Thank you BART!
Rewarding Pre-Trained Models Improves Formality Style Transfer

Huiyuan Lai, Antonio Toral, Malvina Nissim
CLCG, University of Groningen / The Netherlands
{h.lai, a.toral.ruiz, m.nissim}@rug.nl

Abstract

Scarcity of parallel data causes formality style transfer models to have scarce success in preserving content. We show that fine-tuning pre-trained language (GPT-2) and sequence-to-sequence (BART) models boosts content preservation, and that this is possible even with limited amounts of parallel data. Augmenting these models with rewards that target style and content –the two core aspects of the task– we achieve a new state-of-the-art.

1 Introduction and Background

Style transfer is the task of automatically converting a text of one style into another, such as turning the formal "*I viewed it and I believe it is a quality program.*" into the informal "*I've watched it and it is AWESOME!!!!*". This task, which can be used for, e.g., personalised response generation, translation of ancient text into modern text, and text simplification, is particularly challenging since style must be changed while ensuring that content is preserved. Accordingly, the performance of style transfer systems is commonly assessed on both style strength and content preservation.

Due to the general scarcity of parallel data, unsupervised approaches are popular. These include disentangling style and content by learning a distinct representation for each (Shen et al., 2017; Fu et al., 2018; John et al., 2019), and back translation (Zhang et al., 2018; Lample et al., 2019; Luo et al., 2019; Prabhumoye et al., 2018). A common strategy to enhance style accuracy is to introduce a reward in the form of a style classifier (Lample et al., 2019; Gong et al., 2019; Luo et al., 2019; Wu et al., 2019; Sancheti et al., 2020). As a result, unsupervised models achieve good accuracy in style strength. Content preservation is however usually unsuccessful (Rao and Tetreault, 2018).

Parallel data can help to preserve content, but is limited. Niu et al. (2018) combine the train sets of two different domains and incorporate machine translation to train their models with a multi-task learning schema, plus model ensembles. Sancheti et al. (2020) use it to train a supervised sequence-to-sequence model, and in addition to the commonly used style strength reward, they include a reward based on BLEU (Papineni et al., 2002) to enhance content preservation. Shang et al. (2019) propose a semi-supervised model combining parallel data with large amounts of non-parallel data.

Pre-trained models, successful in a variety of NLP tasks, have recently been used in formality style transfer. Zhang et al. (2020) propose several data augmentation methods for pre-training a transformer-based (Vaswani et al., 2017) model and then used gold data for fine-tuning. Using GPT-2 (Radford et al., 2019), Wang et al. (2019) and Wang et al. (2020) propose a harness-rule-based preprocessing method, and joint training of bi-directional transfer and auto-encoding with two auxiliary losses. Contemporary work by Chawla and Yang (2020) develops a semi-supervised model based on BART large (Lewis et al., 2020).

Contributions Focusing specifically on *formality transfer*, for which parallel data is available, (i) we take the contribution of pre-trained models a step further by augmenting them with reward strategies that target content and style, thereby achieving new state-of-the-art results. (ii) We analyse separately the contribution of pre-trained models on content and style, showing that they take care of preserving content (the hardest part of style transfer to date), while ensuring style strength. (iii) Moreover, experimenting with training size, we show that while parallel data contributes to content preservation, fine-tuning pre-trained models with 10% of parallel data is more successful than training on 100% of data from scratch. Reducing the need for parallel data opens up the applicability of

484

(a) Architecture of the GPT-2-based model (b) Architecture of the BART-based model
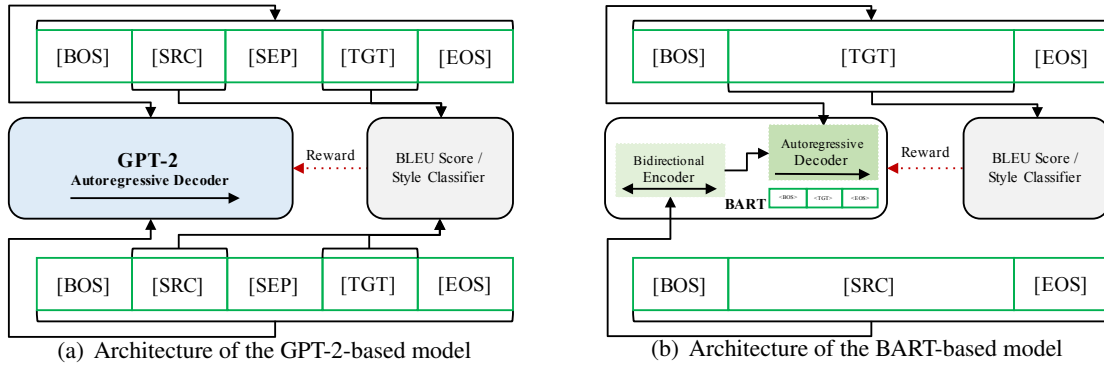
Figure 1: Model architectures. We use three special symbols: [BOS] in front of every source sentence, [SEP] between the source and target sentences (only in GPT-2), and [EOS] at the end of every target sentence.

supervised style transfer to new scenarios: tasks, domains, languages.[1]

2 Method

We propose a framework to control the style of output text for style transfer atop pre-trained models. Given a source sentence $x = \{x_1, \cdots, x_n\}$ of length n with style s_1 and a target style sentence $y = \{y_1, \cdots, y_m\}$ of length m with style s_2, our model aims to learn two conditional distributions, altering the style of a sentence while preserving its original content. Our framework consists of (i) fine-tuning pre-trained models on a formality transfer parallel corpus; (ii) incorporating rewards to enhance style change and content preservation.

2.1 Models

GPT-2 This model (Radford et al., 2019) is a transformer-based network (Vaswani et al., 2017). Given a sentence of tokens $x = \{x_1, \cdots, x_l\}$, the standard language modeling objective is to minimize the following negative log likelihood:

$$L_{(\phi)} = -\Sigma_i \log(p(x_i | x_{i-k:i-1}; \phi)) \quad (1)$$

where k is the size of the context window.

To make GPT-2 rephrase a text in the target style, the input pair ⟨Source Sentence, Target Sentence⟩ is represented as a single sequence with three special tokens to mark beginning [BOS] and end [EOS] of every sequence, and to separate source and target sentences [SEP] (Fig. 1(a)). During inference, we feed to GPT-2 the source sentence with [BOS] and [SEP] to infer the target sentence.

BART This is a denoising autoencoder for pre-training sequence-to-sequence models (Lewis et al., 2020). Given a source sentence x and a target sentence y, the loss function is the cross-entropy between the decoder's output and the target sentence:

$$L_{(\phi)} = -\Sigma_i \log(p(y_i | y_{1:i-1}, x; \phi)) \quad (2)$$

2.2 Rewards

Atop the models, we implement two rewards, used in isolation and together, to enhance style strength (Style Classification Reward) and content preservation (BLEU Score Reward).

Style Classification Reward As often done in previous work (see Section 1), we use a classification confidence reward to encourage larger change in the confidence of a style classifier (SC). We pre-train the binary style classifier TextCNN (Kim, 2014) and use it to evaluate how well the transferred sentence y' matches the target style. SC's confidence is formulated as

$$p(s_i | y') = softmax_i(TextCNN(y', \theta)) \quad (3)$$

where i = {1,2}, and represent source and target style respectively. θ are the parameters of the style classifier, fixed during fine-tuning. The reward is

$$R_{cls} = \lambda_{cls}[p(s_2|y') - p(s_1|y')] \quad (4)$$

where y' is the generated target sentence sampled from the model's distribution at each time step in decoding. For the GPT-2 based model, we also add a classification confidence reward to the source sentence, similar to Eq. 4, since the model generates sentence x' with the original style while generating the target sentence:

$$R_{cls_{source}} = \lambda_{cls}[p(s_1|x') - p(s_2|x')] \quad (5)$$

485

		$0 \longrightarrow 1$		$1 \longrightarrow 0$	
Domain	Train	Valid	Test	Valid	Test
F&R	51,967	2,788	1,332	2,247	1,019
E&M	52,595	2,877	1,416	2,356	1,082

Table 1: GYAFC dataset. 0 = informal; 1 = formal.

BLEU Score Reward Following Sancheti et al. (2020), we introduce a BLEU-based reward to foster content preservation as in Eq. 6, where $\boldsymbol{y}'$ is the target style text obtained by greedily maximizing the distribution of model outputs at each time step, and $\boldsymbol{y}^s$ is sampled from the distribution.

$$R_{bleu} = \lambda_{bleu}[bleu(\boldsymbol{y}', \boldsymbol{y}) - bleu(\boldsymbol{y}^s, \boldsymbol{y})] \quad (6)$$

Gradients and Objectives The rewards are used for policy learning. The policy gradient[2] is

$$\nabla_\phi J(\phi) = E[R \cdot \nabla_\phi log(P(\boldsymbol{y}^s|\boldsymbol{x};\phi))] \quad (7)$$

where R is the SC reward and/or the BLEU reward, $\boldsymbol{y}^s$ is sampled from the distribution of model outputs at each decoding time step, and ϕ are the parameters of the model. Similarly, we add the policy gradient regarding the source sentence for the SC reward (only for the GPT-2-based model).

The overall objectives for ϕ are the loss of the base model (Eq. 1 or Eq. 2) and the policy gradient of the different rewards (Eq. 7).

3 Experiments

Dataset Grammarly's Yahoo Answers Formality Corpus (GYAFC) (Rao and Tetreault, 2018) is a formality style transfer dataset with parallel formal and informal sentences from two domains: Entertainment & Music (E&M) and Family & Relationships (F&R). Table 1 shows the number of sentences in train, validation, and test. Four human references exist for every valid/test sentence.

Setup All experiments are implemented atop Huggingface's transformers (Wolf et al., 2020). Our base models are the GPT-2-based model (117M parameters) and BART-based model (base with 139M parameters and large with 406M). We fine-tune them with the Adam optimiser (Kingma and Ba, 2015) with batch size 32; the initial learning rates are $5e^{-5}$ (GPT-2) and $3e^{-5}$ (BART). The final values for λ are set to 1 for SC and 0.2 for BLEU based on validation results. We use early

---

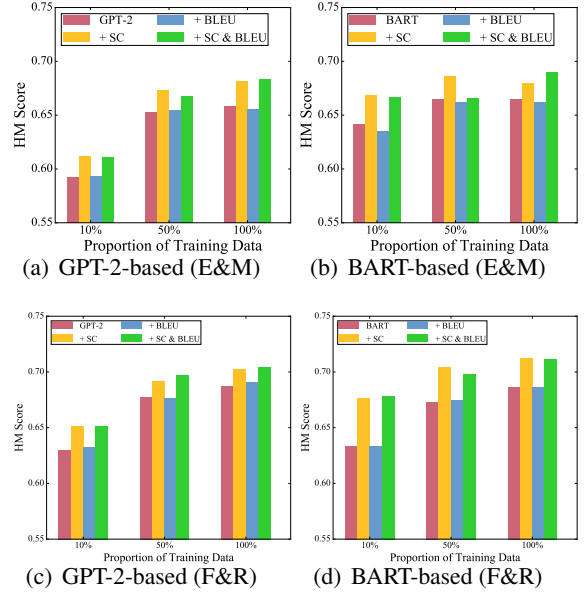[2]Additional details are provided in the Appendix.

Figure 2: HM score of x%-sized training sets of GPT-2-/BART-based models with different rewards (none, +SC, +BLEU, +SC & BLEU) for the two domains (E&M and F&R).

stopping (patience 3) if validation performance does not improve. Test results are reported with the best validation settings.

Evaluation Following previous work (Luo et al., 2019; He et al., 2020; Sancheti et al., 2020), we adopt the following strategies. The binary classifier TextCNN (Kim, 2014) is pre-trained to evaluate style strength; on the human references it has an accuracy of 87.0% (E&M) and 89.3% (F&R). Based on the four human references, we calculate BLEU[3] for content preservation. As overall score we compute the harmonic mean (HM) of style accuracy and BLEU. For our evaluation we also test BLEURT, a recent metric for content preservation which correlates better with human judgments than other metrics that take semantic information into account, e.g. METEOR (Sellam et al., 2020).

Baselines We train a basic supervised model (a Bi-LSTM with attention from OpenNMT (Klein et al., 2017)), to assess the impact of the size of parallel training data. We compare our models to the five baselines from Rao and Tetreault (2018), and to the best performing formality style transfer methods that report results on the datasets we use. These are mentioned in Section 1 and summarised as follows: Bi-directional FT (Niu et al.,

---

[3]We use `multi-bleu.perl` with default settings.

Domain	Model	BLEURT	BLEU	ACC	HM	Model	BLEURT	BLEU	ACC	HM
	OpenNMT + SC & BLEU (10% data)	-0.919	0.231	0.886	0.366	OpenNMT + SC & BLEU (100% data)	-0.420	0.403	0.804	0.537
	(A) INFORMAL ↔ FORMAL					**(B) INFORMAL ⟶ FORMAL**				
	NMT-Combined (Rao and Tetreault, 2018)	-0.100	0.501	0.797	0.615	GPT-CAT (train on E&M and F&R, Wang et al. (2019))	0.176	0.725	0.876	0.793
	GPT-2 + SC & BLEU (10% data, **Ours**)	-0.058	0.495	0.799	0.611	Chawla's (Chawla and Yang (2020))	0.260	0.762	0.910	0.829
E&M	GPT-2 + SC & BLEU (100% data, **Ours**)	-0.007	0.542	**0.923**	0.683	BART + SC & BLEU (train on E&M, **Ours**)	0.218	0.730	0.887	0.801
	BART + SC & BLEU(10% data, **Ours**)	-0.030	0.547	0.855	0.667	BART + SC & BLEU (train on E&M and F&R, **Ours**)	0.236	0.745	**0.937**	0.830
	BART + SC & BLEU (100% data, **Ours**)	**0.044**	**0.577**	0.859	**0.690**	BART large + SC & BLEU (train on E&M and F&R, **Ours**)	**0.274**	**0.765**	0.929	**0.839**
	(C) INFORMAL ↔ FORMAL & COMBINED DOMAINS					**(D) BLEU EVALUATED AGAINST THE FIRST REFERENCE**				
	Bi-directional FT (Niu et al., 2018)	0.023	0.554	0.818	0.661	*TS→CP (Sancheti et al. (2020))	-	0.292	-	-
	BART large + SC & BLEU (100% data, **Ours**)	**0.078**	**0.596**	**0.905**	**0.719**	BART + SC & BLEU (100% data, **Ours**)	-	**0.306**	-	-
	OpenNMT + SC & BLEU (10% data)	-0.706	0.303	0.859	0.448	OpenNMT + SC & BLEU (100% data)	-0.304	0.477	0.789	0.595
	(A) INFORMAL ↔ FORMAL					**(B) INFORMAL ⟶ FORMAL**				
	NMT-Combined (Rao and Tetreault, 2018)	-0.089	0.527	0.798	0.635	*GPT-CAT (train on E&M and F&R, Wang et al. (2019))	-	0.769	-	-
	GPT-2 + SC & BLEU (10% data, **Ours**)	-0.027	0.528	0.849	0.651	Chawla's (Chawla and Yang (2020))	0.302	**0.799**	0.910	0.851
F&R	GPT-2 + SC & BLEU (100% data, **Ours**)	0.038	0.572	**0.915**	0.704	BART + SC & BLEU (train on F&R, **Ours**)	0.271	0.770	0.897	0.829
	BART + SC & BLEU (10% data, **Ours**)	0.039	0.571	0.833	0.678	BART + SC & BLEU (train on F&R and E&M, **Ours**)	0.270	0.777	0.912	0.839
	BART + SC & BLEU (100% data, **Ours**)	0.068	0.595	0.882	**0.711**	BART large + SC & BLEU (train on F&R and E&M, **Ours**)	**0.324**	0.793	**0.920**	**0.852**
	(C) INFORMAL ↔ FORMAL & COMBINED DOMAINS					**(D) 10% PARALLEL TRAINING DATA**				
	Bi-directional FT (Niu et al. (2018)	0.037	0.568	0.839	0.677	*CPLS (Shang et al., 2019)	-	0.379	-	-
	BART large + SC & BLEU (100% data, **Ours**)	**0.100**	**0.611**	**0.900**	**0.728**	BART + SC & BLEU (**Ours**)	-	**0.571**	-	

Table 2: Comparison of our models to previous work. The best score for each metric in each block is boldfaced. Notes: (i) if the output of previous work is available, we re-calculate the scores using our evaluation metrics. Otherwise, scores are from the paper and we mark this with (*); (ii) (B) shows our results on informal-to-formal to compare with Wang et al. (2019) and Chawla and Yang (2020), who only transfer in this direction; (iii) in (C) we train on the concatenated data from both domains, to compare against Niu et al. (2018); (iv) in (E&M (D)) we re-evaluate our system against the first reference only, as done by Sancheti et al. (2020).

2018), CPLS (Shang et al., 2019), GPT-CAT (Wang et al., 2019), S2S-SLS (GPT-2) (Wang et al., 2020), Transformer (data augmentation) (Zhang et al., 2020), TS→CP (Sancheti et al., 2020), and Chawla's (Chawla and Yang, 2020). Since supervised methods significantly outperform unsupervised approaches, results for the latter are not considered as the baseline in our experiment.. Disentanglement-based methods are not included since Lample et al. (2019) provide evidence that they are surpassed.

Results Figure 2 shows the HM score of x%-sized training sets on the E&M and the F&R domains. Increasing train set size from 10% to 50% has a greater boost on GPT-2-based models than BART's. However, BART-based models obtain the highest results. Table 2 reports a selection of our models [4] and previous state-of-the-art work. Zooming in on the single measures, we see in Table 2 how varying training size reveals the impact of parallel data on content preservation: OpenNMT's BLEU score on E&M increases from 0.231 with 10% of the data to 0.403 with 100%. Style accuracy appears instead easier to achieve even with limited supervision. Increasing training size for fine-tuning either pre-trained model does not however yield dramatic improvements in content preservation (e.g. from 0.547 to 0.577 BLEU for BART

base on E&M). In fact, fine-tuning a pre-trained model (either GPT-2 or BART) with just 10% of parallel data, leads to better content preservation (0.547 BLEU with BART on E&M) than Open-NMT with 100% (0.403). This suggests that content preservation is largely taken care of by the pre-trained models, already, and can explain why the BLEU-based reward does not help too much in isolation (see Fig. 2). Conversely, the SC reward consistently boosts style accuracy in both BART and GPT-2. Nevertheless, combining rewards can be beneficial. Overall, BART-based models perform better on content preservation while results on style strength are mixed.

Given the experimental setup of some previous work, we ran additional comparisons (blocks (B), (C), and (D) of Table 2). In all cases, our results are higher than the previous state-of-the-art. For example, in F&R (D) our model with 10% parallel data outperforms Shang et al. (2019)'s semi-supervised model, which uses about 9.5% parallel data and large amounts of non-parallel data (BLEU 0.571 vs 0.379). Fine-tuning BART on both domains (C)[5] leads to the best results to date on both datasets (E&M: 0.719; F&R: 0.728).

With respect to the two evaluation metrics used for content preservation (BLEU and BLEURT), we can observe in Table 2 that they follow a similar trend. In fact, they correlate very highly (Pearson's $r = .951$, p<.001, $n = 14$ for E&M, and $r = .951$,

---

[4] In the table we report results for the models that use both rewards (BLEU and SC) since this setting mostly leads to best results. Complete results for all models (and sample outputs) are in the Appendix.

[5] Following Kobus et al. (2017), we add a token to each training instance that specifies its domain.

System	Sentence	BLEURT	BLEU	ACC
	FROM INFORMAL TO FORMAL			
Source	i say omarion.he has the hair clothes and body,a triple deal on one person.		-	
Reference 1	My choice is Omarion as he has high quality, hair, clothes, and body to create a triple deal in one person.		-	
Reference 2	I would say Omarion because he has the hair, clothes, and body; A triple deal on a single person.		-	
Reference 3	I pick Omarion, he has the hair, the clothes, and the body. A triple deal on one person.		-	
Reference 4	Omarion has the hair, clothes, and the body.		-	
PBMT-Combined (Rao and Tetreault, 2018)	I say omarion. he has the hair clothes and body, the deal on one person.	-0.153	0.509	0.946
Bi-directional FT (Niu et al., 2018)	I say Omarion, he has the hair clothes and body, and a triple deal on one person.	-0.149	0.510	0.953
GPT-CAT (Wang et al., 2019)	I say Omarion. He has the hair, clothes, and body, a triple deal on one person.	0.044	0.585	**1.000**
S2S-SLS (Wang et al., 2020)	I say Omarion. He has the hair clothes and body, a triple deal on one person.	-0.035	0.350	**1.000**
Transformer (Zhang et al., 2020)	I say omarionhe has the hair clothes and body, a triple deal on one person.	-0.255	0.462	0.892
Chawla's (Chawla and Yang, 2020)	I say Marion because he has the hair, clothes and body, a triple deal on one person.	-0.538	0.534	0.989
OpenNMT + SC & BLEU (**Ours**)	I say Omarion. He has the hair clothes and body.	-0.325	0.147	**1.000**
GPT-2 + SC & BLEU (**Ours**)	I say Omarion. He has the hair, clothes, and body, a triple deal on one person.	-0.035	0.350	**1.000**
BART base + SC & BLEU (**Ours**)	I would say Omar. He has the hair, clothes, and body. It is a triple deal on one person.	-0.012	0.589	**1.000**
BART large + SC & BLEU (**Ours**)	I would say Omarion. He has the hair, clothes, and body, a triple deal on one person.	**0.096**	**0.657**	**1.000**
	FROM FORMAL TO INFORMAL			
Source	I suggest avoiding hot dogs, and not watching this movie with your little sister.		-	
Reference 1	Don't eat hot dogs, or watch this movie with your little sister!		-	
Reference 2	Don't do hot dogs or this movie with your kid sister.		-	
Reference 3	don't eat hot dogs and don't watch it w/ ur lil sis!		-	
Reference 4	Don't eat hot dogs or watch this flick with your lil sis!		-	
PBMT-Combined (Rao and Tetreault, 2018)	I suggest avoiding hot dogs, and not watching this movie with your little sister.	-0.298	0.417	0.004
Bi-directional FT (Niu et al., 2018)	I suggest avoiding hot dogs and not watching this movie with your little sister.	-0.233	0.437	0.009
OpenNMT with SC & BLEU	Can't watch this movie with your little sister.	-0.521	0.542	0.783
GPT-2 + SC & BLEU	don't watch this movie with your little sister.	-0.415	0.599	**1.000**
BART + SC & BLEU	avoid hot dogs and not watch this movie with your little sister.	**-0.016**	0.610	0.925
BART large + SC & BLEU	Avoid hot dogs and don't watch this movie with your little sister.	-0.171	**0.800**	0.825

Table 3: Sample model outputs and their sentence-level scores on the E&M domain, where red denotes improperly generated words or content. Note that ACC indicates style confidence here.

p<.001, $n = 13$ for F&R).

Finer-grained Analysis Table 3 shows example outputs and their evaluation according to the metrics we use; the outputs are produced by existing systems we compare to, and our own models.[6]

In the "Informal to Formal" example, we can see that text generated by most systems is assessed with a high confidence in style conversion, except for PBMT-Combined (Rao and Tetreault, 2018) and Transformer (Zhang et al., 2020) (the name "omarionhe" should be "Omarion", and the word "he" at the beginning of the sentence should be "He"). However, the sentences generated by previous systems are not so fluent, and some of them fail in preserving content (Transformer (Zhang et al., 2020) ("omarionhe") and Chawla's (Chawla and Yang, 2020) ("Marion")). For our models, the Bi-LSTM based model fails in content preservation while the systems based on pre-trained models are much better at this task. Our model based on BART Large generates this specific sentence accurately in terms of content preservation, style strength, and fluency.

When looking at the "Formal to Informal" example in Table 3, we observe that the two previously existing systems replace very little (one comma by the Bi-directional FT (Niu et al., 2018)) or nothing at all (PBMT-Combined (Rao and Tetreault, 2018)). Conversely, our systems make substantial modifications, resulting in output sentences that are noticeably more informal than the input sen-

tence. OpenNMT and the GPT-2-based models lose part of the content (the suggestion to avoid hot dogs) while the two BART-based systems manage to preserve the whole message.

4 Conclusions

Fine-tuning pre-trained models proves a successful strategy for formality style transfer, especially towards content preservation, thereby reducing the need for parallel data. A sequence-to-sequence pre-trained model (BART) outperforms a language model (GPT-2) in content preservation, and overall, and with the addition of rewards achieves new state-of-the-art results. The fact that GPT-2 is instead often better at style strength could be (partly) due to how the style reward is implemented in the two models (Eq. 4 and 5), and will need further investigation. For a better understanding of the different behaviour of BART and GPT-2 for this task, the next natural step is to include human evaluation.

---

[6]More examples are in Appendix.

Impact Statement

All work that automatically generates and/or alters natural text could unfortunately be used maliciously. While we cannot fully prevent such uses once our models are made public, we do hope that writing about risks explicitly and also raising awareness of this possibility in the general public are ways to contain the effects of potential harmful uses. We are open to any discussion and suggestions to minimise such risks.

References

Kunal Chawla and Diyi Yang. 2020. Semi-supervised formality style transfer using language model discriminator and mutual information maximization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2340–2354.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 663–670.

Hongyu Gong, Suma Bhat, Lingfei Wu, JinJun Xiong, and Wen-mei Hwu. 2019. Reinforcement learning based text style transfer without parallel training corpus. 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 3168–3180.

Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A probabilistic formulation of unsupervised text style transfer. In *Proceedings of Ninth International Conference on Learning Representations*.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72.

Catherine Kobus, Josep Maria Crego, and Jean Senellart. 2017. Domain control for neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 372–378.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Zhifang Sui, and Xu Sun. 2019. A dual reinforcement learning framework for unsupervised text style transfer. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 5116–5122.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 129–140.

Abhilasha Sancheti, Kundan Krishna, Balaji Vasan Srinivasan, and Anandhavelu Natarajan. 2020. Reinforced rewards framework for text style transfer. In *Advances in Information Retrieval*, pages 545–560.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Mingyue Shang, Piji Li, Zhenxin Fu, Lidong Bing, Dongyan Zhao, Shuming Shi, and Rui Yan. 2019. Semi-supervised text style transfer: Cross projection in latent space. 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 4937–4946.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6833–6844.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *In Advances in Neural Information Processing Systems*, pages 6000–6010.

Yunli Wang, Yu Wu, Lili Mou, Zhoujun Li, and Wenhan Chao. 2019. Harnessing pre-trained neural networks with rules for formality 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 3573–3578.

Yunli Wang, Yu Wu, Lili Mou, Zhoujun Li, and Wen-Han Chao. 2020. Formality style transfer with shared latent space. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2236–2249, Barcelona, Spain (Online).

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Chen Wu, Xuancheng Ren, Fuli Luo, and Xu Sun. 2019. A hierarchical reinforced sequence operation method for unsupervised text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4873–4883.

Yi Zhang, Tao Ge, and Xu Sun. 2020. Parallel data augmentation for formality style transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3221–3228.

Zhirui Zhang, Shuo Ren, Shujie Liu, Jianyong Wang, Peng Chen, Mu Li, Ming Zhou, and Enhong Chen. 2018. Style transfer as unsupervised machine translation. *arXiv preprint, arXiv: 1808.07894*.

A Appendices

This Appendices include: 1) detailed results for all experiments (A.1); 2) more details on policy gradient (A.2); 3) some example outputs of various models and their sentence-level scores, to give an idea of what the generated sentences look like when style transfer is applied. We specifically focus on the 100% parallel data settings for our models (A.3).

A.1 Detailed Results of Models

We report here the full set of results for all our models and previous work.

(a) Detailed Results of Our Models

Model	BLEURT	BLEU	ACC	HM	BLEURT	BLEU	ACC	HM	BLEURT	BLEU	ACC	HM
Proportion of parallel training data		10%				50%				100%		
OpenNMT (Bi-LSTM)	-0.919	0.231	0.886	0.366	-0.489	0.392	0.789	0.524	**-0.420**	0.403	0.804	0.537
OpenNMT + SC	**-0.902**	**0.238**	**0.893**	**0.376**	-0.500	0.386	**0.821**	0.526	-0.451	0.399	0.789	0.530
OpenNMT + BLEU	-0.926	0.232	0.888	0.368	**-0.485**	0.389	0.800	0.523	-0.485	**0.412**	0.767	0.536
OpenNMT + SC & BLEU	-0.903	0.234	0.890	0.371	-0.497	**0.391**	0.813	**0.528**	-0.442	0.403	**0.810**	**0.538**
GPT-2 base	-0.042	0.492	0.741	0.592	0.004	**0.541**	0.825	0.653	**0.006**	**0.549**	0.821	0.658
GPT-2 + SC	-0.048	0.492	**0.810**	**0.612**	-0.014	0.531	**0.919**	**0.673**	-0.001	0.543	0.917	0.682
GPT-2 + BLEU	**-0.041**	**0.497**	0.735	0.593	**0.006**	0.539	0.833	0.655	0.005	0.546	0.822	0.656
GPT-2 + SC & BLEU	-0.058	0.495	0.799	0.611	-0.014	0.530	0.903	0.668	-0.007	0.542	**0.923**	**0.683**
BART base	**0.035**	**0.547**	0.776	0.642	0.036	**0.572**	0.794	0.665	0.048	**0.578**	0.784	0.665
BART + SC	0.021	0.539	**0.882**	**0.669**	0.035	0.566	**0.872**	**0.686**	0.045	0.571	0.841	0.680
BART + BLEU	0.034	0.541	0.769	0.635	0.040	0.567	0.796	0.662	**0.050**	0.576	0.777	0.662
BART + SC & BLEU	0.030	**0.547**	0.855	0.667	**0.042**	0.562	0.817	0.666	0.044	0.577	**0.859**	**0.690**
BART large + SC & BLEU	0.035	0.560	0.847	0.674	0.070	0.585	0.900	0.709	0.072	0.584	0.886	0.704
	COMBINED TWO DOMAINS WITHOUT DOMAIN TAG											
BART base	**0.038**	**0.559**	0.731	0.634	0.050	**0.581**	0.795	0.671	**0.054**	**0.585**	0.809	0.679
BART + SC	0.031	0.546	**0.830**	0.659	0.043	0.575	**0.865**	**0.691**	0.039	**0.585**	**0.884**	**0.704**
BART + BLEU	0.033	0.555	0.743	0.635	0.042	0.575	0.810	0.673	**0.054**	0.583	0.814	0.679
BART + SC & BLEU	0.024	0.556	0.815	**0.661**	**0.054**	0.578	0.845	0.685	0.050	0.580	0.859	0.692
BART large + sc & BLEU	0.071	0.576	0.867	0.692	0.075	0.593	0.887	0.711	0.086	0.597	0.888	0.714
	COMBINED TWO DOMAINS WITH DOMAIN TAG											
BART base	**0.042**	0.552	0.754	0.637	0.054	0.579	0.748	0.653	**0.060**	0.582	0.787	0.669
BART + SC	0.035	0.555	0.831	0.666	0.039	0.571	0.833	0.678	0.046	0.579	**0.895**	**0.703**
BART + BLEU	0.039	0.554	0.745	0.635	**0.056**	0.578	0.745	0.651	0.049	**0.588**	0.825	0.685
BART + SC & BLEU	0.039	**0.556**	**0.845**	**0.671**	0.046	**0.580**	**0.834**	**0.684**	0.047	0.583	0.883	0.702
BART large + SC & BLEU	0.077	0.575	0.793	0.667	0.073	0.587	0.870	0.701	0.078	0.596	0.905	0.719

Table A.1.1: Evaluation results of x%-sized training sets (10%, 50% and 100%) on the E&M domain. The best score for each metric in each table section is boldfaced. BLEURT scores are calculated based on the BLEURT-base model with 128 tokens. Note that (i) Both BLEURT and BLEU are calculated against the four human references; (ii) ACC is the accuracy of the output labeled as the target style by the binary classifier; and (iii) HM is the harmonic mean of ACC and BLEU.

Model	BLEURT	BLEU	ACC	HM	BLEURT	BLEU	ACC	HM	BLEURT	BLEU	ACC	HM
Proportion of parallel training data		10%				50%				100%		
OpenNMT (Bi-LSTM)	-0.706	0.303	0.859	0.448	-0.304	0.449	0.792	0.573	-0.304	0.477	0.789	0.595
OpenNMT + SC	**-0.695**	**0.322**	**0.860**	**0.469**	-0.337	0.447	0.838	**0.583**	-0.289	0.466	0.824	0.595
OpenNMT + BLEU	-0.712	0.311	0.829	0.452	**-0.292**	**0.455**	0.808	0.582	**-0.246**	**0.478**	0.789	0.595
OpenNMT + SC & BLEU	-0.699	0.320	0.828	0.462	-0.332	0.444	**0.847**	**0.583**	-0.288	0.472	**0.848**	**0.606**
GPT-2 base	-0.020	**0.531**	0.775	0.630	**0.027**	**0.567**	0.841	0.677	**0.046**	0.576	0.850	0.687
GPT-2 + SC	-0.031	0.529	0.847	**0.651**	0.020	0.563	0.897	0.692	0.031	0.569	**0.916**	0.702
GPT-2 + BLEU	**-0.016**	0.529	0.786	0.632	0.026	0.566	0.838	0.676	0.041	**0.577**	0.860	0.691
GPT-2 + SC & BLEU	-0.027	0.528	**0.849**	**0.651**	0.015	0.562	**0.917**	**0.697**	0.038	0.572	0.915	**0.704**
BART base	**0.045**	0.565	0.719	0.633	0.071	0.589	0.786	0.673	**0.080**	0.600	0.801	0.686
BART + SC	0.041	0.569	**0.833**	0.676	0.061	**0.592**	0.869	0.704	0.067	0.601	0.874	**0.712**
BART + BLEU	0.041	0.566	0.719	0.633	**0.072**	0.590	0.789	0.675	0.078	**0.602**	0.798	0.686
BART + SC & BLEU	0.039	**0.571**	**0.833**	**0.678**	0.057	0.589	0.858	0.698	0.068	0.595	**0.882**	0.711
BART large + SC & BLEU	0.095	0.585	0.816	0.681	0.087	0.604	0.891	0.720	0.095	0.615	0.876	0.722
COMBINED TWO DOMAINS WITHOUT DOMAIN TAG												
BART base	**0.035**	**0.572**	0.734	0.643	0.060	0.592	0.821	0.688	**0.074**	0.604	0.807	0.691
BART + SC	0.026	0.563	**0.821**	0.668	0.056	0.592	**0.890**	**0.711**	0.054	0.602	**0.877**	**0.714**
BART + BLEU	0.033	0.568	0.732	0.640	**0.064**	0.593	0.834	0.693	0.073	**0.606**	0.831	0.701
BART + SC & BLEU	0.028	**0.572**	0.812	**0.671**	0.054	**0.596**	0.843	0.698	0.063	0.601	0.872	0.712
BART large + SC & BLEU	0.087	0.598	0.869	0.708	0.094	0.607	0.871	0.715	0.100	0.610	0.889	0.724
COMBINED TWO DOMAINS WITH DOMAIN TAG												
BART base	0.042	0.570	0.779	0.658	**0.072**	0.592	0.768	0.669	**0.078**	0.604	0.801	0.689
BART + SC	0.035	**0.574**	0.849	**0.685**	0.058	0.586	**0.861**	0.697	0.059	0.599	0.892	0.718
BART + BLEU	**0.047**	0.572	0.761	0.653	0.071	0.591	0.772	0.669	0.077	**0.605**	0.817	0.695
BART + SC & BLEU	0.043	0.573	**0.850**	**0.685**	0.057	**0.595**	0.849	**0.700**	0.064	0.603	**0.896**	**0.721**
BART large + SC & BLEU	0.089	0.590	0.801	0.679	0.099	0.604	0.869	0.713	0.100	0.611	0.900	0.728

Table A.1.2: Evaluation results of x%-sized training sets (10%, 50% and 100%) on the F&R domain. The best score for each metric in each table section is boldfaced. BLEURT scores are calculated based on the BLEURT-base model with 128 tokens. Note that (i) Both BLEURT and BLEU are calculated against the four human references; (ii) ACC is the accuracy of the output labeled as the target style by the binary classifier; and (iii) HM is the harmonic mean of ACC and BLEU.

(b) Comparison of our models with the other models

Domain	Model	BLEURT	BLEU	ACC	HM	Model	BLEURT	BLEU	ACC	HM
		(A) INFORMAL ↔ FORMAL				**(B) INFORMAL ⟶ FORMAL**				
	Rule-based (Rao and Tetreault, 2018)	-0.221	0.420	0.704	0.526	GPT-CAT (train on E&M, Wang et al. (2019))	0.170	0.713	0.905	0.801
	NMT-baseline (Rao and Tetreault, 2018)	-0.267	0.437	0.851	0.577	GPT-CAT (train on E&M and F&R, Wang et al. (2019))	0.176	0.725	0.876	0.793
	NMT-copy (Rao and Tetreault, 2018)	-0.269	0.441	0.808	0.571	S2S-SLS(Wang et al. (2020))	0.173	0.711	0.919	0.802
	NMT-Combined (Rao and Tetreault, 2018)	-0.100	0.501	0.797	0.615	Transformer (Zhang et al. (2020))	0.191	0.734	0.887	0.803
	PBMT-Combined (Rao and Tetreault, 2018)	-0.088	0.502	0.753	0.602	Chawla's (Chawla and Yang, 2020)	0.260	0.762	0.910	0.829
	GPT-2 + SC & BLEU (10% data, **Ours**)	-0.058	0.495	0.799	0.611	GPT-2 + SC & BLEU (train on E&M, **Ours**)	0.159	0.701	0.927	0.798
E&M	GPT-2 + SC & BLEU (100% data, **Ours**)	-0.007	0.542	**0.923**	0.683	BART + SC & BLEU (train on E&M, **Ours**)	0.218	0.730	0.887	0.801
	BART + SC & BLEU (10% data, **Ours**)	0.030	0.547	0.855	0.667	BART + SC & BLEU (train on E&M and F&R, **Ours**)	0.236	0.745	**0.937**	0.830
	BART + SC & BLEU (100% data, **Ours**)	**0.044**	**0.577**	0.859	**0.690**	BART large + SC & BLEU (train on E&M and F&R, **Ours**)	0.274	0.765	0.929	**0.839**
		(C) INFORMAL ↔ FORMAL & COMBINED DOMAINS				**(D) BLEU EVALUATED AGAINST THE FIRST REFERENCE**				
	Bi-directional FT (Niu et al., 2018)	0.023	0.554	0.818	0.661	*TS→CP (Sancheti et al., 2020)	-	0.292	-	-
	BART large + SC & BLEU (10% data, **Ours**)	0.077	0.575	0.793	0.667	GPT-2 + SC & BLEU (100% data, **Ours**)	-	0.296	-	-
	BART large + SC & BLEU (100% data, **Ours**)	**0.078**	**0.596**	**0.905**	**0.719**	BART + SC & BLEU (100% data, **Ours**)	-	**0.306**	-	-
		(A) INFORMAL ↔ FORMAL				**(B) INFORMAL ⟶ FORMAL**				
	Rule-based (Rao and Tetreault, 2018)	-0.226	0.450	0.738	0.559	*GPT-CAT (train on F&R, Wang et al. (2019))	-	0.773	-	-
	NMT-baseline (Rao and Tetreault, 2018)	-0.183	0.500	0.818	0.621	*GPT-CAT (train on E&M and F&R, Wang et al. (2019))	-	0.769	-	-
	NMT-copy (Rao and Tetreault, 2018)	-0.186	0.492	0.807	0.611	S2S-SLS(GPT-2, Wang et al. (2020))	0.244	0.766	0.857	0.809
	NMT-Combined (Rao and Tetreault, 2018)	-0.089	0.527	0.798	0.635	Transformer (Zhang et al. (2020))	0.246	0.770	0.890	0.827
	PBMT-Combined (Rao and Tetreault, 2018)	-0.062	0.517	0.788	0.624	Chawla's (Chawla and Yang, 2020)	0.302	**0.799**	0.910	0.851
F&R	GPT-2 + SC & BLEU (10% data, **Ours**)	-0.027	0.528	0.849	0.651	GPT-2 + SC & BLEU (train on F&R, **Ours**)	0.226	0.747	0.921	0.825
	GPT-2 + SC & BLEU (100% data, **Ours**)	0.038	0.572	**0.915**	0.704	BART + SC & BLEU (train on F&R, **Ours**)	0.271	0.770	0.897	0.829
	BART + SC & BLEU (10% data, **Ours**)	0.039	0.571	0.833	0.678	BART + SC & BLEU (train on F&R and E&M, **Ours**)	0.270	0.777	0.912	0.839
	BART + SC & BLEU (100% data, **Ours**)	**0.068**	**0.595**	0.882	**0.711**	BART large + SC & BLEU (train on F&R and E&M, **Ours**)	**0.324**	0.793	**0.920**	**0.852**
		(C) INFORMAL ↔ FORMAL & COMBINED DOMAINS				**(D) 10% PARALLEL TRAINING DATA (FROM PAPER)**				
	Bi-directional FT (Niu et al. (2018))	0.037	0.568	0.839	0.677	*CPLS (Shang et al., 2019)	-	0.379	-	-
	BART large + SC & BLEU (10% data, **Ours**)	0.089	0.590	0.801	0.679	GPT-2 + SC & BLEU (**Ours**)	-	0.528	-	-
	BART large + SC & BLEU (100% data, **Ours**)	**0.100**	**0.611**	**0.900**	**0.728**	BART + SC & BLEU (**Ours**)	-	**0.571**	-	-

Table A.1.3: Comparison of our models with the other models. The best score for each metric in each block is boldfaced. BLEURT scores are calculated based on the BLEURT-base model with 128 tokens. Notes: (i) if the output of a previous work is available, we re-calculate the scores using our evaluation metrics. Otherwise we take the scores from the paper and mark this with a (*); (ii) in (B) we report our results on informal-to-formal alone to compare with several systems which only transfer in this direction; (iii) in (C) we train systems on the concatenated data from both domains, to compare against Niu et al. (2018); (iv) in (E&M (D)) we re-evaluate our system against the first reference only, as this is what Sancheti et al. (2020) do.

A.2 Policy Gradient

Reinforcement learning (RL) is a sub-field of machine learning that is concerned with how intelligent agents ought to take actions in an environment in order to maximize the cumulative reward. Here, we employ the policy gradient algorithm (Williams, 1992) to maximize the expected reward (style strength and/or content preservation) of the generated sequence $\boldsymbol{y}^s$, whose gradient with respect to the parameters ϕ of the neural network model is estimated by sampling as:

$$
\begin{aligned}
\nabla_\phi J(\phi) &= R \cdot \nabla_\phi \sum_i P(\boldsymbol{y}_i^s | \boldsymbol{x}_i; \phi) \\
&= \sum_i P(\boldsymbol{y}_i^s | \boldsymbol{x}_i; \phi) R_i \nabla_\theta \log(P(\boldsymbol{y}_i^s | \boldsymbol{x}_i; \phi)) \\
&\simeq \frac{1}{N} \sum_{i=1}^N R_i \nabla_\phi \log(P(\boldsymbol{y}_i^s | \boldsymbol{x}_i; \phi)) \\
&= E[R \cdot \nabla_\phi log(P(\boldsymbol{y}^s | \boldsymbol{x}; \phi))]
\end{aligned}
\tag{8}
$$

where $J(\cdot)$ is the objective function, $\nabla_\phi J(\cdot)$ is the gradient of $J(\cdot)$ with respect to ϕ, R_i is the reward of the i_{th} sequence $\boldsymbol{y}^s$ that is sampled from the distribution of model outputs at each decoding time step, ϕ are the parameters of the model, N is the sample size, and $E(\cdot)$ is the expectation.

Regarding the reward of style classification for GPT-2 based model, we design two rewards (Eq. 4 and Eq. 5) for source sentence and target sentence, respectively. The policy gradient is then

$$
\begin{aligned}
\nabla_\phi J(\phi) = & E[R_{cls_{source}} \cdot \nabla_\phi log(P(\boldsymbol{y}_{source}^s | \boldsymbol{x}_{source}; \phi))] \\
& + E[R_{cls_{target}} \cdot \nabla_\phi log(P(\boldsymbol{y}_{target}^s | \boldsymbol{x}_{source,target}; \phi))]
\end{aligned}
\tag{9}
$$

A.3 Example Outputs of Various Models

System	From informal to formal	BLEURT	BLEU	ACC
Source	i say omarion.he has the hair clothes and body,a triple deal on one person.		-	
Reference 1	My choice is Omarion as he has high quality, hair, clothes, and body to create a triple deal in one person.		-	
Reference 2	I would say Omarion because he has the hair, clothes, and body; A triple deal on a single person.		-	
Reference 3	I pick Omarion, he has the hair, the clothes, and the body. A triple deal on one person.		-	
Reference 4	Omarion has the hair, clothes, and the body.		-	
PBMT-Combined (Rao and Tetreault, 2018)	I say omarion. he has the hair, clothes and body, the deal on one person.	-0.153	0.509	0.946
Bi-directional FT (Niu et al., 2018)	I say Omarion, he has the hair clothes and body, and a triple deal on one person.	-0.149	0.510	0.953
GPT-CAT (Wang et al., 2019)	I say Omarion. He has the hair, clothes, and body, a triple deal on one person.	0.044	0.585	**1.000**
S2S-SLS (Wang et al., 2020)	I say Omarion. He has the hair clothes and body, a triple deal on one person.	-0.035	0.350	**1.000**
Transformer (Zhang et al., 2020)	I say omarionhe has the hair clothes and body, a triple deal on one person.	-0.255	0.462	0.892
Chawla's (Chawla and Yang, 2020)	I say Marion because he has the hair, clothes and body, a triple deal on one person.	-0.538	0.534	0.989
OpenNMT	He has the hair clothes and body.	-0.540	0.139	0.998
OpenNMT with SC	I say Omarion, he has the hair clothes and body.	-0.389	0.558	0.969
OpenNMT with BLEU	I say Omarion. He has the hair clothes and body.	-0.325	0.147	**1.000**
OpenNMT with SC & BLEU	I say Omarion. He has the hair clothes and body.	-0.325	0.147	**1.000**
GPT-2 base	I say Omarion. He has the hair and body, a triple deal on one person.	-0.087	0.342	**1.000**
GPT-2 + SC	I say Omarion because he has the hair clothes and body.	-0.264	0.634	0.976
GPT-2 + BLEU	I say Omarion. He has the hair clothes and body, a triple deal on one person.	-0.035	0.350	**1.000**
GPT-2 + SC & BLEU	I say Omarion. He has the hair clothes and body, a triple deal on one person.	-0.035	0.350	**1.000**
BART base	I would say Omar. He has the hair, clothes, and body. It is a triple deal on one person.	-0.012	0.589	**1.000**
BART + SC	I would say Omar. He has the hair, clothes, and body. It is a triple deal on one person.	-0.012	0.589	**1.000**
BART + BLEU	I would say Omar. He has the hair, clothes, and body of a triple deal on one person.	-0.230	0.600	**1.000**
BART + SC & BLEU	I would say Omar. He has the hair, clothes, and body. It is a triple deal on one person.	-0.012	0.589	**1.000**
BART large + SC & BLEU	I would say Omarion. He has the hair, clothes, and body, a triple deal on one person.	**0.096**	**0.657**	**1.000**

System	From formal to informal	BLEURT	BLEU	ACC
Source	I suggest avoiding hot dogs, and not watching this movie with your little sister.		-	
Reference 1	Don't eat hot dogs, or watch this movie with your little sister!		-	
Reference 2	Don't do hot dogs or this movie with your kid sister.		-	
Reference 3	don't eat hot dogs and don't watch it w/ ur lil sis!		-	
Reference 4	Don't eat hot dogs or watch this flick with your lil sis!		-	
PBMT-Combined (Rao and Tetreault, 2018)	I suggest avoiding hot dogs, and not watching this movie with your little sister.	-0.298	0.417	0.004
Bi-directional FT (Niu et al., 2018)	I suggest avoiding hot dogs and not watching this movie with your little sister.	-0.233	0.437	0.009
OpenNMT	hott dogs and not watching this movie with ur little sister	-0.885	0.118	**1.000**
OpenNMT with SC	Im not watching this movie with your little sister...I suggest him hot dogs.	-0.765	0.349	0.981
OpenNMT with BLEU	Well, and not watching this movie with your little sister.	-0.826	0.445	0.633
OpenNMT with SC & BLEU	Can't watch this movie with your little sister.	-0.521	0.542	0.783
GPT-2 base	Don't watch this movie with your little sister.	-0.415	0.573	0.851
GPT-2 + SC	don't watch this movie with your little sister.	-0.415	0.599	**1.000**
GPT-2 + BLEU	Don't watch this movie with your little sister!	-0.360	0.634	0.919
GPT-2 + SC & BLEU	don't watch this movie with your little sister.	-0.415	0.599	**1.000**
BART base	avoid hot dogs and not watch this movie with your little sister.	**-0.016**	0.610	0.925
BART + SC	avoid hot dogs and not watch this movie with your little sister.	**-0.016**	0.610	0.925
BART + BLEU	avoid hot dogs and not watching this movie with your little sister.	-0.034	0.514	0.910
BART + SC & BLEU	avoid hot dogs and not watch this movie with your little sister.	**-0.016**	0.610	0.925
BART large + SC & BLEU	Avoid hot dogs and don't watch this movie with your little sister.	-0.171	**0.800**	0.825

Table A.3.1: Sample model outputs and their sentence-level scores on the E&M domain, where red denotes improperly generated words or content. Note that ACC indicates style confidence here.