TextSETTR: Label-Free Text Style Extraction and Tunable Targeted Restyling

https://arxiv.org/abs/2010.03802

Ukjae Jeong, jeongukjae@gmail.com

2021.03.21

Introduction

- Recent interest of text style transfer: Modify specific attributes (e.g., sentiment or formality)
- Approaches
  - **Supervised**: Rely on aligned parallel data. Very limited by the availability of parallel corpora.
  - **Unsupervised**: Require labeled training set of each style. Limited to transfer a pre-specified set of styles.
  - **Label-free**: Remove the needs for any training labels. Transfer arbitrary styles at inference time.

Introduction

- In this paper, the authors propose **targeted restyling** besides the **tunable inference** technique.
- Main contribution
  - demonstrate the viability of label-free style transfer
  - use sentence adjacency as a means for inducing text style representations
  - reframe style transfer as **“targeted restyling” directional operations** in style space
  - introduce **“tunable inference” for finer-grained control of transfers**
  - show the effectiveness of “noisy” back-translation training
  - illustrate **few-shot generalization** to a range of style attributes including dialect, emotiveness, formality, politeness, and sentiment.

Method

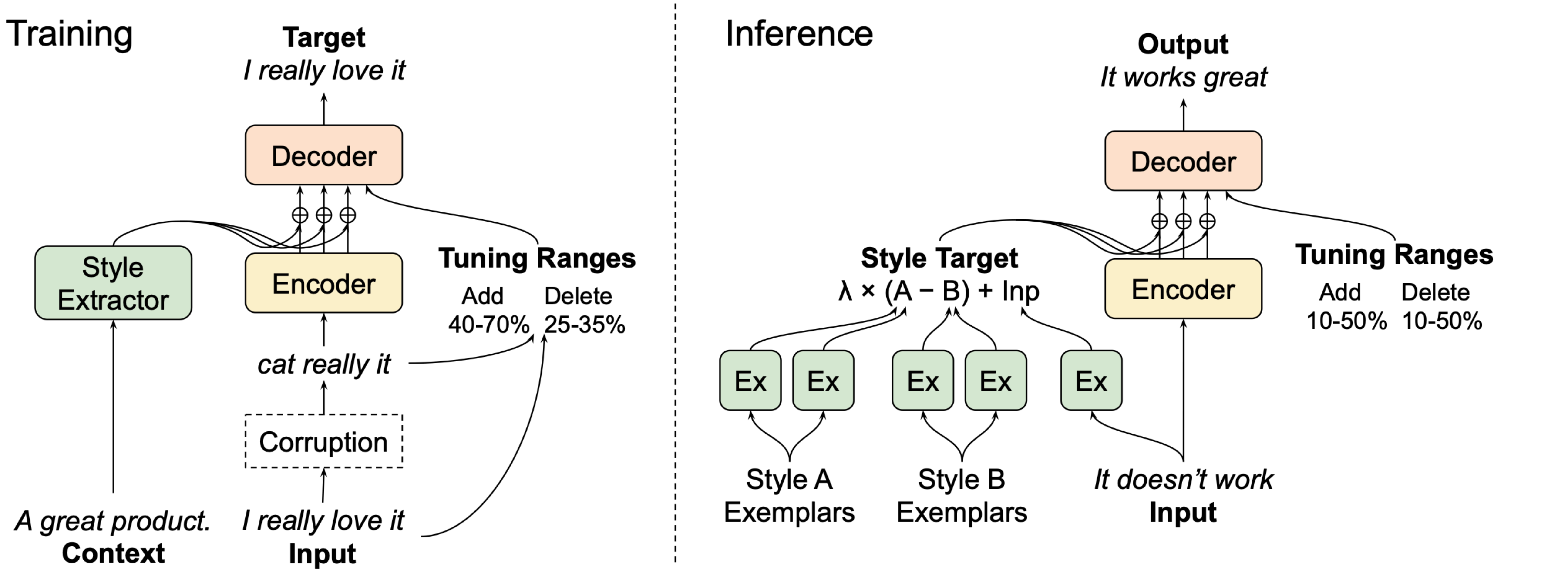

Figure 1: TextSETTR architecture for label-free style transfer. The Encoder, Decoder and Style Extractor (Ex) are transformer stacks initialized from pretrained T5. During training, the model reconstructs a corrupted input, conditioned on a fixed-width "style vector" extracted from the preceding sentence. 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.

Method - Model architecture

- Follow Lampple et al. (2019)[1] at the high-level.
  - Train denoising auto-encoder conditioned on the style vector.
  - The difference is that true style is unknown at training time in this paper.
  - To address this problem, the style extractor is jointly trained with nearby context sentences.
- Model architecture
  - Reconstruct the input text based on pre-trained T5(Text-to-Text Transfer Transformer) model architecture, and extract the style vectors using a pre-trained T5 transformer encoder.
  - The difference between style extractor and encoder is that style extractor is mean-pooling the hidden states into a fixed-length vector.

[1] Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. Multiple-attribute text rewriting. In International Conference on Learning Representations, 2019.

Method - Corruption Strategies

- To implement reconstruction task, sentence s_i in dataset is corrupted by some function f to produce $\hat{s}_i = f(s_i)$.
- The cross entropy loss is calculated using uncorrupted text s_i as the targets, and corrupted text $\hat{s}_i$ and context text s_{i-1} as model inputs.
- Corruption Strategies.
  - **Noise:** Corrupts the inputs by dropping, replacing, and/or shuffling.
  - **Back Translation(BT):** Corrupts s_i using style-transfer using sampled random context s_j.
  - **Noisy Back Translation:** Noise is first applied and result is used as input of BT.

Method - Inference Procedure

- Tunable Add/Delete Rates.
  - Style-transfer has recurring problem that the model would often **failed to achieve the target style** or **failed to preserve the input content**.
  - To address above problem, “add rates range”(the proportion of output tokens absent from the input) and “delete rates range”(the proportion of input tokens absent from the output) values are passed to decoder.
- Targeted Restyling
  - To transfer input sentence x, a small set of exemplar sentences for both source and target values are provided.
  - Compute target style vector: $v^x + \lambda(v^{target} - v^{source})$

Method

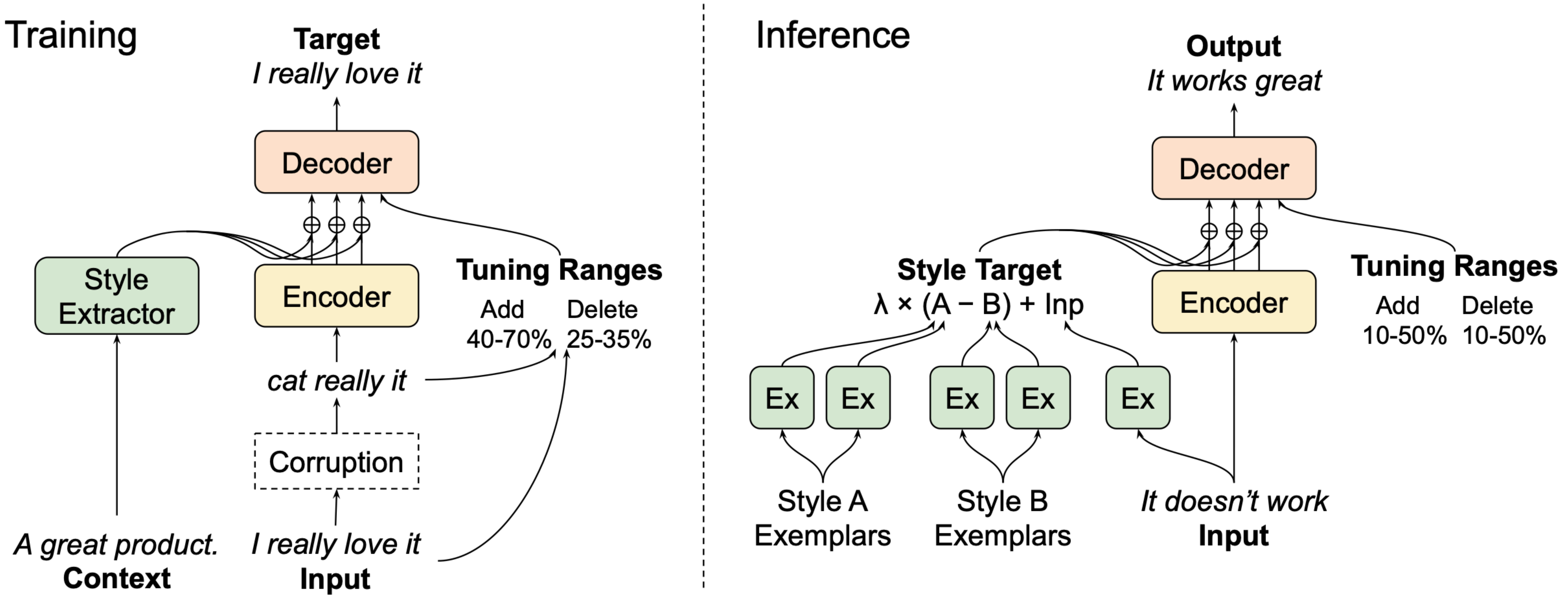

Figure 1: TextSETTR architecture for label-free style transfer. The Encoder, Decoder and Style Extractor (Ex) are transformer stacks initialized from pretrained T5. During training, the model reconstructs a corrupted input, conditioned on a fixed-width "style vector" extracted from the preceding sentence. 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.

Experiments

- **Training Settings**
  - Unlabeled data comes from 233.1M Amazon reviews (Ni et al., 2019[1]), and after preprocessing, 23.6M examples remain.
  - Use pre-trained T5(t5.1.1.large)
- **Evaluation Settings**
  - To estimate transferred sentiment, BERT-Large is fine-tuned on (Li et al., 2018[2]), and scoring 87.8% accuracy on dev split.
  - To estimate content preservation, SacreBLEU is used.
  - To perform transfer, 100 exemplars for each style is sampled. Also experimented using 1000 sampled exemplars and 4 manually chosen exemplars.

[1] Jianmo Ni, Jiacheng Li, and Julian McAuley. 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)
[2] Juncen Li, Robin Jia, He He, and Percy Liang. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In Proceedings of the 2018 Conference of the North American Chap- ter of the Association for Computational Linguistics: Human Language Technologies

Experiments - Results

Model	Acc.	Content
TextSETTR	73.7	34.7
N	23.4	84.4
NBT	70.0	27.8
N + BT	13.3	98.7
−replace noise	66.1	42.1
+shuffle noise	70.3	34.1
manual exemplars	52.4	44.2
1000 exemplars	74.5	37.2
−tunable inference	71.5	39.4
overwrite style	25.3	55.8
small train set	74.5	33.4
CP-G	51.1	35.5
CP-B	36.3	39.8
CrossAligned	68.2	2.9
Delete&Retrieve	49.4	56.9
B-GST	60.2	54.2

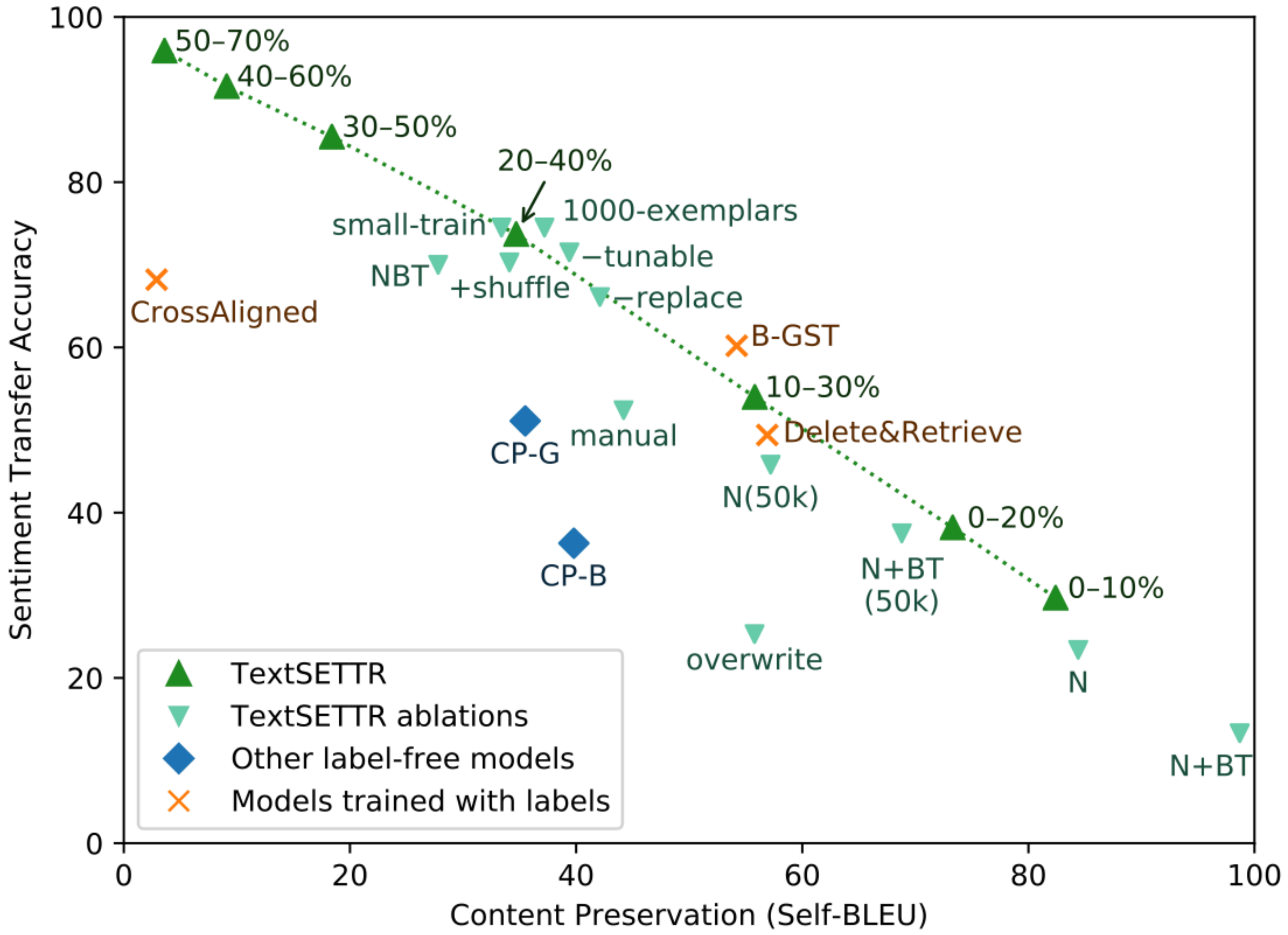

Figure 2: Automatic evaluation metrics comparing our TextSETTR model, ablations, and previous work. Up-and-right is better. We train for 10k steps and use add/delete:20–40% unless otherwise specified. We recalculate metrics for previous approaches, using our BERT classifier for accuracy, ensuring direct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

Experiments - Results

Model	Accuracy	Content
TextSETTR (0–20%)	63.4	76.9
TextSETTR (10–30%)	72.7	60.2
TextSETTR (20–40%)	83.6	39.4
TextSETTR (30–50%)	89.7	21.5
TextSETTR (40–60%)	94.3	11.3
TextSETTR (50–70%)	96.6	5.0
Lample et al. 2019	82.6	54.8

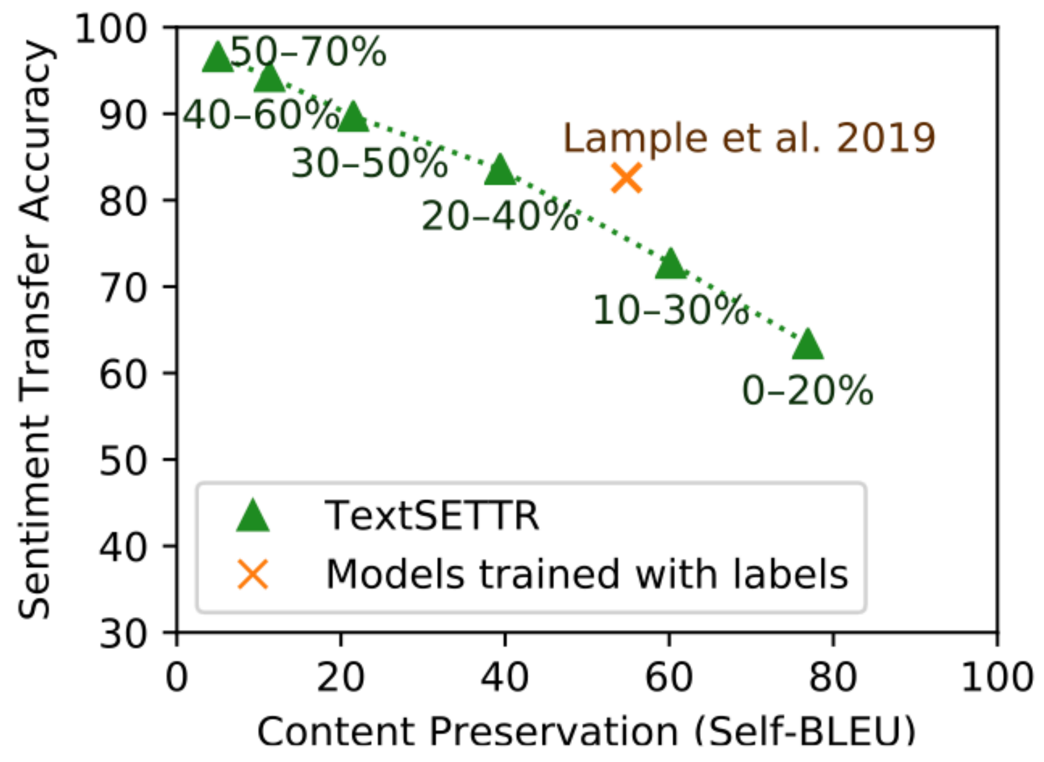

Figure 3: Comparison with Lample et al. (2019) on the evaluation setting that includes pos→pos and neg→neg transfers. Note, a model that simply copies its input can achieve 50% accuracy.

	Negative → Positive			Positive → Negative		
Model	Sentiment	Preservation	Fluency	Sentiment	Preservation	Fluency
TextSETTR	2.8	2.4	4.2	2.3	2.8	3.8
Delete&Retrieve	2.7	2.9	3.2	2.3	3.4	3.4
B-GST	2.3	2.8	3.6	2.1	3.0	3.6

Table 1: Human evaluations on sentiment, content preservation, and fluency.

Experiments - Results - Samples

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

Table 2: 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.

Experiments - Results - Embedding Visualization

Before TextSETTR training (pretrained T5 initialization)

After TextSETTR training

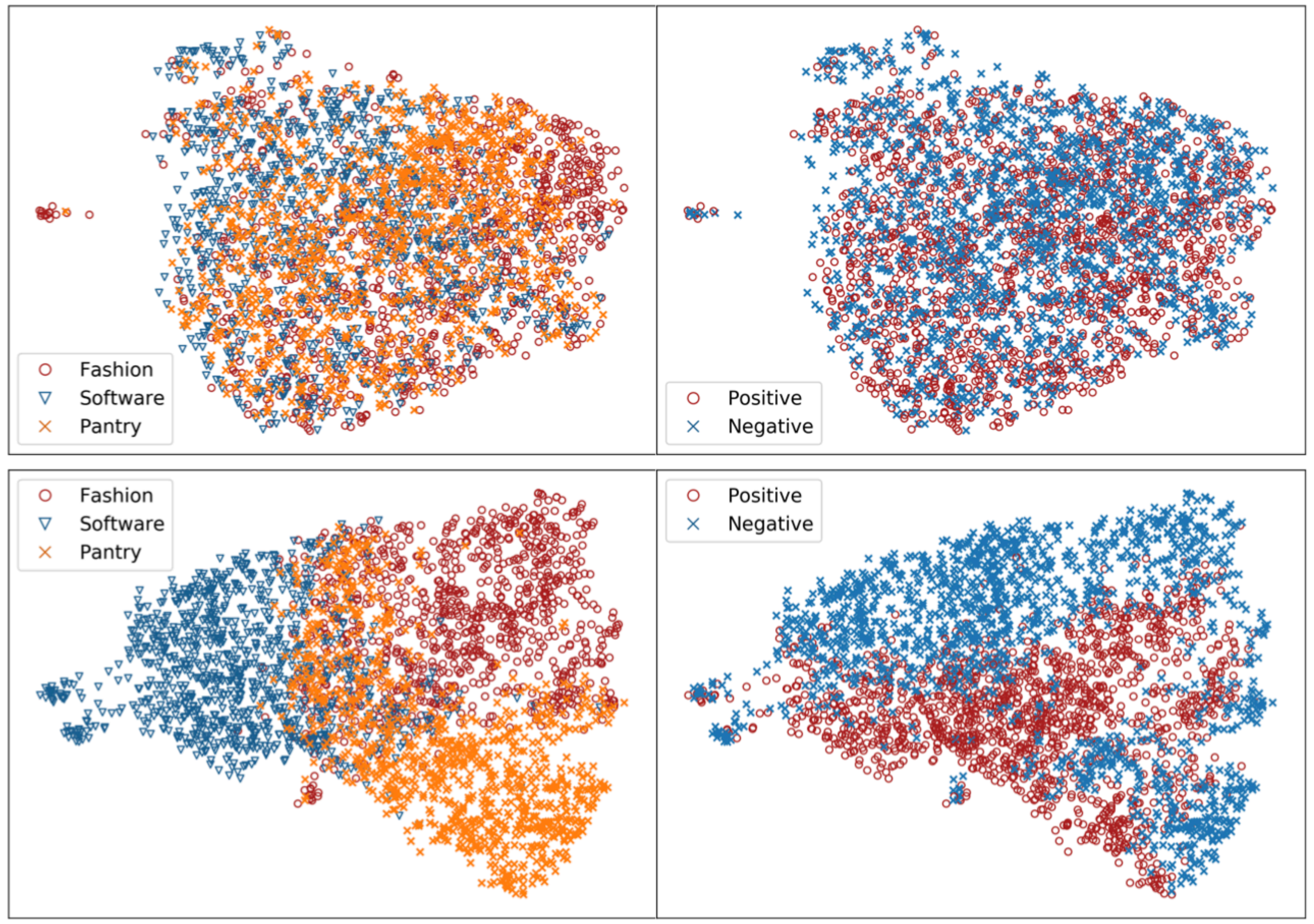

Figure 4: 2D UMAP embeddings of the style vectors extracted by our TextSETTR model before and after training, for text inputs from Amazon reviews covering three product categories and two sentiment labels. Within each row, the same embeddings are visualized with product category labels (left) and sentiment labels (right).

Experiments - Results - Beyond Style Transfer

- **Style-sensitive Completion**
  - If add rate range is set like 40-70%, and delete rate range 0%, the model completes the inputs style-sensitively. (e.g., *My favorite hot drink* -> *Starbucks coffee* (American) or *a mug of tea* (British))
- **Coherent Text Shortening**
  - If add rate range is set like 0-5%, and delete rate range 40-90%, the model performs text shortening.
- **Random augmentation**
  - By using small random delta style vector instead of targeted style vector, model can augment input sentences. (e.g., *What'll the weather be tomorrow?* -> *What's the weather forecast for tomorrow?* or *What's the weather like for the next day?*)

Appendix - Beyond Style Transfer

American ⇒ British	**British ⇒ 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 3: Examples of dialect-sensitive completion (λ=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 2, as listed in Table 7.

Appendix - Beyond Style Transfer

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 4: 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.

Appendix - Beyond Style Transfer

Input Sentence: “What’ll the weather be tomorrow?”

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