A Appendices for "Reformulating Unsupervised Style Transfer as Paraphrase Generation"

A.1 PARANMT-50M Filtering Details

We train our paraphrase model in a seq2seq fashion using the PARANMT-50M corpus (Wieting and Gimpel, 2018) which was constructed by backtranslating (Sennrich et al., 2016) the Czech side of the CzEng parallel corpus (Bojar et al., 2016). This corpus is large and noisy and we aggressively filter it to encourage content preservation and diversity maximization. We use the following filtering,

Content Filtering: We remove all sentence pairs which score lower than 0.5 on a strong paraphrase similarity model from Wieting et al. (2019)[20]. We perform a length filtering and allow a maximum length difference of 5 words in sentence pairs. Finally, we remove very short and long sentences by only keeping sentence pairs with an average token length between 7 and 25.

Lexical Diversity Filtering: We only preserve backtranslated pairs with sufficient unigram distribution difference. We filter all pairs where more than 50% words in the backtranslated sentence can be found in the source sentence. This is computed using the SQuAD evaluation scripts (Rajpurkar et al., 2016). Additionally, we remove sentences with more than 70% trigram overlap.

Syntactic Diversity Filtering: We discard all paraphrases which have a similar word ordering. We compare the relative ordering of the words shared between the input and backtranslated sentence by measuring the Kendall tau distance (Kendall, 1938) or the "bubble-sort" distance. We keep all backtranslated pairs which are at least 50% shuffled.[21]

LangID Filtering: Finally, we discard all sentences where both the input and backtranslated sentence are classified as non-english using `langdetect`.[22]

Effect of each filter: We adopt a pipelined approach to filtering. The PARANMT-50M corpus size after each stage of filtering is shown in Table 8.

	Filter Stage	Corpus Size
0.	Original	51.41M
1.	Content Similarity	30.49M
2.	Trigram Diversity	9.03M
3.	Unigram Diversity	1.96M
4.	Kendall-Tau Diversity	112.01K
5.	Length Difference	82.64K
6.	LangID	74.55K

Table 8: Steps of filtering conducted on PARANMT-50M along with its effect on corpus size.

A.2 Generative Model Details

This section provides details of our seq2seq model used for both paraphrase model and style-specific inverse paraphrase model. Recent work (Radford et al., 2019) has shown that GPT2, a massive transformer trained on a large corpus of unlabeled text using the language modeling objective, is very effective in performing more human-like text generation. We leverage the publicly available GPT2-large checkpoints by finetuning it on our custom datasets with a small learning rate. However, GPT2 is an unconditional language model having only a decoder network, and traditional seq2seq setups use separate encoder and decoder neural network (Sutskever et al., 2014) with attention (Bahdanau et al., 2014). To avoid training an encoder network from scratch, we use the encoder-free seq2seq modeling approach described in Wolf et al. (2018). where both input and output sequences are fed to the decoder network separated with a special token, and use separate segment embeddings. Our model is implemented using the `transformers` library[23] (Wolf et al., 2019). We use encoder-free seq2seq modeling (Wolf et al., 2018) which feeds the input into the decoder neural network, separating it with segment embeddings. We fine-tune GPT2-large to perform encoder-free seq2seq modeling.

Architecture: Let $\mathbf{x} = (x_1, ..., x_n)$ represent the tokens in the input sequence and let $\mathbf{y} = (y_{bos}, y_1, ..., y_m, y_{eos})$ represent the tokens of the output sequence, where y_{bos} and y_{eos} corresponds to special beginning and end of sentence tokens. We feed the sequence $(x_1, ..., x_n, y_{bos}, y_1, ..., y_m)$ as input to GPT2 and train it on the next-word prediction objective for the tokens $y_1, ..., y_m, y_{eos}$

---

[20] We use the SIM model from Wieting et al. (2019), which achieves a strong performance on the SemEval semantic text similarity (STS) benchmarks (Agirre et al., 2016)

[21] An identical ordering of words is 0% shuffled whereas a reverse ordering is 100% shuffled.

[22] This is using the Python port of Nakatani (2010), `https://github.com/Mimino666/langdetect`.

[23] `https://github.com/huggingface/transformers`

using the cross-entropy loss. During inference, the sequence $(x_1, ..., x_n, y_{bos})$ is fed as input and the tokens are generated in an autoregressive manner (Vaswani et al., 2017) until y_{eos} is generated.

Every token in $\mathbf{x}$ and $\mathbf{y}$ is passed through a shared input embedding layer to obtain a vector representation of every token. To encode positional and segment information, learnable positional and segment embeddings are added to the input embedding consistent with the GPT2 architecture. Segment embeddings are used to denote whether a token belongs to sequence $\mathbf{x}$ or $\mathbf{y}$.

Other seq2seq alternatives: Note that our unsupervised style transfer algorithm is agnostic to the specific choice of seq2seq modeling. We wanted to perform transfer learning from massive left-to-right language models like GPT2, and found the encoder-free seq2seq approach simple and effective. Future work includes finetuning more recent models like T5 (Raffel et al., 2019) or BART (Lewis et al., 2019). These models use the standard seq2seq setup of separate encoder / decoder networks and pretrain them jointly using denoising autoencoding objectives based on language modeling.

Hyperparameter Details: We finetune GPT2-large using NVIDIA TESLA M40 GPUs for 2 epochs using early stopping based on validation set perplexity. The models are finetuned using a small learning rate of 5e-5 and converge to a good solution fairly quickly as noticed by recent work (Li et al., 2020; Kaplan et al., 2020). Specifically, each experiment completed within a day of training on a single GPU, and many experiments with small datasets took a lot lesser time. We use a minibatch size of 10 sentence pairs and truncate sequences which are longer than 50 subwords in the input or output space. We use the Adam optimizer (Kingma and Ba, 2015) with the weight decay fix and using a linear learning rate decay schedule, as implemented in the `transformers` library. Finally, we left-pad the input sequence to get a total input length of 50 subwords and right-pad output sequence to get a total output length of 50 subwords. This special batching is necessary to use minibatches during inference time. Special symbols are used to pad the sequences and they are not considered in the cross-entropy loss. Our model has 774M trainable parameters, identical to the original GPT2-large.

A.3 Classifier Model Details

We fine-tune RoBERTa-large to build our classifier, using the official implementation in `fairseq`. We use a learning rate of 1e-5 for all experiments with a minibatch size of 32. All models were trained on a single NVIDIA RTX 2080ti GPU, with gradient accumulation to allow larger batch sizes. We train models for 10 epochs and use early stopping on the validation split accuracy. We use the Adam optimizer (Kingma and Ba, 2015) with modifications suggested in the RoBERTa paper (Liu et al., 2019). Consistent with the suggested hyperparameters, we use a learning rate warmup for the first 6% of the updates, and then decay the learning rate.

A.4 OpenNMT Model Details

We train sequence-to-sequence models with attention based on LSTMs using OpenNMT (Klein et al., 2017) using their PyTorch port.[24] We mostly used the default hyperparameter settings of `OpenNMT-py`. The only hyperparameter we modified was the learning rate schedule, since our datasets were small and overfit quickly. For the paraphrase model, we started decay after 11000 steps and halved the learning rate every 1000 steps. For Shakespeare, we started the decay after 3000 steps and halved the learning rate every 500 steps. For Formality, we started the decay after 6000 steps and halved the learning rate every 1000 steps. These modifications only slightly improved validation perplexity (by 3-4 points in each case).

We used early stopping on validation perplexity and checkpoint the model every 500 optimization steps. The other hyperparameters are the default `OpenNMT-py` settings — SGD optimization using learning rate 1.0, LSTM seq2seq model with global attention (Luong et al., 2015), 500 hidden units and embedding dimensions and 2 layers each in the encoder and decoder.

A.5 More Comparisons with Prior Work

Please refer to Table 12 for an equivalent of Table 1 using BLEU scores.

We present more comparisons with prior work in Table 13. We use the generated outputs for the Formality test set available in the public repository of Luo et al. (2019) (including outputs from the algorithms described in Prabhumoye et al., 2018 and Li et al., 2018) and run them on our evaluation pipeline. We compare the results

---

[24] `https://github.com/OpenNMT/OpenNMT-py`

with our formality transfer model used in Table 1 and Table 2. We note significant performance improvements, especially in the fluency of the generated text. Note that there is a domain shift for our model, since we trained our model using the splits of He et al. (2020) which use the Entertainment & Music splits of the Formality corpus. The outputs in the repository of Luo et al. (2019) use the Family & Relationships split. It is unclear in the paper of Luo et al. (2019) whether the models were trained on the Family & Relationships training split or not.

Other Comparisons: We tried to compare against other recent work in style transfer based on Transformers, such as Dai et al. (2019) and Sudhakar et al. (2019). Both papers do not evaluate their models on datasets we use (Shakespeare and Formality), where parallel sentences preserve semantics.

The only datasets used in Dai et al. (2019) were sentiment transfer benchmarks, which modify semantic properties of the sentence. We attempted to train the models in Dai et al. (2019) using their codebase on the Shakespeare dataset, but faced three major issues 1) missing number of epochs / iterations. The early stopping criteria is not implemented or specified, and metrics were being computed on the *test set* every 25 training iterations, which is invalid practice for choosing the optimal checkpoint; 2) specificity of the codebase to the Yelp sentiment transfer dataset in terms of maximum sequence length and evaluation, making it non-trivial to use for any other dataset; 3) Despite our best efforts we could not get the model to converge to a good minima which would produce fluent text (besides word-by-word copying) when trained on the Shakespeare dataset.

Similarly, the datasets used in Sudhakar et al. (2019) modify semantic properties (sentiment, political slant etc.). On running their codebase on the Shakespeare dataset using the default hyperparameters, we achieved a poor performance of 53.1% ACC, 55.2 SIM and 56.5% FL, aggregating to a J(A,S,F) score of 18.4. Similarly on the Formality dataset, performance was poor with 41.7% ACC, 67.8 SIM and 67.7% FL, aggregating to J(A,S,F) score of 18.1. A qualitatively inspection showed very little abstraction and nearly word-by-word copying from the input (due to the delete & generate nature of the approach), which explains the higher SIM score but lower ACC score (just like COPY baseline in Table 1). Fluency was low despite GPT pretraining, perhaps due to the token deletion step in the algorithm.

A.6 Details of our Dataset, CDS

We provide details of our sources, the sizes of indivdual style corpora and examples from our new benchmark dataset CDS in Table 14. We individually preprocessed each corpus to remove very short and long sentences, boilerplate text (common in Project Gutenberg articles) and section headings. We have added some representative examples from each style in Table 14. More representative examples (along with our entire dataset) will be provided in the project page `http://style.cs.umass.edu`.

Style Similarity: In Figure 4 we plot the cosine similarity between styles using the averaged `[CLS]` vector of the trained RoBERTa-large classifier (inference over validation set). The off-diagonal elements show intuitive domain similarities, such as (Lyrics, Poetry); (AAE, Tweets); (Joyce, Shakespeare) or among classes from the Corpus of Historical American English.

A.7 Diverse Paraphrasing on CDS

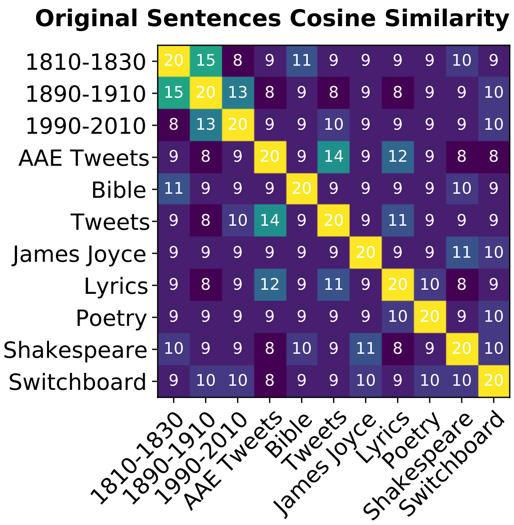
We compare the quality and diversity of the paraphrases generated by our diverse and non-diverse paraphrasers on our dataset CDS in Table 16. Note that this is the pseudo parallel training data for the inverse paraphrase model (described in Section 2.1 and Section 2.4) and not the actual style transferred sentences. Overall, the diverse paraphraser achieves high diversity, with 51% unigram change and 27% word shuffling,[25] compared to 28% unigram and 6% shuffling for non-diverse paraphraser, while maintaining good semantic similarity (SIM= 72.5 vs 83.9 for non-diverse) even in complex stylistic settings.

A.8 Style Transfer Performance on CDS

We provide a detailed breakdown of performance in different styles of CDS in Table 15. For each of the 11 target styles, we style transferred 1,000 sentences from every other style and jointly evaluated the 10,000 generations. Some styles are more successfully transferred than others, such as Switchboard, Lyrics and James Joyce. While wearing the

---

[25] The "unigram change" and "word shuffling" refer to the unigram F1 word overlap and Kendall's τ_B scores.



	1810-1830	1890-1910	1990-2010	AAE Tweets	Bible	Tweets	James Joyce	Lyrics	Poetry	Shakespeare	Switchboard
1810-1830	20	15	8	9	11	9	9	9	9	10	9
1890-1910	15	20	13	8	9	8	9	8	9	9	10
1990-2010	8	13	20	9	9	10	9	9	9	9	10
AAE Tweets	9	8	9	20	9	14	9	12	9	8	8
Bible	11	9	9	9	20	9	9	9	9	10	9
Tweets	9	8	10	14	9	20	9	11	9	9	9
James Joyce	9	9	9	9	9	9	20	9	9	11	10
Lyrics	9	8	9	12	9	11	9	20	10	8	9
Poetry	9	9	9	9	9	9	9	10	20	9	10
Shakespeare	10	9	9	8	10	9	11	8	9	20	10
Switchboard	9	10	10	8	9	9	10	9	10	10	20

Figure 4: Cosine similarities between styles in CDS using the `[CLS]` vectors of the RoBERTa-large classifier (normalized to $[0, 20]$). The off-diagonal elements show intuitive domain similarities, such as (Lyrics, Poetry); (AAE, Tweets); (Joyce, Shakespeare) or among classes from the COHA corpus.

p value for nucleus sampling, we notice a trend similar to the **Nucleus sampling trades off ACC for SIM** experiment in Section 5. Increasing the p value improves ACC at the cost of SIM. However unlike the Shakespeare and Formality dataset, we find $p = 0.6$ the optimal value for the best ACC-SIM tradeoff.

Note that Fluency scores on this dataset could be misleading since even the original sentences from some styles are often classified as disfluent (Orig. FL). Qualitatively, this seems to happen for styles with rich lexical and syntactic diversity (like Romantic Poetry, James Joyce). These styles tend to be out-of-distribution for the fluency classifier trained on the CoLA dataset (Warstadt et al., 2019).

A.9 A Survey of Evaluation Methods

We present a detailed breakdown of evaluation metrics used in prior work in Table 10 and the implementations of the metrics in Table 11. Notably, only 3 out of 23 prior works use an absolute sentence-level aggregation evaluation. Other works either perform "overall A/B" testing, flawed corpus-level aggregation or don't perform any aggregation at all. Note that while "overall A/B" testing cannot be gamed like corpus-aggregation, it has a few issues — (1) it is a *relative* evaluation and does not provided an *absolute* performance score for future reference; (2) "A/B" testing requires human evaluation, which is expensive and noisy; (3) evaluating overall performance will require human annotators to be familiar with the styles and style transfer task setup; (4) Kahneman (2011) has shown that asking humans to give a single number for "overall score" is biased when compared to an aggregation of *independent* scores on different metrics. Luckily, the sentence-level aggregation in Li et al. (2018) does the latter and is the closest equivalent to our proposed $J(\cdot)$ metric.

A.10 Details on Human Evaluation

We conduct experiments of Amazon Mechanical Turk, annotating the paraphrase similarity of 150 sentences with 3 annotators each. We report the label chosen by two or more annotators, and collect additional annotations in the case of total disagreement. We pay workers 5 cents per sentence pair ($10-15 / hr). We only hire workers from USA, UK and Australia with a 95% or higher approval rating and at least 1000 approved HITs. Sentences where the input was exactly copied (after lower-casing and removing punctuation) are automatically assigned the option **2** paraphrase and grammatical. Even though these sentences are clearly not style transferred, we expect them to be penalized in J(ACC,SIM,FL) by poor ACC. We found that every experiment had a Fleiss kappa (Fleiss, 1971) of at least 0.13 and up to 0.45 (slight to moderate agreement according to (Landis and Koch, 1977)). A qualitative inspection showed that crowdworkers found it easier to judge sentence pairs in the Formality dataset than Shakespeare, presumably due to greater familiarity with modern English. We also note that crowdworkers had higher agreement for sentences which were clearly not paraphrases (like the UNMT / DLSM generations on the Formality dataset).

Calculating Metrics in Table 2: To calculate SIM, we count the percentage of sentences which humans assigned a label **1** (ungrammatical paraphrase) or **2** (grammatical paraphrase). This is used as a binary value to calculate J(ACC, SIM). To calculate J(ACC, SIM, FL), we count sentences which are correctly classified as well as humans assigned a label of **2** (grammatical paraphrase). We cannot calculate FL alone using the popular 3-way evaluation, since the fluent sentences which are not paraphrases are not recorded.

A.11 More Example Generations

More examples are provided in Table 9. All of our style transferred outputs on CDS will be available in the project page of this work, `http://style.cs.umass.edu`.

A.12 More Related Work

Our inverse paraphrase model is a style-**controlled text generator** which automatically learns lexical and syntactic properties prevalent in the style's corpus. Explicit syntactically-controlled text generation has been studied previously using labels such as constituency parse templates (Iyyer et al., 2018; Akoury et al., 2019) or learned discrete latent templates (Wiseman et al., 2018). Syntax can also be controlled using an exemplar sentence (Chen et al., 2019; Guu et al., 2018; Peng et al., 2019). While style transfer requires the underlying content to be provided as input, another direction explores attribute-controlled *unconditional* text generation (Dathathri et al., 2020; Keskar et al., 2019; Zeng et al., 2020; Ziegler et al., 2019).

Diversity in text generation is often encouraged during inference time via heuristic modifications to beam search (Li et al., 2016; Vijayakumar et al., 2018), nucleus sampling (Holtzman et al., 2020) or submodular optimization (Kumar et al., 2019); in contrast, we simply filter our training data to increase diversity. Other algorithms learn to condition generation on latent variables during training (Bowman et al., 2016), which are sampled from at inference time to encourage diversity (Jain et al., 2017; Gupta et al., 2018; Park et al., 2019). Relatedly, Goyal and Durrett (2020) promote syntactic diversity of paraphrases by conditioning over possible syntactic rearrangements of the input.

Direction	Input	Output
1810. → Swit.	She forgot her gold pieces lying on the table in front of her.	and uh she forgot the gold pieces on the desk
1810. → AAE	This story is not well devised, my lord.	Lord this story ain’t no good
Poetry → Swit.	Why should our young Endymion pine away!	i don’t know why we should have such a letdown for our young Endymion
Tweet → Shak.	This is an insult to the Munster’s.	Tis an affront to the Munster.
Swit. → 1810.	and my lawn is very green and my flowers are still blooming	I have plenty of green turf and flowers here.
1890. → AAE	I want no pursuit when I do my work.	I dont wanna chase when im at work
1990. → Joyce	“You’re scaring me,” Clara says .	’Tis scare me, says Clara.
Tweet → Bible	Can we build AI without losing control over it?	Can a man build a computer without losing his power?
AAE → 1810.	I Use To Think They Were Ugly.	I thought them odious–and hateful.
Bible → 1990.	hath thy soul lothed Zion?	Is Zion a place where you have your soul?
Bible → Tweet	This persuasion cometh not of him that calleth you.	Its not him who comes from. Its who you call
Swit. → AAE	um yeah just a second	yea just a lil bit
Joyce → Shak.	Smell of burn.	This is the smoke of a burning.
1990. → Joyce	The Second World War, son, ” he answered.	SECOND WORLD WAR, son, says he.
1990. → AAE	no, I implore you not to consider this option.	Naw I asked u not to think bout that option
Tweet → Joyce	i wanted to sleep but now i cant	sleep is my preferred repose.
1990. → 1810.	She ninety-three and she can’t hear too good.	She is 93 years of age and cannot hear well yet.
AAE → Poetry	y’all play? we spray	And weel weel, we’ll spray a’ that.
Joyce → Poetry	Now, don’t tell a big fib!	And say not so big a lie!
1810. → Bible	without first veiling that, I could not sin.	And I could not sin without veil.
1890. → Shak.	When did we ever find Gaston de Marignac objectionable?	When was ’t that we found Gaston de Marignac’s foul faults?
Shak. → 1990.	O, much I fear some ill unlucky thing.	I was afraid of a lot of misadventures.
Shak. → 1810.	By Jesu, a very good blade!	The very knife of heaven, Jesu!
Lyrics → 1890	For a thing to explode	Oh, oh, oh, for something to blow up!
Lyrics → Bible	You are the sun that lights my world	Thou art the sun that lighteth mine earth: thou art my righteousness.
Lyrics → Joyce	But now it shouldn’t have to be this way	But oughtn’t to be like that.
Swit. → 1810.	oh that poor child	(sighing) the poor dear little thing.
Swit. → Lyrics	you know Karen i wonder if we are recording	Karen, I wonder if we’re tapein’
Tweet → Bible	Ok will get back to you ASAP!	And he said, All right; I will return.
Tweet → Shak.	Movies are always funnier in the cinema.	It always is more humorous in the theatre.
Lyrics → Poetry	You know I still fear the bomb	I was yet a-fear’d at the bomb,
Joyce → 1890.	Just missed that by a hair.	It was a hair-spitting thing I missed.
Joyce → 1810.	Me seemeth a dragon man.	Hark ye, I’m a dragon man, Me.
Poetry → Bible	Of deities or mortals, or of both,	Of the gods, and of the people, or of the both of them together.
Poetry → Lyrics	Makes countless thousands mourn!	Gonna be a lot of people mourning!
Poetry → Bible	My helpless lambs, I trust them wi’ him.	Surely he will feed my lambs in safety.
1890. → Poetry	Stark was sentenced to ten years’ imprisonment, Gibbon to five.	Stark was sentence’d ten yeares, Gibbon five;
1890. → Swit.	I want no pursuit when I do my work.	i don’t want to chase after when i’m at work
Tweet → Bible	The way this is set up has me dying	This death cometh because of this.
Tweet → Swit.	Yay! Ill order them then! I dont have patience to wait on them!	now i’m gonna order those
Lyrics → Joyce	And all the guns is in Kesha’s name	All your gunsees belong to Kesha.
AAE → 1810.	boredom gonna be the death of me one day	I shall die, one day, of the insupportable want of amusement.
AAE → 1890.	That’s just what I needed to see.... Thank Ya Lord	Thank you, Lord; that is just what I was expecting.
AAE → Swit.	okay ii will see you later	yeah see you later bye
Poetry → Tweet	Fam’d heroes! had their royal home:	royal bloods heroes:
Tweet → Bible	Check out this new painting that I uploaded to!	Look upon my new picture that I have set before thee!
Swit. → Shak.	so uh what do you wear to work	And what dost thou wear for thy work?
Tweet → Poetry	Now I gotta delete it	O now, must I part? And can I now erase
Tweet → 1810.	#India is now producing the worlds cheapest solar power #energy	Now is India’s solar power cheapest of all the world.
Poetry → Joyce	Away, away, or I shall dearly rue	O offside, away, or do I am rather sad.
Tweet → Swit.	Oh shit ima be a senior	so uh i got to the senior level of the business

Table 9: More example outputs from our model STRAP trained on our dataset CDS. Our project page will provide all 110k style transferred outputs generated by STRAP on CDS.

Paper	Automatic					Human				
	ACC	SIM	FL	CA	SA	ACC	SIM	FL	CA	SA
Hu et al. (2017)	✓									
Shen et al. (2017)	✓					✓		✓		A/B
Shetty et al. (2018)	✓						A/B			
Fu et al. (2018)	✓	✓					✓			
Li et al. (2018)	✓	✓				✓	✓	✓		✓
Zhang et al. (2018)	✓	✓				✓	✓	✓		✓
Nogueira dos Santos et al. (2018)	✓	✓	✓							
Prabhumoye et al. (2018)	✓						A/B	✓		
Xu et al. (2018)	✓	✓		✓		✓	✓		✓	
Logeswaran et al. (2018)	✓	✓	✓			✓	✓	✓		
Yang et al. (2018)	✓	✓	✓							
Subramanian et al. (2019)	✓	✓	✓			✓	✓	✓		A/B
Luo et al. (2019)	✓	✓		✓		✓	✓	✓	✓	✓
Pang and Gimpel (2019)	✓	✓	✓	✓		A/B	A/B	A/B		A/B
Ma et al. (2019)	✓	✓	✓			✓	✓	✓		
Dai et al. (2019)	✓	✓	✓			A/B	A/B	A/B		
Sudhakar et al. (2019)	✓	✓	✓			A/B	A/B	A/B		A/B
Mir et al. (2019)	✓	✓	✓			✓	✓	✓		
Gröndahl and Asokan (2019)	✓	✓					✓			
Tikhonov et al. (2019)	✓	✓								
Syed et al. (2020)	✓	✓								
Madaan et al. (2020)	✓	✓				✓	✓	✓		
He et al. (2020)	✓	✓	✓							
Ours	✓	✓	✓	✓	✓		✓	✓		✓

Table 10: Survey of evaluation methods used in 23 prior papers. We check whether prior work evaluate their algorithm on transfer accuracy (ACC), semantic similarity (SIM), fluency (FL), corpus-level aggregation (CA) and sentence-level aggregation (SA). We use the “A/B” to denote relative comparisons via A/B testing between generations from the baseline and the proposed system, rather than absolute performance numbers. Specific implementations of the metrics have been provided in Table 11. We do not include Pang (2019) since it’s a survey of existing evaluation methods.

Paper	Automatic			Human		
	ACC	SIM	FL	ACC	SIM	FL
Hu et al. (2017)	L-CNN					
Shen et al. (2017)	CNN			Likert-4		Likert-4
Shetty et al. (2018)	RNN/CNN	METEOR			A/B	
Fu et al. (2018)	LSTM	GloVE			Likert-3	
Li et al. (2018)	LSTM	BLEU		Likert-5	Likert-5	Likert-5
Zhang et al. (2018)	GRU	BLEU		Likert-5	Likert-5	Likert-5
Nogueira dos Santos et al. (2018)	SVM	GloVE	PPL			
Prabhumoye et al. (2018)	CNN				A/B	Likert-4
Xu et al. (2018)	CNN	BLEU		Likert-10	Likert-10	
Logeswaran et al. (2018)	CNN	BLEU	PPL	Likert-5	Likert-5	Likert-5
Yang et al. (2018)	CNN	BLEU	PPL			
Subramanian et al. (2019)	fastText	BLEU	PPL	Binary	Likert-5	Likert-5
Luo et al. (2019)	CNN	BLEU		Likert-5	Likert-5	Likert-5
Pang and Gimpel (2019)	CNN	GloVE	PPL	A/B	A/B	A/B
Ma et al. (2019)	CNN	BLEU	PPL	Likert-5	Likert-5	Likert-5
Dai et al. (2019)	fastText	BLEU	PPL	A/B	A/B	A/B
Sudhakar et al. (2019)	fastText	GLEU	PPL	A/B	A/B	A/B
Mir et al. (2019)	EMD	GloVE*	Classify	Likert-5*	Likert-5*	Binary*
Gröndahl and Asokan (2019)	LSTM/CNN	METEOR				
Tikhonov et al. (2019)	CNN	BLEU				
Syed et al. (2020)	FineGrain	BLEU				
Madaan et al. (2020)	AWD-LSTM	METEOR		Likert-5	Likert-5	Likert-5
He et al. (2020)	CNN	BLEU	PPL			
Ours	RoBERTa-L	SIM-PP	Classify		Binary	Binary

Table 11: Survey of implementations of evaluation metrics to measure Accuracy (ACC), Similarity (SIM) and Fluency (FL) used in 23 prior papers. For a cleaner version of this table with aggregation information, see Table 10. The * marks in Mir et al. (2019) denote a carefully designed unique implementation. We do not include Pang (2019) since it's a survey of existing evaluation methods.

Model	Formality					Shakespeare				
	ACC	SIM	FL	GM(A,S,F)	J(A,S,F)	ACC	SIM	FL	GM(A,S,F)	J(A,S,F)
COPY	5.2	41.8	88.4	26.8	0.2	9.6	20.1	79.1	24.8	0.1
NAÏVE	49.7	22.1	89.4	44.4	2.4	49.9	10.5	78.9	34.6	1.1
REF	93.3	100	89.7	94.2	88.2	90.4	100	79.1	89.4	67.2
UNMT	78.5	15.1	52.5	39.7	11.7	70.5	7.9	49.6	30.2	1.7
DLSM	78.0	18.5	53.7	42.6	9.5	71.1	12.5	49.4	35.2	2.0
STRAP ($p = 0.0$)	67.7	28.8	90.4	56.1	19.3	71.7	10.3	85.2	39.8	5.9
STRAP ($p = 0.6$)	70.7	25.3	88.5	54.1	17.2	75.7	8.8	82.7	38.1	5.4
STRAP ($p = 0.9$)	76.8	17.0	77.4	46.6	12.2	79.8	6.1	71.7	32.7	3.4

Table 12: A table equivalent to Table 1 but using BLEU scores for SIM instead of the paraphrase similarity model from Wieting et al. (2019). The Formality dataset had 4 available reference sentences whereas the Shakespeare dataset had only 1 available reference sentence. Our system STRAP significantly beats prior work (UNMT, DLSM) on $J(\cdot)$ metrics even with BLEU scores.

Model	ACC (A)	SIM (S)		FL (F)	J(A,S)		J(A,S,F)	
		BL	PP		BL	PP	BL	PP
COPY	8.0	32.6	80.9	90.1	0.4	7.1	0.3	6.4
REF	87.8	100	100	90.1	91.1	87.8	83.5	78.9
NAÏVE	67.9	10.7	32.0	91.5	1.7	9.3	1.5	8.5
BT (Prabhumoye et al., 2018)	47.4	1.3	21.1	8.0	0.7	11.4	0.0	1.3
MultiDec (Fu et al., 2018)	26.0	12.0	36.9	15.1	1.4	8.9	0.0	1.5
Del. (Li et al., 2018)	24.2	30.1	53.5	20.8	3.1	10.2	0.0	1.6
Unpaired (Xu et al., 2018)	53.9	1.6	16.3	34.9	0.4	10.9	0.0	2.2
DelRetri. (Li et al., 2018)	52.8	21.9	47.6	16.3	11.9	23.4	0.2	4.2
CrossAlign. (Shen et al., 2017)	59.0	3.3	25.0	31.7	2.0	14.9	0.3	5.2
Retri. (Li et al., 2018)	90.0	0.5	9.0	62.1	0.5	8.3	0.3	5.5
Templ. (Li et al., 2018)	37.1	36.4	67.8	32.3	11.9	23.7	1.3	7.8
DualRL (Luo et al., 2019)	51.8	45.0	65.1	59.0	14.6	29.9	8.1	21.7
UNMT (Zhang et al., 2018)	64.5	34.4	64.8	45.9	28.2	41.2	14.7	22.1
STRAP ($p = 0.0$)*	57.7	31.1	69.7	93.8	19.5	40.8	**18.3**	38.7
STRAP ($p = 0.6$)*	63.4	26.5	66.7	91.4	18.3	**43.0**	17.1	**40.0**
STRAP ($p = 0.9$)*	70.3	17.3	59.0	81.4	13.6	41.6	11.8	34.3

Table 13: More comparisons against prior work on the Formality dataset (Rao and Tetreault, 2018) using the outputs provided in the publicly available codebase of Luo et al. (2019) using both BLEU score (BL) and paraphrase similarity (PP). This model uses the Family & Relationships split of the Formality dataset whereas (He et al., 2020) used the Entertainment & Music split. Hence, we have retrained our RoBERTa-large classifiers to reflect the new distribution. ***Note**: While our system significantly outperforms prior work, we re-use the formality system used in Table 1 and Table 2 for these results, which was trained on Entertainment & Music (consistent with He et al. (2020)). There could be a training dataset mismatch between our model and the models from Luo et al. (2019), since the Formality dataset has two domains. This is not clarified in Luo et al. (2019) to the best of our knowledge.

Style	Train	Dev	Test	Source	Examples
Shakespeare	24,852	1,313	1,293	Shakespeare split of Xu et al. (2012).	1. *Why, Romeo, art thou mad?* 2. *I beseech you, follow straight.*
English Tweets	5,164,874	39,662	39,690	A random sample of English tweets collected on 8th-9th July, 2019 using Twitter APIs.	1. *Lol figures why I dont wanna talk to anyone rn* 2. *omg no problem i felt bad holding it! i love youuuu*
Bible	31,404	1,714	1,714	The English Bible collected from Project Gutenberg (Hart, 1992) (link).	1. *Jesus saith unto her, Woman, what have I to do with thee?* 2. *Wherefore it is lawful to do well on the sabbath days.*
Romantic Poetry	26,880	1,464	1,470	The Romantic section of the Poetry bookshelf on Project Gutenberg (link).	1. *There in that forest did his great love cease;* 2. *But, oh! for Hogarth's magic pow'r!*
Switchboard	145,823	1,487	1,488	Conversational speech transcripts (link) from the Switchboard speech recognition corpus (Godfrey et al., 1992).	1. *uh-huh well we're not all like that um* 2. *well yes i i well i- i don't think i have the time to really become a student in every article*
AAE (African American English) Tweets	717,634	7,316	7,315	Using the geo-located tweet corpus collected by Blodgett et al. (2016).	1. *ay yall everything good we did dat...* 2. *I know data right, it don't get more real than that.*
James Joyce	37,082	2,054	2,043	Two novels (Ulysses, Finnegans) of James Joyce from Project Gutenberg (link) and the Internet Archive (link).	1. *At last she spotted a weeny weeshy one miles away.* 2. *chees of all chades at the same time as he wags an antomine art of being rude like the boor.*
Lyrics	4,588,522	252,368	252,397	Music lyrics dataset from MetroLyrics, used in a Kaggle competition (link).	1. *I gotta get my mind off you,* 2. *This is it, we are, baby, we are one of a kind*
1810-1830 historical English	205,286	5,340	5,338	1810-1830 in the Corpus of Historical American English (Davies, 2012) using fiction, non-fiction and magazine domains (link).	1. *The fulness of my fancy renders my eye vacant and inactive.* 2. *What then do you come hither for at such an hour?*
1890-1910 historical English	1,210,687	32,024	32,018	1890-1910 in the Corpus of Historical American English using fiction, non-fiction and magazine domains (link).	1. *Nor shall I reveal the name of my friend; I do not wish to expose him to a torrent of abuse.* 2. *You know olive oil don't give the brightest illumination.*
1990-2010 historical English	1,865,687	48,985	48,982	1990-2010 in the Corpus of Historical American English using fiction, non-fiction and magazine domains (link).	1. *They were, in fact, tears of genuine relief.* 2. *I don't know why, but I sensed there was something wrong.*
Total	14,018,731	393,727	393,748		

Table 14: Details of our new benchmark dataset CDS along with representative examples. Our dataset contains eleven lexically and syntactically diverse styles and has a total of nearly 15M sentences, an order of magnitude larger than previous datasets. We will provide more representative examples along with our entire dataset in the project page `http://style.cs.umass.edu`.

Split	Orig. ACC	Orig. FL	Model	ACC (A)	SIM (S)	FL (F)	J(A,S)	J(A,S,F)
AAE Tweets	87.6	56.4	Ours ($p = 0.0$)	21.0	70.1	71.6	12.6	8.3
			Ours ($p = 0.6$)	32.5	65.7	63.5	18.3	**10.2**
			Ours ($p = 0.9$)	46.1	57.8	45.9	**23.6**	9.8
Bible	98.3	87.5	Ours ($p = 0.0$)	48.0	58.4	81.2	24.7	20.9
			Ours ($p = 0.6$)	52.5	55.1	79.8	**25.7**	**21.3**
			Ours ($p = 0.9$)	56.9	49.4	74.0	25.3	19.3
COHA 1810s-1820s	83.0	89.1	Ours ($p = 0.0$)	25.9	66.5	84.5	16.4	13.7
			Ours ($p = 0.6$)	34.0	63.0	81.5	20.1	16.0
			Ours ($p = 0.9$)	42.7	57.3	73.6	**22.9**	**16.5**
COHA 1890s-1900s	76.5	91.2	Ours ($p = 0.0$)	36.1	68.9	86.7	23.7	21.2
			Ours ($p = 0.6$)	41.1	65.7	83.8	**25.5**	**22.1**
			Ours ($p = 0.9$)	44.3	59.4	72.0	25.0	19.2
COHA 1990s-2000s	86.9	96.8	Ours ($p = 0.0$)	40.4	69.0	87.7	26.6	24.4
			Ours ($p = 0.6$)	46.1	65.6	86.0	**28.9**	**26.3**
			Ours ($p = 0.9$)	46.1	59.4	76.1	26.1	21.7
English Tweets	80.7	79.9	Ours ($p = 0.0$)	20.0	71.0	79.1	13.5	11.0
			Ours ($p = 0.6$)	28.9	67.5	72.2	18.1	**13.7**
			Ours ($p = 0.9$)	40.8	60.0	55.5	**22.7**	13.4
James Joyce	87.1	48.2	Ours ($p = 0.0$)	43.0	69.6	79.8	28.7	22.0
			Ours ($p = 0.6$)	52.2	63.7	62.8	32.0	**29.6**
			Ours ($p = 0.9$)	63.6	54.8	40.5	**33.5**	11.3
Lyrics	88.7	78.9	Ours ($p = 0.0$)	51.9	71.6	79.4	**35.6**	**29.0**
			Ours ($p = 0.6$)	53.4	68.6	71.4	34.8	26.0
			Ours ($p = 0.9$)	53.3	62.1	51.9	31.4	18.1
Romantic Poetry	93.8	40.2	Ours ($p = 0.0$)	55.0	63.8	58.9	33.5	**17.2**
			Ours ($p = 0.6$)	62.4	60.3	51.8	35.6	16.2
			Ours ($p = 0.9$)	69.8	55.3	40.3	**36.8**	13.0
Shakespeare	86.1	59.9	Ours ($p = 0.0$)	36.8	65.5	76.9	21.7	15.4
			Ours ($p = 0.6$)	52.1	58.6	65.4	28.2	**16.6**
			Ours ($p = 0.9$)	63.7	48.9	44.2	**29.3**	11.3
Switchboard	99.7	63.1	Ours ($p = 0.0$)	62.9	67.4	77.0	40.8	32.0
			Ours ($p = 0.6$)	77.2	63.7	64.2	**47.5**	**30.2**
			Ours ($p = 0.9$)	84.9	56.6	44.0	46.8	20.1
Overall	88.0	71.9	Ours ($p = 0.0$)	40.1	67.4	78.4	25.3	19.6
			Ours ($p = 0.6$)	48.4	63.4	71.1	28.6	**20.7**
			Ours ($p = 0.9$)	55.7	56.5	56.2	**29.4**	15.8

Table 15: A detailed performance breakup when transferring **to** each style in CDS from the other 10 styles. We test three nucleus sampling (Holtzman et al., 2020) strategies with our trained model by varying the p value between 0.0 (greedy) and 1.0 (full sampling). For reference, the classification accuracy (Orig. ACC) and fluency (Orig. FL) of original sentences in the target style corpus are provided.

Split	Diverse Paraphraser Similarity (↑)	Lexical (↓)	Syntactic (↓)	Non-Diverse Paraphraser Similarity (↑)	Lexical (↓)	Syntactic (↓)
AAE Tweets	65.1	44.7	0.43	74.3	66.4	0.82
Bible	74.6	48.5	0.55	88.3	73.5	0.92
COHA 1810s-1820s	74.0	50.6	0.51	86.3	71.8	0.92
COHA 1890s-1900s	75.3	52.0	0.50	88.2	75.3	0.93
COHA 1990s-2000s	77.6	57.4	0.53	89.9	80.7	0.95
English Tweets	73.1	52.4	0.50	82.8	75.7	0.91
James Joyce	71.5	47.8	0.35	82.4	69.8	0.82
Lyrics	74.5	52.8	0.52	86.7	78.6	0.92
Romantic Poetry	72.3	46.3	0.44	81.3	67.1	0.86
Shakespeare	67.9	38.7	0.23	81.4	63.4	0.75
Switchboard	71.6	50.1	0.55	81.1	72.4	0.90
Overall	72.5	49.2	0.46	83.9	72.3	0.88

Table 16: A detailed style-wise breakup of the **diverse paraphrase quality** in CDS (the training data for the inverse paraphrase model, described in Section 2.1 and Section 2.4). The ideal paraphraser should score lower on "Lexical" and "Syntactic" overlap and high on "Similiarity". Overall, our method achieves high diversity (51% unigram change and 27% word shuffling, compared to 28% unigram and 6% shuffling for non-diverse), while maintaining good semantic similarity (SIM= 72.5 vs 83.9 for non-diverse) even in complex stylistic settings. We measure lexical overlap in terms of unigram F1 overlap using the evaluation scripts from Rajpurkar et al. (2016). Syntactic overlap is measured using Kendall's τ_B (Kendall, 1938) of shared vocabulary. A $\tau_B = 1.0$ indicates no shuffling whereas a value of $\tau_B = -1.0$ indicates 100% shuffling (complete reversal). Finally, the SIM model from Wieting et al. (2019) is used for measuring similarity.