text style transfer

@altsoph

me

what is text style transfer?

why transfer text style?

text style transfer

Seems to be related to:

- image style transfer
- NMT
- paraphrase generation
- summarization

style definition

- sentiments / dialects / author’s style / …
- style is non-orthogonal to content
- no good definitions
- typically defined by explicit examples

[arXiv:1808.04365]

style definition by examples

- **non parallel data**
  - YELP [https://www.yelp.com/dataset]
  - politeness,
  - emojis,
  - ...
- **parallel data**
  - Bibles [arXiv:1711.04731]
  - Shakespeare [https://github.com/cocoxu/Shakespeare]
  - GYAFC [arXiv:1803.06535]
  - YELP-aug [arXiv:1810.06526]

no style for token

- latent variable classification
- gumbel trick
- reinforcement learning
- non-autoregressive generation
- ...

goals and metrics

[arXiv:1904.02295, arXiv:1908.06809]

- style accuracy
  - classifiers
  - human eval
- fluency
  - LM PPL
  - human eval
- content preservation
  - syntax similarity (BLEU-mods, ROUGE, METEOR, …)
  - embedding based (w2v, FT, ELMo, BERT_score…)
  - human eval
  - learnable (VERTa, SimiLe, BLEURT, …)

style matching

- **cross-entropy**

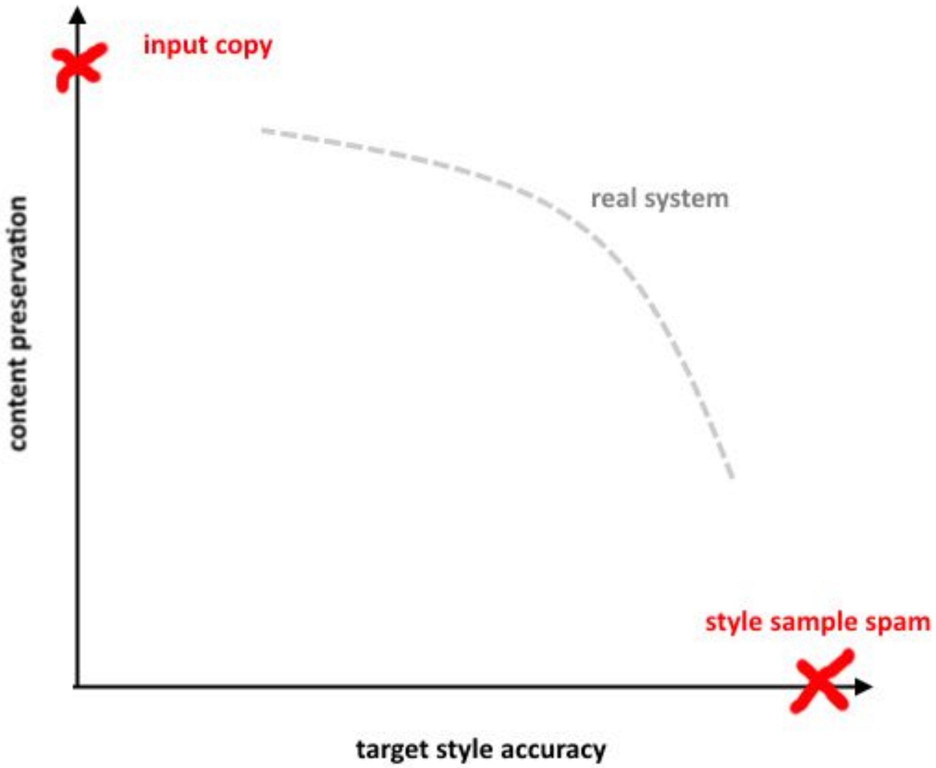
Model $G(A_i)$/ author	Shakespeare	Poe	Carroll	Wilde	Marley	Nirvana	MUSE
Generated-Shakespeare	**19.0****	21.6	**18.5***	19.9	21.8	22.0	22.4
Generated-Poe	22.0	**20.4****	21.2	**19.0***	26.0	25.4	26.0
Generated-Carroll	22.2	23.6	**18.9***	22.5	22.4	**21.8****	23.8
Generated-Wilde	21.2	20.9	**20.5****	**18.4***	24.5	24.8	26.4
Generated-Marley	24.1	26.5	22.0	27.0	**15.5***	**15.7****	16.0
Generated-Nirvana	23.7	26.2	20.0	26.6	19.3	**18.3***	**19.1****
Generated-MUSE	21.1	23.9	18.5	23.4	17.4	**16.0****	**14.6***
Uniform Random	103.1	103.0	103.0	103.0	103.5	103.3	103.6
Weighted Random	68.6	68.8	67.4	68.5	68.5	68.0	68.0
SELF	23.4	21.8	25.1	27.3	20.8	17.8	13.3

Table 3. Sample cross entropy between generated texts $\{T_i^G|A_i\}$ and actual texts for different authors.

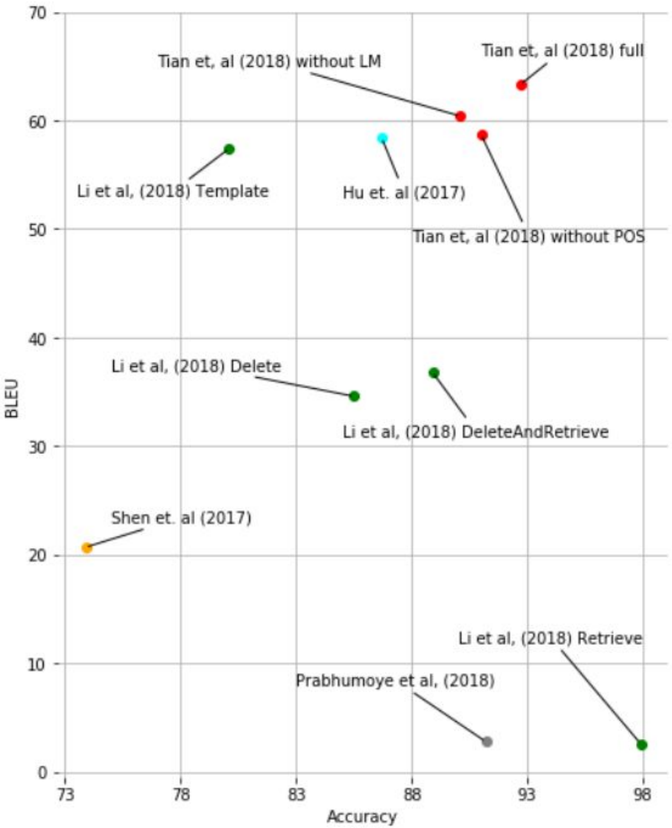
- **classification**

truth \ pred	Brodskiy	Pushkin	Esenin	Pasternak	Tsvetaeva	Mayakovskiy	Akhmatova	Tyutchev	Mandelshtam	Lermontov
Brodskiy	77,2%	1,7%	2,3%	4,3%	2,3%	1,5%	4,0%	1,3%	3,6%	1,7%
Pushkin	1,1%	77,0%	8,0%	0,3%	0,0%	0,3%	1,9%	3,3%	0,6%	7,5%
Esenin	3,9%	4,9%	73,8%	3,0%	1,3%	1,6%	5,9%	0,7%	1,6%	3,3%
Pasternak	16,3%	2,6%	10,7%	54,9%	2,1%	1,7%	3,9%	1,3%	6,0%	0,4%
Tsvetaeva	9,1%	2,8%	5,1%	4,0%	51,1%	1,7%	18,2%	1,1%	5,7%	1,1%
Mayakovskiy	8,2%	2,9%	11,7%	5,8%	3,5%	59,1%	0,6%	1,2%	7,0%	0,0%
Akhmatova	4,5%	4,5%	17,0%	3,4%	3,4%	0,0%	59,7%	1,1%	1,7%	4,5%
Tyutchev	3,0%	14,1%	3,7%	3,0%	0,7%	0,7%	5,9%	55,6%	2,2%	11,1%
Mandelshtam	9,2%	6,6%	9,2%	11,8%	1,3%	5,3%	15,8%	1,3%	35,5%	3,9%
Lermontov	2,6%	15,8%	9,2%	0,0%	2,6%	0,0%	9,2%	9,2%	2,6%	48,7%

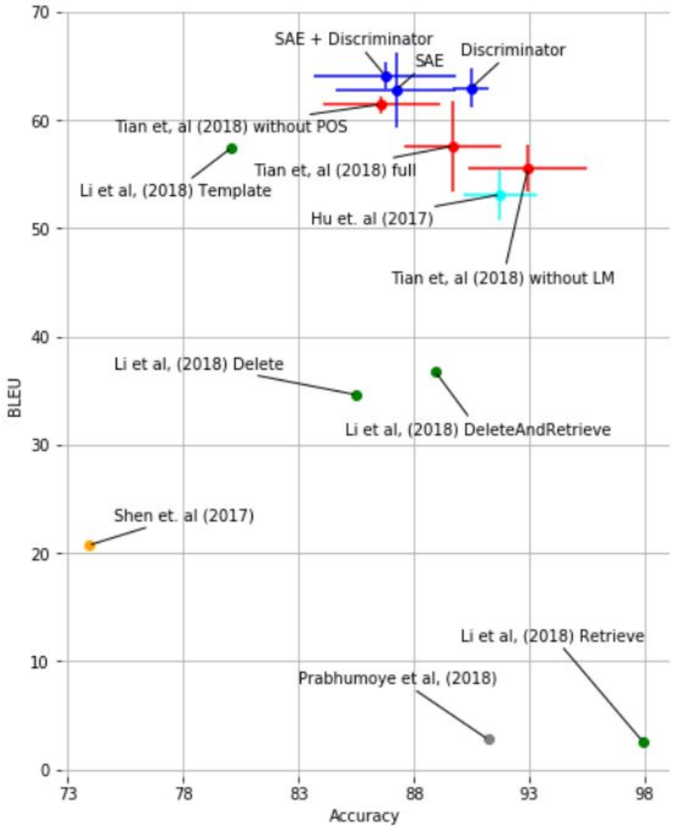
goals trade-off



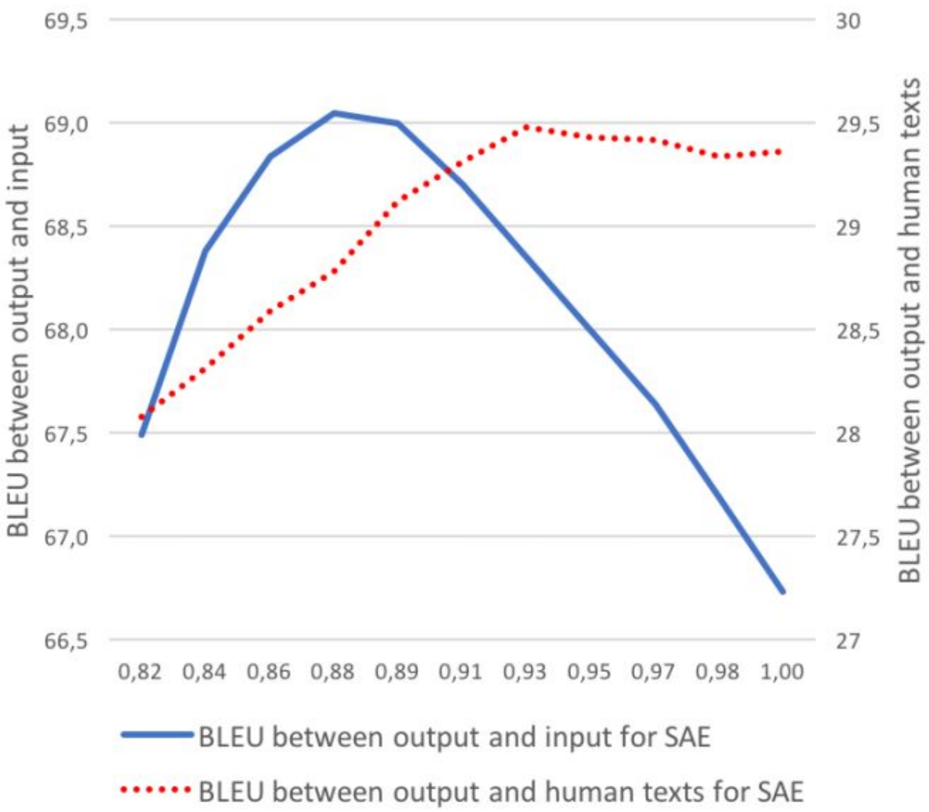
unfair reporting



unstable balance



self-BLEU is evil



content preservation

Premise:

[arXiv:2004.05001]

dist(random pair) > dist(style transfer pair) > dist(paraphrase)

Metric	Bibles random	Paralex random	Paraphrase random	Yelp! random rewrite	GYAFC random rewrite	GYAFC random informal	GYAFC random formal	Yelp! rewrite	GYAFC rewrite	GYAFC informal	GYAFC formal	Bibles	Paralex	Paraphrase
POS	14	10	8	9	11	12	13	1	4	7	2	5	6	3
Word overlap	10	9	14	11	12	13	8	4	3	6	1	2	5	7
chrF	9	10	14	11	12	13	8	4	2	7	3	1	5	6
Word2Vec	8	12	14	11	7	10	9	4	2	5	3	1	6	13
FastText	7	12	14	11	9	10	8	4	3	6	2	1	5	13
WMD	8	13	14	11	10	9	12	4	1	6	3	2	5	7
ELMo L2	8	13	14	12	11	10	9	4	3	5	2	1	6	7
ROUGE-1	10	9	14	11	13	12	8	5	3	6	1	2	4	7
ROUGE-2	10	9	14	13	12	8	11	4	2	6	1	3	5	7
ROUGE-L	9	10	14	11	13	12	8	4	3	7	2	1	5	6
BLEU	10	11	14	12	13	8	9	4	3	5	2	1	6	7
Meteor	10	9	14	11	12	13	8	4	3	7	2	1	5	6
BERT score	10	9	14	8	12	13	11	3	4	7	1	2	5	6
Human Labeling	9	14	13	8	12	10	11	7	1	5	2	4	6	3

Table 3: Different semantic similarity metrics sort the paraphrase datasets differently. Cosine similarity calculated with Word2Vec or FastText embeddings do not comply with Inequality $M(D_r) < M(D_p)$. All other metrics clearly distinguish randomized texts from style transfers and paraphrases and are in line with Inequalities 1. However, none of the metrics is completely in line with human evaluation.

content preservation

Premise: [arXiv:2004.05001]

dist(random pair) > dist(style transfer pair) > dist(paraphrase)

	POS-distance	Word overlap	chrF	Word2Vec	FastText	WMD	ELMO L2	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	Meteor	BERT score	Human score
POS-distance	1,00	0,73	0,71	0,45	0,44	0,69	0,66	0,71	0,72	0,71	0,68	0,74	0,82	0,72
Word overlap	0,73	1,00	0,98	0,80	0,84	0,86	0,92	0,99	0,91	0,98	0,92	0,99	0,95	0,80
chrF	0,71	0,98	1,00	0,79	0,83	0,89	0,93	0,97	0,89	0,99	0,92	0,99	0,93	0,83
Word2Vec	0,45	0,80	0,79	1,00	0,98	0,87	0,88	0,78	0,79	0,78	0,82	0,77	0,73	0,64
FastText	0,44	0,84	0,83	0,98	1,00	0,86	0,90	0,83	0,81	0,83	0,85	0,81	0,76	0,65
WMD	0,69	0,86	0,89	0,87	0,86	1,00	0,96	0,86	0,92	0,89	0,92	0,86	0,85	0,89
ELMO L2	0,66	0,92	0,93	0,88	0,90	0,96	1,00	0,92	0,92	0,94	0,96	0,92	0,87	0,86
ROUGE-1	0,71	0,99	0,97	0,78	0,83	0,86	0,92	1,00	0,93	0,98	0,93	0,98	0,94	0,82
ROUGE-2	0,72	0,91	0,89	0,79	0,81	0,92	0,92	0,93	1,00	0,91	0,96	0,90	0,87	0,81
ROUGE-L	0,71	0,98	0,99	0,78	0,83	0,89	0,94	0,98	0,91	1,00	0,94	0,99	0,94	0,83
BLEU	0,68	0,92	0,92	0,82	0,85	0,92	0,96	0,93	0,96	0,94	1,00	0,92	0,87	0,84
Meteor	0,74	0,99	0,99	0,77	0,81	0,86	0,92	0,98	0,90	0,99	0,92	1,00	0,95	0,80
BERT score	0,82	0,95	0,93	0,73	0,76	0,85	0,87	0,94	0,87	0,94	0,87	0,95	1,00	0,82
Human score	0,72	0,80	0,83	0,64	0,65	0,89	0,86	0,82	0,81	0,83	0,84	0,80	0,82	1,00

Figure 1: Pairwise correlations of the orders induced by the metrics of semantic similarity.

approaches

Simple:

- NMT-like on parallel corpora [arXiv:1707.01161, ...]
- template / lexical based … [arXiv:2005.12086, ...]

More interesting:

- Z-space search [arXiv:1905.12926, 1905.12304, 2004.04092]
- disentangled representations [...]
- UNMT-like [arXiv:1711.00043, arXiv:2002.03912]
- TextSETTR [...]

NMT-like

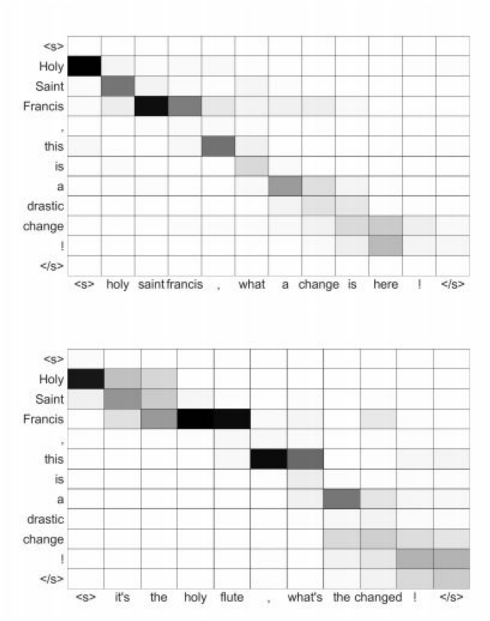


Figure 2: Attention matrices from a *Copy* (top) and a *simple S2S* (bottom) model respectively on the input sentence *"Holy Saint Francis, this is a drastic change!"* . $< s >$ and $< /s >$ are start and stop characters. Darker cells are higher-valued.

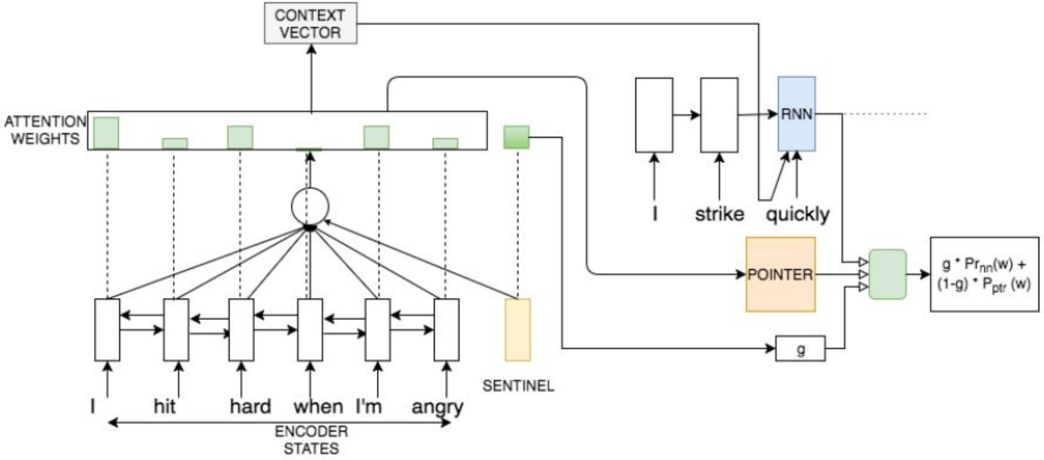


Figure 1: Depiction of our overall architecture (showing decoder step 3). Attention weights are computed using previous decoder hidden state h_2, encoder representations, and sentinel vector. Attention weights are shared by decoder RNN and pointer models. The final probability distribution over vocabulary comes from both the decoder RNN and the pointer network. Similar formulation is used over all decoder steps

[arXiv:1707.01161]

token / lexical / template replacement

```
A quick brown [ fox ] runs over lazy dog
               eye                     0.185885
               ##ie                    0.175180
               cat                     0.035072
               bear                    0.032281
               streak                  0.023462
               fox                     0.017081
               coat                    0.015879
```

```
is slow but there was great [ attention ] to detail .
                              attention          0.9986
                              regard             0.0002
                              time               0.0001
                              effort             0.0001
                              access             0.0001
                              care               0.0001
                              eye                0.0001
                              loss               0.0000
                              work               0.0000
```

token / lexical / template replacement

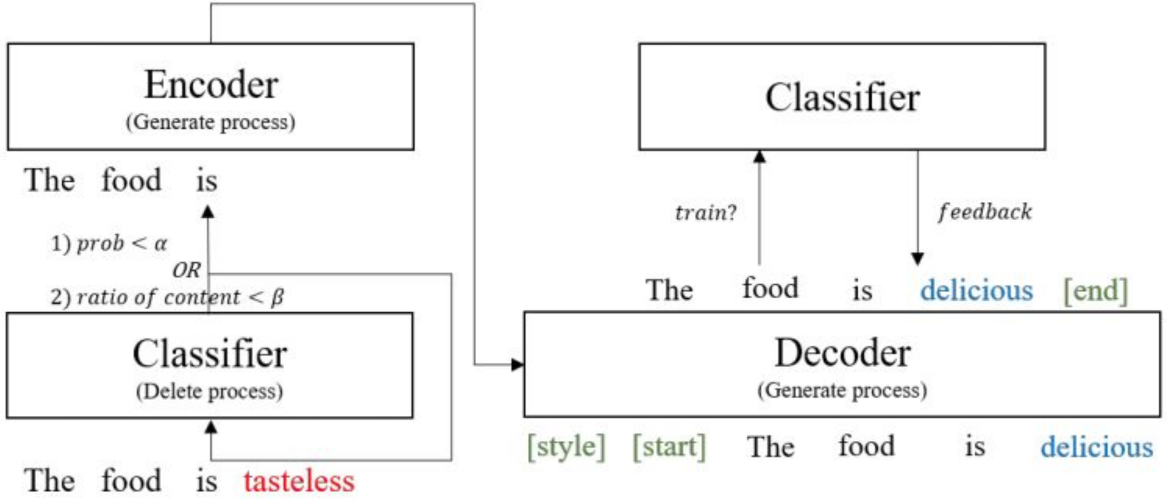


Figure 1: The proposed model framework consists of Delete and Generate process. Delete process is a method using a pre-trained classifier, and the Generate process consists of an encoder and a decoder. In the training time, our model receives feedback from the classifier's probability of the generated sentence.

[arXiv:2005.12086]

token / lexical / template replacement

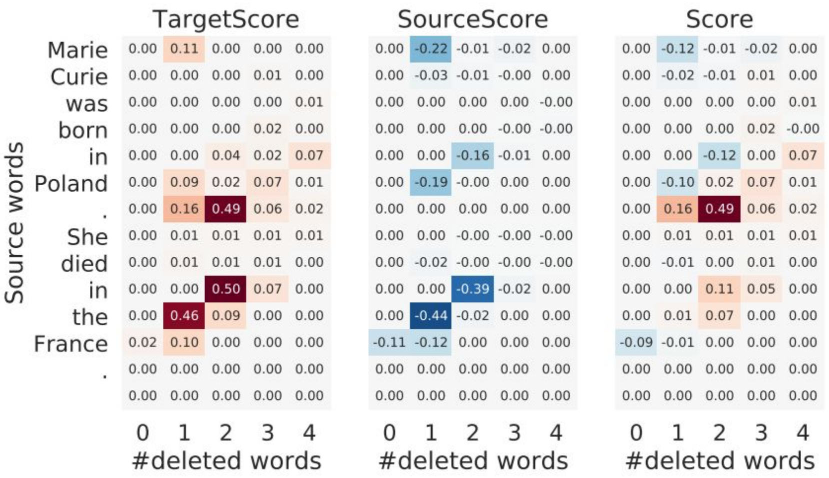


Figure 1: MASKER replaces span ". *She*" by "*and* `[PAD] [PAD] [PAD]`", resulting in the following fused sentence: *Marie Curie was born in Poland and died in the France .*

Random Sample of Correct MASKER Predictions	
Source	so far i 'm not really impressed .
Prediction	so far i 'm really impressed .
Source	either way i would never recommend buying from camping world .
Prediction	either way i would recommend buying from camping world .
Source	this is a horrible venue .
Prediction	this is a great venue .
Source	this place is a terrible place to live !
Prediction	this place is a great place to live !
Source	i 'm not one of the corn people .
Prediction	i 'm one of the corn people .
Source	this is easily the worst greek food i 've had in my life .
Prediction	this is easily the best greek food i 've had in my life .
Source	the sandwich was not that great .
Prediction	the sandwich was great .
Source	its also not a very clean park .
Prediction	its also a very clean park .

[arXiv:2010.01054]

Z-space search

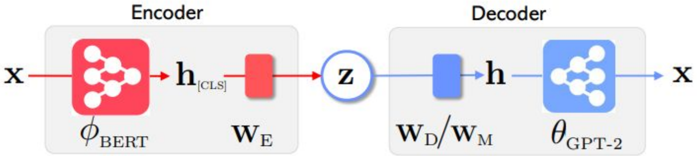


Figure 1: Illustration of OPTIMUS architecture.

0.0	children are looking for the water to be clear.
0.1	children are looking for the water.
0.2	children are looking at the water.
0.3	the children are looking at a large group of people.
0.4	the children are watching a group of people.
0.5	the people are watching a group of ducks.
0.6	the people are playing soccer in the field.
0.7	there are people playing a sport.
0.8	there are people playing a soccer game.
0.9	there are two people playing soccer.
1.0	there are two people playing soccer.

Table 3: Interpolating latent space. Each row shows τ, and the generated sentence (in blue) conditioned on z_τ.

$$\boldsymbol{x}_D \approx \boldsymbol{x}_B - \boldsymbol{x}_A + \boldsymbol{x}_C$$

Source $\boldsymbol{x}_A$	**Target** $\boldsymbol{x}_B$
a girl makes a silly face	two soccer players are playing soccer
Input $\boldsymbol{x}_C$	**Output** $\boldsymbol{x}_D$
• a girl poses for a picture	• two soccer players are at a soccer game.
• a girl in a blue shirt is taking pictures of a microscope	• two football players in blue uniforms are at a field hockey game
• a woman with a red scarf looks at the stars	• two men in white uniforms are field hockey players
• a boy is taking a bath	• two baseball players are at the baseball diamond
• a little boy is eating a bowl of soup	• two men are in baseball practice

Table 2: Sentence transfer via arithmetic operation in the latent space. The output sentences are in blue.

Z-space search

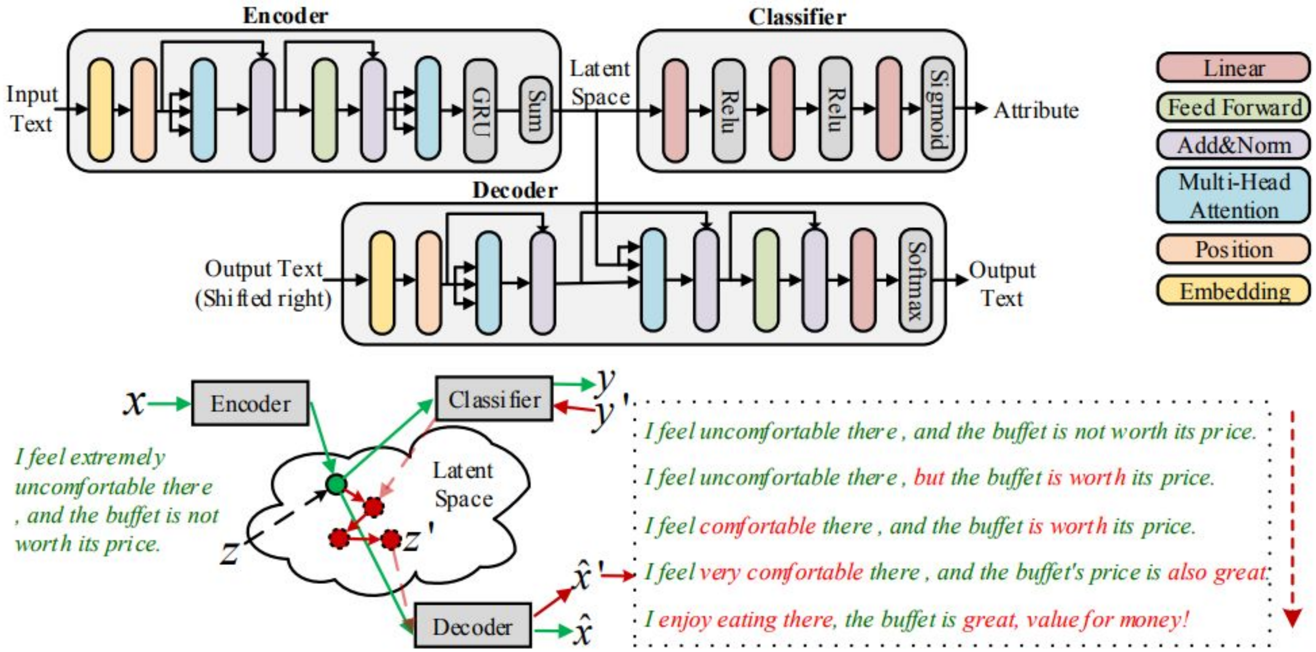


Figure 1: Model architecture.

Z-space search

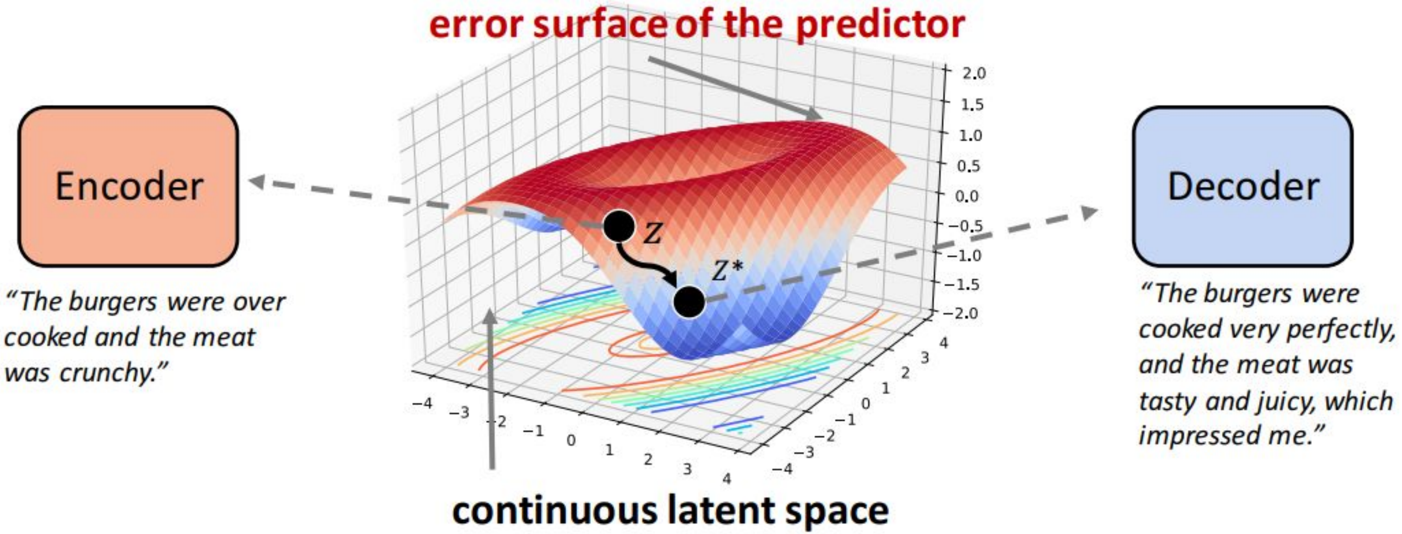


Figure 1: There is an example of content-preserving text sentiment transfer, and we hope to further increase the length of the target sentence compared with the original sentence. The original sentence x with negative sentiment is mapped to continuous representation z via encoder. Then z is revised into z^* by minimizing the error $\mathcal{L}_{\text{Attr},s_1}(\theta_{s_1}; s_1 = \{\text{sentiment} = \textit{positive}\}) + \mathcal{L}_{\text{Attr},s_2}(\theta_{s_2}; s_2 = \{\text{length} = \textit{20}\}) + \lambda_{\text{bow}}\mathcal{L}_{\text{BOW}}(\theta_{\text{bow}}; x_{\text{bow}} = [\textit{burgers}, \textit{meat}])$ with the sentiment predictor f_1, length predictor f_2, and the content predictor f_{bow}. Afterwards the target sentence x^* is generated by decoding z^* with beam search via decoder [best viewed in color].

disentangled representations

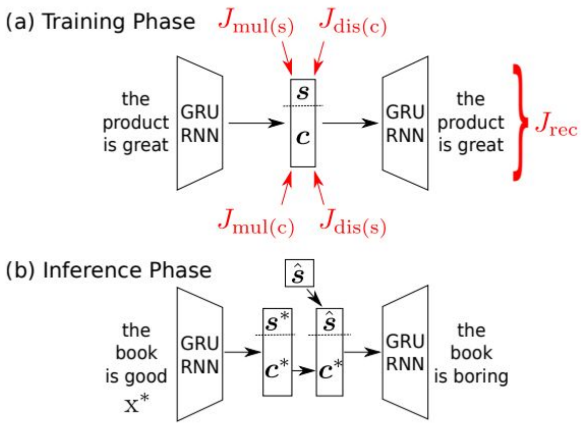


Figure 1: Overview of our approach.

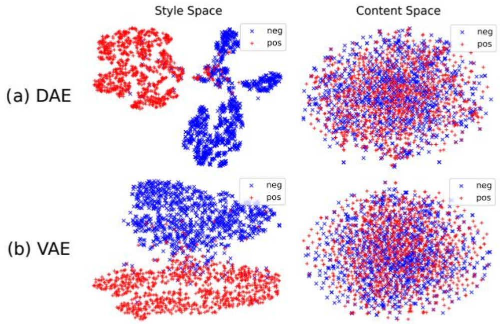
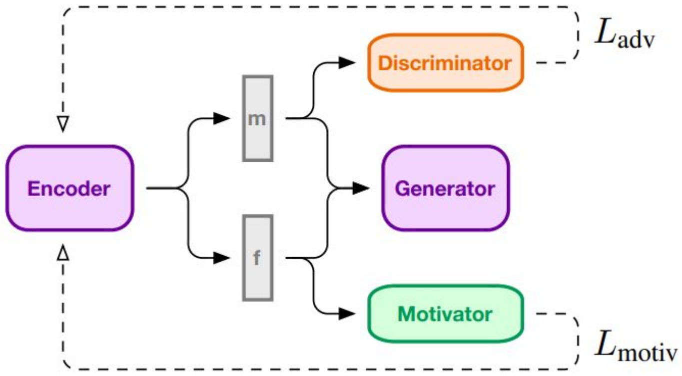


Figure 2: t-SNE plots of the disentangled style and content spaces (with all auxiliary losses on the Yelp dataset).



[arXiv:1808.04339]

[arXiv:1808.09042]

disentangled representations

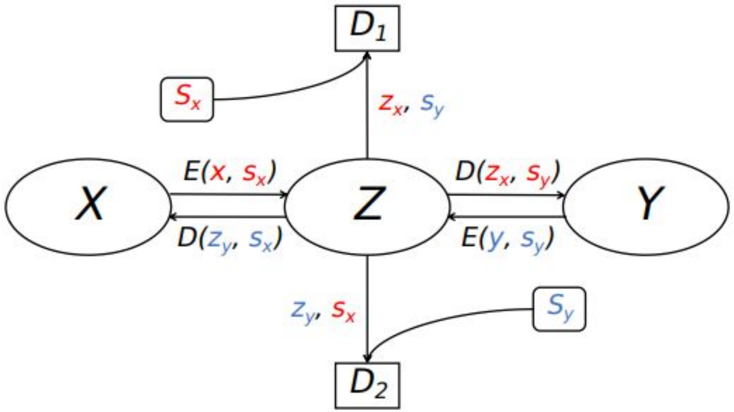


Figure 1: CrossAlign architecture

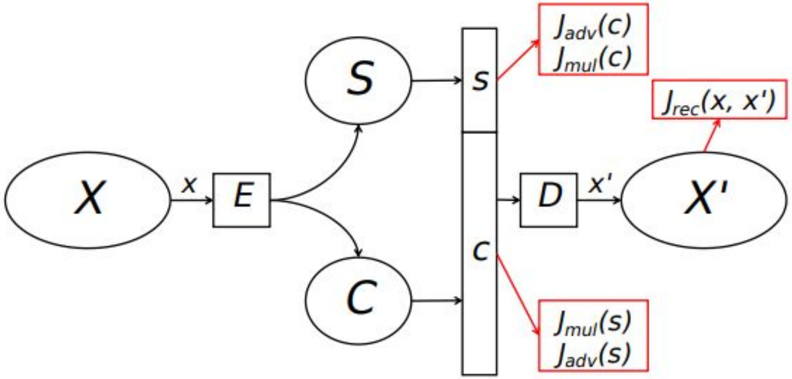


Figure 2: VAE architecture

[arXiv:2004.11742]

disentangled representations

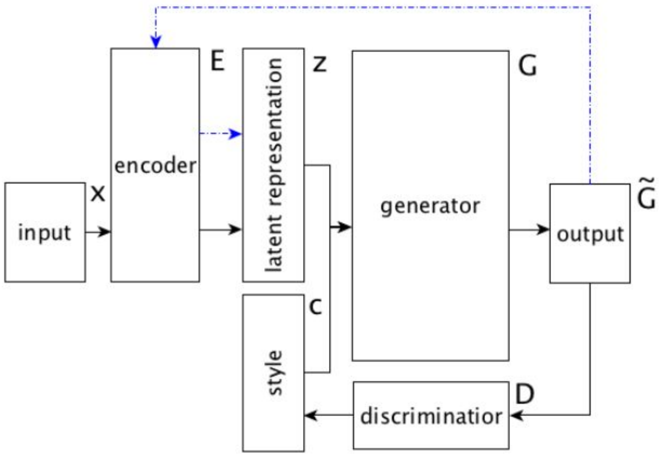
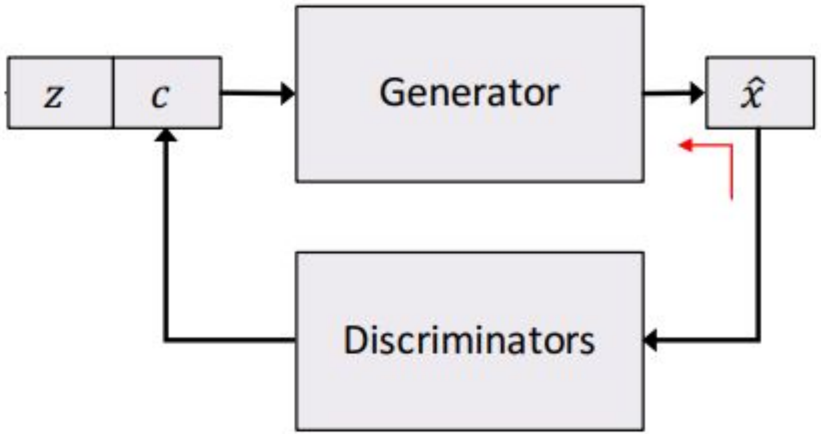


Figure 3: The generative model, where style is a structured code targeting sentence attributes to control. Blue dashed arrows denote the proposed independence constraint of latent representation and controlled attribute, see (Hu et al., 2017a) for the details.



[arXiv:1703.00955, arXiv:1809.00794]

disentangled representations

Are Adversarial Models really Doing Disentanglement?

λ_{adv}	Discriminator Acc (Train)	Post-fit Classifier Acc (Test)
0	89.45%	93.8%
0.001	85.04%	92.6%
0.01	75.47%	91.3%
0.03	61.16%	93.5%
0.1	57.63%	94.5%
1.0	52.75%	86.1%
10	51.89%	85.2%
fastText	-	97.7%

[arXiv:1811.00552]

disentangled representations

$$\mathcal{L}_{cos}(x,c) = \cos\left(E(\tilde{G}(E(x),c)), E(x)\right),$$
$$\mathcal{L}_{cos^-}(x,c) = \cos\left(E(\tilde{G}(E(x),\bar{c})), E(x)\right). \quad (8)$$

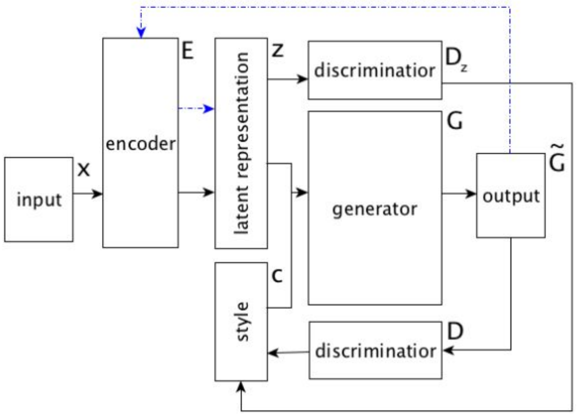


Figure 4: The generative model with dedicated discriminator introduced to ensure that semantic part of the latent representation does not have information on the style of the text.

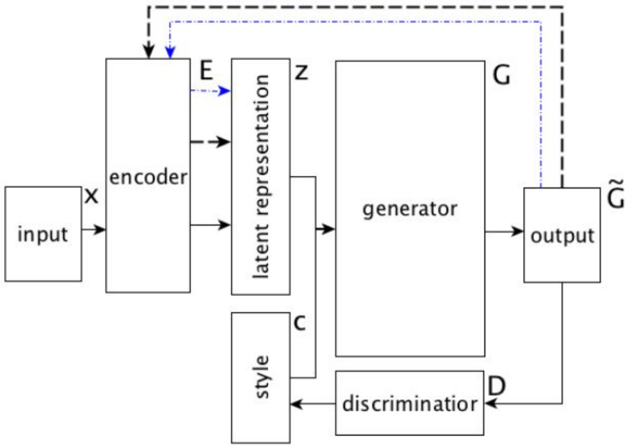
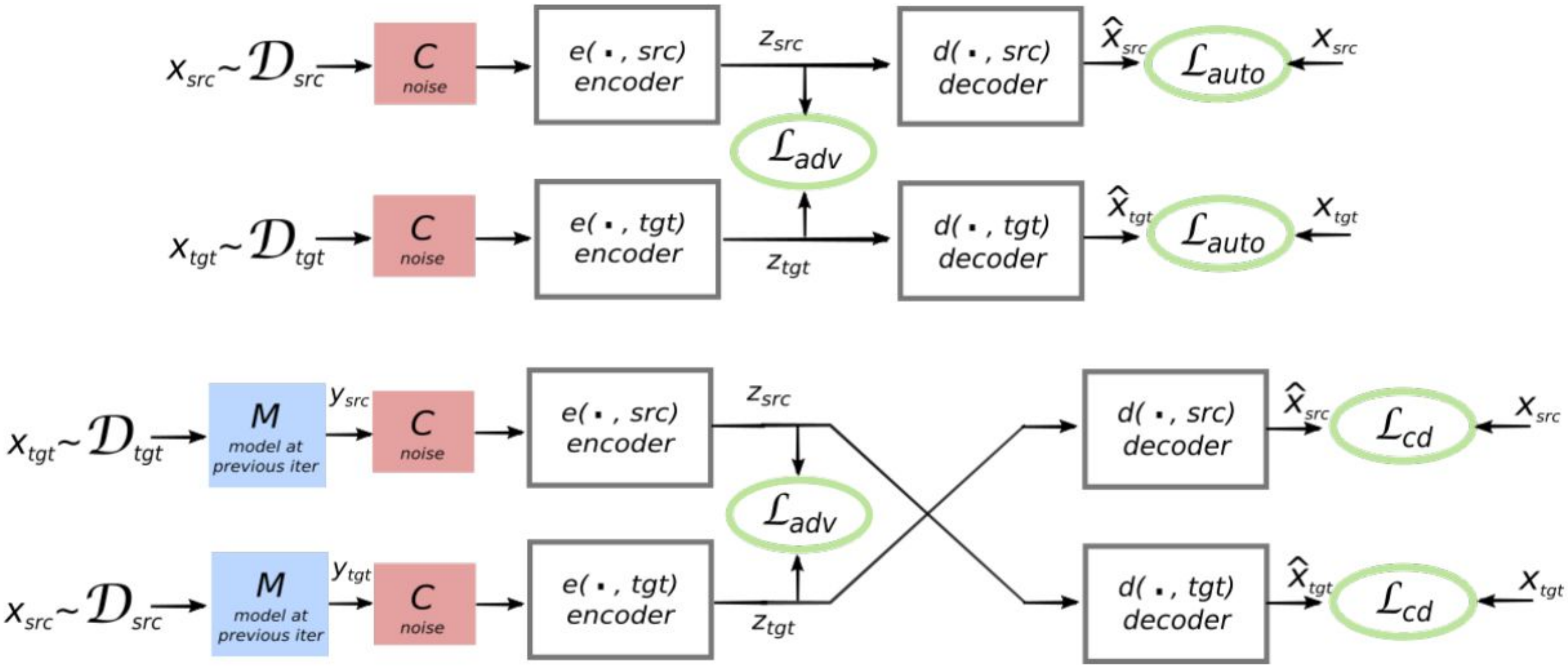


Figure 5: The generative model with a dedicated loss added to control that semantic representation of the output, when processed by the encoder, is close to the semantic representation of the input.

[arXiv:1908.06809]

UNMT-like



[arXiv:1711.00043]

UNMT-like

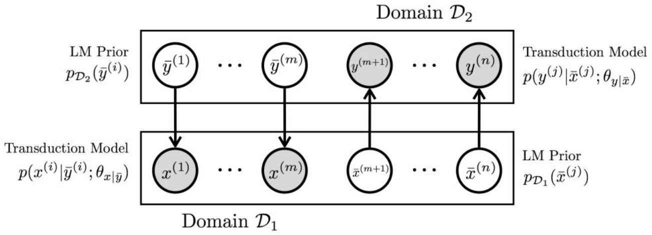


Figure 1: Proposed graphical model for style transfer via bitext completion. Shaded circles denote the observed variables and unshaded circles denote the latents. The generator is parameterized as an encoder-decoder architecture and the prior on the latent variable is a pretrained language model.

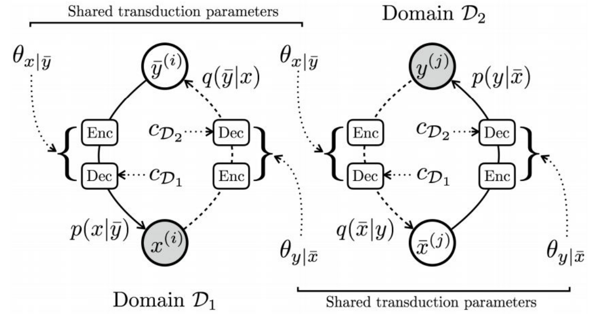


Figure 2: Depiction of amortized variational approximation. Distributions $q(\bar{y}|x)$ and $q(\bar{x}|y)$ represent inference networks that approximate the model's true posterior. Critically, parameters are shared between the generative model and inference networks to tie the learning problems for both domains.

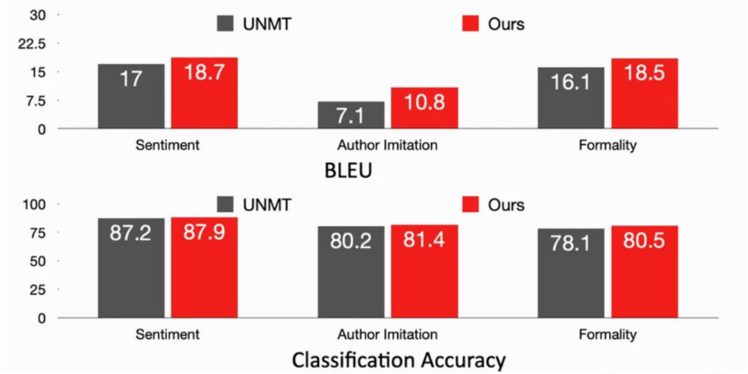


Table 3: Examples for author imitation task

Methods	Shakespeare to Modern
Source	Not to his father's .
Reference	Not to his father's house .
UNMT	Not to his brother .
Ours	Not to his father's house .
Source	Send thy man away .
Reference	Send your man away .
UNMT	Send an excellent word .
Ours	Send your man away .
Source	Why should you fall into so deep an O ?
Reference	Why should you fall into so deep a moan ?
UNMT	Why should you carry so nicely , but have your legs ?
Ours	Why should you fall into so deep a sin ?

[arXiv:2002.03912]

unsupervised style learning

Under review as a conference paper at ICLR 2021

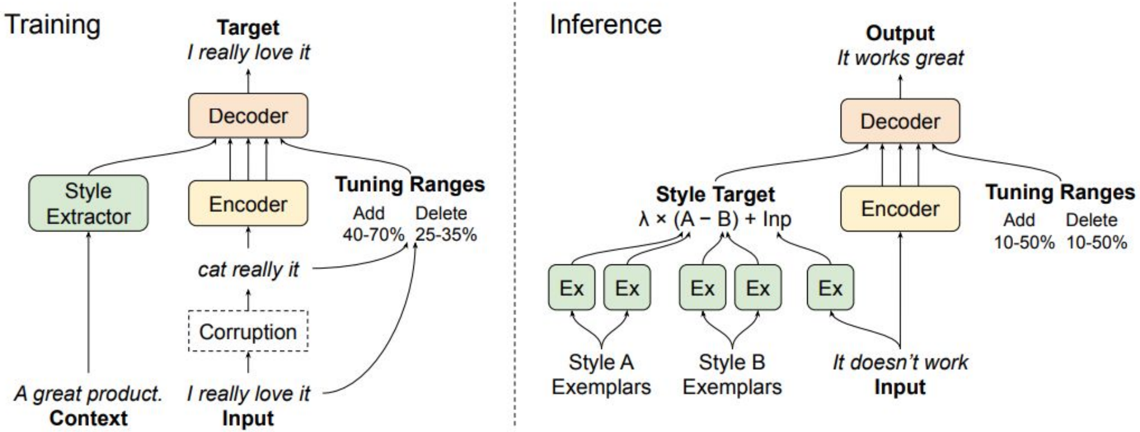


Figure 1: TextSETTR architecture for label-free style transfer. The Encoder, Decoder and Style Extractor (Ex) are transformer stacks initialized from pretrained T5. During training, the model reconstructs a corrupted input, conditioned on a fixed-width "style vector" extracted from the preceding sentence.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

unsupervised style learning

Model	Acc.	Content
TextSETTR	73.3	34.7
N	23.4	84.4
N + BT	13.3	98.7
−replace noise	66.1	42.1
+shuffle noise	70.3	34.1
manual exemplars	52.4	44.2
−tunable inference	71.5	39.4
CP-G	60.1	35.4
CP-B	40.0	39.7
CrossAligned	83.1	15.2
Delete&Retrieve	50.9	16.1
B-GST	60.0	73.6

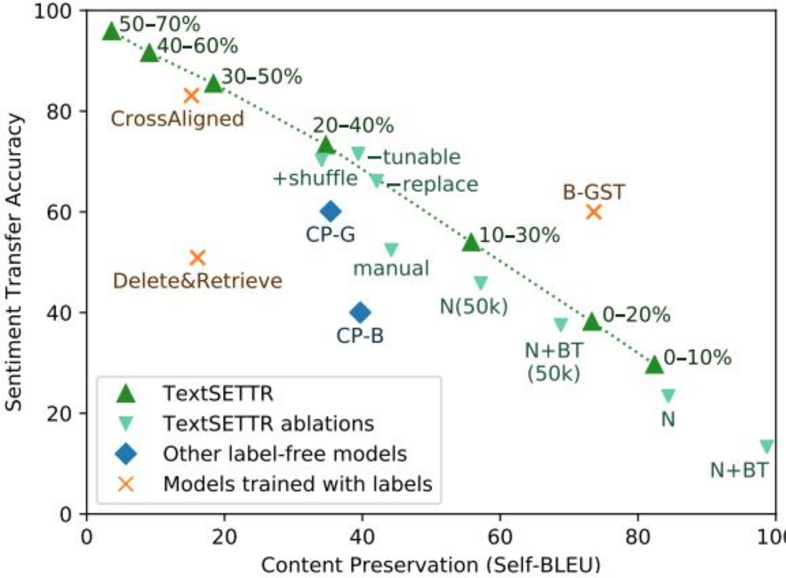


Figure 2: Automatic evaluation metrics comparing our TextSETTR model, ablations, and previous work. Up-and-right is better. We train for 10k steps and use add/delete:20–40% unless otherwise specified. Scores for CrossAligned, Delete&Retrieve and B-GST are from Sudhakar et al. (2019).

unsupervised style learning

Model	Accuracy	Content
TextSETTR	83.6	39.4
add/del: 0–20%	63.4	76.9
add/del: 10–30%	72.7	60.2
add/del: 30–50%	89.7	21.5
Lample et al. 2019	82.6	54.8

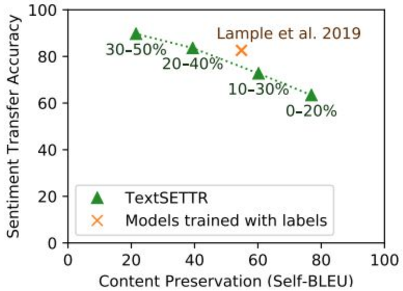


Figure 3: Comparison with Lample et al. (2019) on the evaluation setting that includes pos→pos and neg→neg transfers.

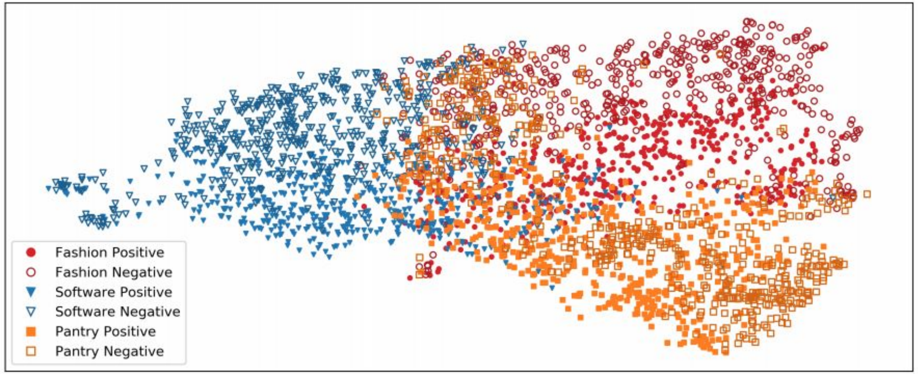


Figure 4: 2D UMAP embedding of the style vectors extracted by our TextSETTR model for text inputs from Amazon reviews covering three product categories and two sentiment labels.

unsupervised style learning

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.

takeaways

- style transfer is ill-defined problem
- no good content preservation metrics yet
- remember content/style trade-off
- know your error margins
- sometimes it works :)

thanks for attention!