On Text Style Transfer via Style-Aware Masked Language Models

Sharan Narasimhan, Pooja Shekar, Suvodip Dey, Maunendra Sankar Desarkar

Indian Institute of Technology Hyderabad, India

Text Style Transfer

⇒ Text Style Transfer (TST) involves transforming a source sentence with a given style label to an output with another target style meanwhile **preserving content and fluency**.
⇒ Sentiment, Formality, Author Style datasets etc. are usually partial i.e. **the target style domain is missing and we cannot treat it as a supervised task.**
⇒ SOTA models are **generative**, highly engineered and fine-tuned for simple styles such as **sentiment and formality**. They employ cycle consistency losses, latent space disentanglement, prompt engineering, and reinforcement learning.

Contributions

⇒ We introduce a **decoder-less, non-generative approach where a single self-attention encoder block trained to reconstruct style-masked sentences** can perform as well as highly engineered heavy-weight SOTA models.
⇒ We introduce an **accurate and efficient O(1) method** for style masking sentences rather than traditional sorting-based methods.
⇒ We show that this simple approach trained on this style reconstruction task even works on **tougher style transfer tasks like Contradiction to Entailment and vice versa.**

Direction	Entailment to Contradiction	Contradiction to Entailment
Input	a guy in a red jacket is snowboarding in midair. a guy is outside in the snow	a woman is sitting outside at a table using a knife to cut into a sandwich. a woman is sitting inside
Style Masked	a guy in a red jacket is snowboarding in midair. a guy is <mask> in the <mask>	a woman is sitting outside at a table using a knife to cut into a sandwich. a woman <mask> <mask> <mask>
Output	a guy in a red jacket is snowboarding in midair. a guy is swimming in the park	a woman is sitting outside at a table using a knife to cut into a sandwich. a woman is there outside

Table 1. Examples of Sentiment and Discourse style transfer by the SA-MLM on the IMDb and SNLI datasets respectively.

At a glance

A reconstruction task: Assuming we could accurately style mask every instance in the dataset, we can now train a model to perform a simple reconstruction task using this partial dataset.
How do you control style?: We append a meta token to the style masked input to inform the model what style to reconstruct the masked version into. This way we can force the model to estimate the opposite style of a given source sentence not seen during training.

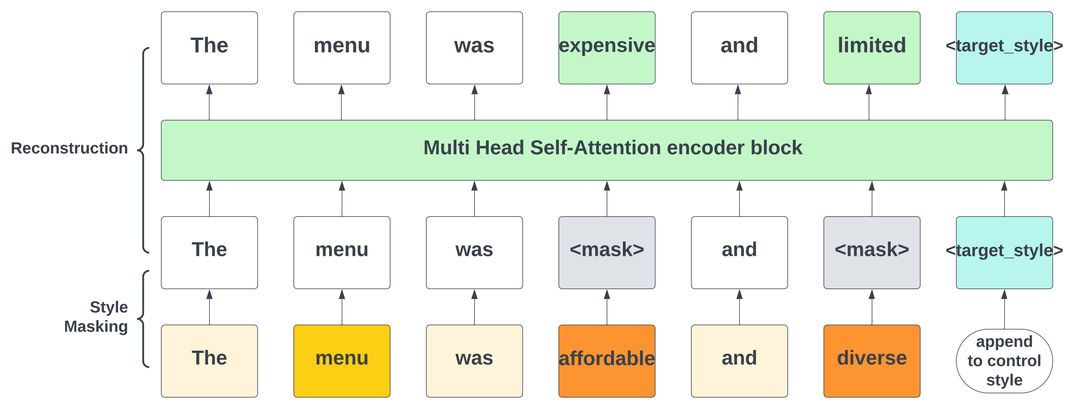

Figure 1. Probabilistic overview of our style transfer method.

Style Masking Tokens via "Attention Surplus"

Intuition: Assume we have accurate attribution scores from a style classifier. We can reason that tokens with **'greater than average'** score contribute to style. We can apply this batchwise in O(1) time.

$$A^{baseline} = (1 + \lambda_\epsilon) * A^{mean} \tag{1}$$

$$\text{Mask}[A_i \geq A^{baseline}] = 1 \tag{2}$$

	Yelp		IMDb		Amazon		SNLI	
Attribution Model	**Acc.%**	**s-BLEU**	**Acc.%**	**s-BLEU**	**Acc.%**	**s-BLEU**	**Acc.%**	**s-BLEU**
Vanilla Attention (VA)	73.8	62.41	**69.8**	62.4	**70**	57.54	**50.76**	66
Explainable Attention (EA)	**71.3**	**64.32**	75.25	**70**	77.36	**73.21**	66.5	**85.14**
Vanilla Gradients	74.2	38.8	81.5	54.47	74.64	44.19	61.36	39
Gradients * X	97.2	37	93	50.35	84.92	40.37	70.14	39
Integrated Gradients	77.7	37.29	81.75	42.42	71	40.77	74.73	43
No Masking	100	100	100	100	100	100	100	100

Table 2. Comparison of quality of style-masking produced using various attribution models. We found that $\lambda_\epsilon = 0.0$ worked best with all gradient-based methods. For attention based methods (VA and EA), we found that $\lambda_\epsilon = 0.15, 0.5$ worked best for {Yelp, IMDb, Amazon}, SNLI respectively.

The Style Masked Language Model

⇒ We train a **self-attention block with 8 heads and 2 layers for 15 epochs on the reconstruction task.**
⇒ By doing so, the model also learns to perform style transfer by simply changing the target style meta token.
⇒ The model **naturally excels in fluency and content preservation** because of the nature of this style masked language task.
⇒ To also ensure the target style is present, we **fine-tune the model using signals from a trained classifier for 10 epochs**.

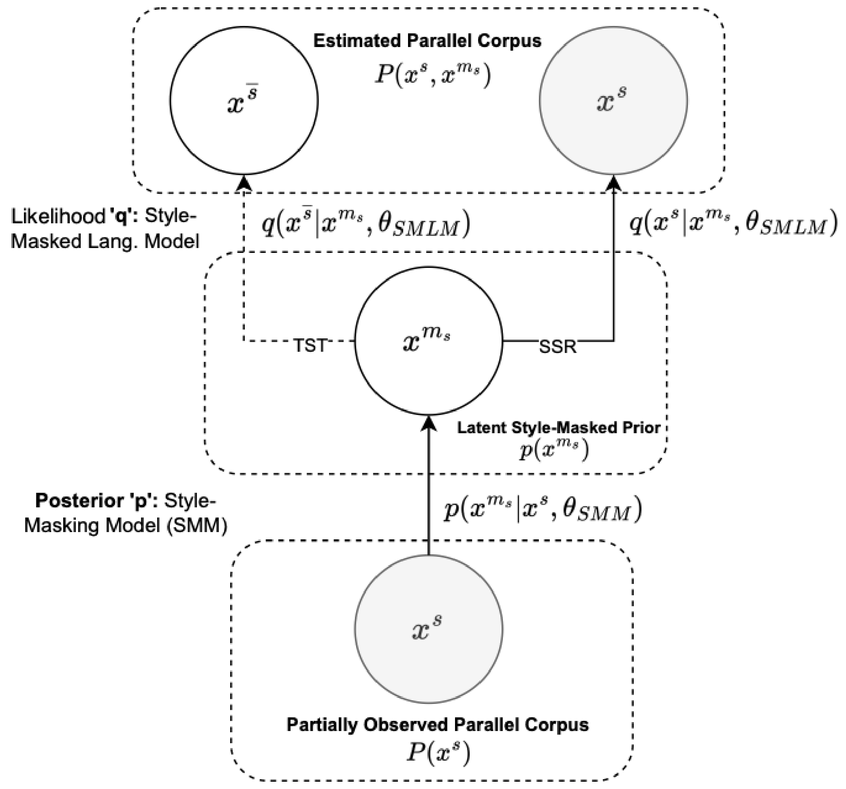

Figure 2. Probabilistic overview of our style transfer method.

Experiments

⇒ We perform sentiment transfer on the Yelp dataset and discourse transfer on the SNLI dataset and compare our model with strong baselines.
⇒ Style transfer quality is measured by **content preservation (source BLEU), Naturalness/Fluency measured by a fluency classifier and style transfer accuracy measured by a style classifier**. We take the mean of these normalised scores as the primary metric.
⇒ We observe that **our encoder-only approach beats/ competes with strong baselines** in quantitative experiments as well as human evaluations.

Model	TST%	s-BLEU	Nat.	Mean
Tag	48.3	90.2	0.98	78.83
Ensemble	52.2	88.5	0.98	79.57
Generational	58	86.7	0.98	80.9
Encoder-only	76.3	86.3	0.94	**85.53**

Table 3. Quantitative metrics for the SNLI dataset.

Model	TST%	r-BLEU	s-BLEU	Nat.	Mean
DirR	92.9	23.5	60.8	0.84	**79.27**
Stable	81.6	15.6	39.2	0.73	64.6
Transforming	84.8	18.1	44.7	0.83	70.9
Tag	87.7	16.9	47	0.83	72.57
CrossAligned	74.4	6.8	20.2	0.68	54.2
CycleRL	51.1	14.8	46.1	0.86	61.07
StyleEmbedding	8.59	16.7	67.6	0.87	54.4
D&R	88	12.6	36.8	0.89	71.27
CycleMulti	83.8	22.5	63	0.86	77.6
Lewis	93.1	-	58.5	0.84	**78.53**
Ensemble	56.5	20.5	63.2	0.85	68.23
Generational	63.4	20.3	61.3	0.83	69.23
Encoder-only	91.2	18.3	53.4	0.88	**77.6**

Table 4. Quantitative metrics for the Yelp Dataset.

Task	Positive to Negative	Contradiction to Entailment
Input	This movie is by far one of the best urban crime dramas i've seen.	a woman is sitting outside at a table using a knife to cut into a sandwich. a woman is sitting inside
Style Masked	This movie is by <mask> one of the <mask> urban crime <mask> i've seen .	a woman is sitting outside at a table using a knife to cut into a sandwich. a woman <mask> <mask> <mask>
Output	This movie is by far one of the worst urban crime garbage i 've seen .	a woman is sitting outside at a table using a knife to cut into a sandwich. a woman is there outside

Table 5. Examples of Sentiment and Discourse style transfer by the SA-MLM on the IMDb and SNLI datasets respectively.

⇒ **Limitation**: Though the model performs well in the discourse transfer task, it could be argued that this style masking approach might fail for other styles wherein the **style cannot be modified with token manipulations.**

⇒ **Code** available at *github.com/sharan21/Style-Masked-Language-Model*

Email: sharan.n21@gmail.com