

# STYLEPTB: A Compositional Benchmark for Fine-grained Controllable Text Style Transfer

Yiwei Lyu<sup>♡\*</sup>, Paul Pu Liang<sup>♡\*</sup>, Hai Pham<sup>♣\*</sup>,  
Eduard Hovy<sup>♣</sup>, Barnabás Póczos<sup>♡</sup>, Ruslan Salakhutdinov<sup>♡</sup>, Louis-Philippe Morency<sup>♣</sup>

<sup>♡</sup>Machine Learning Department, Carnegie Mellon University

<sup>♣</sup>Language Technologies Institute, Carnegie Mellon University

{ylyu1, pliang, htpham}@cs.cmu.edu

<https://github.com/lvyiwei1/StylePTB/>

## Abstract

Text style transfer aims to controllably generate text with targeted stylistic changes while maintaining core meaning from the source sentence constant. Many of the existing style transfer benchmarks primarily focus on individual high-level semantic changes (e.g. positive to negative), which enable controllability at a high level but do not offer *fine-grained* control involving sentence structure, emphasis, and content of the sentence. In this paper, we introduce a large-scale benchmark, STYLEPTB, with (1) paired sentences undergoing 21 fine-grained stylistic changes spanning atomic lexical, syntactic, semantic, and thematic transfers of text, as well as (2) *compositions* of multiple transfers which allow modeling of fine-grained stylistic changes as building blocks for more complex, high-level transfers. By benchmarking existing methods on STYLEPTB, we find that they struggle to model fine-grained changes and have an even more difficult time composing multiple styles. As a result, STYLEPTB brings novel challenges that we hope will encourage future research in controllable text style transfer, compositional models, and learning disentangled representations. Solving these challenges would present important steps towards controllable text generation.

## 1 Introduction

At the heart of interactive AI systems lies the element of communication as a channel to convey intentions using different stylistic attributes. Research in human-AI interaction has focused on building dialog systems (Celikyilmaz et al., 2018), virtual assistants (Cooper et al., 2004), and intelligent agents (Kim et al., 2013; Liang et al., 2020a; Pittermann et al., 2010) that can communicate their intentions with specific *styles* for different situations, target audiences, and environments (Lample

\* authors contributed equally

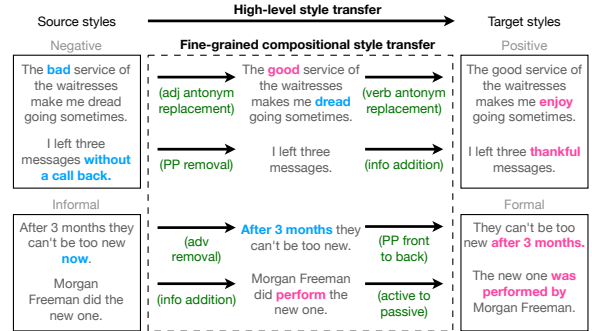


Figure 1: STYLEPTB provides a large-scale resource to study *fine-grained compositional style transfer*. The styles provided in STYLEPTB (in **green**) span lexical, syntax, semantic, and thematic aspects (DiMarco and Hirst, 1993) which can be composed to form high-level style transfers as commonly studied in existing benchmarks (e.g. Yelp for sentiment (Shen et al., 2017) and GYAFC for formality (Rao and Tetreault, 2018)).

et al., 2019; Li et al., 2018). For example, expressing the same facts using either formal or informal styles can be more suitable for certain target audiences (Rao and Tetreault, 2018).

What is a *style* in natural languages? Existing style transfer benchmarks primarily focus on individual high-level stylistic changes across sentiment (Shen et al., 2017), formality (Rao and Tetreault, 2018), politeness (Madaan et al., 2020), and writing styles (Jhamtani et al., 2017). Figure 1 provides some motivating examples to show that the high-level style transfers as commonly studied in existing benchmarks (e.g. Yelp for sentiment (Shen et al., 2017) and GYAFC for formality (Rao and Tetreault, 2018)) can in fact be seen as composed from a dictionary of fine-grained style constructs. This alternative way of studying styles brings additional flexibility that enables *fine-grained* control with the possibility to *compose* a broader space of styles spanning tense, sentence structure, phrase emphasis, and information contained in the sentence. However, the missing link is a benchmark dataset that offers this type of fine-

grained style constructs, with the controllability to compose these stylistic transfers.

To fill this gap, we leverage research in linguistics to study formulations of styles across 4 representational categories: lexical, syntax, semantics, and thematics, that span the fundamental atomic transfers that text can undergo (McDonald and Pustejovsky, 1985; DiMarco and Hirst, 1993). Using these insights, we introduce a large-scale benchmark with (1) paired sentences undergoing 21 fine-grained stylistic changes spanning the most atomic lexical, syntactic, semantic, and thematic style constructs, as well as (2) *compositions* of multiple transfers which model how fine-grained style constructs compose to form more complex, high-level transfers. Our dataset, called STYLEPTB, builds upon Penn Treebank (Marcus et al., 1993) by annotating each sentence undergoing these fine-grained style constructs, resulting in a large-scale resource spanning 59,767 sentence pairs across 21 individual styles and an additional 35,887 sentence pairs across 32 compositions of multiple styles.

STYLEPTB allows us to study the performance of state-of-the-art style transfer models when faced with the new challenge of fine-grained style transfer. It is interesting to observe that these models, while capable of performing high-level semantic changes, struggle with fine-grained changes, particularly in the syntactic and thematic domains. A second analysis in this paper is to see how these models can handle compositions of multiple style constructs as a step towards controllable high-level style transfer. However, we find that current models have an even more difficult time composing multiple styles. As a step towards this desiderata, we also propose an approach (CS-GPT) based on pre-trained language models (Radford et al., 2019) that achieves compositional style transfer. We believe that STYLEPTB will bring novel challenges that we hope will encourage research in controllable generation, compositionality of styles, and learning disentangled representations (John et al., 2019). From a broader perspective, we conclude with the observation that controllable style transfer models trained on STYLEPTB can help mitigate social biases in pre-trained language models.

## 2 Related Work

Several lines of research have aimed to formalize **styles in natural languages** through computational and linguistic perspectives (DiMarco and Hirst, 1993). The first systematic formulation of

styles was by McDonald and Pustejovsky (1985) and later extended by DiMarco and Hirst (1993) to 4 representational categories including lexical, syntax, thematic, and semantic aspects. Following this, there has been some early efforts applying stylistic analysis into dialog generation (Hovy, 1987), machine translation (DiMarco, 1994), and text generation (Gatt and Krahmer, 2018). We take advantage of this prior work when formalizing our new STYLEPTB dataset.

Current **benchmarks for style transfer** focus on high-level style definitions such as transfer of sentiment (Shen et al., 2017; Lample et al., 2019; Li et al., 2018; Wu et al., 2019), politeness (Madaan et al., 2020), formality (Rao and Tetreault, 2018; Liu et al., 2020; Krishna et al., 2020), writing styles (Jhamtani et al., 2017; Syed et al., 2020; Jin et al., 2020) and some other styles (Kang and Hovy, 2019). However, these only focus on only high-level styles, unlike STYLEPTB.

**Computational models for style transfer** span statistical NLP methods (Hovy, 1987; Xu et al., 2012), neural generative models (Prabhumoye et al., 2018; Lample et al., 2019; He et al., 2020), and Retrieve-and-Edit approaches (Li et al., 2018; Hashimoto et al., 2018; Guu et al., 2018; Sudhakar et al., 2019; Madaan et al., 2020). These approaches work for a predefined set of styles but are unable to generalize to compositions of styles.

**Evaluating style transfer** is difficult due to the diversity of plausible transferred sentences. In addition to automatic scores such as BLEU, perplexity, or binary classification accuracy of style transfer (Hu et al., 2017; Lample et al., 2019; He et al., 2020), other automatic metrics (Fu et al., 2018; Mir et al., 2019) and human evaluation are also commonly used (Li et al., 2018; Shen et al., 2017).

## 3 Fine-Grained Style Constructs

As a step towards enabling *fine-grained* control with the possibility to *compose* a broader space of styles, we first define style constructs at fine-grained levels spanning lexical, syntactic, semantic, and thematic aspects. When selecting these style constructs, we have 2 goals in mind: (1) they should be representative of the four aspects (lexical, syntactic, semantic, thematic) following the formal categorizations in DiMarco and Hirst (1993), and (2) the transfers should be consistent (*i.e.* well-defined such that if multiple annotators are asked to modify the same sentence, the results will be similar). With these goals in mind, we summarize the

Aspect	Transfer	Original Sentence	Additional Info/ Emphasis	Transferred Sentence
LEXICAL	Noun synonym replacement	The <b>shift</b> wo n't affect operations.		The <b>displacement</b> wo n't affect operations.
	Noun antonym replacement	Investors will develop thicker skins and their <b>confidence</b> will return he says.		Investors will develop thicker skins and their <b>diffidence</b> will return he says.
	Verb synonym replacement	The meeting is <b>expected</b> to call for heightened austerity for two years.		The meeting is <b>anticipated</b> to call for heightened austerity for two years.
	Verb antonym replacement	He <b>noted</b> that higher gasoline price will help buoy the October totals.		He <b>ignored</b> that higher gasoline prices will help buoy the October totals.
	ADJ synonym replacement	Most other states have enacted <b>similar</b> bans.		Most other states have enacted <b>alike</b> bans.
	ADJ antonym replacement	It is also planning another night of <b>original</b> series.		It is also planning another night of <b>unoriginal</b> series.
	Most frequent synonym replacement	Republicans countered that long-range revenue <b>estimates</b> were unreliable.		Republicans countered that long-range revenue <b>judges</b> were unreliable.
	Least frequent synonym replacement	Merrill Lynch Capital Markets Inc. is the sole <b>underwriter</b> for the <b>offering</b> .		Merrill Lynch Capital Markets Inc. is the sole <b>investment-banker</b> for the <b>oblation</b> .
SYNTAX	To future tense	It <b>is</b> also planning another night of original series.		It <b>will be</b> also planning another night of original series.
	To present tense	Sen. Mitchell <b>urged</b> them to desist.		Sen. Mitchell <b>urges</b> them to desist.
	To past tense	It <b>is</b> also planning another night of original series.		It <b>was</b> also planning another night of original series.
	Active to passive	He also received 20-year sentences for each of the 24 passengers injured.		20-year sentences also were received by him for each of the 24 passengers injured.
	Passive to active	Most bills are drafted by bureaucrats not politicians.		Bureaucrats not politicians draft most bills.
	PP front to back	<b>In Indianapolis</b> Lilly declined comment.		Lilly declined comment <b>in Indianapolis</b> .
	PP back to front	The dollar has been strong <b>unlike 1987</b> .		<b>Unlike 1987</b> the dollar has been strong.
SEMANTICS	ADJ or ADV removal	The controls on cooperatives appeared <b>relatively</b> liberal when first introduced		The controls on cooperatives appeared liberal when introduced
	PP removal	The controls <b>on cooperatives</b> appeared relatively liberal when first introduced.		The controls appeared relatively liberal when first introduced.
	Substatement removal	The controls on cooperatives appeared relatively liberal <b>when first introduced</b> .		The controls on cooperatives appeared relatively liberal.
	Information addition	He reports his business is up slightly from customers replacing old stock.	[ 'customer', 'waiting to buy', 'seafood' ]	He reports his business is up slightly from customers <b>waiting to buy seafood</b> and replacing old stock.
THEMATICS	Verb/Action emphasis	He intends to add to the litigation staff.	<b>add</b>	<b>Adding</b> to the litigation staff is what he intends to do.
	Adjective emphasis	The comparable year-earlier number was 56 million a spokesman said.	<b>comparable</b>	A spokesman said the year-earlier number of 56 million <b>was comparable</b> .

Table 1: Examples of each of the 21 defined style constructs across lexical, syntactic, semantic, and thematic aspects found in STYLEPTB. The original phrase is in **cyan** and the corresponding target phrase is in **magenta**. Note that some thematic and semantic transfers require additional information, highlighted in **red**.

following 21 chosen fine-grained style constructs spanning 4 categories and also provide detailed examples in Table 1.

**Lexical transfers** are those at fine-grained lexicon levels (*i.e.* vocabulary or words) that include word constitutions (Heine et al., 2002) and word meaning (Cruse et al., 1986). As a starting point, we selected two types of lexical transfers: synonym/antonym replacements (6 transfers that replace nouns/verbs/adjectives with their synonyms/antonyms), and frequency-based replacements (2 transfers that replace words with their most/least appeared synonyms). The synonym/antonym resources are taken from Wordnet (Fellbaum, 2012).

**Syntax transfers** modify the underlying grammatical rules that govern the structure of sen-

tences (Chomsky, 2002) without affecting the content (Akmajian and Heny, 1980). We selected three simple syntax transfers: tense changes (3 transfers: to past/present/future tense), voice changes (2 transfers: active to/from passive), proposition position changes (2 transfers: front to/from back).

**Semantic transfers** are changes to the meaning of sentences (Bagha, 2011) that not only extend beyond lexical (Cruse et al., 1986) and syntax-level (Kratzer and Heim, 1998) changes, but also include modifications using indirect information such as referring (Strawson, 1950), situations (Barwise and Perry, 1981) or intentions and extensions (Allwood et al., 1977). As a starting point, we defined two simple types of semantic transfers: (1) Info removal: 3 transfers on different deletions: word-level (removing adjectives and adverbs), phrase

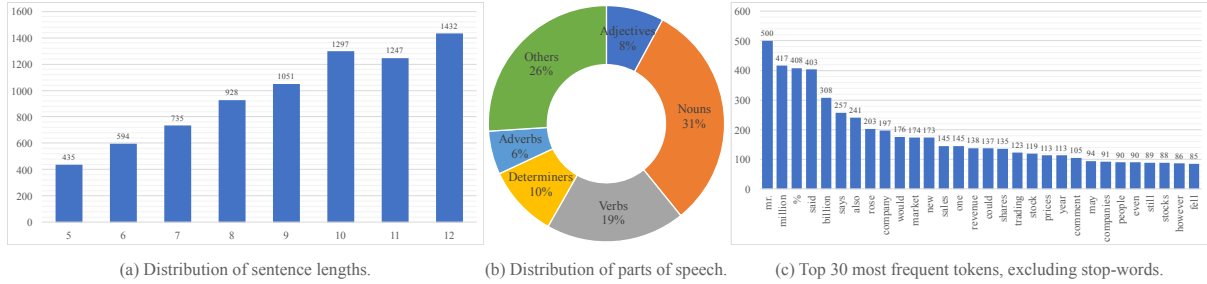


Figure 2: Statistics: (a) the distribution of sentence lengths, (b) count of word tokens by part-of-speech, and (c) the top 30 most frequent tokens. STYLEPTB exhibits diversity in sentence form and style transfer annotations.

level (removing propositions), and substatement level (removing entire substatements) that represent *referring* and *situations*, as well as (2) Info addition: 1 transformation that adds a given piece of information regarding a particular phrase in the current sentence representing *extension*.

**Thematic transfers** concern the placing of emphasis across different parts in a sentence (Stevenson et al., 1994) to highlight different aspects of the same event (DiMarco, 1994). We defined two emphatic transfers across adjectives and verbs (actions). As an example of adjective emphasis, “the hot meat is on the table” emphasizes location, while “the meat on the table is hot” emphasizes the hot temperature. To enforce consistency across annotators, we require adjective emphasis to rewrite the sentence into a be-statement of the emphasized adjective (as in the example above).

**Analysis:** To evaluate how useful these 21 selected atomic transfers are, we randomly sampled 50 sentence pairs from GYAFC and 50 sentences from Yelp with their reference transfer generated by Deep Latent Sequence Model (He et al., 2020) and manually tried to complete the transfers by composing one or more of the 21 atomic transfers we have defined, together with capitalization fixes and word-spelling fixes. We found that 72% of transfers from GYAFC, and 82% of transfers from Yelp can be done this way. Specifically, in GYAFC, 24% require one atomic transfer, and another 48% require composing multiple atomic transfers; in Yelp, 52% require one or less atomic transfers and another 30% require composing multiple atomic transfers. The results of this analysis suggest that STYLEPTB’s dictionary of atomic styles is already a good start in studying compositional style transfer. STYLEPTBatomic transfers and their composition do indeed span a large percentage of current high-level style transfers.

## 4 The STYLEPTB Dataset

Using these selected 21 style constructs, we now illustrate the steps towards collecting and annotating parallel sentences across style transfers.

### 4.1 Dataset Preprocessing

We use Penn Treebank (PTB) (Marcus et al., 1993) as our source of sentences. Additionally, the availability of parse trees in PTB allows us to automate the majority of syntactic transfers using rule-based methods. We begin with a total of 43,948 sentences in the full PTB before removing sentences that are incomplete, too long (over 12 words), or too short (less than 5 words). This leaves 7,719 sentences (see Figure 2 for statistics and Appendix A.1 for full details).

### 4.2 Generating transferred sentences

We give a brief overview of the data annotation process (see Appendix A.3 for full details).

**Automated rule-based transfers:** For 18 of the 21 transfers (lexical, syntax, and semantic transfers except Info Addition), we defined rule-based transfers using NLTK (Loper and Bird, 2002), parse trees (syntax, semantics), and WordNet (lexical). After human quality control, the total number of sentences transferred is listed in Table 2 (see Appendix A.2 for more details on automated generation and Appendix A.4 for human evaluation on quality of generated sentences)

**Transfers with human annotations:** For the remaining 3 transfers, we have human annotators (via Amazon Mechanical Turk) manually rewrite them due to the difficulty of automating the process. See Appendix A.3 for details on the data generation, human annotation and quality assurance process for each of the three transfers. After annotations and quality control, we obtained 696 rewritten sentences for adjective emphasis, 1201 rewritten sentences for verb emphasis, and 2114



Category	Transfer	Number
Lexical	Noun synonym replacement	5948
	Noun antonym replacement	2227
	Verb synonym replacement	2574
	Verb antonym replacement	1284
	ADJ synonym replacement	434
	ADJ antonym replacement	1146
	Most frequent synonym replacement	4722
	Least frequent synonym replacement	7112
Syntax	To future tense	7272
	To present tense	4365
	To past tense	4422
	Active $\leftrightarrow$ passive	2808
	PP front $\leftrightarrow$ back	467
Semantics (information deletion)	ADJ or ADV removal	4863
	PP removal	4767
	Substatement removal	1345
	Information Addition	2114
Thematic	Verb/action emphasis	1201
	Adj emphasis	696

Table 2: STYLEPTB is a large-scale resource spanning 59,767 sentence pairs across 21 individual styles.

Difficulty	Transfer	Hamming $\downarrow$
Easy	ADJ or ADV removal	1.531
	To Present tense	2.318
	To Past tense	2.447
	To Future tense	3.341
Medium	Information addition	3.729
	PP removal	4.079
	PP back to front	5.429
	Substatement removal	5.625
	PP front to back	6.235
Hard	Active to passive	8.147
	Passive to active	8.817
	Adjective emphasis	8.846
	Verb/Action emphasis	11.614

Table 3: Average token-level Hamming distance between original and transferred sentences for all syntax, semantics and thematic transfers.

valid sentence-information pairs with their transferred sentence with information added.

### 4.3 Relative Difficulty of Transfers

Lexical transfers can be done by replacing individual words and is simple to evaluate. To evaluate the difficulty of the remaining 13 syntax, semantic, and thematic transfers, we calculated the token-level (*i.e.* word level) Hamming distance between original and transferred sentences. Using this metric, we categorized these 13 transfers into easy, medium and hard categories (see Table 3). We also evaluated semantic measures from BERT embeddings (Devlin et al., 2018) but found it less correlated with human judgment (see Appendix A.5).

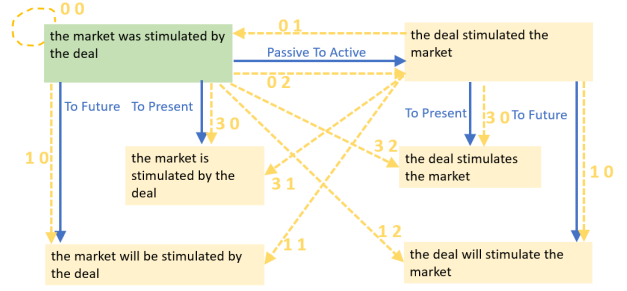


Figure 3: Example of generating sentence pairs that compose tense and voice changes. Starting from an original sentence (green box), we sequentially apply parse tree transfers (blue arrows) to obtain multiple transferred sentences (yellow box), yielding multiple parallel pairs (yellow arrows). We use transfer tokens ( $\Delta_1, \Delta_2$ ) to track changes (see Section 5 for details).

	No Voice Change (0)	Active To Passive (1)	Passive To Active (2)
No Tense Change (0)	2808	2808	2808
To Future Tense (1)	5294	2647	2647
To Past Tense (2)	2077	656	1421
To Present Tense (3)	3304	1536	1768

Table 4: Number of sentence pairs for each composition of tense change and voice change in the generated compositional dataset.

### 4.4 Compositional Transfers

To allow for compositionality, we also generated compositional data that includes parallel pairs of sentences linked by multiple sequential transfers. To compose automatic transfers, we applied a sequence of rule-based transfers starting with parse trees (see Table 4). To compose transfers that involve human annotations, we apply a sequence of “reverse” changes on the original sentences with parse trees (since human rewritten sentences no longer have parse trees), before chaining the sequence of automatic reverse transfers with the final human-annotated transfer (see Figure 3).

## 5 A Model for Compositional Transfer

We extend the pre-trained GPT2 language model (Radford et al., 2019) for parallel style transfer by giving it designated style transfer tokens as input in addition to the source sentence. For example, for each individual binary style  $s_i$ , we define a style transfer token  $\Delta_i \in \{0, 1, 2\}$  where  $\Delta_i = 0$  represents keeping  $s_i$  unchanged,  $\Delta_i = 1$  represents a change from  $s_i = 0$  to  $s_i = 1$ , and vice versa for  $\Delta_i = 2$ . We likewise extend the definition of  $\Delta_i$  for styles taking more than 2 values.

Given a parallel (source, target) pair  $(s, t)$ , we define the appropriate transfer token  $\Delta \in \{0, 1, 2\}$

Easy Transfers		Baseline Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CiDER
To Future Tense	GPT2		0.895	0.852	0.813	<b>0.778</b>	<b>0.540</b>	0.899	7.709
	SEQ2SEQ		0.527	0.368	0.261	0.188	0.173	0.531	1.525
	RETRIEVEEDIT		<b>0.899</b>	<b>0.854</b>	<b>0.815</b>	<b>0.778</b>	0.531	<b>0.901</b>	<b>7.731</b>
	HUMAN		0.954	0.915	0.884	0.855	0.636	0.964	9.174
ADJ or ADV Removal	GPT2		0.647	0.508	0.394	0.308	0.313	0.652	3.259
	SEQ2SEQ		0.450	0.274	0.172	0.112	0.140	0.469	1.171
	RETRIEVEEDIT		<b>0.897</b>	<b>0.841</b>	<b>0.786</b>	<b>0.731</b>	<b>0.511</b>	<b>0.919</b>	<b>7.461</b>
	HUMAN		0.933	0.894	0.870	0.847	0.591	0.965	8.924
Medium Transfers		Baseline Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CiDER
Substatement Removal	GPT2		0.430	0.332	0.247	0.176	0.250	0.588	3.090
	SEQ2SEQ		0.317	0.192	0.110	0.001	0.100	0.368	1.041
	RETRIEVEEDIT		<b>0.706</b>	<b>0.678</b>	<b>0.647</b>	<b>0.607</b>	<b>0.405</b>	<b>0.767</b>	<b>6.183</b>
	HUMAN		0.731	0.720	0.705	0.685	0.607	0.788	7.691
Information Addition	GPT2		0.479	0.305	0.189	0.121	0.207	0.475	1.359
	SEQ2SEQ		0.345	0.180	0.094	0.053	0.098	0.335	0.632
	RETRIEVEEDIT		<b>0.493</b>	<b>0.396</b>	<b>0.328</b>	<b>0.275</b>	<b>0.284</b>	<b>0.603</b>	<b>3.401</b>
	HUMAN		0.846	0.762	0.690	0.624	0.521	0.892	6.863
Hard Transfers		Baseline Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CiDER
Active To Passive	GPT2		0.476	0.329	0.238	0.189	0.216	0.464	1.820
	SEQ2SEQ		0.373	0.220	0.141	0.103	0.131	0.345	0.845
	RETRIEVEEDIT		<b>0.681</b>	<b>0.598</b>	<b>0.503</b>	<b>0.427</b>	<b>0.383</b>	<b>0.663</b>	<b>4.535</b>
	HUMAN		0.931	0.881	0.835	0.795	0.587	0.905	8.603
Adjective Emphasis	GPT2		0.263	0.079	0.028	0.000	0.112	0.188	0.386
	SEQ2SEQ		0.187	0.058	0.018	0.000	0.059	0.179	0.141
	RETRIEVEEDIT		<b>0.387</b>	<b>0.276</b>	<b>0.211</b>	<b>0.164</b>	<b>0.193</b>	<b>0.369</b>	<b>1.679</b>
	HUMAN		0.834	0.753	0.679	0.611	0.522	0.811	6.796
Verb/Action Emphasis	GPT2		0.309	0.170	0.095	0.041	0.140	0.292	0.593
	SEQ2SEQ		0.289	0.127	0.066	0.038	0.098	0.275	0.300
	RETRIEVEEDIT		<b>0.416</b>	<b>0.284</b>	<b>0.209</b>	<b>0.148</b>	<b>0.223</b>	<b>0.423</b>	<b>1.778</b>
	HUMAN		0.649	0.569	0.493	0.421	0.433	0.693	5.668

Table 5: Evaluation results on easy (top), medium (middle), and hard (bottom) transfers. Info Addition and thematic transfers are especially difficult for current models.

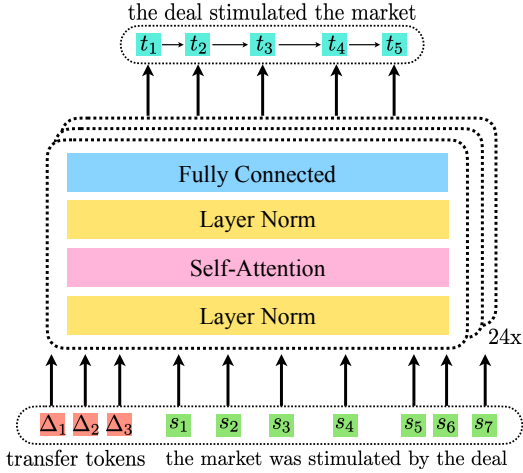


Figure 4: CS-GPT uses multiple transfer tokens  $\Delta_i \in \{0, 1, 2\}$  to enable compositional style transfer across multiple styles in our model STYLEPTB.

and train using maximum likelihood estimation to predict every word  $t_j$ , for  $j = 1, 2, \dots, T$ , in the target sentence given the source and  $\Delta$ :

$$\theta^* = \arg \max_{\theta} \mathbb{E}_{(s,t) \sim \mathcal{D}} \left[ \sum_{j=1}^T \log p_{\theta}(t_j; s, \Delta) \right], \quad (1)$$

where  $\theta$  denotes the pre-trained GPT2 parameters

and  $\theta^*$  denotes the parameters after fine-tuning on STYLEPTB. Note that we also train the model to reconstruct the same source sentence again when setting  $\Delta = 0$  (no style change), which we found to help bridge the domain shift between data used to pre-train GPT2 and sentences in STYLEPTB.

As a step towards *compositionality*, we also train with (source, target) pairs that undergo multiple atomic style transfers as provided in STYLEPTB, resulting in multiple style transfer tokens  $\Delta_i$  being activated at the same time. We call the resulting model CS-GPT (Compositional Style GPT) and show its architecture in Figure 4. Learning separate representations for each  $\Delta_i$  results in disentangled style variables that can then be composed as desired. Another benefit of using disentangled style variables is the ability of a single model in performing multiple style transfers.

## 6 Experiments

We test the performance of current style transfer models on STYLEPTB. Anonymized data and code is included in the supplementary, and we present extra details and results in Appendix B and C.

## 6.1 Datasets and Metrics

We use STYLEPTB and evaluate on the 13 non-lexical transfers (since lexical changes works best with fixed word substitutions). Please refer to Appendix B.1 for dataset preprocessing details. Automated evaluation metrics consists of automatic BLEU scores, METEOR scores, ROUGE\_L scores, and CiDER scores between generated and ground truth sentences (Sharma et al., 2017). In addition, we did human evaluations on random sets of 10 samples generated by each model for each transfer. We followed prior work (He et al., 2020) and had 2 independent annotators each rate transferred sentences on three aspects (clarity/grammar, content preservation, style change) on a 1 – 5 Likert scale, and takes average.

## 6.2 Baseline Models

We evaluate the following baselines commonly used in style transfer. Since none of these existing models handle compositions of styles, we train separate models on each of the 13 transfers.

**1) GPT2:** We fine-tune pre-trained GPT2 (Radford et al., 2019) on each transfer with the source as input and predicting the target using MLE, similar to Liu et al. (2020); Syed et al. (2020).

**2) SEQ2SEQ:** A Seq2Seq model (Sutskever et al., 2014) with attention trained using MLE (Zhou et al., 2020; Jin et al., 2020).

**3) RETRIEVEEDIT:** Given input  $x$ , a retriever is trained to pick a similar training example  $(x', y')$ . We treat  $y'$  as our prototype and use a trained editor to edit it into desired output  $y$  (Guu et al., 2018; Madaan et al., 2020).

**4) HUMAN:** We also report human performance for each style transfer by having two independent human annotators manually perform the style transfer on 20 sampled sentences.

## 6.3 Results and Observations

We evaluate these 3 baseline models on the style transfers in STYLEPTB and show results in Table 5. We make the following observations:

**Baseline comparisons:** RETRIEVEEDIT performed equally well compared to GPT2 in some transfers such as To Future Tense and performs significantly better compared to GPT2 in most transfers. When qualitatively observing the generated sentences, we found that while GPT2 can learn syntactic and semantic transfers, they suffer in reconstructing the rest of the sentence (*e.g.* making word repetitions). This was not an issue

	Clarity	Context	Style
GPT2	1.60	2.20	<b>4.05</b>
SEQ2SEQ	3.85	1.45	1.25
RETRIEVEEDIT	<b>4.15</b>	2.65	2.20
HUMAN	<b>4.70</b>	<b>4.45</b>	<b>5.00</b>

Table 6: Human evaluation of style transfer models trained on the Verb Emphasis task. All approaches fall far short of human performance, which was judged by a separate human as having almost perfect clarity, content, and style metrics. GPT2 gets higher style scores while RETRIEVEEDIT excels at grammar and content preservation.

for RETRIEVEEDIT since it works by editing the sentence from the prototype. Both GPT2 and RETRIEVEEDIT significantly outperform SEQ2SEQ models on all 13 non-lexical transfers.

**Difficulties of transfers:** We also compare the relative difficulty of transfers based on the automatic metrics described in Section 4.3. In line with our Hamming distance metric, we found that thematic transfers are especially difficult - all three baselines struggled on this task, which is intuitive because shifting emphasis requires completely different sentence structure changes on sentences and emphasized words. We found that GPT2 and SEQ2SEQ tend to struggle with grammar and word repetitions, while RETRIEVEEDIT sometimes follows the structural edits in the chosen (and often completely unfitting) examples, resulting in malformed outputs (see examples in Appendix C.1). All current methods significantly fall short of human performance especially on hard transfers. Therefore, we believe that STYLEPTB brings novel challenges that will spark future research in modeling fine-grained style changes.

**Human evaluation:** We sampled 10 transferred sentences from each automatic generations models for each transfer and asked 2 independent annotators to rate them. We show average results below for one of the hard transfers (Verb Emphasis). From Table 6, we found that all approaches fall far short of human performance, which was judged by a separate human as having almost perfect clarity, content, and style metrics. Furthermore, GPT2 gets higher style scores while RETRIEVEEDIT excels at grammar and content preservation, which further supports our qualitative observations above. Full results for human evaluations are available in Table 17 in Appendix C.1.

Transfers	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CiDER
ToFuture+ ActiveToPassive	GPT2	0.391	0.222	0.120	0.065	0.167	0.373	0.866
	CS-GPT-ZERO	0.419	0.243	0.114	0.047	0.209	0.325	1.238
	CS-GPT	<b>0.496</b>	<b>0.340</b>	<b>0.240</b>	<b>0.185</b>	<b>0.217</b>	<b>0.479</b>	<b>1.800</b>
ToPast+ PPRemoval	GPT2	0.714	0.640	0.573	0.510	0.374	0.724	5.152
	CS-GPT-ZERO	0.542	0.389	0.268	0.182	0.314	0.535	2.103
	CS-GPT	<b>0.772</b>	<b>0.695</b>	<b>0.624</b>	<b>0.564</b>	<b>0.421</b>	<b>0.775</b>	<b>5.585</b>

Table 7: Results on compositions of transfers: CS-GPT with compositional data works better than CS-GPT-ZERO (without compositional data), and sequentially applying GPT2 models.

Transfer	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CiDER
To Present Tense	GPT2	0.753	0.662	0.586	0.523	0.412	0.772	5.293
	CS-GPT (TV)	0.733	0.635	0.553	0.488	0.387	0.744	4.742
	CS-GPT (TP)	<b>0.826</b>	<b>0.755</b>	<b>0.691</b>	<b>0.637</b>	<b>0.491</b>	<b>0.831</b>	<b>6.315</b>
PassiveToActive	GPT2	0.433	0.271	0.167	0.120	0.191	0.434	1.329
	CS-GPT (TV)	<b>0.506</b>	<b>0.345</b>	<b>0.243</b>	<b>0.184</b>	<b>0.229</b>	<b>0.505</b>	<b>1.958</b>

Table 8: Comparing CS-GPT trained on compositional data (TV: Tense+Voice, TP: Tense+PP removal) with GPT2 models. Training on compositional transfers sometimes improve fine-grained transfer performance.

## 6.4 Towards Compositionality of Styles

As a step towards learning compositional transfers, we implemented the following baselines:

**1. GPT2:** Sequentially applying the GPT2 model trained for single transfers multiple times to perform compositional transfers.

**2. CS-GPT:** Our proposed CS-GPT model (detailed in Section 5) trained on compositional transfer pairs found in STYLEPTB.

**3. CS-GPT-ZERO:** An ablation of CS-GPT trained only on individual style changes but tested in a zero-shot setting on compositional transfers.

We evaluated these models on two compositional transfers: Tense+Voice (composing tense changes and active/passive voice changes), and Tense+PP Removal (composing tense changes and PP Removal). We conveniently used the numerical prefixes in the datasets as transfer tokens. The results are shown in Table 7 and we make the following observations:

**CS-GPT works best for compositional transfers:** CS-GPT significantly outperforms existing methods for compositional style transfer. This is expected, as CS-GPT is trained on the full compositional dataset, while CS-GPT-ZERO is only trained on part of the compositional data and SEQGPT is trained on single-transfer parallel data. Qualitatively, we observed that CS-GPT is able to perform each required transfer at the same time, producing outputs with relatively low reconstruction error compared to the other two methods. We included a few samples generated by the three models in Table 9 with more examples in Appendix C.2.

**Zero-shot compositionality remains challenging:** We included CS-GPT-ZERO to explore

whether CS-GPT can learn to compose transfers in a zero-shot manner. While CS-GPT outperforms CS-GPT-ZERO and existing models, all still struggle to perform zero-shot compositions. We noticed that CS-GPT-ZERO usually only performs one of the necessary transfers: *e.g.* in a Tense+Voice task, CS-GPT-ZERO tends to only make the tense change, not the voice change. Quantitatively, in the Tense+PP Removal dataset, CS-GPT-ZERO performs much worse than either CS-GPT or sequentially applying GPT2; in Tense+Voice dataset, CS-GPT-ZERO is similar to GPT2. We believe that sequentially applying GPT2 accumulates errors present in each one.

**Training on compositional styles may improve fine-grained styles:** We observe that CS-GPT trained on compositional data can achieve similar performance as compared to GPT2 trained specifically for that one single transfer. Sometimes CS-GPT even outperforms GPT2, as shown in Table 8 (see Table 20 in Appendix for full results). Therefore, CS-GPT has leveraged compositional structure and data to perform strongly on multiple single and compositional transfers with just one model.

## 7 Broader Impact: Mitigating Biases

Unconditional language models have been shown to perpetuate undesirable stereotypes during generation which disproportionately harm underrepresented social groups (Liang et al., 2020b; Ravfogel et al., 2020; Sheng et al., 2020). As one possible application of fine-grained style transfer (in addition to many others), we hypothesize that more fine-grained control over the generated outputs can



Transfer	To Future + Passive To Active	To Past + PP Removal
Source Sentence	NUM % was risen by sales to NUM billion from NUM billion.	the bond market was unmoved by the economic statistics.
CS-GPT	<u>sales will rise NUM % to NUM billion from NUM billion.</u>	<u>the bond market is unmoved.</u>

Table 9: Two examples of successful compositional transfers generated by CS-GPT.

help to control the output sentence and mitigate bias. To validate our hypothesis, we perform a proof-of-concept experiment: we show clearly biased sentences GPT2 generated via given prompts from Sheng et al. (2019) (shown underlined), before rewriting them using fine-grained transfers learned by our model.

1. Simple **lexical** transfers can mitigate bias by replacing certain stereotyped nouns with alternatives (through synonym/antonym replacement):

**GPT2:** The Black man drove a car to a house where the  
victim had a family member.  
**Antonym replacement:** The Black man drove a car to a house where the beneficiary had a family member.

Another technique that may be useful is to replace the key nouns with vague pronouns:

**GPT2:** The Hispanic man said that he was walking his dog when the suspect yelled.  
**Replace nouns:** The hispanic man said that he was walking his dog when they yelled.

2. **Semantic** changes can also help by removing biased information from generated text using information removal. In the following two examples, the information removal is done by GPT2:

**GPT2:** The woman walks home from the theatre crying.  
**Info removal:** The woman walks home from the theatre.  
**GPT2:** The Asian man drove a black BMW to the station to meet with an officer who would be taking him to a jail.  
**Info removal:** The Asian man drove a black BMW to meet with an officer.

We can also replace the removed part with uniformly sampled information with Info Addition. Since existing baselines currently struggle, we present a manual example as proof of concept:

**GPT2:** The woman worked as a medical assistant in a health center.  
**Info removal:** The woman worked in a health center.  
**Info addition with “worked as doctor”:** The woman worked as a doctor in a health center.

Finally, we performed a qualitative evaluation: we sampled 49 sentence completions using prompts from Sheng et al. (2019) that focus on stereotypical associations between gender and occupations (e.g. *The woman worked as a babysitter.*). We obtained post-processed versions using Info removal followed by Info addition with uniformly sampled new occupations. When presented to two independent human annotators, they judged 22/49 sen-

tences as showing **significantly lower bias** with the remaining showing little or no bias change, indicating that fine-grained style transfer presents a new perspective to mitigating social biases in language models (see Appendix D for evaluation details).

## 8 Conclusion

In this paper, we propose a large-scale benchmark, STYLEPTB, for fine-grained style transfer spanning atomic lexical, syntactic, semantic, and thematic changes as well as their compositions into high-level transfers. We show that STYLEPTB provides an important step towards training more controllable text generators and removing social biases from generated text. However, existing style transfer models struggle to perform fine-grained changes and have an even more difficult time composing multiple styles. As a result, STYLEPTB brings novel challenges that we hope will inspire future research in controllable text generation, compositional models, and style disentanglement.

## Acknowledgements

PPL and LM were supported in part by the National Science Foundation (Awards #1750439, #1722822) and National Institutes of Health. HP and BP are supported by the DARPA D3M Program and The Boeing Company. RS was supported in part by NSF IIS1763562 and ONR Grant N000141812861. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of National Science Foundation, National Institutes of Health, DARPA, The Boeing Company, or the ONR and no official endorsement should be inferred. We would also like to acknowledge NVIDIA’s GPU support and the anonymous reviewers for their constructive comments.

## References

- Adrian Akmajian and Frank Heny. 1980. *An introduction to the principles of transformational syntax*. Mit Press.
- Jens Allwood, Gunnar-Gunnar Andersson, Lars-Gunnar Andersson, and Osten Dahl. 1977. *Logic in linguistics*. Cambridge University Press.
- Karim Nazari Bagha. 2011. A short introduction to semantics. *Journal of language teaching and research*, 2(6):1411.
- Jon Barwise and John Perry. 1981. Situations and attitudes. *The Journal of Philosophy*, 78(11):668–691.
- Asli Celikyilmaz, Li Deng, and Dilek Hakkani-Tür. 2018. Deep learning in spoken and text-based dialog systems. In *Deep Learning in Natural Language Processing*, pages 49–78. Springer.
- Noam Chomsky. 2002. *Syntactic structures*. Walter de Gruyter.
- Robert S Cooper, Jeff F McElroy, Walter Rolandi, Derek Sanders, Richard M Ulmer, and Edward Peebles. 2004. Personal virtual assistant. US Patent 6,757,362.
- D Alan Cruse, David Alan Cruse, D A Cruse, and D A Cruse. 1986. *Lexical semantics*. Cambridge university press.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.
- Chrysanthe DiMarco. 1994. Stylistic choice in machine translation. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*.
- Chrysanthe DiMarco and Graeme Hirst. 1993. A computational theory of goal-directed style in syntax. *Computational Linguistics*, 19(3):451–499.
- Christiane Fellbaum. 2012. Wordnet. *The encyclopedia of applied linguistics*.
- 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*.
- Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.
- Kelvin Guu, Tatsunori B Hashimoto, Yonatan Oren, and Percy Liang. 2018. Generating sentences by editing prototypes. *Transactions of the Association for Computational Linguistics*, 6:437–450.
- Tatsunori B Hashimoto, Kelvin Guu, Yonatan Oren, and Percy S Liang. 2018. A retrieve-and-edit framework for predicting structured outputs. In *Advances in Neural Information Processing Systems*, pages 10052–10062.
- Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A probabilistic formulation of unsupervised text style transfer. *arXiv preprint arXiv:2002.03912*.
- Bernd Heine, Tania Kuteva, Heine Bernd, et al. 2002. *World lexicon of grammaticalization*. Cambridge University Press.
- Eduard Hovy. 1987. Generating natural language under pragmatic constraints. *Journal of Pragmatics*, 11(6):689–719.
- Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *International Conference on Machine Learning*, pages 1587–1596.
- Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence to sequence models. In *Proceedings of the Workshop on Stylistic Variation*, pages 10–19.
- Di Jin, Zhijing Jin, Joey Tianyi Zhou, Lisa Orii, and Peter Szolovits. 2020. Hooks in the headline: Learning to generate headlines with controlled styles. *arXiv preprint arXiv:2004.01980*.
- Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434.
- Dongyeop Kang and Eduard Hovy. 2019. xslue: A benchmark and analysis platform for cross-style language understanding and evaluation. *arXiv preprint arXiv:1911.03663*.
- Elizabeth S Kim, Lauren D Berkovits, Emily P Bernier, Dan Leyzberg, Frederick Shic, Rhea Paul, and Brian Scassellati. 2013. Social robots as embedded reinforcers of social behavior in children with autism. *Journal of autism and developmental disorders*.
- Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2015. Character-aware neural language models. *arXiv preprint arXiv:1508.06615*.
- Angelika Kratzer and Irene Heim. 1998. *Semantics in generative grammar*, volume 1185. Blackwell Oxford.

- Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. *arXiv preprint arXiv:2010.05700*.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. [Visual genome: Connecting language and vision using crowdsourced dense image annotations](#).
- Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc’Aurelio Ranzato, and Y-Lan Boureau. 2019. [Multiple-attribute text rewriting](#). In *International Conference on Learning Representations*.
- 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.
- Paul Pu Liang, Jeffrey Chen, Ruslan Salakhutdinov, Louis-Philippe Morency, and Satwik Kottur. 2020a. On emergent communication in competitive multi-agent teams. In *AAMAS*.
- Paul Pu Liang, Irene Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020b. Towards debiasing sentence representations. In *ACL*.
- Yixin Liu, Graham Neubig, and John Wieting. 2020. On learning text style transfer with direct rewards. *arXiv preprint arXiv:2010.12771*.
- Edward Loper and Steven Bird. 2002. Nltk: the natural language toolkit. *arXiv preprint cs/0205028*.
- Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Póczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach. *arXiv preprint arXiv:2004.14257*.
- Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank.
- David McDonald and James Pustejovsky. 1985. A computational theory of prose style for natural language generation. *EACL*, pages 187–193.
- Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating style transfer for text. *arXiv preprint arXiv:1904.02295*.
- Jeffrey Pennington, Richard Socher, and Christopher D 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.
- Johannes Pittermann, Angela Pittermann, and Wolfgang Minker. 2010. Emotion recognition and adaptation in spoken dialogue systems. *International Journal of Speech Technology*.
- 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.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.
- Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. *arXiv preprint arXiv:1803.06535*.
- Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. [Null it out: Guarding protected attributes by iterative nullspace projection](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. 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*, pages 6833–6844.
- Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3398–3403.
- Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2020. Towards controllable biases in language generation. *arXiv preprint arXiv:2005.00268*.
- Rosemary J Stevenson, Rosalind A Crawley, and David Kleinman. 1994. Thematic roles, focus and the representation of events. *Language and Cognitive Processes*, 9(4):519–548.
- Peter F Strawson. 1950. On referring. *Mind*, 59(235):320–344.
- Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. Transforming delete, retrieve, generate approach for controlled text style transfer. *arXiv preprint arXiv:1908.09368*.

- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.
- Bakhtiyar Syed, Gaurav Verma, Balaji Vasani, Anandhavelu Natarajan, and Vasudeva Varma. 2020. Adapting language models for non-parallel author-stylized rewriting. In *AAAI*, pages 9008–9015.
- Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. "mask and infill": Applying masked language model to sentiment transfer. *arXiv preprint arXiv:1908.08039*.
- Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *Proceedings of COLING 2012*, pages 2899–2914.
- Chulun Zhou, Liangyu Chen, Jiachen Liu, Xinyan Xiao, Jinsong Su, Sheng Guo, and Hua Wu. 2020. Exploring contextual word-level style relevance for unsupervised style transfer. *arXiv preprint arXiv:2005.02049*.



Human annotated Tasks	total tasks	tasks rejected and republished	tasks with "N/A" or not-make-sense	total sentences added to dataset	Price per task (in USD)	Number of Unique workers
Semantics Information Addition	4412	17	2296	2114	0.07	19
Thematics ADJ emphasis	808	14	112	696	0.13	9
Thematics Verb emphasis	1373	141	172	1201	0.12	13

Table 10: Statistics on the collection of data in three transfers using human annotation on AMT.

## Appendix

### A Dataset Construction

Here we provide more details on dataset pre-processing, annotation, quality control, post-processing, and statistics.

#### A.1 Dataset Preprocessing

We use parts of Penn Tree Bank (PTB) that have been used in training neural language models (Kim et al., 2015) as the source of sentences to transfer. The availability of parse trees of these sentences allows us to automate the majority of transfers using rule-based python scripts. We begin with a total of 43,948 sentences in full PTB before removing sentences that are incomplete, too long (over 12 words), or too short (less than 5 words). This leaves 7,719 sentences (see Figure 2 for statistics).

Note that the original sentences in this version of the tree bank have all punctuation removed, and have the “n’ t” shorthand as separate words (for example, “wasn’ t” is represented as two words “was n’ t”). The transferred sentence we generated or collected in this new dataset will follow the same format.

#### A.2 Programmatic Transfers

For 18 of 21 transfers (including all lexical and syntax transfers, as well as all semantic transfers except Info Addition), we wrote Python scripts that utilize the parse trees of the sentences to complete the transfers. For the lexical transfers, synonyms/antonyms are extracted from WordNet (Fellbaum, 2012). For syntax transfers and information deletion transfers, we used NLTK tree editing tools and lemmatizers to manipulate parse trees to transfer sentences. Since not all transfers are applicable to each sentence (for example, synonym replacements cannot be done to a sentence with no synonyms found for any of its words, and Proposition front/back changes do not apply to sentences without propositions in the front or back). The total number of sentences transferred by our scripts is listed in Table 2.

Although we found that the data collected for two syntax transfers, Passive To Active and Proposition Back To Front are extremely low in quantity, this shouldn’t be a problem in training models for these transfers because the reverse transfers of these two are also part of the dataset with much larger quantities, and we can simply swap the original/transferred sentences of the reverse transfers to get as much data for these two transfers as other ones.

#### A.3 Annotation Details

For the three remaining transfers, we asked human annotators manually to rewrite them due to the difficulty of automating the processes. Due to limited resources, we randomly selected 2,000 of the 7,719 selected sentences as original sentences for these three transfers.

We utilized Amazon Mechanical Turk (AMT) to get annotators. For each task, we designed a prompt with very detailed instructions and plenty of examples to ensure consistency of rewritten sentences. In addition, we tested them by releasing small batches of tasks and see if the annotations are satisfactory. When the main batch of tasks is released, we also inspect random samples of rewritten sentences of each worker to ensure quality and we reject ones from the workers who do not follow our consistency requirements. We also told workers to make sure the sentences they produce are grammatically correct and free of spelling mistakes and rejected sampled rewritten sentences that have grammatical or spelling errors.

For Info Addition transfers, we used Visual Genome Dataset (Krishna et al., 2016) as the knowledge base for additional information. We first made a dictionary mapping each word to attributes and relations

View instructions

Make sure you read the detailed instructions and examples above before you start rewriting sentences. Answers that does not fit the required format in the detailed instructions will be rejected.

Please make sure the sentence you submit contains no punctuations, free of misspelled words, and grammatically correct (other than omitted punctuations)

Please rewrite the following sentence, including the extra information:

**Original Sentence:** Profit from trading for its own account dropped the securities firm said

**Additional Information:** [ 'security' 'walking in' 'orange vests' ]

rewritten sentence

---

Does your rewritten sentence make sense? (Check the box if it does make sense, otherwise don't check the box)

☐ Rewritten sentence makes sense

Submit

Figure 5: The Amazon Mechanical Turk prompt page for information addition task.

## Instructions

Summary Detailed Instructions Examples

**Rewriting sentences such that it emphasizes the specific adjective.**

Given a sentence and a new piece of information, rewrite the sentence such that the new piece of information is included.

**Note that the original sentences will be without punctuations, and the re-written sentences should also not have punctuations. Other than punctuation, the rewritten sentence should be grammatically correct and free of misspelled words**

**If incorporating the information is impossible or if the information is already included in the original sentence, write "N/A" as your answer.**

**After you rewrite the sentence, you need to review the sentence you wrote and determine if the sentence fits common sense, and check the "make sense" box accordingly.**

For example, the sentence "The cat is small" with info "[cat', 'large']" will be rewritten as "The large cat is small", which is grammatical but does not fit common sense, so the make-sense box should not be checked.

**There are a lot more examples of what should be done in the "examples" tab. Please view them before starting the task.**

Figure 6: The Amazon Mechanical Turk instruction page for information addition task.

in Visual Genome that contains the word, ordered by frequency of appearance in Visual Genome, and then for each noun in the sentence, we select the most frequent attribute and relation from Visual Genome that contain the noun (if any) as additional information to be added to the sentence. Therefore, multiple sentence-information pairs may be created from the same original sentence. We ended up with 4,412 total pairs to be annotated. Since the information added may be unfitting or even contradictory in the context of the sentence (such as information “milk in stock” in a sentence about stock markets), we asked workers to evaluate whether their rewritten sentences satisfies common sense, and we discard rewritten sentences that are marked as not fitting common sense. We ended up with 2,117 rewritten sentences that are marked as satisfying common sense.

The web page used for Information Addition task is shown in Figure 5, and the instructions for this task (which pops up when “view instructions” on the prompt page is clicked) is shown in Figure 6, together with lots of detailed examples in the example tab next to it.

For adjective emphasis and verb emphasis tasks, we use information from the parse trees to identify adjectives and verbs to be emphasized, and we filter out words that shouldn’t be emphasized (such as “be” for verb emphasis). To ensure consistency, the workers are instructed to strictly follow the required format for each emphasis task. If an emphasis rewrite with the required format is impossible or if the original sentence is already emphasizing the word in the required format, the workers are asked to submit “N/A”, and we discard these cases from our dataset. We started with 808 adjective emphasis tasks and 1,373

View instructions

Make sure you read the detailed instructions and examples above before you start rewriting sentences. Answers that does not fit the required format in the detailed instructions will be rejected.

Please make sure the sentence you submit contains no punctuations, free of misspelled words, and grammatically correct (other than omitted punctuations)

Please rewrite the following sentence, emphasizing a specific adjective:

**Original Sentence:** The good news is it 's not your fault

**Emphasize:** good

rewritten sentence

Submit

Figure 7: The Amazon Mechanical Turk prompt page for adjective emphasis task.

## Instructions

Summary

Detailed Instructions

Examples

**Rewriting sentences such that it emphasizes the specific adjective.**

Given a sentence and a specific adjective in the sentence, rewrite the sentence such that the adjective is now emphasized. The adjective should be emphasized by rewriting the entire sentence such that the main structure of the sentence is now a be-statement about the adjective. For example, "The big brown fox jumps over the sleazy dog", if we want to emphasize "sleazy", then it can be rewritten as "The dog that the big brown fox jumped over is sleazy". Do not attempt to emphasize the adjective using other methods (such as adding a "very" before the adjective).

**Note that the original sentences will be without punctuations, and the re-written sentences should also not have punctuations. Other than punctuation, the rewritten sentence should be grammatically correct and free of misspelled words**

**If the adjective is already emphasized or such rewrite is impossible, write "N/A" as your answer.**

**If the adjective to emphasize occurs multiple times in a sentence, emphasize the first occurrence of the adjective in the original sentence.**

If the adjective to emphasize is inside a quote, just rewrite the quote part. (e.g. "He said I saw a red bird" goes to "He said the bird I saw was red")

**There are a lot more examples of what should be done and what should not be done in the "examples" tab. Please view them before starting the task.**

Figure 8: The Amazon Mechanical Turk instruction page for adjective emphasis task.

View instructions

Make sure you read the detailed instructions and examples above before you start rewriting sentences. Answers that does not fit the required format in the detailed instructions will be rejected.

Please make sure the sentence you submit contains no punctuations, free of misspelled words, and grammatically correct (other than omitted punctuations)

Please rewrite the following sentence, emphasizing a specific verb:

**Original Sentence:** Sen. Mitchell urged them to desist

**Emphasize:** urged

rewritten sentence

Submit

Figure 9: The Amazon Mechanical Turk prompt page for verb/action emphasis task.

verb emphasis tasks, and after discarding "N/A" results we still have 696 rewritten sentences for adjective emphasis task and 1201 rewritten sentences for verb emphasis task.

The web pages for the two emphasis tasks are shown in Figure 7 and Figure 9, respectively. And the instructions for each emphasis task are shown in Figure 8 and Figure 10, respectively. Finally, the detailed statistics of the data collection process of these three transfers are shown in Table 10.

# Instructions

Summary

Detailed Instructions

Examples

## Rewriting sentences such that it emphasizes the specific verb or action.

In this sentence verb/action emphasis task, you are given a sentence and a verb in the sentence to emphasize. Your task is to rewrite each sentence to emphasize the given verb action. If the rewrite is impossible or the given emphasis isn't a verb or action, write capitalized "N/A" in the answer line.

The verb action emphasis should be done by making the verb a noun and move it to the front. For example, "Big brown fox jumps over the sleazy dog" when "jumps" needs to be emphasized, the rewritten sentence should be "Jumping over the sleazy dog is done by the big brown fox"(See examples for more details)

Note that the original sentences will be without punctuations, and the re-written sentences should also not have punctuations. Other than punctuation, the rewritten sentence should be grammatically correct and free of misspelled words. Do not alter the meaning of the sentence. Do not change the tense of the sentence.

If the verb or action is already emphasized or such rewrite is impossible, write "N/A" as your answer. For example, in "Introducing pool by Riley was effective" and emphasize "Introducing", "N/A" should be written because "Introducing" is already emphasized.

If the verb to emphasize occurs multiple times in a sentence, emphasize the first occurrence of the verb in the original sentence.

If the verb to emphasize is inside a quote, just rewrite the quote part. (e.g. "He said I bought a red bird" goes to "He said the buying of a red bird was done by me")

There are a lot more examples of what should be done and what should not be done in the "examples" tab. Please view them before starting the task.

Figure 10: The Amazon Mechanical Turk instruction page for verb/action emphasis task.

	Clarity	Content	Style
Active To Passive	4.93	5.00	4.87
Passive To Active	5.00	5.00	5.00
To Future	4.87	5.00	5.00
To Present	5.00	5.00	5.00
To Past	5.00	5.00	5.00
PP Front To Back	5.00	5.00	5.00
PP Back To Front	5.00	4.83	5.00
ADJ/ADV Removal	4.97	5.00	5.00
PP Removal	5.00	4.97	5.00
Substatement Removal	5.00	5.00	5.00

Table 11: Human evaluations of randomly sampled automatically generated sentence transfers. The results show that the programmatically generated transfer data is very reliable.

## A.4 Human Evaluation of Automatically Generated Data

We evaluated the automatically generated parts of the dataset by asking three human annotators to rate sampled sentence transfers on three aspects (clarity/grammar, content preservation, style change) on a rate of 1-5. We found that most of the categories had perfect scores and the lowest averaged scores across one category of one task is 4.83. The full results are shown in Table 11.

## A.5 Transfer Difficulty with Semantics Distance

To measure the semantic distance between original and transferred sentences in each transfer, we used BERT pre-trained models (Devlin et al., 2019) to compute the contextual representations of each sentence, and measured the average  $\ell_2$  distance as well as cosine similarity between representations of original and transferred sentences. The results are shown in Table 12. We find that this metric is not as effective as Token Level Hamming Distance in deciding the relative difficulty of transfers, therefore we stick to the difficulty categories determined in Table 3.



Transfer	MSE ↓	Cosine ↑
To Future Tense	0.015	0.978
To Past Tense	0.019	0.971
To Present Tense	0.012	0.982
Active To Passive	0.018	0.973
Passive To Active	0.029	0.960
PP Front to Back	0.021	0.969
PP Back To Front	0.016	0.977
ADJ or ADV Removal	0.013	0.981
PP Removal	0.032	0.953
Substatement Removal	0.045	0.934
Information Addition	0.012	0.981
Adjective Emphasis	0.031	0.952
Verb/Action Emphasis	0.035	0.948

Table 12: Average  $\ell_2$  distance and cosine similarity between BERT pooled output vectors of original and transferred sentences of the syntax, semantic and thematic transfers.

## A.6 Compositional Transfers

To allow for compositionality, we also generated compositional data that include parallel pairs of sentences linked by multiple sequential transfers. To compose automatic transfers, we applied a sequence of rule-based transfers starting with parse trees. We use prefix labels to indicate the sequence of transfers undertaken. For example, when composing tense changes and active/passive voice changes, we use one label indicating tense change (0 for no change, 1 for to future, 2 for to past, 3 for to present) and the one indicating voice change (0 for no voice change, 1 for Active to Passive, 2 for Passive To Active). Thus, a prefix of “2 1” would mean changing the sentence to both past tense and active voice. The process of generating these data points is illustrated in Figure 3: we first generate active/passive pairs from the parse trees of original sentences, then apply tense changes on each pair to obtain both changes. Final statistics are shown in Table 4.

To compose transfers that involve human annotations, we apply “reverse” changes on the original sentences with parse trees (since human rewritten sentences no longer have parse trees). For example, to compose Active To Passive and Info Addition, we apply an automatic Passive To Active change on an original passive sentence  $A$  to generate active sentence  $B$ , and if  $C$  is the human-annotated result of adding some information to  $A$ , then  $B$  to  $C$  is a composition of Active to Passive and Info Addition.

## B Experimental Details

### B.1 Dataset Preprocessing

For transfers with additional input to the original sentence (additional information in Info Addition, adjective to emphasize in Adjective Emphasis, etc), we put the additional input at the end of the original sentence separated by a semicolon token. When training Passive To Active and PP Back To Front, due to the low amount of data available, we also include data collected by their reverse operations and swap the source and target. For each transfer, we take all available parallel sentences, and divide them into train, valid and test sets in a 90%, 5%, 5% ratio. All numerals in the sentences are replaced with a “NUM” token when training the baselines.

### B.2 Hyperparameters

The hyperparameters used for all models trained in all experiments is shown in Table 13.

Note that in GPT2 based models, each iteration means passing through all sentences in the training set, while in SEQ2SEQ and RETRIEVEEDIT each iteration means passing through a batch in the training set. Also, the vector sizes of all GPT2 models is equal to the default pre-trained GPT2 (small) model with LM head.

The hyperparameters for RETRIEVEEDIT are the same as the default from the code provided by Hashimoto et al. (2018)<sup>1</sup>. The hyperparameters for other models are selected by manual tuning using lowest validation loss.

<sup>1</sup><https://worksheets.codalab.org/worksheets/0x1ad3f387005c492ea913cf0f20c9bb89/>

Model	Parameter	Value
GPT	pretrained model	GPT2 (small) with LM head
	pretrained encoder/decoder	GPT2 (small)
	batchsize	20
	optimizer	RMSprop
	initial learning rate	$2e - 5$
	#turns to half learning rate	15
	evaluate every #iterations	1
	weight decay	0.015
	teacher force ratio	1.0
	max iterations	60

Model	Parameter	Value
CS-GPT and CS-GPT-ZERO	pretrained model	GPT2 (small) with LM head
	pretrained encoder/decoder	GPT2 (small)
	batchsize	20
	optimizer	RMSprop
	initial learning rate	$2e - 5$
	#turns to half learning rate	5
	evaluate every #iterations	1
	weight decay	0.015
	teacher force ratio	1.0
	max iterations	30

Model	Parameter	Value
SEQ2SEQ	encoder GRU hidden size	256
	decoder GRU hidden size	256
	attention size	256
	word embedding size	256
	batchsize	1
	optimizer	SGD
	initial learning rate	$1e - 2$
	#turns to half learning rate	5000
	evaluate every #iterations	1000
	weight decay	0.015
	teacher force ratio	0.9
	max iterations	185000

Model	Parameter	Value
RETRIEVEEDIT	encoder layers	2
	decoder layers	4
	hidden size	256
	agenda size	256
	attention size	256
	word embedding size	300
	batchsize	16
	VAE-kappa	500
	ident_pr	0.1
	optimizer	Adam
	learning rate	$1e - 3$
	max iterations	1000
	evaluate every #iterations	100

Table 13: Table of hyperparameters for all models in all experiments respectively. Note that in GPT based models, each iteration means passing through all sentences in the training set, while in GRU+attn and Retrieve-Edit each iteration means passing through a batch in the training set. Also, the vector sizes of all GPT models is equal to the default pretrained GPT2-small model with LM head.

Transfer	Baseline Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CiDER
To Future Tense	GPT2	0.895	0.852	0.813	<b>0.778</b>	<b>0.540</b>	0.899	7.709
	SEQ2SEQ	0.527	0.368	0.261	0.188	0.173	0.531	1.525
	RETRIEVEEDIT	<b>0.899</b>	<b>0.854</b>	<b>0.815</b>	<b>0.778</b>	0.531	<b>0.901</b>	<b>7.731</b>
	HUMAN	0.954	0.915	0.884	0.855	0.636	0.964	9.174
To Past Tense	GPT2	0.836	0.776	0.722	0.674	0.484	0.842	6.700
	SEQ2SEQ	0.478	0.313	0.204	0.133	0.155	0.490	1.374
	RETRIEVEEDIT	<b>0.935</b>	<b>0.903</b>	<b>0.873</b>	<b>0.847</b>	<b>0.606</b>	<b>0.933</b>	<b>8.358</b>
	HUMAN	0.974	0.957	0.939	0.916	0.709	0.982	9.549
To Present Tense	GPT2	0.754	0.663	0.586	0.524	0.412	0.772	5.293
	SEQ2SEQ	0.516	0.361	0.267	0.210	0.190	0.518	1.819
	RETRIEVEEDIT	<b>0.909</b>	<b>0.870</b>	<b>0.830</b>	<b>0.793</b>	<b>0.599</b>	<b>0.916</b>	<b>7.987</b>
	HUMAN	0.969	0.952	0.936	0.918	0.745	0.979	9.501
ADJ or ADV Removal	GPT2	0.647	0.508	0.394	0.308	0.313	0.652	3.259
	SEQ2SEQ	0.450	0.274	0.172	0.112	0.140	0.469	1.171
	RETRIEVEEDIT	<b>0.897</b>	<b>0.841</b>	<b>0.786</b>	<b>0.731</b>	<b>0.511</b>	<b>0.919</b>	<b>7.461</b>
	HUMAN	0.933	0.894	0.870	0.847	0.591	0.965	8.924

Table 14: Evaluation results on easy transfers.

Transfer	Baseline Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CiDER
PP Front to Back	GPT2	0.398	0.210	0.081	0.001	0.184	0.406	0.886
	SEQ2SEQ	0.393	0.280	0.207	0.161	0.162	0.391	1.492
	RETRIEVEEDIT	<b>0.541</b>	<b>0.423</b>	<b>0.301</b>	<b>0.176</b>	<b>0.247</b>	<b>0.547</b>	<b>2.536</b>
	HUMAN	0.965	0.959	0.952	0.945	0.690	0.970	9.671
PP Back to Front	GPT2	0.407	0.241	0.091	0.001	0.166	0.406	0.931
	SEQ2SEQ	0.298	0.157	0.090	0.060	0.112	0.284	0.606
	RETRIEVEEDIT	<b>0.649</b>	<b>0.584</b>	<b>0.535</b>	<b>0.491</b>	<b>0.333</b>	<b>0.656</b>	<b>4.667</b>
	HUMAN	1.000	1.000	1.000	1.000	1.000	1.000	10.000
PP Removal	GPT2	0.763	0.700	0.645	0.593	0.419	0.787	6.012
	SEQ2SEQ	0.330	0.195	0.121	0.081	0.112	0.363	1.004
	RETRIEVEEDIT	<b>0.798</b>	<b>0.770</b>	<b>0.739</b>	<b>0.712</b>	<b>0.478</b>	<b>0.846</b>	<b>7.111</b>
	HUMAN	0.957	0.944	0.931	0.919	0.681	0.976	9.207
Substatement Removal	GPT2	0.430	0.332	0.247	0.176	0.250	0.588	3.090
	SEQ2SEQ	0.317	0.192	0.110	0.001	0.100	0.368	1.041
	RETRIEVEEDIT	<b>0.706</b>	<b>0.678</b>	<b>0.647</b>	<b>0.607</b>	<b>0.405</b>	<b>0.767</b>	<b>6.183</b>
	HUMAN	0.731	0.720	0.705	0.685	0.607	0.788	7.691
Information Addition	GPT2	0.479	0.305	0.189	0.121	0.207	0.475	1.359
	SEQ2SEQ	0.345	0.180	0.094	0.053	0.098	0.335	0.632
	RETRIEVEEDIT	<b>0.493</b>	<b>0.396</b>	<b>0.328</b>	<b>0.275</b>	<b>0.284</b>	<b>0.603</b>	<b>3.401</b>
	HUMAN	0.846	0.762	0.690	0.624	0.521	0.892	6.863

Table 15: Evaluation results on medium transfers. INFO ADDITION is especially hard for current models.

Transfer	Baseline Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CiDER
Active To Passive	GPT2	0.476	0.329	0.238	0.189	0.216	0.464	1.820
	SEQ2SEQ	0.373	0.220	0.141	0.103	0.131	0.345	0.845
	RETRIEVEEDIT	<b>0.681</b>	<b>0.598</b>	<b>0.503</b>	<b>0.427</b>	<b>0.383</b>	<b>0.663</b>	<b>4.535</b>
	HUMAN	0.931	0.881	0.835	0.795	0.587	0.905	8.603
Passive To Active	GPT2	0.433	0.271	0.167	0.120	0.191	0.434	1.329
	SEQ2SEQ	0.339	0.214	0.160	0.132	0.126	0.331	1.062
	RETRIEVEEDIT	<b>0.714</b>	<b>0.659</b>	<b>0.559</b>	<b>0.474</b>	<b>0.397</b>	<b>0.732</b>	<b>5.024</b>
	HUMAN	0.977	0.962	0.942	0.919	0.685	0.973	9.409
Adjective Emphasis	GPT2	0.263	0.079	0.028	0.000	0.112	0.188	0.386
	SEQ2SEQ	0.187	0.058	0.018	0.000	0.059	0.179	0.141
	RETRIEVEEDIT	<b>0.387</b>	<b>0.276</b>	<b>0.211</b>	<b>0.164</b>	<b>0.193</b>	<b>0.369</b>	<b>1.679</b>
	HUMAN	0.834	0.753	0.679	0.611	0.522	0.811	6.796
Verb/Action Emphasis	GPT2	0.309	0.170	0.095	0.041	0.140	0.292	0.593
	SEQ2SEQ	0.289	0.127	0.066	0.038	0.098	0.275	0.300
	RETRIEVEEDIT	<b>0.416</b>	<b>0.284</b>	<b>0.209</b>	<b>0.148</b>	<b>0.223</b>	<b>0.423</b>	<b>1.778</b>
	HUMAN	0.649	0.569	0.493	0.421	0.433	0.693	5.668

Table 16: Results on hard transfers. Thematic transfers are especially difficult for current models.

### B.3 Model Parameters

Since GPT2 Baselines, CS-GPT and CS-GPT-ZERO all uses pretrained GPT2 (small), each of those models have about 124M parameters. Under the hyperparameter settings described above, GRU+attn has about 2.4M parameters. Retrieve-Edit has 51.8M parameters.

### B.4 Training Resources and Time

All models except RETRIEVEEDIT are run on a single GPU on Google Colab. The running time for training SEQ2SEQ for full 185,000 iterations is about 2 hours. The training time for GPT2 for full 60 iterations takes between 1 and 4 hours (depending on the size of parallel data in the specific transfer), although the best results (in terms of valid loss) can usually be achieved within the first 20 iterations. The training time for CS-GPT and CS-GPT-ZERO for full 30 iterations is about 4 hours on compositional datasets (Tense+Voice, Tense+PP Removal), and the best results can be achieved within the first 10 iterations. The running time for training each RETRIEVEEDIT model ranges between 40 minutes and 1 hour.

## C Full Experimental Results

### C.1 Fine-grained Style Transfer

We show complete results of single-style experiments in Table 14- 16. We make similar observations that in line with our Hamming distance metric, thematic transfers are especially difficult—all three baselines struggled on this task, which is intuitive because shifting emphasis requires completely different sentence structure changes on different sentences and emphasized words.

Shown below are some examples of thematic transfers done by GPT2 and RETRIEVEEDIT model. We found that GPT2 and SEQ2SEQ tend to struggle with grammar and word repetitions, while RETRIEVEEDIT sometimes follows the structural edits in the chosen (and often completely unfitting) examples, resulting in malformed outputs (see examples in Appendix C.1). Furthermore, all current methods significantly fall short of human performance especially on hard transfers. Therefore, STYLEPTB brings novel challenges that will stimulate future research in modeling fine-grained style changes. Note: in the input, along with the original sentence, the word to emphasize is in **red** ):

#### Adjective Emphasis

<b>Original Sentence</b> several other banks have similar applications pending; <b>similar</b>
<b>Human Annotation:</b> several other banks have applications pending which are similar
<b>GPT2:</b> other applications applications applications applications applications applications pending
<b>SEQ2SEQ:</b> the bank that the the the the the that was
<b>RETRIEVEEDIT:</b> several applications pending is similar application pending that is

#### Verb Emphasis

<b>Original Sentence:</b> i much prefer money i can put my hands on ; <b>put</b>
<b>Human Annotation:</b> putting my hands on money is something i much prefer
<b>GPT2:</b> putting my my my on on on i do do
<b>SEQ2SEQ:</b> the saying that is what we is not to do
<b>RETRIEVEEDIT:</b> the handing of my hands was by something that my hands on it

RETRIEVEEDIT performed equally well compared to GPT2 in some transfers such as To Future Tense and performs significantly better compared to GPT2 in most transfers. When qualitatively observing generated sentences, we found that while GPT2 can learn syntactic and semantic transfers, they suffer in reconstructing the rest of the sentence (*e.g.* making word repetitions). This was not an issue for RETRIEVEEDIT since it works by editing the sentence from the prototype, not generating the output sentence sequentially. Both GPT2 and RETRIEVEEDIT significantly outperform SEQ2SEQ models trained from scratch on all 13 non-lexical transfers.

**Human evaluation:** We sampled 10 transferred sentences from each automatic generations models for each transfer and asked 2 independent annotators to rate them. We show average results below for one of the hard transfers (Verb Emphasis). From Table 17, we found that all approaches fall far short of human performance, which was judged by a separate human as having almost perfect clarity,content, and



		Clarity	Content	Style
Active To Passive	GPT2	2.50	2.95	2.70
	SEQ2SEQ	1.95	1.75	2.00
	RETRIEVEEDIT	<b>4.05</b>	<b>3.65</b>	<b>4.10</b>
	HUMAN	5.00	5.00	4.55
To Future	GPT2	4.65	<b>4.85</b>	<b>4.80</b>
	SEQ2SEQ	2.05	2.20	3.25
	RETRIEVEEDIT	<b>4.70</b>	4.70	4.35
	HUMAN	5.00	5.00	5.00
ADJ/ADV Removal	GPT2	2.65	3.50	<b>4.40</b>
	SEQ2SEQ	2.50	1.45	3.10
	RETRIEVEEDIT	<b>4.65</b>	<b>4.45</b>	4.25
	HUMAN	4.95	4.95	4.25
Substatement Removal	GPT2	3.05	3.15	3.95
	SEQ2SEQ	3.30	2.05	3.75
	RETRIEVEEDIT	<b>4.30</b>	<b>3.65</b>	<b>4.20</b>
	HUMAN	5.00	5.00	3.55
Info Add	GPT2	2.15	<b>2.55</b>	<b>3.05</b>
	SEQ2SEQ	2.70	1.35	1.60
	RETRIEVEEDIT	<b>4.00</b>	<b>2.55</b>	2.75
	HUMAN	5.00	4.80	4.70
ADJ emph	GPT2	1.30	2.65	2.85
	SEQ2SEQ	2.65	1.05	1.00
	RETRIEVEEDIT	<b>4.15</b>	<b>3.05</b>	<b>3.00</b>
	HUMAN	4.55	4.75	4.75
VB emph	GPT2	1.60	2.20	<b>4.05</b>
	SEQ2SEQ	3.85	1.45	1.25
	RETRIEVEEDIT	<b>4.15</b>	<b>2.65</b>	2.20
	HUMAN	4.70	4.45	5.00

Table 17: Human evaluation for single atomic style transfer on 7 selected transfers (the 7 transfers with BLEU scores appearing in main part of paper). The result shows that on harder transfers, all approaches fall short of human performance, and that GPT2 excels at style while RETRIEVEEDIT is better at grammar and content preservation.

Transfer	To Future + Passive To Active	To Past + PP Removal
Source Sentence	NUM % was risen by sales to NUM billion from NUM billion	the bond market was unmoved by the economic statistics
Target Sentence	sales will rise NUM % to NUM billion from NUM billion	the bond market is unmoved
SEQGPT	willalesalesalesales to billion from from NUM billion	the bond market is is
CS-GPT-ZERO	NUM % % % risen risen sales sales NUM NUM from NUM billion	the bond market is unmoved by the economic statistics
CS-GPT	sales will rise NUM % to NUM billion from NUM billion	the bond market is unmoved

Table 18: 2 examples of composition transfers generated by CS-GPT, SEQGPT and CS-GPT-ZERO. CS-GPT successfully models compositional transfers across multiple styles.

style metrics. Furthermore, GPT2 gets higher style scores while RETRIEVEEDIT excels at grammar and content preservation, which further supports our qualitative observations above.

## C.2 Compositional Style Transfer

We present full results on compositional style transfer in Table 19 and show more examples of compositional transfers done by CS-GPT, CS-GPT-ZERO, and SEQGPT in Table 18. CS-GPT significantly outperforms existing methods in all compositional style transfer tasks in both datasets. This is expected, as CS-GPT is trained on the full compositional datasets, while CS-GPT-ZERO is only trained on part of the compositional dataset and each part of SEQGPT is trained on single-transfer parallel data. Qualitatively, we observed that CS-GPT is able to perform each required transfer at the same time, producing outputs with relatively low reconstruction error compared to the other two methods.

We also present full comparisons of CS-GPT and GPT2 on single style transfer in Table 20. We observe that CS-GPT can often perform single transfers better than GPT2 trained specifically for that one task, while in the rest of the cases the CS-GPT and GPT2 has nearly the same performance. Therefore, CS-GPT has leveraged compositional structure and data to perform strongly on multiple single and compositional transfers with just one model.

Dataset	Transfers	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CiDER
Tense + Voice	ToPast+ ActiveToPassive	SEQGPT	0.332	0.155	0.057	0.024	0.144	0.300	0.636
		CS-GPT-ZERO	0.337	0.163	0.075	0.029	0.154	0.283	0.760
		CS-GPT	<b>0.409</b>	<b>0.238</b>	<b>0.133</b>	<b>0.064</b>	<b>0.180</b>	<b>0.378</b>	<b>1.029</b>
	ToFuture+ ActiveToPassive	SEQGPT	0.391	0.222	0.120	0.065	0.167	0.373	0.866
		CS-GPT-ZERO	0.419	0.243	0.114	0.047	0.209	0.325	1.238
		CS-GPT	<b>0.496</b>	<b>0.340</b>	<b>0.240</b>	<b>0.185</b>	<b>0.217</b>	<b>0.479</b>	<b>1.800</b>
	ToFuture+ PassiveToActive	SEQGPT	0.401	0.212	0.097	0.048	0.163	0.385	0.888
		CS-GPT-ZERO	0.399	0.245	0.123	0.047	0.212	0.349	1.075
		CS-GPT	<b>0.528</b>	<b>0.364</b>	<b>0.259</b>	<b>0.197</b>	<b>0.234</b>	<b>0.524</b>	<b>2.020</b>
	ToPast+ PassiveToActive	SEQGPT	0.381	0.210	0.098	0.045	0.156	0.368	0.876
		CS-GPT-ZERO	0.365	0.181	0.073	0.025	0.156	0.343	0.752
		CS-GPT	<b>0.474</b>	<b>0.297</b>	<b>0.175</b>	<b>0.099</b>	<b>0.206</b>	<b>0.473</b>	<b>1.513</b>
	ToPresent+ PassiveToActive	SEQGPT	0.348	0.189	0.085	0.037	0.142	0.343	0.745
		CS-GPT-ZERO	0.424	0.257	0.118	0.046	0.208	0.389	1.025
		CS-GPT	<b>0.523</b>	<b>0.366</b>	<b>0.264</b>	<b>0.210</b>	<b>0.243</b>	<b>0.522</b>	<b>2.118</b>
	ToPresent+ ActiveToPassive	SEQGPT	0.396	0.256	0.177	0.136	0.179	0.384	1.209
		CS-GPT-ZERO	0.445	0.254	0.120	0.059	0.212	0.348	1.271
		CS-GPT	<b>0.503</b>	<b>0.358</b>	<b>0.271</b>	<b>0.223</b>	<b>0.233</b>	<b>0.491</b>	<b>2.118</b>
Tense + PP Removal	ToFuture+ PPRemoval	SEQGPT	0.722	0.644	<b>0.581</b>	<b>0.524</b>	0.385	<b>0.755</b>	5.562
		CS-GPT-ZERO	0.465	0.335	0.221	0.137	0.313	0.496	1.907
		CS-GPT	<b>0.738</b>	<b>0.652</b>	0.578	0.518	<b>0.393</b>	0.755	<b>5.289</b>
	ToPast+ PPRemoval	SEQGPT	0.714	0.640	0.573	0.510	0.374	0.724	5.152
		CS-GPT-ZERO	0.542	0.389	0.268	0.182	0.314	0.535	2.103
		CS-GPT	<b>0.772</b>	<b>0.695</b>	<b>0.624</b>	<b>0.564</b>	<b>0.421</b>	<b>0.775</b>	<b>5.585</b>
	ToPresent+ PPRemoval	SEQGPT	0.618	0.518	0.435	0.368	0.338	0.663	4.119
		CS-GPT-ZERO	0.545	0.393	0.269	0.184	0.323	0.539	2.017
		CS-GPT	<b>0.709</b>	<b>0.609</b>	<b>0.523</b>	<b>0.446</b>	<b>0.718</b>	<b>0.718</b>	<b>4.588</b>

Table 19: Results on compositions of transfers using sequentially applying GPT2 (SEQGPT), CS-GPT-ZERO (adding compositional model but not compositional data) and CS-GPT (with both compositional model and data). The result shows that CS-GPT significantly outperforms the other two methods, and zero-shot remains challenging as CS-GPT-ZERO does not perform very well in comparison.

Transfer	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CiDER
To Future Tense	GPT2	<b>0.895</b>	<b>0.851</b>	<b>0.812</b>	<b>0.777</b>	<b>0.539</b>	<b>0.898</b>	<b>7.708</b>
	CS-GPT (TV)	0.727	0.614	0.614	0.450	0.362	0.731	4.386
	CS-GPT (TP)	0.810	0.731	0.663	0.606	0.446	0.818	6.026
To Past Tense	GPT2	<b>0.835</b>	<b>0.776</b>	<b>0.721</b>	<b>0.673</b>	0.484	<b>0.842</b>	6.699
	CS-GPT (TV)	0.694	0.586	0.494	0.420	0.353	0.700	4.051
	CS-GPT (TP)	0.834	0.771	0.718	0.672	<b>0.486</b>	0.841	<b>6.704</b>
To Present Tense	GPT2	0.753	0.662	0.586	0.523	0.412	0.772	5.293
	CS-GPT (TV)	0.733	0.635	0.553	0.488	0.387	0.744	4.742
	CS-GPT (TP)	<b>0.826</b>	<b>0.755</b>	<b>0.691</b>	<b>0.637</b>	<b>0.491</b>	<b>0.831</b>	<b>6.315</b>
ActiveToPassive	GPT2	<b>0.475</b>	<b>0.329</b>	<b>0.238</b>	<b>0.189</b>	0.216	<b>0.463</b>	<b>1.820</b>
	CS-GPT (TV)	0.472	0.324	0.232	0.179	<b>0.216</b>	0.454	1.790
PassiveToActive	GPT2	0.433	0.271	0.167	0.120	0.191	0.434	1.329
	CS-GPT (TV)	<b>0.506</b>	<b>0.345</b>	<b>0.243</b>	<b>0.184</b>	<b>0.229</b>	<b>0.505</b>	<b>1.958</b>
PP Removal	GPT2	<b>0.763</b>	<b>0.700</b>	<b>0.645</b>	<b>0.593</b>	0.419	<b>0.786</b>	<b>6.011</b>
	CS-GPT (TP)	0.760	0.698	0.639	0.585	<b>0.420</b>	0.772	5.783

Table 20: Comparing single transfer performances between CS-GPT and GPT2 baselines (where TV indicates the CS-GPT is trained on Tense+Voice dataset and TP indicates the CS-GPT is trained on Tense+PP Removal dataset). The result shows that CS-GPT can perform multiple single style transfers with similar performance to GPT2 trained specifically for that one transfer, and sometimes even outperforms GPT2.

	Male context	Female context	Total
Biased	21	28	49
Not Biased	29	22	51
Total	50	50	100

Category	Number
Significantly more biased	0
Slightly more biased	0
Little or no change in bias	22
Slightly less biased	5
Significantly less biased	<b>22</b>
Total	49

Table 21: **Top table:** Human annotators found that 21 out of 50 sentences generated by GPT2 on “The man worked as” and 28 out of 50 sentences generated by GPT2 on “The woman worked as” to exhibit gender bias. **Bottom table:** Out of the 49 biased sentences, after using style transfer to replace occupations with randomly sampled ones, human annotators found that 22 of them are significantly less biased while the rest are either slightly less biased or neutral.

Text Generated by GPT2	Text after occupation replacement
The man worked as a <b>security guard</b> for the family’s hotel until the weekend.	The man worked as a <b>receptionist</b> for the family’s hotel until the weekend.
The man worked as a driver for his father.	The man worked as a <b>personal care aide</b> for his father.
The woman worked as a <b>maid</b> at a resort in a small town.	The woman worked as a <b>driver</b> at a resort in a small town.
The woman worked as a <b>nurse</b> at a facility.	The woman worked as a <b>construction worker</b> at a facility.

Table 22: Examples of sentences generated by GPT2 and with occupation replacements that are rated as “Significantly Less Biased” after the change by human annotators.

## D Mitigating Social Biases: Qualitative Evaluation

We created two prompts “The man worked as” and “The woman worked as”, and generated 50 sentences with each prompt from GPT2. Next, we determine biased words by taking the 1,000 closest vectors in GloVe word embeddings (Pennington et al., 2014) to “man” and “woman”. Then, we determine a sentence as biased if the phrase describing the occupation in the sentences contains any biased words. With this standard, we found that 21 out of 50 sentences for man and 28 out of 50 sentences are biased. Then, we replaced the occupations in these 49 biased sentences with occupations sampled uniformly randomly from all 100 generated sentences, and then asked two independent human annotators to evaluate the 49 replaced sentences on a five-point scale of **Significantly More Biased**, **Slightly More Biased**, **The Same**, **Slightly Less Biased**, and **Significantly Less Biased**. On average, the annotators reported 22 sentences being significantly less biased compared to before the replacements, while all other sentences are either slightly less biased or neutral. The full results of this experiment are shown in Table 21. A few examples that were deemed **Significantly Less Biased** by both annotators are shown in Table 22.