures of human judgment for aspects of interest, and automated methods more strongly correlated with human judgment than previously used methods. Lastly, in Section 5, we show that the three examined models exhibit aspect tradeoffs, highlighting the importance of evaluating style transfer models at specific points of their tradeoff plots. We release software with our evaluation metrics at https://github.com/passeul/style-transfer-model-evaluation.

2 Aspects of Evaluation

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

1. *style transfer intensity* $STI(SC(x), SC(x'))$ quantifies the difference in style, where $SC(\cdot)$ maps an input to a style distribution
2. *content preservation* $CP(x, x')$ quantifies the similarity 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 characterize differences. For instance, if a model transfers from negative to positive sentiment, but alters content such as place names, it preserves content poorly.

[1]This definition of style transfer makes a simplifying assumption that "style" words can be distinguished from "content" words, or words carrying relatively less or no stylistic weight, such as "cafè" in "What a nice cafè." The definition is motivated by penalizing unnecessary changes to content words, e.g. "What a nice cafè" to "This is an awful cafè."

Proceedings of NAACL-HLT 2019, pages 495–504
Minneapolis, Minnesota, June 2 - June 7, 2019.

	Style Transfer			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 1: Summary of past evaluation techniques. HRC is human rating on a continuous scale (e.g. 1 to 5). HRD is on discrete options (e.g. positive/negative). HRR is human ranking (most to least similar) of outputs, with respect to given input x. $\{x'\}$ is the set of x' from models trained on different parameters. SC is a style classifier. PPL is perplexity. Superscripts denote that evaluation is done for fluency (F) or grammar (G), which we consider subsets of naturalness. Readers can see the original papers for details on methods falling under these techniques.

If it preserves content well, but sequentially repeats words such as "the", the output is unnatural. Conversely, a model that overemphasizes text reconstruction would yield high content preservation and possibly high naturalness, but little to no style transfer. All three aspects are thus critical to analyze in a system of style transfer evaluation.

3 Related Work

We review previously used approaches for evaluating the outputs of style transfer models. Due to the high costs related to obtaining human evaluations, we focus on three models: the cross-aligned autoencoder (CAAE), adversarially regularized autoencoder (ARAE), and delete-and-retrieve (DAR) models (Shen et al., 2017; Zhao et al., 2018; Li et al., 2018). Table 1 illustrates the spread of evaluation practices in these papers using our notation from Section 2, showing that they all rely on a different combination of human and automated evaluation. For human evaluation, the papers use different instruction sets and scales, making it difficult to compare scores. Below we describe the automated metrics used for each aspect. Some rely on training external models on the corpus of input texts, X, and/or the corpus of output texts, X'. We encourage readers seeking details on how to compute the metrics to reference the algorithms in the original papers.

Style Transfer Previous work has trained classifiers on X and corresponding style labels, and measured the number of outputs classified as having a target style (Shen et al., 2017; Zhao et al., 2018; Li et al., 2018). Results from this *target style scoring* approach may not be directly comparable across papers due to different classifiers used in evaluations.

Content Preservation To evaluate content preservation between x and x', previous work has used BLEU (Zhao et al., 2018; Li et al., 2018), an n-gram based metric originally designed to evaluate machine translation models (Papineni et al., 2002). BLEU does not take into account the aim of style transfer models, which is to alter style by necessarily changing words. Intended differences between x and x' are thus penalized.

Naturalness Past evaluations of naturalness have relied largely on human ratings on a variety of scales under different names: grammaticality, fluency/readability, and naturalness itself (Table 1). An issue with measuring grammaticality is that text with proper syntax can still be semantically nonsensical, e.g. "Colorless green ideas sleep furiously" (Chomsky, 1957). Furthermore, input texts may not demonstrate perfect grammaticality or readability, despite being written by humans and thus being natural by definition (Section 2). This undermines the effectiveness of measures for such specific qualities of output texts.

Zhao et al. (2018) used perplexity to evaluate fluency, which, like grammaticality, we consider a subset of naturalness itself. Low perplexity signifies less uncertainty over which words can be used to continue a sequence, quantifying the ability of a language model to predict gold or reference texts (Brown et al., 1992; Young et al., 2006). However, style transfer outputs are not necessarily gold standard, and the correlation between perplexity and human judgments of those outputs is unknown in the style transfer setting.

4 Methods

We describe how to construct a style lexicon for use in human and automated evaluations. We also describe best practices that we recommend for obtaining scores of those evaluations, as well as how they can be used for evaluating other datasets. Please refer to Section 5 for experimental results.

Negative Sentiment	Positive Sentiment
ruined	mouthwatering
worst	delightfully
failure	wonderfully
lackluster	marvelous
horrible	refreshing

Table 2: Sample of words in a sentiment style lexicon.

4.1 Construction of Style Lexicon

Because the process of style transfer may result in the substitution or removal of more stylistically weighted words, it is ideal to have a lexicon of style-related words to reference. Words in x and/or x' that also appear in the lexicon can be ignored in evaluations of content preservation.

While building a new style lexicon or an extension of existing ones like WordNet-Affect (Strapparava and Valitutti, 2004) may be feasible with binary sentiment as the style, it may not be scalable to manually do so for various other types of styles. Static lexica also might not take context into account. This is an issue for text with words or phrases that are ambiguous in terms of stylistic weight, e.g. "dog" in "That is a man with a dog" vs. "That man is a dog."

It is more appropriate to automate the construction of a style lexicon per dataset of interest. While multiple options may exist for doing so, we emphasize the simplicity and replicability of training a logistic regression classifier on X and corresponding style labels. We populate the lexicon with features having the highest absolute weights, as those have the most impact on the outcome of the style labels. (Table 2 shows sample words in the lexicon constructed for the dataset used in our experiments.) While sentiment datasets have been widely used in the literature (Shen et al., 2017; Zhao et al., 2018; Li et al., 2018), a lexicon can be constructed for other datasets in the same manner, as long as the dataset has style labels.

Given existing NLP techniques, it may not be possible to correctly identify all style-related words in a text. Consequently, there is a tradeoff between identifying more style-related words and incorrectly marking some other (content) words as style-related. We opt for higher precision and lower recall to minimize the risk of removing content words, which are essential to evaluations of content preservation. This issue is not critical because researchers can compare their style transfer methods using our lexicon.

4.2 Human Evaluation

As seen in Table 1, past evaluations of both style transfer and naturalness consider only output text x'. Existing work from other fields have, however, shown that asking human raters to evaluate two relative comparisons provides more accurate scores than asking them to provide a numerical score for a single observation (Stewart et al., 2005; Bijmolt and Wedel, 1995). With this knowledge, we construct more reliable ways of obtaining human evaluations via *relative scoring* instead of *absolute scoring*.

Style Transfer Intensity Past evaluations have raters mark the degree to which x' exhibits a target style (Li et al., 2018). We instead ask raters to score the difference in style between x and x', on a scale of 1 (identical styles) to 5 (completely different styles). This approach can also used for non-binary cases. Consider text modeled as a distribution over multiple emotions (e.g. happy, sad, scared, etc.), where each emotion can be thought of as a style. One task could be to make a scared text more happy. Presented with x and x', raters would still rate the degree to which they differ in style.

Content Preservation We consider the difficulty of asking raters to ignore style-related words as done in (Shen et al., 2017). Because not all raters may identify the same words as stylistic, their evaluations may vary substantially from one another. To account for this, we ask raters to evaluate content preservation on the same texts, but where we have masked style words using our style lexicon. Under this new "masking" approach, raters have a simpler task, as they are no longer responsible for taking style into account when they rate the similarity of two texts on a scale of 1 to 5.

Naturalness We ask raters to determine whether x or x' (they are not told which is which) is more natural. An x' marked as more natural indicates some success on the part of the style transfer model, as it is able to fool the rater. This is in contrast to previous work, where raters score the naturalness of x' on a continuous scale without taking x into account at all, even though x serves as the basis for comparison of what is considered natural.

4.3 Automated Evaluation

In this section, we describe our approaches to automating the evaluation of each aspect of interest.

No modification Input: the girls up front *incompetent* . Output: the girls up front are *amazing* .
Style removal Input: the girls up front . Output: the girls up front are .
Style masking Input: the girls up front $\langle customstyle \rangle$. Output: the girls up front are $\langle customstyle \rangle$.

Table 3: Text under different settings of style-based modification, as used in evaluations of content preservation. The sample output is from ARAE ($\lambda = 1$).

Style Transfer Intensity Rather than count how many output texts achieve a target style, we can capture more nuanced differences between the style distributions of x and x', using Earth Mover's Distance (Rubner et al., 1998; Pele and Werman, 2009). $EMD(SC(x), SC(x'))$ is the minimum "cost" to turn one distribution into the other, or how "intense" the transfer is. Distributions can have any number of values (styles), so EMD handles binary and non-binary datasets.

Note that even if $argmax(SC(x'))$ is not the target style class, EMD still acknowledges movement towards the target style with respect to $SC(x)$. However, we penalize (negate) the score if $SC(x')$ displays a relative change of style in the wrong direction, away from the target style.

Depending on x, not a lot of rewriting may be necessary to achieve a different style. This is not an issue, as STI relies on a style classifier to quantify not the difference between the content of x and x', but their style distributions. For the style classifier, we experiment with textcnn (Kim, 2014; Lee, 2018) and fastText (Joulin et al., 2017).

Content Preservation We first subject texts to different settings of modification: style removal and style masking. This is to address undesired penalization of metrics on texts expected to demonstrate changes after style transfer (Section 3). For style removal, we remove style words from x and x' using the style lexicon. For masking, we replace those words with a $\langle customstyle \rangle$ placeholder. Table 3 exemplifies these modifications.

For measuring the degree of content preservation, in addition to the widely used BLEU, we consider METEOR and embedding-based metrics. METEOR is an n-gram based metric like BLEU, but handles sentence-level scoring more robustly, allowing it to be both a sentence-level and corpus-level metric (Banerjee and Lavie, 2005).

For the embedding-based metrics, word embeddings can be obtained with methods like Word2Vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014). Sentence-level embeddings can be comprised of the most extreme values of word embeddings per dimension (*vector extrema*) (Forgues et al., 2014), or word *embedding averages* (Sharma et al., 2017). *Word Mover's Distance* (WMD), based on EMD, calculates the minimum "distance" between word embeddings of x and of x', where smaller distances signify higher similarity (Kusner et al., 2015). *Greedy matching* greedily matches words in x and x' based on their embeddings, calculates their similarity (e.g. cosine similarity), and averages all the similarities. It repeats the process in the reverse direction and takes the average of those two scores (Rus and Lintean, 2012).

We evaluate with all these metrics to identify the one most strongly correlated with human judgment of content preservation.

Naturalness For a baseline understanding of what is considered "natural," any method used for automated evaluation of naturalness requires the human-sourced input texts. We train unigram and neural logistic regression classifiers (Bowman et al., 2016) on samples of X and X' for each transfer model. Via adversarial evaluation, these classifiers must distinguish human-generated inputs from machine-generated outputs. The more natural an output is, the likelier it is to fool a classifier (Jurafsky and Martin, 2018). We calculate agreement between each type of human evaluation (Section 4.2) and each classifier AC. Agreement is the ratio of instances where humans and AC rate a text as more natural than the other.

We also train LSTM language models (Hochreiter and Schmidhuber, 1997) on X and compute sentence-level perplexities for each text in X' in order to determine the relative effectiveness of adversarial classification as a metric.

5 Experiments and Results

Due to high costs of human evaluation, we focus on CAAE, ARAE, and DAR models with transfer tasks based on samples from the Yelp binary sentiment dataset (Shen et al., 2017).[2] Below we detail

[2]Like most literature, including the papers on CAAE, ARAE and DAR, we focus on the binary case. Creating a high-quality, multi-label style transfer dataset for evaluation is a demanding task, which is out of scope for this paper.

Input	would *n't recommend* until management works on friendliness and communication with residents .
ARAE ($\lambda = 1$)	highly *recommend* this place while living in tempe and management .
CAAE ($\rho = 0.5$)	would highly *recommend* management on duty and staff on business .
DAR ($\gamma = 500$)	until management works on friendliness and is a *great* place for communication with residents .

Table 4: Sample outputs of a negative to positive sentiment style transfer task. Italicized words are style-related, according to a style lexicon. They can be masked or removed in evaluations of content preservation (Section 4.3).

the range of parameters each model is trained on in order to compare evaluation practices and generate aspect tradeoff plots. Each of three Amazon Turk raters evaluated 244 texts per aspect, per model. Of those texts, half are originally of positive sentiment transferred to negative, and vice versa.

For brevity, we reference average scores (correlation, kappa, and agreement, each of which is described below) from across all models in our analysis of results. For detailed scores per model, please refer to the corresponding tables.

5.1 Style Transfer Models

For each style transfer model, we choose a wide range of training parameters to allow for variation of content preservation, and indirectly, of style transfer intensity, in X. We show sample outputs from the models for a given input text in Table 4.

CAAE uses autoencoders (Vincent et al., 2008) that are cross-aligned, assuming that texts already share a latent content distribution (Shen et al., 2017). It uses latent states of the RNN and multiple discriminators to align distributions of texts in X' exhibiting one style with distributions of texts in X exhibiting another. Adversarial components help separate style information from the latent space where inputs are represented. We train CAAE on various values (0.01, 0.1, 0.5, 1, 5) of ρ, a weight on the adversarial loss.

CAAE is a baseline for other style transfer models, such as ARAE, which trains a separate decoder per style class (Zhao et al., 2018). We train ARAE on various values (1, 5, 10) of λ, which is also a weight on adversarial loss.

The third model that we evaluate, which also uses CAAE as a baseline, avoids adversarial methods in an approach called Delete-and-Retrieve (DAR) (Li et al., 2018). 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 DAR on $\gamma = 15$, where γ is a threshold parameter for 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) during test time because, by design, the model does not need to be retrained (Li et al., 2018).

Model	Text Modification Setting	
	Unmasked	Style Masked
CAAE	0.158	**0.289**
ARAE	0.201	**0.321**
DAR	0.161	**0.281**
Average	0.173	**0.297**

Table 5: Fleiss' kappas for human judgments of content preservation of unmasked and style-masked texts.

Model	Absolute		Relative
	$\tau = 3$	$\tau = 2$	
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 6: Fleiss' kappas for human judgments of absolute naturalness and relative naturalness of texts.

5.2 Human Evaluation

We use Fleiss' kappa κ of inter-rater reliability (see formula in L. Fleiss and Cohen, 1973) to identify the more effective human scoring task for different aspects of interest. The kappa metric is often levied in a relative fashion, as there are no universally accepted thresholds for agreements that are slight, fair, moderate, etc. For comprehensive experimentation, we compare kappas over the outputs of each style transfer model. The kappa score for ratings of content preservation based on style-masked texts is 0.297. Given the kappa score of 0.173 for unmasked texts, style masking is a more reliable approach towards human evaluation for content preservation (Table 5).

For style transfer intensity, kappas for relative scoring do not show improvement over the previously used approach of absolute scoring of x'. However, we observe the opposite for the aspect of naturalness. Kappas for relative naturalness scoring tasks exceed those of the absolute scoring ones (Table 6). Despite the two types of tasks having

Model	fastText		textcnn	
	Target Style Scores	Earth Mover's Distance	Target Style Scores	Earth Mover's Distance
CAAE	0.566 ± 0.038	$\mathbf{0.573 \pm 0.038}$	0.587 ± 0.037	$\mathbf{0.589 \pm 0.037}$
ARAE	0.513 ± 0.053	$\mathbf{0.516 \pm 0.053}$	0.515 ± 0.053	$\mathbf{0.519 \pm 0.053}$
DAR	0.470 ± 0.049	$\mathbf{0.539 \pm 0.045}$	0.508 ± 0.047	$\mathbf{0.566 \pm 0.043}$
Average	0.516 ± 0.047	$\mathbf{0.543 \pm 0.045}$	0.537 ± 0.046	$\mathbf{0.558 \pm 0.044}$

Table 7: Correlations of automated style transfer intensity metrics with human scores.

Model	BLEU	METEOR	Embed Average	Greedy Match	Vector Extrema	WMD
CAAE	0.458 ± 0.044	$\mathbf{0.498 \pm 0.042}$	0.370 ± 0.048	0.489 ± 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	$\mathbf{0.445 \pm 0.058}$
DAR	0.440 ± 0.051	0.455 ± 0.050	0.379 ± 0.054	0.472 ± 0.049	0.472 ± 0.049	$\mathbf{0.484 \pm 0.048}$
Average	0.412 ± 0.053	0.447 ± 0.051	0.354 ± 0.056	0.460 ± 0.051	0.464 ± 0.050	$\mathbf{0.475 \pm 0.049}$

Table 8: Absolute correlations of content preservation metrics with human scores on texts with style removal.

Model	BLEU	METEOR	Embed Average	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	$\mathbf{0.517 \pm 0.041}$
ARAE	0.356 ± 0.063	0.374 ± 0.062	0.302 ± 0.066	0.405 ± 0.061	0.422 ± 0.060	$\mathbf{0.457 \pm 0.057}$
DAR	0.444 ± 0.050	0.454 ± 0.050	0.370 ± 0.054	0.450 ± 0.050	0.473 ± 0.049	$\mathbf{0.475 \pm 0.049}$
Average	0.429 ± 0.052	0.448 ± 0.051	0.343 ± 0.056	0.448 ± 0.051	0.464 ± 0.050	$\mathbf{0.483 \pm 0.049}$

Table 9: Absolute correlations of content preservation metrics with human scores on texts with style masking.

different numbers of categories (2 vs 5), we can compare them by using a threshold τ to bin the absolute score for each text into a "natural" group (x' is considered to be more natural than x) or "unnatural" one (vice versa), like in relative scoring. For example, $\tau = 2$ places texts with absolute scores greater than or equal to 2 into the natural group. Judgments for relative tasks yield greater inter-rater reliability than those of absolute tasks across multiple thresholds ($\tau \in \{2, 3\}$). This suggests that the relative scoring paradigm is preferable in human evaluations of naturalness.

5.3 Automated Evaluation

Per aspect of interest, we compute Pearson correlations between scores from the existing metric and human judgments. (As there were three raters for any given scoring task, we take the average of their scores.) We do the same for our proposed metrics to identify which metric is more reliable for automated evaluation of a given aspect.

For style transfer intensity, across both the fastText and textcnn classifiers, our proposed direction-corrected Earth Mover's Distance metric has higher correlation with human scores than the past approach of target style scoring (Table 7).

For content preservation, METEOR, shown to have higher correlation with human judgments than BLEU for machine translation (Banerjee and Lavie, 2005), shows the same relationship for style transfer. However, across various text modification settings, WMD generally shows the strongest correlation with human scores (Tables 8 and 9). Because WMD is lower when texts are more similar, it is anti-correlated with human scores. We take absolute correlations to facilitate comparison with other content preservation metrics. With respect to text modification, style masking may be more suitable as it, on average for WMD, exhibits a higher correlation with human judgments.

Model	Unigram Adv. Clf.		Neural Adv. Clf.	
	Absolute	Relative	Absolute	Relative
CAAE	56.07	**64.51**	57.38	**67.87**
ARAE	49.45	**66.67**	50.68	**67.90**
DAR	65.16	**65.57**	61.07	**62.30**
Average	56.89	**65.58**	56.38	**66.02**

Table 10: Percent agreement between adversarial classifiers and human scores on the naturalness of texts.

For naturalness, both unigram and neural classifiers exhibit greater agreement on which texts are considered more natural with the humans given relative scoring tasks than with those given absolute scoring tasks (Table 10), although the neural classifier achieves higher agreements on average. We also confirm that sentence-level perplexity is

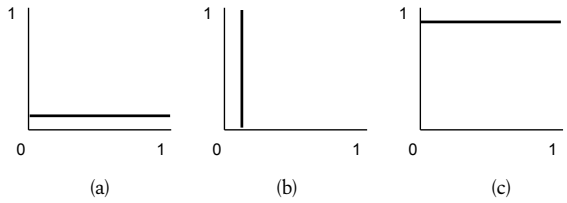
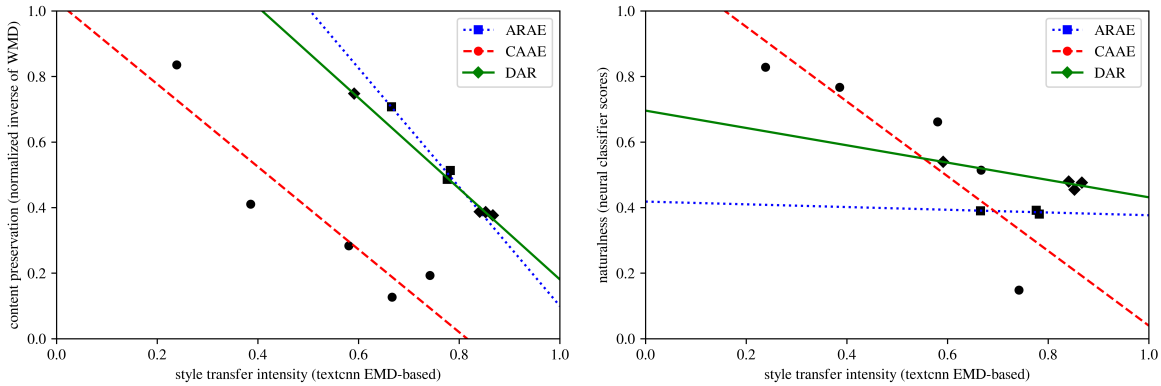


Figure 1: Extreme tradeoff plots, with style transfer intensity on the x-axis and content preservation on the y-axis.



(a) Content vs. Style Tradeoffs

(b) Naturalness vs. Style Tradeoffs

Figure 2: Tradeoffs between aspects of evaluation, using metrics most strongly correlated with human scores.

not an appropriate metric. It exhibits no significant correlation with human scores ($\alpha = 0.05$). These results suggest that adversarial classifiers can be useful for automating measurement of naturalness.

5.4 Aspect Tradeoffs

Previous work has compared models with respect to a single aspect of interest at a time, but has only, to a limited degree, considered how relationships between multiple aspects influence these comparisons. In particular, concurrent work by (Li et al., 2018) examines tradeoff plots, but focuses primarily on variants of its own model, while including only a single point on the plots of style transfer models from other papers. For a comprehensive comparison, it is ideal to have plots for all models.

It is helpful to first understand the tradeoff space. For example, we define extreme cases for style transfer intensity and content preservation, where we assume measurement of the latter ignores stylistic content. Consider two classes of suboptimal models. One class produces outputs with a wide range of style transfer intensity, but poor content preservation (Figure 1a). The other class of models produces outputs with low style transfer intensity, but a wide range of content preservation (Figure 1b).

This is in contrast to a model that yields a wide range of style transfer intensity and consistently high content preservation (Figure 1c). If we take that to be an ideal model for a sentiment dataset, we can interpret 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 suboptimal ones. The plot for an ideal model will likely vary by dataset, especially because the tradeoff between content preservation and style transfer intensity depends on the level of distinction between style words and content words of the dataset.

With this interpretation of the tradeoff space, we construct a plot for each style transfer model (Figure 2), where each point represents a different hyperparameter setting for training (Section 5.1). We collect scores based on the automated metrics most strongly correlated with human judgment: direction-corrected EMD for style transfer intensity, WMD for content preservation, and percent of output texts marked by an adversarial classifier as more natural than input texts. Because WMD scores are lower when texts are more similar, we instead take the normalized inverses of the scores to represent the degree of content preservation.

Across all models, there is a trend of reduction in content preservation and naturalness as style transfer intensity increases. Without the plots, one

might conclude that ARAE and DAR perform substantially differently, especially if hyperparameters are chosen such that ARAE achieves the leftmost point on its plot and DAR achieves the rightmost point on its plot. With the plots, at least for the set of hyperparameters considered, it is evident that they perform comparably (Figure 2a) and do not exhibit the same level of decrease in naturalness as CAAE (Figure 2b).

6 Discussion

Previous work on style transfer models used a variety of evaluation methods (Table 1), making it difficult to meaningfully compare results across papers. Moreover, it is not clear from existing research how exactly to define particular aspects of interest, or which methods (whether human or automated) are most suitable for evaluating and comparing different style transfer models.

To address these issues, we specified 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. Our proposed automated metrics (direction-corrected EMD, WMD on style-masked texts, and adversarial classification) exhibited stronger correlations with human scores than existing automated metrics on a binary sentiment dataset. While human evaluation may still be useful in future research, automation facilitates evaluation when it is infeasible to collect human scores due to prohibitive cost or limited time.

6.1 Human Evaluation

For style transfer intensity, the relative scoring task (rating the degree of stylistic difference between x and x') did not have greater rater reliability than the previously used task of rating output texts on an absolute scale. This is likely due to task complexity or rater uncertainty, which motivates the need for further exploration of task design for this particular aspect of interest.

For content preservation, our form of human evaluation operates on texts whose style words are masked out, unlike the previous approach (no masking). Our approach addresses the unintentional variable of rater-dependent style identification that could lead to noisy, less reliable ratings.

Identification and masking of words was made possible with a style lexicon. We automatically constructed the lexicon in a way that can be done for any style dataset, as long as style labels are available (Section 4.1). We acknowledge a trade-off between filling the lexicon with more style words and being conservative in order to avoid capturing content words. We justify taking a more conservative approach as content words are naturally critical to evaluations of content preservation.

For naturalness, we introduced a paradigm of relative scoring that uses both the output and input texts. This achieved a higher inter-rater reliability than did absolute scoring, the previous approach.

6.2 Automated Evaluation

For style transfer intensity, we proposed using a metric with EMD as the basis to acknowledge the spectrum of styles that can appear in outputs and to handle both binary and non-binary datasets. The metric also accounts for direction by penalizing scores in the cases where the style distribution of the output text explicitly moves away from the target style. Previous work used external classifiers, whose style distributions for x and x' can be used to calculate direction-corrected EMD, making it a simple addition to the evaluation workflow.

For content preservation, WMD (based on EMD) works in a similar fashion, but with word embeddings of x and of x'. BLEU, used widely in previous work, may yield weaker correlations with human judgment in comparison as it was designed to have multiple reference texts per candidate text (Papineni et al., 2002). Several reference texts, which are more common in machine translation tasks, increase the chance of n-gram (such as $n \geq 3$) overlap with the candidate. In the style transfer setting, however, the only reference text for x' is x. Having a single reference text reduces the likelihood of overlap and the overall effectiveness of BLEU.

For naturalness, strong agreement of adversarial classifiers with relative scores assigned by humans suggest that classifiers are suitable for automated evaluation. One might assume input texts would almost always be rated as more natural by both humans and classifiers, biasing the agreement. This is not the case, as we justify our rating scheme with evidence of outputs being rated as more natural across several models (Figure 2b). Output texts classified as more natural indicate some success for a style transfer model, as it can produce texts with a quality like that of human-generated inputs,

which are, by definition, natural.

Finally, with aspect tradeoff plots constructed using scores from the automated metrics, we can directly compare models with respect to multiple aspects simultaneously. Points of intersection, or near intersection, for different models signify that they, at the hyperparameters that yielded those points, can achieve similar results for various aspects. These parameters can be useful for understanding the impact of decisions made during model design and optimization phases.

6.3 Future Research

As we confirmed, sentence-level perplexity of output x' is not meaningful by itself for the automated evaluation of naturalness. The idea of using both x and x', akin to how we train automated classifiers of naturalness (Section 4.3), can be extended to construct a perplexity-based metric that also takes into account the perplexity of input x.

Another avenue for future work could be evaluating on datasets with a different style or number of style classes. It is worth studying the distinction between style words and content words in the vocabulary of each such dataset. Given the definition of style transfer and its simplifying assumption in Section 1, it would be reasonable to expect naturally low content preservation scores for any given style transfer model operating on datasets with less distinction, such as those of formality. This is not so much an issue as it is a dataset-specific trend that can be visualized in corresponding tradeoff plots, which would provide a holistic evaluation of model performance. In any case, results from inter-rater reliability and correlation testing on these additional datasets would overall enable more consistent evaluation practices and further progress in style transfer research.

Acknowledgments

We would like to thank Juncen Li, Tianxiao Shen, and Junbo (Jake) Zhao for guidance in the use of their respective style transfer models. These models serve as markers of major progress in the area of style transfer research, without which this work would not have been possible.

References

Satanjeev Banerjee and Alon Lavie. 2005. 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. Association for Computational Linguistics.

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

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21. Association for Computational Linguistics.

Peter F. Brown, Vincent J. Della Pietra, Robert L. Mercer, Stephen A. Della Pietra, and Jennifer C. Lai. 1992. An estimate of an upper bound for the entropy of english. *Computational Linguistics*, 18(1).

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

Gabriel Forgues, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay. 2014. Bootstrapping dialog systems with word embeddings.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.

Dan Jurafsky and James H Martin. 2018. *Speech and Language Processing*, volume 3.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751. Association for Computational Linguistics.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, pages 957–966.

Joseph L. Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33:613–619.

Dongjun Lee. 2018. text-cnn. https://github.com/DongjunLee/text-cnn-tensorflow.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Ofir Pele and Michael Werman. 2009. Fast and robust earth mover's distances. pages 460–467. IEEE.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876. Association for Computational Linguistics.

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

Vasile Rus and Mihai Lintean. 2012. 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.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *CoRR*, abs/1706.09799.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems 30*, pages 6830–6841.

Neil Stewart, Gordon DA Brown, and Nick Chater. 2005. Absolute identification by relative judgment. *Psychological review*, 112(4):881–911.

Carlo Strapparava and Alessandro Valitutti. 2004. Wordnet affect: an affective extension of wordnet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. European Language Resources Association (ELRA).

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 1096–1103. ACM.

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

Junbo Zhao, Yoon Kim, Kelly Zhang, Alexander Rush, and Yann LeCun. 2018. Adversarially regularized autoencoders. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5902–5911.