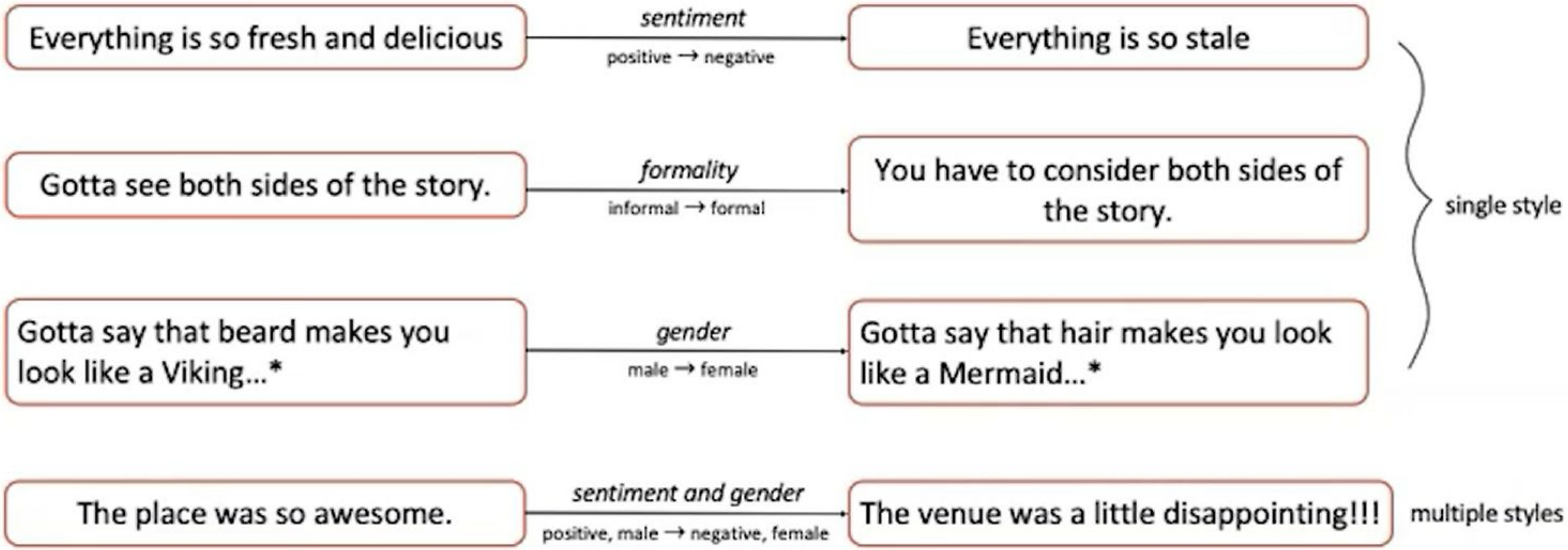
Multi-Style Transfer with Discriminative Feedback on Disjoint Corpus

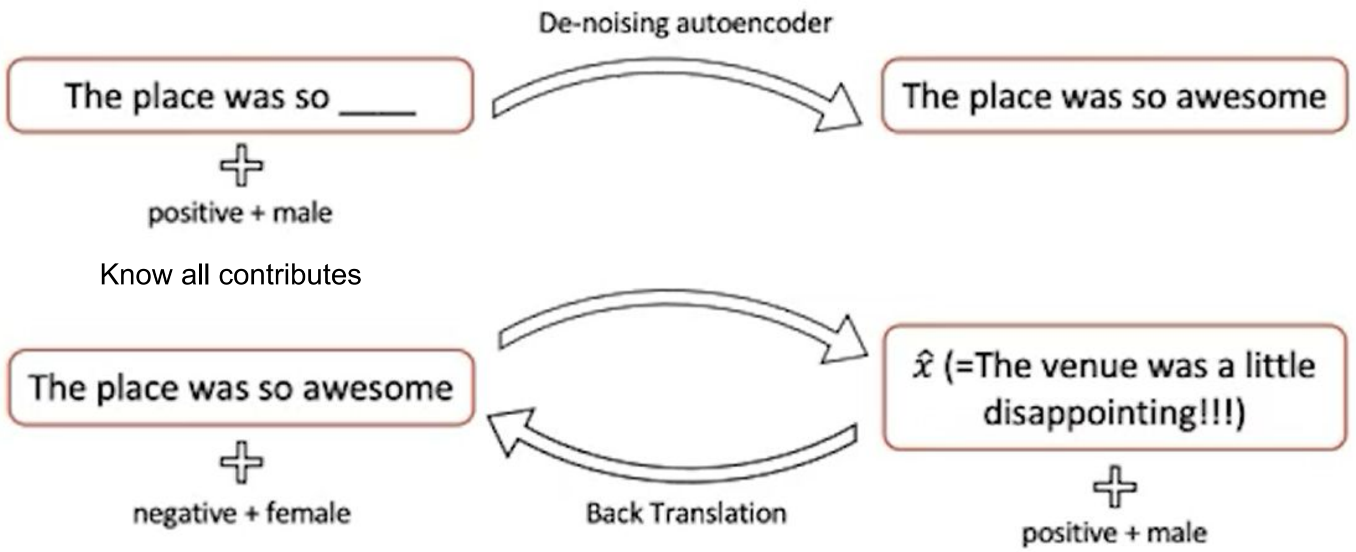
Navita Goyal, Balaji Vasan Srinivasan, Anandhavelu N, Abhilasha Sancheti

Some slides copied from authors' presentation at NAACL-2021

Style Transfer

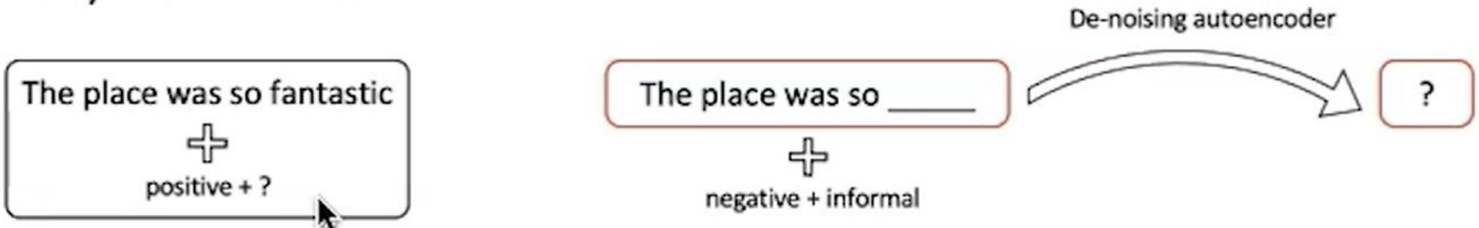


Previous Work

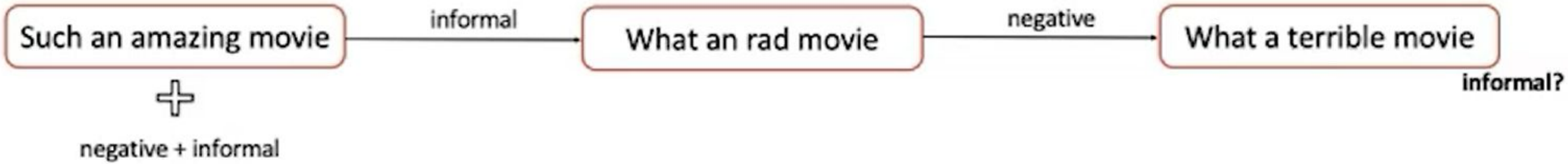


Motivation

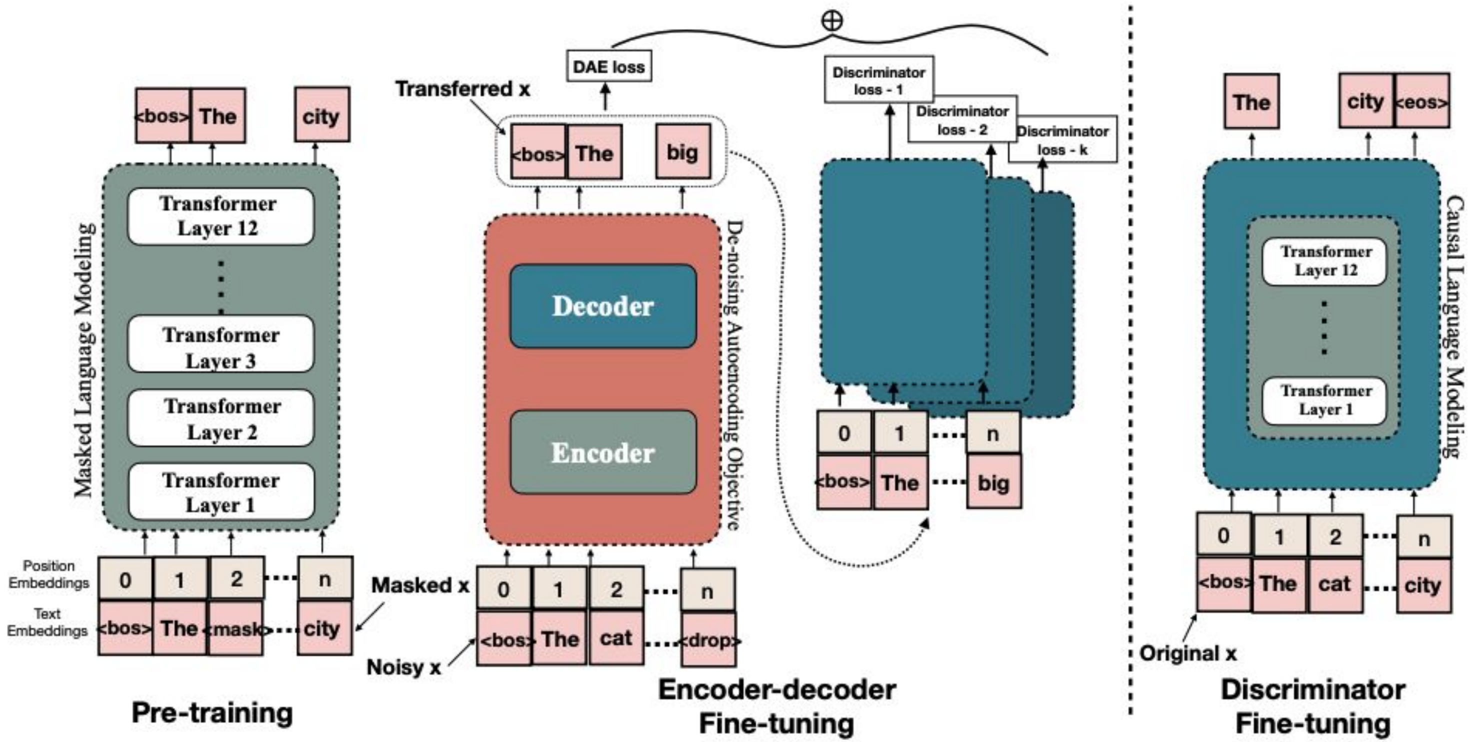
- Jointly annotated dataset



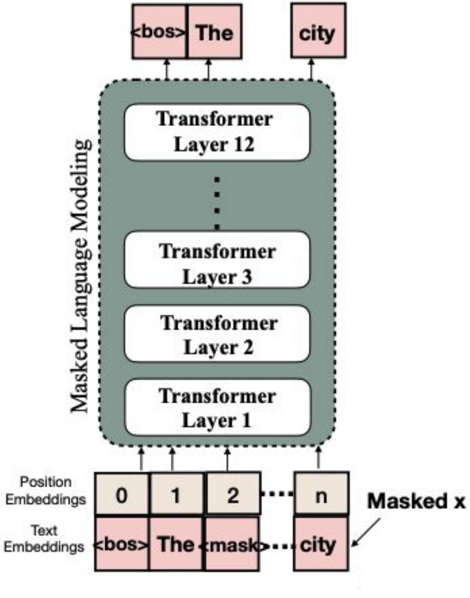
- Style independence (Cascaded system)



Proposed Approach



Language Model Pre-training



Pre-training

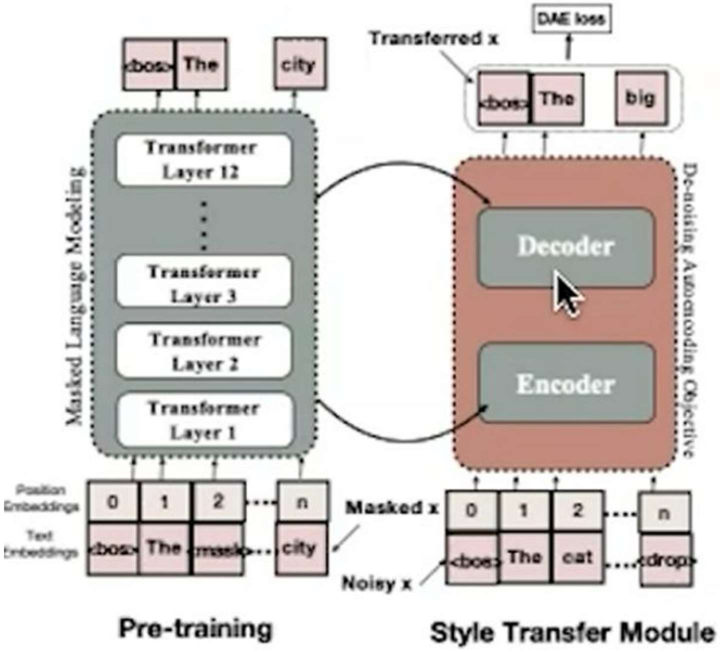
Motivation:

- Common grounding for multiple discriminators and style transfer module

Trained on masked language modeling objectives.

Used Wikipedia data

Pre-trained LM as Encode-Decoder



Motivation:

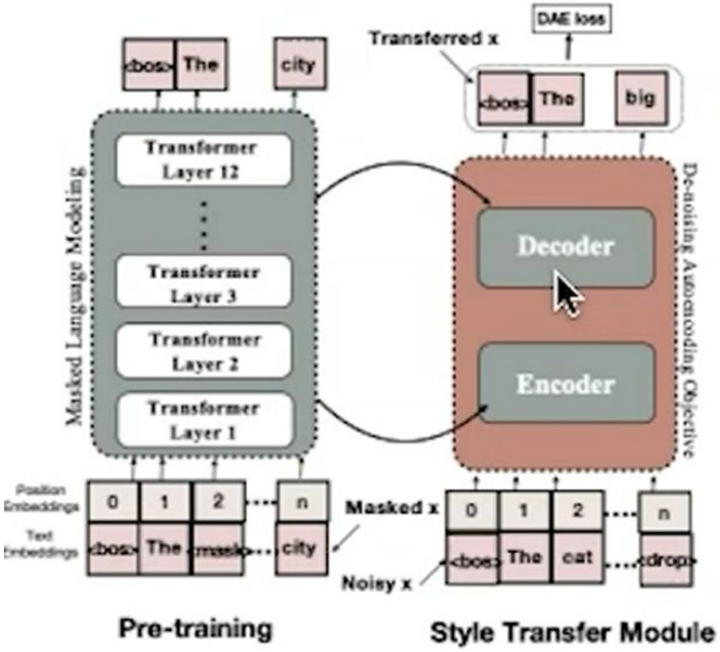
- Lower computational cost
- Allows common initialization as discriminators

Training (Denoising autoencoder)

- $\mathcal{L}_{DAE}(\theta_G) = \mathbf{E}_{x\sim T}[-\log P_{\theta_G}(x|\tilde{x})]$

Trained on a target-domain corpus or mixture of datasets of multiple styles.

Pre-trained LM as Encode-Decoder



Motivation:

- Lower computational cost
- Allows common initialization as discriminators

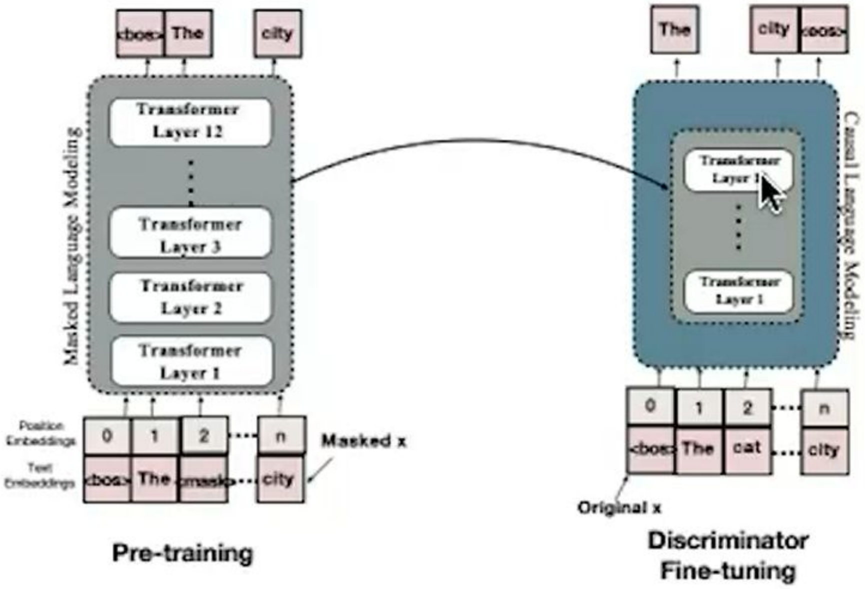
Training (Denoising autoencoder)

- $\mathcal{L}_{DAE}(\theta_G) = \mathbf{E}_{x \sim T}[-\log P_{\theta_G}(x|\tilde{x})]$

Trained on a target-domain corpus or mixture of datasets of multiple styles.

Problem: The model doesn't know the source style and is trained to generate sentences to match the style of the given target-domain corpus.

Discriminator Fine-tuning



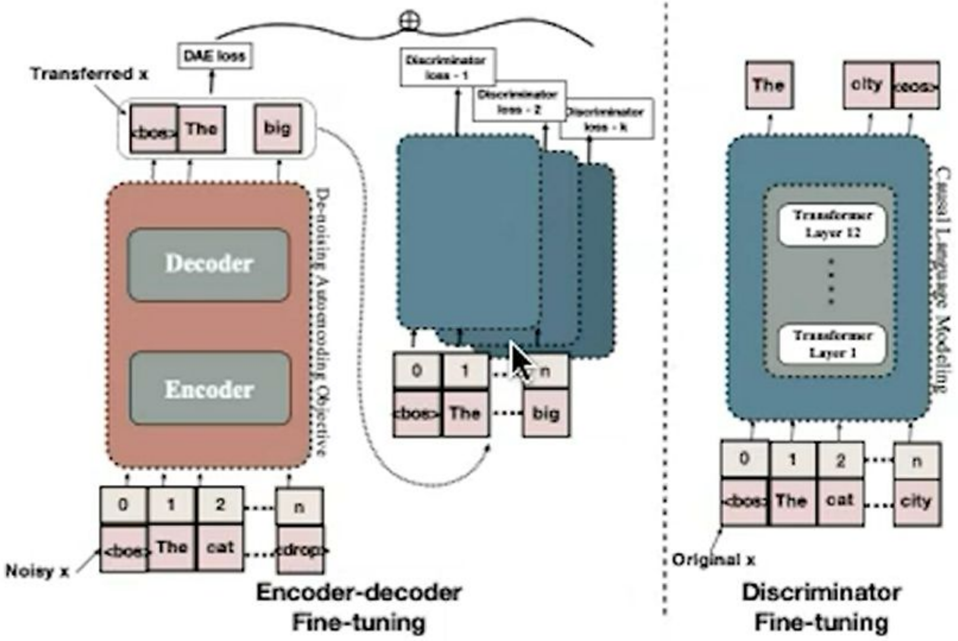
Motivation:

- Imbibe language distribution of style s_i to serve as soft-discriminator
- No adversarial training

Training

- $\mathbf{E}_{x\sim T_i}\sum_{t=1}^{n}[-\log P_{LM}(x_t|x_1,\ldots,x_{t-1})]$

Fine-tuned LM as Discriminator



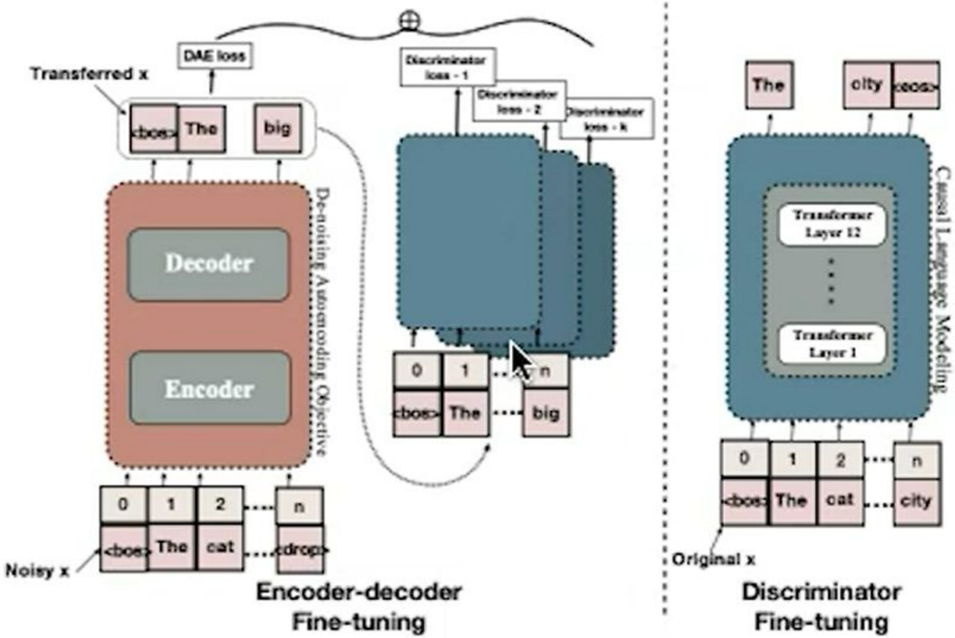
Motivation:

- Provide feedback for partially annotated data
- Soft-signal

The fine-tuned discriminative language model is implicitly capable of assigning **high** perplexity to **negative** samples (out-of-style samples).

$$\underset{\theta_G}{\operatorname{argmin}} \mathcal{L}^{s_i} = \mathbf{E}_{x \sim T, x' \sim P_{\theta_G}(x)} \left[\sum_{t=1}^{n} - \log P_{LM_i}(x'_t | x'_1, .., x'_{t-1}) \right] \quad (3)$$

Fine-tuned LM as Discriminator



$$\underset{\theta_G}{\operatorname{argmin}} \mathcal{L}^{s_i} = \mathbf{E}_{x\sim T, x'\sim P_{\theta_G}(x)} \left[\sum_{t=1}^{n} -\log P_{LM_i}(x'_t|x'_1,..,x'_{t-1})\right] \tag{3}$$

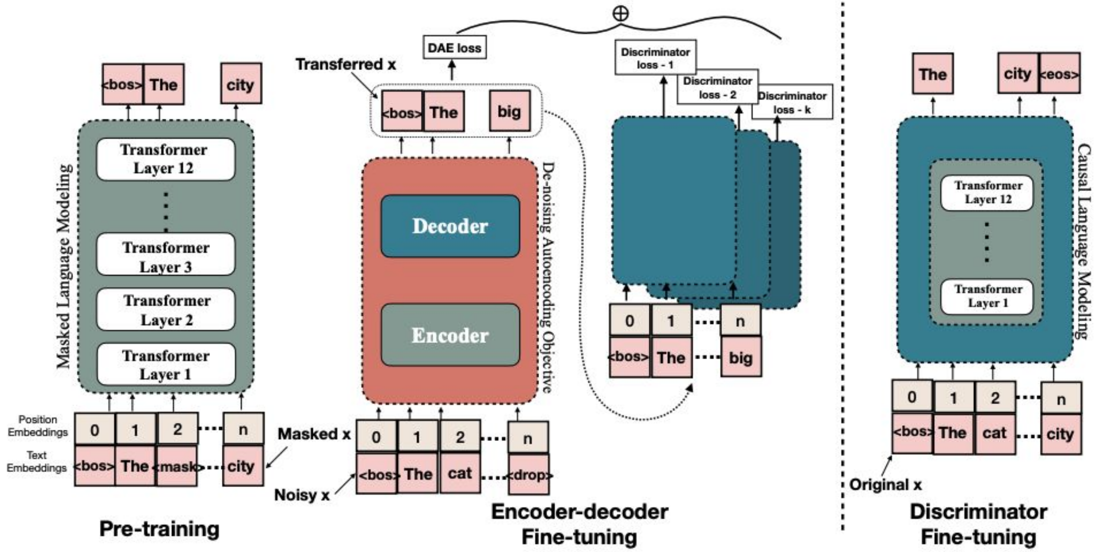
Use a policy gradient reinforcement learning approach using REINFORCE algorithm.

$$r(x) = \sum_{t=1}^{n} \log P_{LM_i}(x_t|x_1,..,x_{t-1}) \tag{4}$$

Using these rewards, the RL objective is to minimize the loss $\mathcal{L}^{s_i}$ given by,

$$\mathcal{L}^{s_i} = \mathbf{E}_{x\sim T, x'\sim P_{\theta_G}(x)}(r(x') - r(x)) [-\log P_{\theta_G}(x'|\tilde{x})] \tag{5}$$

Overall



$$\mathcal{L} = \lambda_{DAE}\mathbf{E}_{x\sim T}[-\log P_\theta(x|\tilde{x})] + \sum_{i=1}^{k}\lambda_i\mathcal{L}^{s_i},$$

Experiment - Data

Style	Dataset	Train	Test
Sentiment	IMDB[1]+Yelp[2]	600k	3000
Formality	GYAFC[3]	208k	4849

Experiment - Style-awareness of LM

Style/Dimension	Sentiment %	Formality %
Positive	71.41	67.09
Negative	76.17	75.59

Table 1: Accuracy of sentences generated by model fine-tuned on style s_i as % of generated sentences labelled as class s_i by the classifier trained on the corresponding style dimension.

Fine-tuning corpus	Test Corpus	
	Same ↓	Opposite ↑
Positive	6.9275	9.6850
Negative	7.7131	9.9637

Table 2: Perplexity of test corpus on models fine-tuned positive and negative corpus (rows). The column *Same* represents that test corpus is same as fine-tuning corpus, leading to lower perplexities and *Opposite* represent test corpus from opposite polarity as fine-tuning corpus leading to higher perplexity.

Results

Model	Style Accuracy			Content Preservation		Fluency
	Classifier ↑		Lexical Scoring ↑	BLEU ↑		Perplexity ↓
	Sentiment	Formality	Formality	*-self*	*-ref*	
Cascaded Style Transformer (Dai et al., 2019)	72.17	64.08	81.29	0.6066	0.3479	8.8657
Adapted Rewriting LM (Syed et al., 2020)	52.59	36.39	72.21	**0.7917**	**0.4259**	6.5963
Cascaded Discriminative LM	69.30	48.18	83.02	0.6634	0.3579	6.6846
Joint Discriminative LM	**79.78**	**65.33**	**85.39**	0.7710	0.4136	**6.4574**

Table 3: Quantitative Comparison of our proposed approach (Joint Discriminative LM) against Cascaded Style Transformer (Dai et al., 2019), Cascaded Discriminative LM method and multi-style transfer using Adapted Rewriting LM (Syed et al., 2020). The upward arrow signifies that higher is better and vice versa. Score of near 100 on formality lexical scoring imply the transferred text is close in formality to the target corpus.

Results

Model	Style Accuracy Sentiment	Style Accuracy Formality	Content Preservation	Fluency	Transfer Quality
Cascaded Style Transformer (Dai et al., 2019)	3.5909	2.7424	3.2803	2.7424	2.9318
Joint Discriminative LM (Our Model)	**3.8561**	**3.0379**	**4.1061**	**4.1894**	**4.1091**

Table 5: Results for Human Evaluation across different metrics. Each value represents the average of rating between 1 (Very bad) and 5 (Very good).

Target style	Source sentence	Transferred Sentence Style Transformer	 Our model (multi-style)
Positive+Formal	That's not funny. I don't think she'll like it.	So funny movie. I really like it.	That was very funny. I am sure she will **appreciate it**.
	Give your brother some money and tell him to take a hike.	Just give your brother some time and it will be good again.	Give your brother some money and **request him to leave**.
Negative+Formal	An intelligent, rewarding film that I look forward to watching again.	ludicrous, shallow film that look forward to watching again.	An unintelligent, poor film that **I would not look forward** to watching again.
	super friendly staff, quick service and amazing and simple food was done right!	says wait staff, quick not amazing before overcooked food done were okay.	dirty staff and slow service and simple food was not done right.
Positive+Informal	You need to separate the bad thing and move on.	need to the great thing and move on.	You need to enjoy the good stuff and move on.
	The evening started out slow.	The evening spent in professional show.	The evening **began** amazing.
Negative+Informal	Great food recommendations steak and tuna were both great.	terrible food 9am steak and were both terrible.	Disappointing food recommendations steak and tuna were horrible.
	That person in hilarious.	You person in worse!	**That guy** in so boring.