

TextSETTR: Few-Shot Text Style Extraction and Tunable Targeted Restyling

Parker Riley^{a*}, Noah Constant^b, Mandy Guo^b,
Girish Kumar^{c*}, David Uthus^b, Zarana Parekh^b

^aUniversity of Rochester

^bGoogle Research

^cStanford University

Abstract

We present a novel approach to the problem of text style transfer. Unlike previous approaches requiring style-labeled training data, our method makes use of readily-available unlabeled text by relying on the implicit connection in style between adjacent sentences, and uses labeled data only at inference time. We adapt T5 (Raffel et al., 2020), a strong pre-trained text-to-text model, to extract a style vector from text and use it to condition the decoder to perform style transfer. As our label-free training results in a style vector space encoding many facets of style, we recast transfers as “targeted restyling” vector operations that adjust specific attributes of the input while preserving others. We demonstrate that training on *unlabeled* Amazon reviews data results in a model that is competitive on sentiment transfer, even compared to models trained fully on labeled data. Furthermore, applying our novel method to a diverse corpus of unlabeled web text results in a single model capable of transferring along multiple dimensions of style (dialect, emotiveness, formality, politeness, sentiment) despite *no additional training* and using only a handful of exemplars at inference time.

1 Introduction

There has been a recent surge of interest in text style transfer, with the aim of training models able to modify specific attributes of input text (e.g., sentiment or formality) while preserving the remaining content. For example, a sentiment transfer model might transform the input “best book ever!” into “worst book ever!”, while a formality transfer model might change the same input into “This is the best book I have ever read.” In these contexts, we define “style” as the attributes intended to be changed,

while “content” consists of the attributes intended to be preserved.¹

Work in this area falls into three categories. **Supervised** approaches like Jhamtani et al. (2017) transfer between pre-selected styles, and rely on parallel training data to learn the desired input/output correspondence. This method is limited by the availability of parallel corpora. So-called “**unsupervised**” approaches like Li et al. (2018) and Lample et al. (2019) remove the need for parallel data, but still require that all training examples have style labels, and are limited to transfer between a pre-specified set of styles. **Few-shot** approaches like that of Xu et al. (2020) remove the need for any training labels, instead using a small number of labeled examples during inference. While the most challenging, this offers the potential for transferring between arbitrary styles at inference time and has significant value, as curated datasets are not available for many style attributes.

In this work, we explore the hypothesis that large pretrained text-to-text models like T5 (Raffel et al., 2020) already contain a strong representation of textual style, which can be extracted and used to condition the decoder of a style transfer model through a relatively lightweight fine-tuning procedure. To isolate style information in the absence of labels, we rely on the observation that style is a “slow-moving” feature, which tends to be consistent over large spans of text. Specifically, given two adjacent sentences from an unlabeled corpus, we train our model to extract a “style vector” from the first and use that vector to perform denoising and other reconstruction tasks on the second. This technique extends the approach of Lample et al. (2019) to the few-shot setting, and is loosely reminiscent of the work of Akama et al. (2018), who found

* Work done while at Google Research. Please direct correspondence to priley3@cs.rochester.edu, nconstant@google.com and xyguo@google.com.

¹ Krishna et al. (2020) use a different definition of style, under which certain transfers such as sentiment would instead be examples of *attribute* transfer.

large context windows useful for encoding style information in word embeddings. Our approach also allows us to reformulate the style transfer operation as a directional operation in style vector space using the difference between target and source style vectors; we call this “targeted restyling”. When combined with a novel “tunable inference” technique for controlling token add/delete rates, this gives our final model: **Text Style Extraction and Tunable Targeted Restyling (TextSETTR)**.

Our main contributions are to: (1) present a new, flexible approach to few-shot style transfer, (2) use sentence adjacency as a means for inducing text style representations, (3) reframe style transfer as “targeted restyling” directional operations in style space, (4) introduce “tunable inference” for finer-grained control of transfers, (5) show the effectiveness of “noisy” back-translation training, and (6) illustrate few-shot generalization to a range of style attributes including dialect, emotiveness, formality, politeness, and sentiment.

2 Method

Figure 1 illustrates our proposed TextSETTR architecture. At a high level, our approach follows Lample et al. (2019), who train a denoising auto-encoder conditioned on a fixed-width style vector. The key difference in our case is that the true style is unknown at training time. To overcome this, we jointly train a “style extractor” component to induce a useful style representation (that can aid in reconstruction) from text in the nearby context. We describe this in more detail below.

2.1 Model Architecture

We conduct our experiments using a modified version of the Text-to-Text Transfer Transformer (T5) (Raffel et al., 2020). Like T5, our model includes a transformer-based encoder and decoder. As in T5 pretraining, the input to the encoder is a corrupted version of the target, resulting in a reconstruction task. Our goal is to design a type of corruption that results in this training task resembling style transfer, despite the lack of labeled training data.

Our core addition to T5 is the style extractor. This component’s architecture is based on that of the encoder, and its input is an uncorrupted sentence in the same style as the target; relying on our assumption that style is a slow-moving feature, we use the sentence preceding the target (the “context”) for this. This encourages extracting a

style representation that is useful for repairing the corrupted input. We note that this can result in a representation that encodes slow-moving attributes in general, which may include some features that do not fit an intuitive definition of textual style (such as topic).

The only architectural difference between the encoder and style extractor is that we mean-pool the style extractor’s hidden state sequence into a single fixed-width “style vector”; in our experiments, the dimensionality of this vector and the encoder hidden states is 1024. To incorporate the style vector into the rest of the model, we simply add it to each of the final encoder hidden states.

We initialize the weights of our model with those of a pretrained T5 model. We initialize both the style extractor and encoder from the pretrained encoder, but the weights are *not* tied during training.

2.2 Corruption Strategies

We experiment with combinations of three different reconstruction tasks, each contributing a loss term. All three share the same overall structure, where a sentence s_i in the dataset is corrupted by some function f to produce $\tilde{s}_i = f(s_i)$. The cross-entropy loss is calculated using the uncorrupted sentence s_i as the target, the corrupted sentence \tilde{s}_i as the input, and the uncorrupted preceding sentence s_{i-1} as the context. The three choices of f are Noise (N), Back-Translation (BT), and Noisy Back-Translation (NBT), described below.

Noise (N) This function corrupts the input by (i) dropping, (ii) replacing, and/or (iii) shuffling tokens, in that order. For each example we sample a separate noise probability p for each sub-type of noise from a uniform distribution in the range 20–60%; doing so should widen the model’s range of possible style transfers at test time.

For *drop* noise, we drop each token in s_i with probability p . For *replace* noise, let s_{ik} be the k -th token within s_i . For each s_i , a random other example s_j is chosen, and then each token s_{ik} is replaced by s_{jk} with probability p . If s_j has fewer than k tokens, then the replacement does not occur. For *shuffle* noise, each token in s_i is chosen with probability p , and then all chosen tokens are randomly shuffled to the position of another chosen token, leaving non-chosen tokens in place.

The use of drop and shuffle noise results in a loss term similar to the denoising loss used by Lample et al. (2019). Their motivation for this loss was

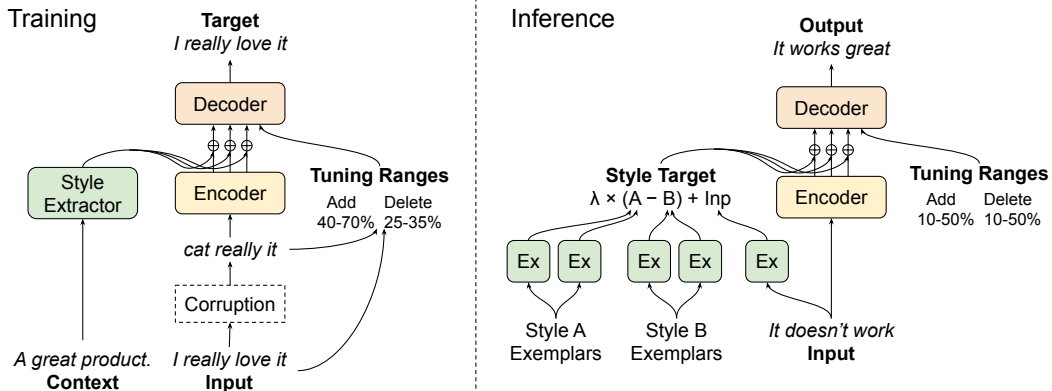


Figure 1: TextSETTR architecture for few-shot style transfer. The Encoder, Decoder and Style Extractor (Ex) are transformer stacks initialized from pretrained T5. During training, the model reconstructs a corrupted input, conditioned on a fixed-width “style vector” extracted from the preceding sentence. At inference time, a new style vector is formed via “targeted restyling”: adding a directional delta to the extracted style of the input text. Stochastic tuning ranges provide extra conditioning for the decoder, and enable fine-grained control of inference.

to encourage language modeling. As we fine-tune an already-strong T5 language model in our experiments, our motivation is rather to introduce a *conditional* element to the language model, in the form of the extracted style vector input.

Back-Translation (BT) This corruption function, used by Lample et al. (2019), runs the current version of the model in inference mode to transfer s_i into a different style, giving the corrupted \tilde{s}_i . In prior work using labels, specifying a different target style was straightforward. In our case, because we do not have access to labels, we simply sample a random sentence s_j to use as the context. To increase diversity of the generated examples, we decode with sampling instead of greedy decoding.

Because \tilde{s}_i is produced by a strong language model, BT should result in examples where both the input and output are coherent sentences, matching our inference setting. By contrast, Noise corruption does not resemble test-time inputs.

Noisy Back-Translation (NBT) This novel corruption function is a composition of the previous two. Noise is first applied to s_i as described above, and the result is used as the input (with randomly-sampled s_j as the context) to the model in inference mode to produce \tilde{s}_i via sampling, as in BT.

Once the model has learned to undo random noise, NBT should produce training examples where some of the tokens are preserved from s_i while others were generated by the model itself under the influence of the “incorrect” context s_j . This is similar to BT, but we hypothesize that it may be better suited to style transfer. BT was origi-

nally used for machine translation (Sennrich et al., 2016), a setting where most or all input tokens need to change. In contrast, style transfer within a single language usually requires only changing a subset of tokens; the training examples resulting from NBT should have this property. We believe that this will encourage the model to identify which tokens in the input do not match the target style indicated by s_{i-1} and change them, which is exactly what we want a style transfer model to do.

Final Loss The final loss term used for training is the sum of the above loss terms, each calculated from the same input s_i . However, not every model we experiment with includes all three losses.

2.3 Inference Procedure

Tunable Add/Delete Rates In preliminary experiments, we observed a recurring problem that the model would often change either far too little (failing to achieve the target style), or far too much (failing to preserve the input content). To address this problem, we introduce a “tunable inference” mechanism to constrain how much content should be added and deleted at inference time.

For every input/output pair during training, we calculate the proportions of tokens that were added and deleted. The “add rate” is the proportion of output tokens absent from the input, and the “delete rate” is the proportion of input tokens absent from the output.² We provide these rates to the decoder as *ranges* covering but not necessarily centered

²This calculation ignores word order. As one example, if a token appears three times in the input and five times in the output, two of the five occurrences are counted as “added”.

on the true rates.³ This approach provides more flexibility at inference time, so we can enforce tight or loose constraints on each rate.

Targeted Restyling While previous work on style transfer has largely assumed a fixed set of discrete styles, we expect our model’s learned style representations to capture a rich summary of the sentence covering many attributes without specifying them beforehand. For example, a given style vector might encode that a sentence is informal, humorous, in British English, and so on.

In this framework, transferring a single attribute (e.g., informal \rightarrow formal) is not as simple as just providing a vanilla “formal” style target, as this would ignore all the other attributes that defined the original input. Rather, we must operate in style space to construct a new target style that is simultaneously formal, humorous, British, and so on.

Concretely, at inference time, we assume access to a small set of “exemplar” sentences (between 1 and 100) for both the source value (e.g., informal) and target value (e.g., formal) of the attribute being modified. We infer style vectors for each exemplar using the style extractor, and take the mean of each class, giving vectors v^{src} and v^{trg} . Assuming the exemplar pools are relatively diverse, this averaging should “wash out” most untargeted attributes.

To transfer an input sentence x , we apply a targeted restyling in the appropriate direction. After extracting the original style from the input itself, v^x , we compute the target output style by moving in the direction of the delta between the source and target attributes values, as in (1), producing the style vector used for decoding. In practice, we find that the delta scale λ is an important hyperparameter to tune. Generally values in the range [1.0, 10.0] work well, with the best values depending on the attribute and the exemplars in question.

$$v^x + \lambda \times (v^{trg} - v^{src}) \quad (1)$$

3 Experiments on Sentiment Transfer

To evaluate our approach and better understand the effects of our various design choices, we test on few-shot sentiment transfer, using the Amazon reviews dataset of Li et al. (2018). However, as their training split doesn’t indicate which sentences

³Specifically, we sample each range width uniformly from [0,1], and uniformly sample the “alignment” of the true rate within the range. The final ranges are clipped to [0,1], and a vector containing the upper and lower bound of each range is prepended to the encoder hidden state sequence.

were adjacent in the original reviews, we make use of a different source of raw review text.

Training Procedure Our unlabeled training data comes from the 233.1M Amazon reviews provided by Ni et al. (2019). Ignoring the star ratings completely, we extract adjacent lines from multi-line reviews to use as the context and input for our training procedure, giving 23.6M examples. We also preprocess all text to match the format of the Li et al. (2018) data, as detailed in Appendix A.4. Initializing our model from pretrained T5 (t5.1.1.large), we fine-tune on these examples, optimizing the joint reconstruction loss from Section 2. Our default TextSETTR configuration is selected based on preliminary experiments (on development data) varying the set of reconstruction tasks and inference procedures. The model uses an equally weighted combination of the Noise (N) and Noisy Back-Translation (NBT) tasks. For both tasks, we use drop and replace noise, but no shuffle noise. We fine-tune for 10k steps, with a batch size of 65,536 tokens, and a fixed learning rate of 1e-3.

Evaluation Procedure Following prior work, we use automatic metrics to assess attribute control (sentiment) and content preservation on the data from Li et al. (2018). To estimate the sentiment of the output, we fine-tune a BERT-Large classifier (Devlin et al., 2019) on the train split, scoring 87.8% accuracy on the dev split. For content preservation, we follow Sudhakar et al. (2019) and Xu et al. (2020) and calculate self-BLEU between the output and input, using SacreBLEU (Post, 2018).^{4,5} Following Xu et al. (2018), we report “G-score” (the geometric mean of accuracy and content) as a summary of overall model quality.

To perform transfers, we follow the procedure from Section 2.3. For our default setup, we sample 100 positive and 100 negative exemplars from the Li et al. (2018) train split. Unless otherwise specified, we use greedy decoding, a delta scale of $\lambda=8$, and add/delete tuning ranges of 20–40%.

Core Results Figure 2 shows our core results. Our default TextSETTR configuration (N+NBT training, tuning ranges 20–40%) achieves 73.7% classifier-judged accuracy at swapping sentiment, while still staying somewhat close to the original

⁴Version string: BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.1.4.13

⁵Some prior work reports instead BLEU scores between outputs and human-generated transfers from Li et al. (2018); we found this to be highly correlated with self-BLEU but report it in Appendix A.3 for completeness.

	Model	Acc.	Content	G
Few-Shot	TextSETTR (10–30%)	54.0	55.8	54.9
	TextSETTR (20–40%)	73.7	34.7	50.6
	N	23.4	84.4	44.4
	NBT	70.0	27.8	44.1
	N + BT	13.3	98.7	36.2
	–replace noise	66.1	42.1	52.8
	+shuffle noise	70.3	34.1	49.0
	manual exemplars	52.4	44.2	48.1
	1000 exemplars	74.5	37.2	52.6
	–tunable inference	71.5	39.4	53.1
	overwrite style	25.3	55.8	37.6
	small train set	74.5	33.4	49.9
	CP-G	51.1	35.5	42.6
	CP-B	36.3	39.8	38.0
Labeled	CrossAligned	68.2	2.9	14.1
	Delete&Retrieve	49.4	56.9	53.0
	B-GST	60.2	54.2	57.1

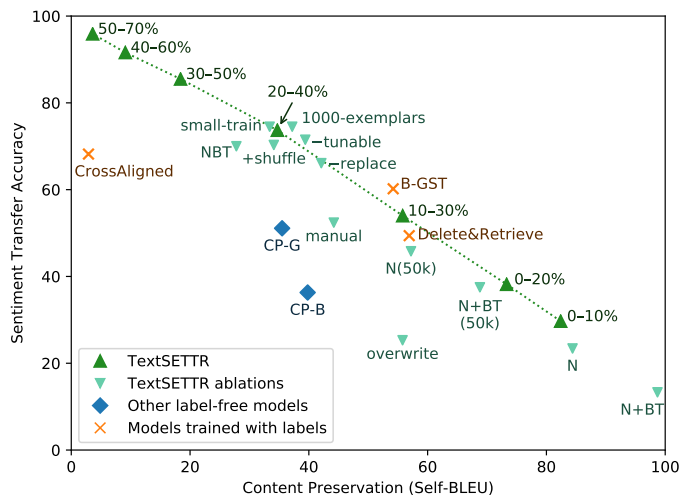


Figure 2: Automatic evaluation metrics comparing our TextSETTR model, ablations, and previous work. Up-and-right is better. We train for 10k steps and use add/delete:20–40% unless otherwise specified. We recalculate metrics for previous approaches, using our BERT classifier for accuracy, ensuring direct comparability.

Model	Accuracy	Content	G
TextSETTR (10–30%)	72.7	60.2	66.2
TextSETTR (20–40%)	83.6	39.4	57.4
Lample et al. 2019	82.6	54.8	67.3

Table 1: Comparison with Lample et al. (2019) on the setting that includes pos→pos and neg→neg transfers.

input text (self-BLEU 34.7). Due to our tunable inference technique, we can also trade off accuracy for content preservation by adjusting the add/delete rates, as seen in the points along the green line. Notably, TextSETTR outperforms the few-shot CP-G and CP-B models of Xu et al. (2020). More remarkably, TextSETTR outperforms several approaches that rely on training labels: CrossAligned (Shen et al., 2017) and Delete&Retrieve (Li et al., 2018). However there is still a small gap between our few-shot approach and the best labeled model, B-GST (Sudhakar et al., 2019).

In Table 1, we compare with Lample et al. (2019) on the evaluation setting including pos→pos and neg→neg transfers. This setting doesn’t match our inference procedure, which assumes that the input and output styles differ. Nevertheless, TextSETTR comes close to the performance of Lample et al. (2019), despite not benefiting from training labels.

As automatic metrics can diverge from human judgment (Sudhakar et al., 2019), we also conduct human evaluations of the three strongest models from Figure 2. We sample 200 examples per transfer direction from the Li et al. (2018) test set, and ask three annotators to evaluate each input/output

Model	Sentiment	Preservation	Fluency
TextSETTR (10–30%)	2.0	3.5	2.9
TextSETTR (20–40%)	2.5	2.6	4.0
Delete&Retrieve	2.5	3.1	3.3
B-GST	2.2	2.9	3.6

Table 2: Human evaluation metrics.

pair on three metrics: sentiment transfer (how well the model changed the sentiment), content preservation, and fluency, on scales of 1–5. The results in Table 2 confirm that TextSETTR achieves similar quality to models that benefit from training labels. Further details are presented in Appendix A.5.

3.1 Ablations

Modifying Inference Procedure To better understand the value of our proposed “targeted restyling” mechanism, we consider an alternative inference procedure where we ignore the style of the input and simply use the average target exemplar style v^{trg} as the style vector. We expect that since our learned style space covers multiple attributes, this will result in setting the target attribute (e.g. sentiment) while simultaneously *overwriting* all other style attributes (e.g. formality) using the average style of the target exemplars. This is borne out in our “overwrite style” ablation, which performs significantly worse than our baseline: accuracy drops from 54.0% to 25.3% with no gain in self-BLEU.

To assess the value of tunable add/delete rates, we also train a model (–tunable) without this feature. While the automatic metrics are slightly above the TextSETTR line, we observe several advan-

tages to the tunable model. For one, we observe it significantly reduces the variance in self-BLEU across different inputs. For example, focusing on the case of overly high self-BLEU, we find that without tunable inference, 14.6% of dev eval outputs are identical to their inputs, whereas with tunable inference, this goes to 0.9%. Additionally, through qualitative analysis in Section 4, we find that tunable inference allows more flexibility for controlling different types of transfer.

Adjusting Data Sizes While our unlabeled training data set consists of 23.6M examples, our model only sees 5.1M of these over its 10k steps of training. Yet this is still nearly $10\times$ more data than the 0.6M examples in the Li et al. (2018) training set used by previous approaches. For a more direct comparison, we experiment with a “small train set”, sampling 0.6M examples from our training set. Remarkably, the results in Figure 2 are nearly identical to our baseline, supporting our hypothesis that a fairly lightweight adaptation is sufficient to allow T5 to extract and transfer textual style.

To test the limits of our model’s generalization, we reduce the set of exemplars to four manually selected examples of each class. In this setting, we also find reducing delta scale to $\lambda=4$ is beneficial. The results, shown as “manual exemplars” in Figure 2, are still competitive, indicating that our approach generalizes well to this very-few-shot inference setting. In the other direction, we find that increasing the number of sampled exemplars from 100 to 1000 only provides small additional gains.

Modifying Training Task Lample et al. (2019) showed promising results by combining noise (N) with back-translation (BT). However we find this combination unstable.⁶ When training for 10k steps, our N and N+BT models nearly always copy their input. Training for 50k steps recovers reasonable performance, but the metrics still fall below the TextSETTR line, using our novel NBT task. We also experiment with using NBT in isolation, but this again underperforms our baseline. We expect that the denoising task helps to ensure the NBT inputs (themselves the outputs of denoising) consist of realistic well-formed text. Finally, while Lample

⁶For all experiments in the paper, we use 0.0 for the add/delete rates during the forward pass of back-translation. However we later found that using *random* add/delete rates in back-translation can improve performance in the N+BT setting. On sentiment transfer, this improved our N+BT ablation to self-BLEU 42.4, accuracy 71.4%, G-score 55.0.

et al. (2019) use drop and shuffle noise, we find that only drop and replace are valuable.

3.2 Embedding Visualization

To demonstrate that our learned style extractor encodes multiple aspects of textual style, we compute style vectors for 12,000 lines of text from three review categories (Fashion, Software, Pantry) from the Ni et al. (2019) Amazon data. Within each category, we sample 2,000 positives (4 or 5 star) and 2,000 negatives (1 or 2 star), filtering examples where our BERT classifier disagrees with the label. Figure 3 (bottom) plots a 2D UMAP dimensionality reduction (McInnes et al., 2018) of the vectors, and shows clear separations among sentiments and product categories. The top row runs UMAP with the same settings, but over style vectors from our model *before* training, where the style extractor is initialized from pretrained T5. The contrast is a clear indication that our training procedure is helping to learn a representation space where sentiment and topic values are well separated.

To confirm that the observed separation isn’t an artifact of dimensionality reduction, we compute the average distance between style vectors (a) *within* a class, and (b) *across* classes. We measure “separation” as the relative increase in mean distance between these two conditions. For product category, we find TextSETTR training improves separation from 1.7% to 8.1%. For sentiment, TextSETTR training improves separation from 0.9% to 4.7%.

4 One Model for All Styles

An advantage of few-shot style transfer is that, in theory, a single model can perform transfer along any “dimension” of style given only a few exemplars, without the need for additional training. In this section, we investigate the degree to which our approach achieves this goal in practice. For this purpose, we train a single general-purpose TextSETTR model, with the same configuration as our model from Section 3, except fine-tuned for 200k steps on English Common Crawl data (the same “C4” data that T5 pretrained on) instead of Amazon reviews.

Qualitative Evaluation Given that our architecture limits the style representation to 1024 dimensions, one may ask how the unsupervised model will make use of this capacity, and which style attributes will be encoded in the learned space. Encouragingly, we find that our model trained on un-

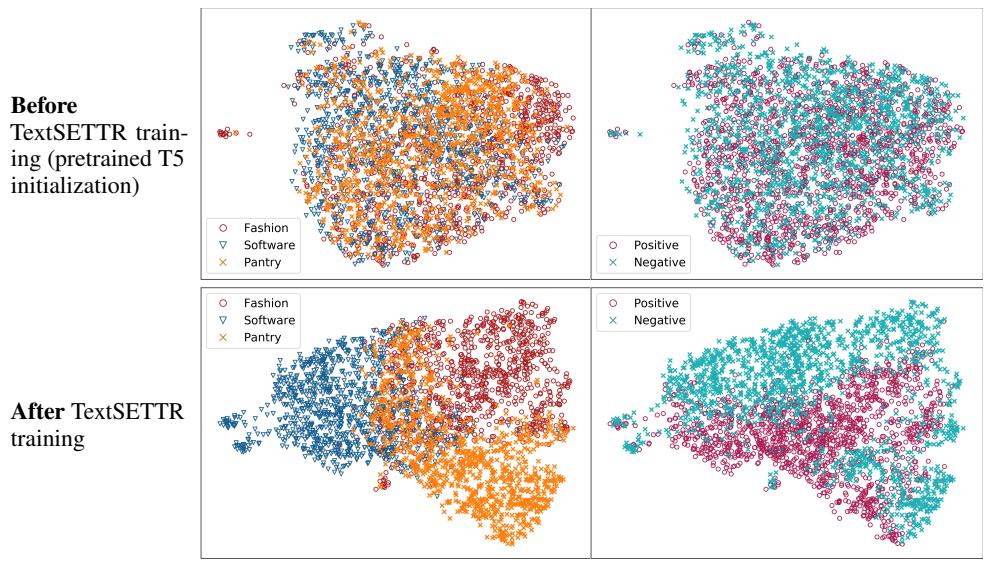


Figure 3: 2D UMAP embeddings of the style vectors extracted by our TextSETTR model before and after training, for text inputs from Amazon reviews covering three product categories and two sentiment labels. Within each row, the same embeddings are labeled with product category (left) and sentiment (right). We sub-sample to 3,000 points after dimensionality reduction. Note, we don't expect perfect separation, as inputs may be underspecified for category ("I love this product") or for sentiment ("I bought this last month"). We also don't expect to see crisp linear separation within each attribute since we aim for the learned embedding space to encode many style attributes simultaneously.

<p>Reserved ⇒ Emotive</p> <p>I <u>liked</u> the movie. ⇒ I <u>cannot even describe how amazing this</u> movie <u>was!!</u></p> <p>I was <u>impressed</u> with the results. ⇒ I was <u>absolutely blown away</u> with the results!!</p>	<p>Emotive ⇒ Reserved</p> <p>I <u>loved every minute of</u> the movie! ⇒ I <u>liked</u> the movie.</p> <p>I was <u>shocked</u> by the <u>amazing</u> results! ⇒ I was <u>surprised</u> by the results.</p>
<p>American ⇒ British</p> <p>The <u>elevator</u> in my <u>apartment</u> isn't working. ⇒ The <u>lift</u> in my <u>flat</u> isn't working.</p> <p>The <u>senators</u> will return to <u>Washington</u> next week. ⇒ The <u>MPs</u> will return to <u>Westminster</u> next week.</p>	<p>British ⇒ American</p> <p>The <u>lift</u> in my <u>flat</u> isn't working. ⇒ The <u>elevator</u> in my <u>apartment</u> isn't working.</p> <p><u>MPs</u> will return to <u>Westminster</u> next week. ⇒ <u>Representatives</u> will return to <u>Washington</u> next week.</p>
<p>Polite ⇒ Rude</p> <p><u>Are you positive</u> you've understood my point? ⇒ you've <u>never</u> understood my point!</p> <p><u>Could</u> you ask <u>before</u> using my phone? ⇒ I ask you <u>to stop</u> using my phone!</p>	<p>Rude ⇒ Polite</p> <p><u>What the hell</u> is <u>wrong</u> with your attitude? ⇒ <u>Perhaps</u> the <u>question</u> is <u>more about</u> your attitude.</p> <p>I could <u>care less</u>, <u>go</u> find somebody else to do this <u>crap</u>. ⇒ I could <u>be wrong</u>, <u>but I would try to</u> find somebody else to do this.</p>
<p>Formal ⇒ Informal</p> <p>I <u>hereby commit to</u> never <u>purchase</u> anything from this <u>institution in the future</u>. ⇒ <u>i gonna</u> never <u>buy</u> anything from this <u>place again</u>.</p> <p>I <u>couldn't figure out</u> what <u>the author was</u> trying to say. ⇒ <u>i dont know</u> what <u>ur</u> trying to say.</p>	<p>Informal ⇒ Formal</p> <p><u>best</u> book <u>ever!!</u> ⇒ <u>The book is highly recommended</u>.</p> <p><u>couldnt</u> figure out what author <u>tryna</u> say ⇒ <u>The reader couldn't</u> figure out what <u>the</u> author <u>was trying to</u> say.</p>
<p>Positive ⇒ Negative</p> <p>I was pretty <u>impressed</u> with the results. ⇒ I was pretty <u>disappointed</u> with the results.</p> <p>I will definitely buy this brand again. ⇒ I will definitely <u>not</u> buy this brand again.</p>	<p>Negative ⇒ Positive</p> <p>I was pretty <u>disappointed</u> with the results. ⇒ I was pretty <u>impressed</u> with the results.</p> <p>I definitely won't buy this brand again. ⇒ I definitely won't <u>hesitate to</u> buy this brand again.</p>

Table 3: Examples of transferring along five different axes of style. The same model is used across all examples, with no additional training. Words deleted from the input are red, and words added in the output are blue. Within each category, a fixed tiny set of exemplars is chosen, and fixed delta scale and tuning rates are used. The exemplars and settings are provided in Appendix A.2.

labeled Common Crawl data is capable of transferring along many independent axes of style. Table 3 shows selected successful examples of our Common Crawl model transferring emotiveness, dialect, politeness, formality and sentiment. The same model is used in each case, with no additional training. At inference time, a tiny set of exemplars (1–5 examples of each class) is the only labeled data used to compute the style vector delta; these exemplars are presented in Appendix A.2.

Across each type of transfer, we see evidence of generalization beyond the specifics of the chosen exemplars. In making text more emotive, the model uses *amazing* and *blown away*, despite these terms not occurring in the exemplars. In making text more polite, the model inserts novel hedges like *perhaps* and *I could be wrong*. In transferring between American and British styles, the model generalizes to unseen vocabulary items (*elevator* \leftrightarrow *lift*) and draws sound analogies (*senators* \leftrightarrow *MPs*). We do note though that the latter case illustrates that the model is willing to change the semantic content of the input in cases where it would otherwise be out-of-place in the target style. Future work includes investigating ways to control this in settings where such behavior is not desired.

Quantitative Evaluation To assess the quality of our general-purpose TextSETTR model, we benchmark the same model on three distinct transfer tasks in Table 4.⁷ The **sentiment transfer** task follows the evaluation procedure from Section 3. While our generic model underperforms our model trained on Amazon reviews, it still outperforms other few-shot methods. For **author transfer**, we use the Shakespeare-to-modern task of Jhamtani et al. (2017). Here, TextSETTR outperforms the previous best model of He et al. (2020) that leveraged 36,790 labeled examples during training. For **personality transfer**, we use the task of Li et al. (2020), which requires transferring between three personalities: angry, happy, malicious. We compare⁸ TextSETTR, which sees no labels in training and only 100 of each class in inference, with CARA (Li et al., 2020), which trained on 2,604 labels.

⁷For each task, we set our tuning ranges to 20–40% and compute target styles using 100 exemplars of each class taken from the train set. We use λ values of sentiment:8, author:16, personality:8. To measure accuracy, we fine-tune BERT-Large classifiers over the training data, reaching validation accuracies of sentiment:87.8%, author:89.7%, personality:81.9%.

⁸Note, as Li et al. (2020) use a different classifier to assess accuracy, those numbers may not be directly comparable.

Task	Model	Acc.	Content	G
Sentiment	CP-G	51.1	35.5	42.6
	CP-B	36.3	39.8	38.0
	TextSETTR	44.9	54.4	49.4
	CrossAligned	68.2	2.9	14.1
	Delete&Retrieve	49.4	56.9	53.0
Author	B-GST	60.2	54.2	57.1
	UNMT	68.5	7.8	23.1
	BT+NLL	59.3	12.4	27.1
	He et al. 2020	68.5	12.5	29.2
	TextSETTR	81.7	13.8	33.5
Personality	CARA	91.6	21.6	44.5
	CARA _{AB}	66.2	29.7	44.3
	Ctrl-Gen	67.6	22.9	39.3
	ARAE-	88.0	20.3	42.3
	TextSETTR	49.3	46.0	47.6

Table 4: Automated metrics comparing our general-purpose TextSETTR model with recent work on three transfer tasks. To enable direct comparison, “content” refers to reference-BLEU for author transfer, and self-BLEU elsewhere. Apart from CP-G/CP-B, all competitors are trained for only one type of transfer using labeled data. Personality transfer results are from Li et al. (2020), while all others are recalculated from scratch.

4.1 Dialect-Sensitive Completion

In addition to performing style and attribute transfer, we find that our system can also be used as a style-aware language model capable of completing prompts in a specified style. Examples of completions in American and British English are given in Table 5. In each case, the input is of the form “My favorite X: ”. Despite the fact that TextSETTR is not trained specifically for completions, we can use the add/delete rates to encourage the model to insert a few additional tokens, while leaving the original prompt largely unchanged.⁹

The completions demonstrate knowledge of stereotypical American and British culture. It is remarkable that the model is able to generalize to “deeper” cultural differences such as music and drink preferences, given only the shallow vocabulary differences (e.g., *neighbor* vs. *neighbour*) presented in the limited set of exemplars in Table 9.

It is also worth highlighting that, thanks to our directional transfer procedure, these completions are not merely “typical American” or “typical British” such as we would expect from a conditional language model trained on each sub-domain of text. Rather, since our inference procedure pushes the style away from one domain and towards the other, the resulting completions are *distinctive* representations of each dialect. As one example, we expect

⁹We note that in transferring American to British, the model prefers to change the prompt from *favorite* to *favourite*.

American \Rightarrow British	British \Rightarrow American
My favourite food: fish and chips.	My favorite food: quinoa.
My favourite hot drink: a mug of tea.	My favorite hot drink: Starbucks Coffee.
My favourite dessert: a scone!	My favorite dessert: a brownie.
My favourite city: Cardiff.	My favorite city: San Diego.
My favourite band: The Beatles.	My favorite band: The Black Keys.
My favourite sports league: the English Premier League.	My favorite sports league: the NFL.
My favourite newspaper: The Daily Telegraph.	My favorite newspaper: The Washington Post.
My favourite museum: the British Museum.	My favorite museum: The National Air and Space Museum.

Table 5: Examples of dialect-sensitive completion ($\lambda=8$, add:40–70%, delete:0%). In each case, the input text consists of an unfinished phrase, for example: “My favorite food: ”. The three exemplars used for each dialect are the same as those used for the transfers in Table 3, as listed in Table 9.

“quinoa” would not only be a common American favorite, but also an uncommon British favorite.

Additional examples of using our model for tasks other than pure style transfer are presented in Appendix A.1.

5 Related Work

As mentioned at the outset, recent work on text style transfer falls into three classes: supervised, “unsupervised”, and few-shot. Supervised style transfer has seen limited research due to the difficulty of obtaining parallel data. Examples include Jhamtani et al. (2017) and Carlson et al. (2018).

Unsupervised Approaches The bulk of research has focused on “unsupervised” approaches, which rely on labeled but non-parallel data. Typically, labels are assumed to be available for both source and target styles (Shen et al. 2017, Li et al. 2018, Niu et al. 2018, and many others). Zhao et al. (2018) explore the case where only the target style is labeled. The use of labels at training time can aid modeling, but limits the applicability of these methods, as labeled datasets are not readily available for many attributes of interest.

Our work differs from the above by removing the need for training labels, and offering a single model that can target an unrestricted set of style attributes. Despite these differences, our work shares some similarities with past work. For example, our encoder-decoder architecture and corruption methods are similar to Lample et al. (2019), and we leverage a strong pretrained language model, as in Sudhakar et al. (2019) and Wu et al. (2019).

Few-Shot Approaches A few-shot approach has recently been explored by Xu et al. (2020). The authors train a variational auto-encoder on unlabeled text, where a “manipulable” portion of the latent representation is constrained to fall on a k -dimensional simplex. To perform transfer,

they identify empirically the basis vector that most strongly corresponds to the target attribute, and manipulate its magnitude. Compared to our approach, a key difference is that the number of latent factors must be chosen ahead of time, which limits the number of attributes that may be controlled. Additionally, there is no guarantee that a single basis of the learned simplex will correspond to a target attribute such as dialect or politeness.

Controlled Generation A separate strand of research explores “controlled generation” methods for supplementing generative language models to allow control of specific attributes of the output text. As with style transfer, this can be achieved either through labeled training examples, as in CTRL (Keskar et al., 2019) and PPLM (Dathathri et al., 2020), or a few-shot approach, as in CoCon (Chan et al., 2020). These models differ from style transfer models in that they aim to generate plausible *continuations* following a prompt, as opposed to transferring attributes of a fully-formed input while preserving as much content as possible. It is not clear if controlled generation models could be used to perform style transfer, and they have not to our knowledge been evaluated in this context.

6 Conclusion

We have presented a unique approach to few-shot text style transfer that is competitive with systems trained with labels (an easier setting), while allowing control of how much of the input is changed. We demonstrate that this approach can produce a single system capable of transferring many different styles while requiring only a handful of exemplars at inference time.

Acknowledgments

We thank Llion Jones, Rami Al-Rfou, and Daniel Gildea for helpful discussion and comments on an earlier draft.

References

- Reina Akama, Kento Watanabe, Sho Yokoi, Sosuke Kobayashi, and Kentaro Inui. 2018. [Unsupervised learning of style-sensitive word vectors](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 572–578, Melbourne, Australia. Association for Computational Linguistics.
- Keith Carlson, Allen Riddell, and Daniel Rockmore. 2018. [Evaluating prose style transfer with the bible](#). *Royal Society Open Science*, 5(10):171920.
- Alvin Chan, Yew-Soon Ong, Bill Pung, Aston Zhang, and Jie Fu. 2020. [CoCon: A self-supervised approach for controlled text generation](#). *CoRR*, abs/2006.03535.
- Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. [Plug and play language models: A simple approach to controlled text generation](#). In *International Conference on Learning Representations*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. [A probabilistic formulation of unsupervised text style transfer](#). In *International Conference on Learning Representations*.
- Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. [Shakespeareizing modern language using copy-enriched sequence to sequence models](#). In *Proceedings of the Workshop on Stylistic Variation*, pages 10–19, Copenhagen, Denmark. Association for Computational Linguistics.
- Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. [CTRL: A conditional transformer language model for controllable generation](#). *CoRR*, abs/1909.05858.
- Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. [Reformulating unsupervised style transfer as paraphrase generation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762, Online. Association for Computational Linguistics.
- 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, New Orleans, Louisiana. Association for Computational Linguistics.
- Yuan Li, Chunyuan Li, Yizhe Zhang, Xiujuan Li, Guoqing Zheng, Lawrence Carin, and Jianfeng Gao. 2020. [Complementary auxiliary classifiers for label-conditional text generation](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8303–8310.
- Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. [UMAP: Uniform manifold approximation and projection](#). *Journal of Open Source Software*, 3(29):861.
- Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. [Justifying recommendations using distantly-labeled reviews and fine-grained aspects](#). 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 188–197, Hong Kong, China. Association for Computational Linguistics.
- Xing Niu, Sudha Rao, and Marine Carpuat. 2018. [Multi-task neural models for translating between styles within and across languages](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1008–1021, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Improving neural machine translation models with monolingual data](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.
- Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. [Style transfer from non-parallel text by cross-alignment](#). In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan,

and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6830–6841. Curran Associates, Inc.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. “Transforming” delete, retrieve, generate approach for controlled text style transfer. 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 3269–3279, Hong Kong, China. Association for Computational Linguistics.

Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Mask and infill: Applying masked language model for sentiment transfer. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5271–5277. International Joint Conferences on Artificial Intelligence Organization.

Jingjing Xu, Xu Sun, Qi Zeng, Xiaodong Zhang, Xuancheng Ren, Houfeng Wang, and Wenjie Li. 2018. Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 979–988, Melbourne, Australia. Association for Computational Linguistics.

Peng Xu, Jackie Chi Kit Cheung, and Yanshui Cao. 2020. On variational learning of controllable representations for text without supervision. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10534–10543. PMLR.

Yanpeng Zhao, Wei Bi, Deng Cai, Xiaojiang Liu, Kewei Tu, and Shuming Shi. 2018. Language style transfer from sentences with arbitrary unknown styles. *CoRR*, abs/1808.04071.

A Appendix

A.1 Beyond Style Transfer

In this section, we provide additional examples illustrating the abilities of our TextSETTR model trained on Common Crawl data, beyond typical style transfer.

Examples of **shortening** are given in Table 6, with inputs taken from the first five sentences of the Wikipedia article “Artificial neural network”. As shortening may require minor rephrases, we set our tuning ranges to add:0–5%, delete:40–90%. Since our intention is to leave the style unchanged (apart from length), we extract the target style directly from the input text, with no delta added. The model is largely successful at identifying and removing “superfluous” content, and finding ways of rephrasing to shorten while preserving meaning.

Examples of **random augmentations** are given in Table 7. In each case, we transfer the input sentence “What’ll the weather be tomorrow?” to a slightly different style. Specifically, for each transfer, we extract this sentence’s style vector and apply a small amount of noise, with each component of the noise vector sampled from a Gaussian $\mathcal{N}(0, 0.08)$. Note that apart from the noise in the style vector, the transfer process is deterministic, as we use greedy decoding.

The cells of Table 7 apply different tuning ranges, conditioning the model to change a little or a lot. Within each cell, we repeatedly sample the noised style, and present the first five unique outputs. The results indicate that many random changes in style are largely meaning preserving, especially when a small change is requested. With larger add/delete rates, the outputs are still closely related in meaning, despite low lexical overlap.

A.2 Settings used for Qualitative Analysis

For each of the transfer types (e.g., formal \leftrightarrow informal) in Table 3, we specify the intended target styles through a tiny set of exemplars. These exemplars are provided in Tables 8–12. Additionally, for each transfer type, we select a delta scale λ and add/delete rates. These settings are selected through initial experiments, and are held fixed across all examples of transfer shown.

A.3 Human Reference BLEU

Li et al. (2018) provide human reference transfers for their Amazon test data, and report BLEU scores of model outputs against these targets. In principle,

we believe this metric is less informative than self-BLEU, as style transfer is a relatively open-ended task, and successful transfers may differ significantly from the single human reference. However, for completeness, we report “reference BLEU” of our models and those of prior work in Figure 4. We observe BLEU and self-BLEU are highly correlated, and the “Accuracy vs. BLEU” plot conveys the same relationships we saw in Figure 2. As before, all BLEU scores are calculated using Sacre-BLEU (Post, 2018).

A.4 Amazon Reviews Preprocessing

We use the code in Figure 5 to process raw Amazon reviews from the Ni et al. (2019) dataset and extract pairs of adjacent lines, preprocessed to have a similar format to Li et al. (2018) dataset. We split reviews on newlines, and clip lines to 100 characters, always ending with a period. This gives results similar to Li et al. (2018), where one line may contain multiple sentences, and may consist of a “half-sentence” ending with “e.g.” or a similar non-sentence-final period. Additionally, we apply various tokenization and normalization operations to roughly match the observed Li et al. (2018) text.

A.5 Human Evaluation Setup

For the human evaluations of our models, we employed 3 in-house annotators. The annotators were paid hourly wages that are competitive for their locale and have standard rights as contractors. They spoke native English.

For the evaluation task, the annotators were shown both the original and transformed pieces of text. They were then asked to evaluate for three metrics: fluency, meaning preservation, and sentiment change.

For *fluency*, they were asked, “For the new text, how do you rate the fluency, i.e., the quality and readability of the text, with 1 being not fluent at all and 5 being very fluent.” For *meaning preservation*, they were asked, “Comparing the new text against the original text, and ignoring the change of style, how well does the new text preserve as much of the original meaning, with 1 being all meaning is lost and 5 being preserving as much as possible given the sentiment change?” And for *sentiment change*, they were asked, “Comparing the new text against the original text, how well did the sentiment of the new text become more positive, with 1 being not more positive and 5 being a lot more positive?”

Artificial neural networks (ANN) or connectionist systems are computing systems that are inspired by, but not identical to, biological neural networks that constitute animal brains.

⇒ Artificial neural networks (ANNs) are computing systems that are inspired by the biological neural networks that constitute animal brains.

Such systems “learn” to perform tasks by considering examples, generally without being programmed with task-specific rules.

⇒ Such systems learn to perform tasks by considering examples, generally without explicit rules.

For example, in image recognition, they might learn to identify images that contain cats by analyzing example images that have been manually labeled as “cat” or “no cat” and using the results to identify cats in other images.

⇒ For example, image recognition systems might learn to identify images that contain cats by analyzing images that have been manually classified as “cat” or “no cat”.

They do this without any prior knowledge of cats, for example, that they have fur, tails, whiskers and cat-like faces.

⇒ They do not know that cats have fur, tails, whiskers and cat-like faces.

Instead, they automatically generate identifying characteristics from the examples that they process.

⇒ Instead, they automatically generate identifying characteristics.

Table 6: Examples of shortening (add:0–5%, delete:40-90%), using the first five sentences from the Wikipedia article “Artificial neural network”. For each sentence, the target style is extracted directly from the input text, and no delta is added.

Add/Delete: 10–30%	Add/Delete: 30–50%
What’ll the weather be like?	What’s the weather like?
What’ll the weather be like tomorrow?	What will the weather be like tomorrow?
What’s the weather like tomorrow?	Will the weather be better tomorrow?
What’ll the weather be tomorrow?	What’s the weather forecast for tomorrow?
What’s the weather supposed to be tomorrow?	How will the weather be tomorrow?
Add/Delete: 50–70%	Add/Delete: 70–90%
Will the weather be perfect tomorrow?	How do you know what the weather will be like?
What’s the weather for tomorrow?	Is it supposed to be cold tomorrow?
What’s the weather like on the course?	What will the weather be like in the South?
Hopefully the weather will be better tomorrow.	I’m not a fan of the weather.
What’s the weather like for the next day?	What is the temperature and what is the humidity.

Table 7: Random augmentations of input text “What’ll the weather be tomorrow?”, using random style vector deltas with components sampled from $\mathcal{N}(0, 0.08)$.

Reserved Exemplars	Emotive Exemplars
1. That is a very pretty painting.	1. OMG, that’s such a beautiful painting!
2. I’m excited to see the show.	2. I’m sooo excited to see the show, it’s going to be stellar!!
3. I’m surprised they rescheduled the meeting.	3. I absolutely can not believe that they rescheduled the meeting!
4. This specimen is an example of the baroque style.	4. This wonderful specimen is a truly spectacular example of the baroque style.
5. After the performance, we ate a meal.	5. After the superb performance, we ate a delicious meal.

Table 8: Emotiveness transfer exemplars. Transfer settings: $\lambda=9$, add/delete rates: 0–100%.

American Exemplars	British Exemplars
1. It cost ten bucks.	1. It cost ten quid.
2. My neighbor apologized.	2. My neighbour apologised.
3. I’m heading out to the bar with some friends.	3. I’m heading out to the pub with some mates.

Table 9: Dialect transfer exemplars. Transfer settings: $\lambda=8$, add/delete rates: 10–30%.

Polite Exemplars	Rude Exemplars
1. No thank you, I'd prefer not to.	1. Hell no, you can't make me do that.
2. This game could have been better designed.	2. This game is such a piece of garbage!
3. Do you know why they might have delayed the launch?	3. Why in god's name would they delay the damn launch?
4. Sorry, I wasn't certain if you were joking.	4. Are you frigging kidding me?

Table 10: Politeness transfer exemplars. Transfer settings: $\lambda=5$, add/delete rates: 20–50%.

Formal Exemplars	Informal Exemplars
1. This was a remarkably thought-provoking read.	1. reading this rly makes u think
2. It is certainly amongst my favorites.	2. Its def one of my favs
3. We humbly request your presence at our gala on the 12th.	3. come swing by our bbq next week if ya can make it

Table 11: Formality transfer exemplars. Transfer settings: $\lambda=4$, add/delete rates: 40–80%.

Positive Exemplars	Negative Exemplars
1. Five stars, I love it.	1. Zero stars, I hate it.

Table 12: Sentiment transfer exemplars. Transfer settings: $\lambda=3$, add/delete rates: 0–100%.

Model	BLEU	Self-BLEU
CrossAligned	2.0	2.9
Delete&Retrieve	29.7	56.9
B-GST	29.0	54.2
CP-G	17.0	35.5
CP-B	19.4	39.8
TextSETTR (0–20%)	39.0	73.3
TextSETTR (10–30%)	30.7	55.8
TextSETTR (20–40%)	20.0	34.7
TextSETTR (30–50%)	10.6	18.4
TextSETTR (40–60%)	5.5	9.1
TextSETTR (50–70%)	2.2	3.6

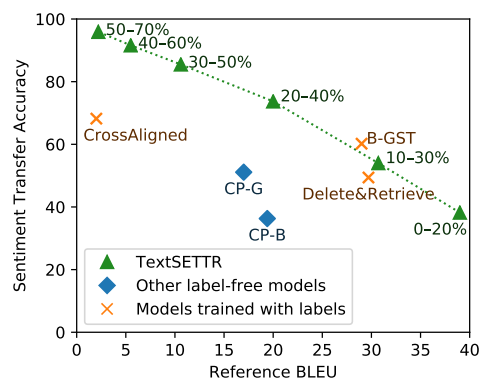


Figure 4: BLEU scores between model outputs and human references provided by Li et al. (2018), along with self-BLEU for comparison. The first group of models in the table had access to labels at training time, while the second group did not. TextSETTR (X–Y%) refers to our model with add/delete rate ranges set to X–Y%.

```

import re
from html.parser import HTMLParser

html_parser = HTMLParser()

def preprocess(line):
    """Simulate Li et al. preprocessing of one review line."""
    # Lowercase.
    line = line.lower()
    # Replace apostrophes, parens and quotes with spaces.
    line = re.sub("'(\)|\|)", " ", line)
    # Replace dollar values ==> $
    line = re.sub("\$[\d.]*", "$", line)
    # Replace percent values ==> %
    line = re.sub("[\d.]*%", "%", line)
    # Replace single digits ==> num_num
    line = re.sub("\d[ ,]", " num_num ", line)
    # Replace multi-digits and codes ==> num_extend
    line = re.sub("\d[^\d]*", " num_extend", line)
    # Remove remaining numbers, including decimals.
    line = re.sub("\d[\d.]*", "", line)
    # Add spaces around certain punctuation marks.
    line = re.sub("([.,?!:])", r" \1 ", line)
    # Remove double spaces after periods before words.
    return re.sub(r"\. ([a-z])", r". \1", line)

def acceptable_line(line):
    """Check if text looks like an acceptable line from Li et al."""
    if not line or len(line) < 30 or len(line) >= 100:
        return False
    # Avoid lines with any char absent from Li et al. train.
    if re.search('[^ !$%+, .:;>?@\^_`a-z{|}]', line):
        return False
    return True

def clip_to_last_period(line):
    return line[:len(line) - line[::-1].index('.')]

def adjacent_lines(review):
    """Extract a list of adjacent line pairs from review text."""
    review = html_parser.unescape(review)
    review = review.replace('\\"', '')
    # Simulate Li et al. splitting and filtering.
    if '\n' not in review:
        return
    lines = review.split('\n')
    lines = [preprocess(clip_to_last_period(l[:100]))
              for l in lines if l and "." in l[:100]]
    lines = [preprocess(l) for l in lines]
    lines = [l for l in lines if acceptable_line(l)]
    if len(lines) < 2:
        return
    return list(zip(lines[:-1], lines[1:]))

```

Figure 5: Python code to extract adjacent lines of text from raw Amazon reviews, producing outputs in a similar style to Li et al. (2018).