Text Style Transfer via Optimal Transport

Nasim Nouri
Raouf Medical Group
Tehran, Iran
nasimnouri@raoufmed.com

Abstract

Text style transfer (TST) is a well-known task whose goal is to convert the style of the text (e.g., from formal to informal) while preserving its content. Recently, it has been shown that both syntactic and semantic similarities between the source and the converted text are important for TST. However, the interaction between these two concepts has not been modeled. In this work, we propose a novel method based on Optimal Transport for TST to simultaneously incorporate syntactic and semantic information into similarity computation between the source and the converted text. We evaluate the proposed method in both supervised and unsupervised settings. Our analysis reveal the superiority of the proposed model in both settings.

1 Introduction

Text style transfer (TST) is an important task in NLP that aims to change the style of a given text from source style to target style (e.g., formal to informal) while preserving its content. For instance, the formal sentence "*However, I do believe it to be punk*" is converted to the informal equivalent sentence "*I'd say it is punk though*". This task could be helpful for downstream applications such as text simplification, information extraction, and question answering.

Due to the importance of TST, this task has been approached with different techniques ranging from feature-based models (Xu et al., 2012) to recent advanced deep learning solutions (Chen et al., 2018; Lee et al., 2021a; Huang et al., 2021). The recent work can be categorized as supervised (i.e., parallel corpus with sentences in source and target style) (Lai et al., 2021), unsupervised (i.e., sentences in source and target style are available but they are not aligned) (Krishna et al., 2020), or semi-supervised (combination of parallel and non-aligned corpora) (Chawla and Yang, 2020) methods. The three critical objectives of any TST system are to (1) generate a text in the target style, (2) keep the content of the source text, and (3) generate fluent sentences (Krishna et al., 2020). It has been shown that fine-tuning transformer-based language models on each of these objectives (i.e., using Reinforcement Learning) can achieve promising results (Lai et al., 2021; Liu et al., 2021). However, one of the limitations of the existing works is that the content preservation (i.e., the second objective) is fulfilled at either the surface-form level (i.e., by encouraging the same words to appear in both texts) (Lai et al., 2021) or at the semantics level (i.e., by encouraging high mutual information between the two texts) (Chawla and Yang, 2020); ignoring the role of syntactic information. Syntactic information (e.g., dependency tree) can be used to explicitly encode the connections between the words of the sentence, thereby playing an important role in the equivalency of two sentences. For instance, consider the source sentence "*a crap touch bar with a nice screen!!!*" and the converted sentence "*The screen is great but the touch bar is terrible*". The corresponding dependency between "*touch bar* $\rightarrow$ *crap*" in the source sentence and "*terrible* $\rightarrow$ *touch bar*" in the target sentence and also "*screen* $\rightarrow$ *nice*" in the source sentence and "*great* $\rightarrow$ *screen*" in the target sentence are helpful to assess the equivalency of the two sentences. Although the pre-trained language models such as BERT have been shown to be able to encode the syntactic information, it is not yet verified that these models can take into account the syntactic dependencies when computing the similarity between two sentences, especially for the TST task. To the best of our knowledge, there is one prior work that shows the importance of the syntactic information for transformer-based TST models (Ma et al., 2019). Specifically, Ma et al. (2019) shows that reconstructing both the words of the source text and their POS tags could boost the performance of TST. However, there are two limi-

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2532 - 2541

tations in this work: (1) the syntactic structure (i.e., dependencies between words) is ignored; (2) the interaction between semantics and the syntax of the sentences is neglected. More specifically, to obtain the most value of the syntactic information, it is crucial to consider the relations between the words and also their semantics as shown in the example above. As such, in this work, we propose a novel method to simultaneously incorporate the interaction between syntax and semantics of the sentences into the content preservation objective of TST training. More specifically, for the first time in text style transfer, we propose to use Optimal Transport (OT) as an efficient method to consider both the syntax and the semantics of the two sentences when computing their content similarity. OT has been shown to be an effective method for image style transferring (Kolkin et al., 2019; Risser, 2020) and our work exhibits its application for the domain of the text. We evaluate the proposed model on three benchmark datasets and two settings, i.e., supervised and unsupervised. Our extensive analysis reveals the effectiveness of the proposed model by establishing new state-of-the-art results.

2 Model

Problem Definition: The task of text style transfer is formally defined as follows: Given the input sentence $D = [w_1, w_2, \ldots, w_n]$ with style s, the goal is to generate a new sentence $D' = [w'_1, w'_2, \ldots, w'_m]$ in the target style t while preserving the content of D in D'. We study both supervised and unsupervised settings. Specifically, in the supervised setting, for every training sample (D, s) there is an aligned sentence $\bar{D}$ in target style t, i.e., $(\bar{D}, t)$, whose content is the same as D. Whereas for the unsupervised setting, there is no equivalent pair for (D, s). Note that in the unsupervised setting, there are sentences for both styles.

In this work, we employ a transformer-based generative language model, i.e., GPT-2 (Radford et al., 2019), for TST and we train the model using REINFORCE algorithm. Specifically, the source sentence D is prompted to the GPT-2 model to generate the target sentence D'. Following the prior work, (Lai et al., 2021), the GPT-2 model is encouraged to generate the sentence D' in the target style t and with the same content as D. Also, in the supervised setting, we use the gold target sentence $\bar{D}$ as an additional supervision signal to train the model. Since $\bar{D}$ is not available in the unsupervised setting, we follow the prior work (Lee et al., 2021a) to use reconstruction loss as an additional training signal. The rest of this section provides details for generating sentences, rewards for generation, and training procedures.

2.1 Generating Target Sentence

Following the prior work (Lai et al., 2021), we employ the input sentence D as a prompt to GPT-2 model to generate the target sentence D'. More specifically, the prompt to GPT-2 consists of the sequence $P = [BOS, w_1, w_2, \ldots, w_n, SEP]$, where BOS and SEP are special token indicating the beginning and the end of the input sentence D. In addition to the input document D, during training of the supervised model, the gold target sentence $\bar{D} = [\bar{w}_1, \bar{w}_2, \ldots, \bar{w}_{n'}]$ is concatenated to the prompt to create the training sequence $S = [BOS, w_1, w_2, \ldots, w_n, SEP, \bar{w}_1, \bar{w}_2, \ldots, \bar{w}_{n'}]$ and the model is trained in an auto-regressive manner:

$$\mathcal{L}_{LM} = \sum_{i}^{n+n'+2} -\log(Q(S_i|S_{<i}, \theta)) \quad (1)$$

where θ is GPT-2 parameters and $Q(\cdot|S_{<i}, \theta)$ is the distribution over vocabulary obtained from the last hidden states of GPT-2 model. During inference, only the prompt P is provided to the GPT-2 model and the words of D' are sampled from the distribution predicted by GPT-2 model until EOS is sampled.

Unlike the supervised setting in which the GPT-2 model is trained for uni-directional style conversion, i.e., from the source style to the target style, in the unsupervised setting, the model is trained for both directions, i.e., from the source to the target and vice versa. Specifically, given a sentence and a style, the GPT-2 model is trained to generate another sentence with the same content in the given style. Formally, the prompt P is concatenated with the style st where $st \in \{s, t\}$, i.e., $S = [BOS, w_1, w_2, \ldots, w_n, SEP, st]$. To train the model, following the prior work (Lee et al., 2021a), two types of reconstruction loss are employed:

Self-Reconstruction: The GPT-2 model is encouraged to reconstruct the original input sentence D when st is s, i.e., the given style to the model is the

same as the input sentence style. Concretely, the loss function $\mathcal{L}_{LM}$ is defined as follows:

$$\mathcal{L}_{LM} = \sum_{i}^{n} -\log(Q(D_i|D, s, \theta)) \quad (2)$$

Cycle Reconstruction: If st is t, i.e., the given style to the model is different from the style of the input sentence, then the GPT-2 model is first employed to generate the sentence $\bar{D}'$ in the style t. Next, the model is encouraged to reconstruct the original input sentence D using the input $S = [BOS, \bar{w}_1, \bar{w}_2, \ldots, \bar{w}_{n'}, SEP, s]$, where $\bar{w}_i$ is the i-th word in the generated sentence $\bar{D}'$. Concretely, the loss function $\mathcal{L}_{LM}$ is defined as follows:

$$\mathcal{L}_{LM} = \sum_{i}^{n} -\log(Q(D_i|\bar{D}', s, \theta)) \quad (3)$$

2.2 Rewarding GPT-2 Model

Prior work shows that rewarding generative models to observe the requirements for TST could improve the performance (Lai et al., 2021; Liu et al., 2021). Hence, we follow this optimization step to update the GPT-2 model based on two different rewards, i.e., Style Conversion and Content Preservation.

Style Conversion: One of the critical objectives of TST is to change the style of the given text. To encourage the model for this objective, prior works commonly use a pre-trained discriminator to predict the style of the generated text. Here, we follow the same approach by pre-training a BERT model (Devlin et al., 2019) on the combination of the training sentences D in both styles to identify the style of the given text (i.e., a binary text classification task). Next, during the training stage of the GPT-2 model, we send the generated sentence D' to the pre-trained BERT model. The probability of the target style is employed as the style conversion reward: $R_{SC}(D') = Q_{BERT}(t|D', \phi)$, where ϕ is the BERT parameters.

Content Preservation Content preservation is an important requirement of TST and prior works use either surface form of D and D' (Lai et al., 2021; Huang et al., 2021), their semantics (Chawla and Yang, 2020), or only shallow syntax (Ma et al., 2019) to compute the content overlap between the source and the generated sentence. None of these works consider the syntactic structure of two sentences and more importantly its interaction with the semantics of the sentence. As the main novelty of the proposed work, inspired by the success of Optimal Transport in image style transfer (Kolkin et al., 2019; Risser, 2020) and other related NLP tasks (Xu et al., 2021), we show that OT is an effective tool for addressing the shortcoming of syntax-semantics interaction for content preservation in prior TST literature.

To represent the semantics of the source and target sentence D and D', we employ the hidden states of the final GPT-2 layer for each word w_i and w'_j, i.e., $H = [h_1, h_2, \ldots, h_n]$ and $H' = [h'_1, h'_2, \ldots, h'_n]$. Moreover, the syntactic structures of the two sentences are obtained from an off-the-shelf dependency tree parser[1], represented by T and T'. The criterion we use to compute the content preservation between two sentences D and D' is that the semantically related words in both sentences should have the same syntactic importance too. In particular, we expect that similar words appear at the same level in the dependency tree of T and T'. However, since the structure of the sentence might change during style conversion and also the number of words might alter, similar words might appear in other levels too. As such, the optimal mapping between similar words in the dependency trees T and T' is not trivial. Fortunately, optimal transport (OT) can be helpful to solve this issue. OT is a mathematical method to compute the cheapest plan for converting one data distribution to another one. We first formally describe OT and then we elaborate on how it is employed for our purpose.

OT is an established method to find the optimal plan to convert (i.e., transport) one distribution to another one. Formally, given the probability distributions $p(x)$ and $q(y)$ over the domains $\mathcal{X}$ and $\mathcal{Y}$, and the cost/distance function $C(x, y) : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ for mapping $\mathcal{X}$ to $\mathcal{Y}$, OT finds the optimal joint alignment/distribution $\pi^*(x, y)$ (over $\mathcal{X} \times \mathcal{Y}$) with marginals $p(x)$ and $q(y)$, i.e., the cheapest transportation from $p(x)$ to $q(y)$, by solving the following problem:

$$\pi^*(x, y) = \min_{\pi \in \Pi(x,y)} \int_{\mathcal{Y}} \int_{\mathcal{X}} \pi(x, y) C(x, y) dx dy \quad (4)$$
$$\textbf{s.t. } x \sim p(x) \text{ and } y \sim q(y),$$

where $\Pi(x, y)$ is the set of all joint distributions with marginals $p(x)$ and $q(y)$. Note that if the distributions $p(x)$ and $q(y)$ are discrete, the integrals in Equation 4 are replaced with a sum and the joint distribution $\pi^*(x, y)$ is represented by a

[1]We employ Stanford dependency parser

		0 → 1		1 → 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: Statistics of GYAFC dataset employed for supervised setting (number of samples). 0=informal, 1=formal

Dataset	Style	Train	Dev	Test
Yelp	Positive	266,041	2,000	500
	Negative	177,218	2,000	500
IMDB	Positive	178,869	2,000	1,000
	Negative	187,597	2,000	1,000

Table 2: Statistics of the IMDB and Yelp datasets employed for unsupervised setting (number of samples).

matrix whose entry (x, y) ($x \in \mathcal{X}, y \in \mathcal{Y}$) represents the probability of transforming the data point x to y to convert the distribution $p(x)$ to $q(y)$. Finally, the cost of optimal conversion (i.e., Wasserstein distance $Dist_W$) is computed by: $Dist_W = \Sigma_{x\in\mathcal{X}}\Sigma_{y\in\mathcal{Y}}\pi^*(x, y)C(x, y)$.

In our method, we use the words $w_i \in D$ as the domain $\mathcal{X}$ and the words $w'_j \in D'$ as the domain $\mathcal{Y}$. In order to define their distance, we use the Euclidean distance between their semantic vector representation $C(w_i, w'_j) = \left\|h_i - h'_j\right\|$. Finally, to define the distributions $p(x)$ and $q(y)$, we use the level of words w_i and w'_j in the dependency tree T and T', respectively. Concretely, $p(w_i) = softmax(M - L_i)$, where M is the maximum depth of T, L_i is the depth of w_i in T and $softmax$ is computed over all words $w_i \in D$. Similarly, $q(w'_j)$ is defined by $q(w'_j) = softmax(M' - L'_j)$. By solving the equation 4[2], the cheapest conversion of the two sentence D and D' is obtained and its cost is equal to Wasserstein distance $Dist_W$. We use this distance as the content preservation penalty, i.e., $R_{CP}(D') = -Dist_w$.

3 Training

To train the model, we combine the content preservation reward $R_{CP}(D')$, with style conversion and the language model loss. We use REINFORCE algorithm (Williams, 1992) to train the model. In particular, the GPT-2 model is trained on the combination of the language model loss, i.e., $\mathcal{L}_{LM}$ and the rewards of style conversion and content preservation. The REINFORCE algorithm is employed to incorporate rewards into fine-tuning of GPT-2. First, the overall reward is computed by $R(D') = R_{SC}(D') + \alpha R_{CP}(D')$, where α is a trade-off hyper-parameter. Next, we seek to minimize the negative expected reward $R(D')$ over the possible choices of D': $\mathcal{L}_R = -\mathbb{E}_{\hat{D}'\sim P(\hat{D}'|D)}[R(\hat{D}')]$. The policy gradient is then estimated by: $\nabla\mathcal{L}_R = -\mathbb{E}_{\hat{D}'\sim P(\hat{D}'|D)}[(R(\hat{D}') - b)\nabla \log P(\hat{D}'|D)]$. Using one roll-out sample, we further estimate $\nabla\mathcal{L}_R$ via the generated sentence D': $\nabla\mathcal{L}_R = -(R(D') - b)\nabla \log P(D'|D)$ where b is the baseline to reduce variance. In this work, we obtain the baseline b via: $b = \frac{1}{|B|}\sum_{i=1}^{|B|} R(D'_i)$, where $|B|$ is the mini-batch size and D'_i is the generated sentence for the i-th sample in the mini-batch.

4 Experiments

Datasets: We evaluate the proposed model, i.e., Optimal Transport-based Text sTyle Transfer (OT4), in two different settings, i.e., supervised and unsupervised. In the supervised setting we employ the Grammarly's Yahoo Answers Formality Corpus (GYAFC) dataset (Rao and Tetreault, 2018). GYAFC is a parallel dataset in two domains Entertainment & Music (E&M) and Family & Relationships (F&R). Table 1 shows the statistics of this dataset.

For the unsupervised setting, we employ two commonly used datasets: Yelp (Li et al., 2018) and IMDB (Dai et al., 2019) review. Both datasets contain sentiment-annotated reviews. The text style transfer on these datasets is defined as sentiment polarity conversion. In particular, given a sentence with a specific sentiment polarity (e.g., positive), the goal is to generate a new sentence with the opposite sentiment polarity (e.g., negative). Note that no parallel data is available for the sentences in these datasets. The statistics of both datasets are provided in Table 2

Evaluations: We validate the model performance using both automatic and human evaluation. For the automatic evaluation in the supervised setting, following the prior work (Lai et al., 2021), we assess the performance of the models based on: (1) Style Strength (ACC): The binary style classifier TexCNN (Kim, 2014) (with 87.0% and 89.3% accuracy on E&M and F&R domains, respectively) is employed to predict the strength of the style conversion; (2) Content Preservation (BLEU): The BLEU

[2]Note that as solving the OT problem in Equation 4 is intractable, we employ the entropy-based approximation of OT and solve it with the Sinkhorn algorithm (Peyre and Cuturi, 2019).

score computed using four reference sentences; (3) HM: The harmonic mean of ACC and BLEU; and (4) BLEURT: A new metric for content preservation proposed by Sellam et al. (2020). For the automatic evaluation in the unsupervised setting, following prior work (Lee et al., 2021b), we use: (1) Style Transfer Accuracy (S-ACC): Following (Lee et al., 2021b), a Bi-GRU layer with attention mechanism, trained for style classification on IMDB and Yelp dataset, is employed to assess the style transfer; (2) Content Preservation (self-BLEU, ref-BLEU, BERT-P, BERT-R, and BERT-F1): To validate the content preservation in the generated sentence, its BLEU score with input sentence, i.e., self-BLEU, and with the human-generated sentence, i.e., ref-BLEU, are used. Moreover, to incorporate contextual semantics, the BERT score proposed by (Zhang et al., 2020) is employed to assess the similarity between the generated sentence and the human reference. Following prior work (Lee et al., 2021b), we report precision, recall, and F1 score for this metric; (3) Fluency (PPL): The fluency of the generated sentences is evaluated based on the perplexity of the sentences using the 5-gram KenLM (Heafield, 2011) model trained on both datasets;

For human evaluation, following prior work (Lee et al., 2021b), we randomly select 150 documents for each test set and we hire 4 annotators to rate model predictions from 1 (Very Bad) to 5 (Very Good) on content preservation, style conversion, and fluency. For each annotator, we provide them with the source text, source style, target style, and model-generated text.

Baselines: We compare our model with the prior state-of-the-art models in each setting. Specifically, for the supervised setting on GYAFC, we compare our model with **GPT-2 + SC & BLEU** (Lai et al., 2021): Similar to our model, this baseline employs GPT-2 to generate the target sentence. The generative model is trained using rewards for style conversion (SC) and content preservation (BLEU); **BART + SC & BLEU** (Lai et al., 2021): The same as the previous baseline with the difference of using BART instead of GPT-2; **NMT-Combined** (Rao and Tetreault, 2018): This baseline casts TST as a machine translation problem and employs attention based BiLSTM encoder-decoder architecture; **Bi-directional FT** (Niu et al., 2018): This model employs BiLSTM encoder to jointly learn text formality style transfer in both direction (from formal to informal and vice versa); **CPLS** (Shang et al., 2019): This baselines employs an encoder-decoder architecture to obtain latent space representation of the styles, then a projection model converts the styles in the latent space; **GPT-CAT** (Wang et al., 2019): This baseline combines rule-based methods with neural-based TST systems. GPT-2 is employed as the neural component; **TS→CP** (Sancheti et al., 2020): This model exploits reinforcement learning to explicitly encourage content preservation and transfer strength. It exerts BLEU score between generated and ground-truth sentence to compute content preservation reward; and **Chawla's** (Chawla and Yang, 2020): This baseline uses a language model discriminator to guide the text formality style transfer. For content preservation, it employs mutual information between source and target sentence.

For the unsupervised setting on IMDB and Yelp, we compare with **Cross-Alignment** (Shen et al., 2017): This baseline is trained to generate a sentence in the target style that could match the example sentences in the source style. To this end, a cross-aligned auto-encoder is utilized; **Controlled-Gen** (Hu et al., 2017): This model employs variational author encoder (VAE) with attribute discriminators to impost semantic structure, including text style; **Style Transformer** (Dai et al., 2019): In this baseline, a transformer model is employed to directly takes the input sentence and target style to generate the target sentence; **Deep Latent** (He et al., 2020): This baseline models the unsupervised text style transfer as the task of inferring latent variables, i.e., target sentences, on the partially observed data of each style. A recurrent language model is employed to fulfill the objective. **RACoLN** (Lee et al., 2021b): In this baseline, the reverse attention technique is employed to remove style information from the representations of the tokens in the source sentence.

4.1 Results

Supervised: Table 3 shows the results of the evaluations on the test set. Following prior work, we compare the performance of the proposed OT4 model in the following settings: (1) **Informal** ↔ **Formal**: In this setting the performance of the baselines for converting a formal to informal text or vice versa is evaluated. From this table, we observe that GPT-2 model has a better capability of style conversion. However, the baseline model using GPT-2

Domain	Model	BLEURT	BLEU	ACC	HM	Model	BLEURT	BLEU	ACC	HM
E&M	(A) INFORMAL ↔ FORMAL					(B) INFORMAL → FORMAL				
	NMT-Combined	-0.100	0.501	0.797	0.615	GPT-CAT (train on E&M and F&R)	0.176	0.725	0.876	0.793
	BART + SC & BLEU	0.044	0.577	0.859	0.690	Chawla's	0.260	0.762	0.910	0.829
	GPT-2 + SC & BLEU	-0.007	0.542	0.923	0.683	BART large + SC & BLEU	0.274	0.765	0.929	0.839
	OT4 (Ours)	**0.102**	**0.602**	**0.949**	**0.736**	OT4 (Ours)	**0.322**	**0.812**	**0.951**	**0.876**
	(C) INFORMAL ↔ FORMAL & COMBINED DOMAINS					(D) BLEU EVALUATED AGAINST THE FIRST REFERENCE				
	Bi-directional FT	0.023	0.554	0.818	0.661	TS→CP	-	0.292	-	-
	BART large + SC BLEU	0.078	0.596	0.905	0.719	BART + SC & BLEU	-	0.306	-	-
	OT4 (Ours)	**0.192**	**0.671**	**0.939**	**0.782**	OT4 (Ours)	-	**0.352**	-	-
F&R	(A) INFORMAL ↔ FORMAL					(B) INFORMAL → FORMAL				
	NMT-Combined	-0.089	0.527	0.798	0.635	GPT-CAT (train on E&M and F&R)	-	0.769	-	-
	BART + SC & BLEU	0.068	0.595	0.882	0.711	Chawla's	0.302	0.799	0.910	0.851
	GPT-2 + SC & BLEU	0.038	0.572	0.915	0.704	BART large + SC & BLEU	0.324	0.793	0.920	0.852
	OT4 (Ours)	**0.112**	**0.618**	**0.942**	**0.746**	OT4 (Ours)	**0.401**	**0.825**	**0.961**	**0.887**
	(C) INFORMAL ↔ FORMAL & COMBINED DOMAINS					(D) 10% PARALLEL TRAINING DATA				
	Bi-directional FT	0.037	0.568	0.839	0.677	CPLS	-	0.379	-	-
	BART large + SC & BLEU	0.100	0.611	0.900	0.728	BART + SC & BLEU	-	0.571	-	-
	OT4 (Ours)	**0.185**	**0.652**	**0.933**	**0.767**	OT4 (Ours)	-	**0.644**	-	-

Table 3: Automatic evaluation result on GYAFC dataset. The performance of the baselines are taken from (Lai et al., 2021).

Model	S-ACC	ref-BLEU	self-BLEU	PPL	G-score	BERT-P	BERT-R	BERT-F1
Cross-Alignment	74.2	4.2	13.2	53.1	32.0	87.8	86.2	87.0
ControlledGen	83.7	16.1	50.5	146.3	65.0	90.6	89.0	89.8
Style Transformer	87.3	19.8	55.2	73.8	69.4	91.6	89.9	90.7
Deep Latent	85.2	15.1	40.7	**36.7**	58.9	89.8	88.6	89.2
RACoLN	91.3	20.0	59.4	60.1	73.6	91.8	90.3	91.0
OT4	**93.4**	**26.7**	**71.2**	42.1	**81.4**	**97.7**	**96.7**	**97.2**

Table 4: Automatic evaluation result on Yelp dataset. G-Score is the geometric mean of self-BLEU and S-ACC. The evaluation results of the baselines are taken from (Lee et al., 2021b)

	S-ACC	self-BLEU	PPL	G-score
Cross-Alignment	63.9	1.1	**29.9**	8.4
ControlledGen	81.2	63.8	119.7	71.2
Style Transformer	74.0	70.4	71.2	72.2
Deep Latent	59.3	64.0	41.1	61.6
RACoLN	83.1	70.9	45.3	76.8
OT4	**86.2**	**80.5**	39.8	**83.30**

Table 5: Automatic evaluation result on IMDB dataset. Since human references are not available for IMDB dataset, self-BLEU and BERT scores are omitted. The evaluation results of the baselines are taken from (Lee et al., 2021b)

employs BLEU score to encourage content preservation. In contrast, we equip our GPT-2 model with OT-based reward that can incorporate both semantics and syntax of the sentences and achieve the best results; (2) **Informal → Formal**: In this setting, only the conversion from informal to formal text is evaluated. compared to the previous setting, we see an improvement in the style conversion and content preservation for the equivalent models. It shows that this direction of conversion is relatively easier. However, the proposed OT4 model still significantly outperform the baselines in this setting too, indicating the importance of content preservation for this setting; (3) **Informal ↔ Formal & Combined Domains**: In this setting, the data from both domains are combined for training and the model is evaluated for conversion in both direction. The results show that in this setting, all baselines benefit from the extra training data from the other domain, however, the proposed OT4 enjoys the largest improvement. Our hypothesis for such improvement in OT4 is that the existence of the other domain data provides more syntactic structure to the model, therefore, compared to the baselines that miss this information, the proposed OT4 baseline can benefit from more training signals. (5) **Evaluation with the first reference**: To conduct a comprehensive comparison, we also compare our model with the baselines that only report the performance of the models evaluated on the first reference sentence. In this setting, we see that the proposed model achieves the highest BLEU score; and finally (6) **10% Parallel Data**: To show the effectiveness of the proposed model in the case of low-resource setting, we compare the performance of the proposed model when only 10% of the training parallel data is employed. We see in this setting the improvement achieved by our proposed model is higher, especially in terms of BLEURT, reflecting its superiority to benefit more

	Yelp			IMDB		
	Style	Content	Fluency	Style	Content	Fluency
Cross-Alignment	2.5	2.1	3.2	2.0	2.0	1.9
ControlledGen	3.1	3.5	3.4	3.0	3.4	3.2
Style Transformer	3.3	3.6	3.5	3.5	3.1	3.5
Deep Latent	3.5	2.9	4.1	2.6	3.1	3.1
RACoLN	3.6	3.7	3.8	3.3	3.5	3.7
OT4 (Ours)	**4.1**	**4.6**	**4.2**	**3.9**	**4.6**	**4.0**

Table 6: Human evaluation of the baseline outputs for Yelp and IMDB datasets. Numbers are the average score of the four annotators.

Model	Style	Content	Fluency
TS→CP	2.1	2.5	2.7
CPLS	3.2	2.6	3.0
NMT-Combined	3.0	2.6	2.2
Bi-directional FT	3.3	3.0	2.0
Chawla's	2.9	2.7	3.4
GPT-CAT	3.4	3.4	3.8
GPT-2 + SC & BLEU	3.1	3.4	3.1
BART + SC & BLEU	3.1	3.3	3.5
OT4 (Ours)	**3.9**	**4.2**	**3.9**

Table 7: Human evaluation of the baseline outputs for GYAFC dataset. Numbers are the average score of the four annotators.

efficiently from training signals.

Unsupervised: The results of the evaluation of the unsupervised model on the Yelp and IMDB datasets are presented in Table 4 and 5, respectively. Note that due to the lack of reference target sentences in the IMDB dataset, we omit self-BLEU and BERT scores in Table 5. There are several observations from these tables. First, the proposed OT4 model outperforms all baselines with respect to style conversion and content preservation. Specifically, for style conversion, our model improves the S-ACC on Yelp and IMDB by 2.1% and 3.1%, respectively. Compared the baselines, we attribute the style conversion improvement to the explicit rewards employed in our model to directly train the model for better style conversion. More importantly, since our model is equipped with OT to improve content preservation, we see a significant improvement for this metric. In particular, on the Yelp dataset, our model improves BERT-F1 score by 6.2% and self-BLEU by 11.8%. Considering the improvement on ref-BLEU on this dataset also indicates that while our model improves the content preservation, it is not repeating the input sentence. Finally, comparison of the fluency of the generated sentences shows that our model is competitive with baselines by achieving the second-lowest PPL on both datasets.

4.2 Ablation Study

In order to shed more light on the contribution of the proposed OT-based content preservation reward, in this section we study the performance of alternative architecture designs: (1) **No Semantics**: Here, the cost function $C(x, y)$ is replaced by the constant function $C(x, y) = 1$, hence removing all information regarding the semantics of the sentence; (2) **No Syntax**: In this baseline, the probability distribution $p(x)$ and $q(y)$ are represented by uniform distribution, thus removing all information about the syntactic structure; (3) **NO OT**: In this baseline, the content preservation reward is completely removed; (4) **Reconstruct**: Following prior work (Ma et al., 2019), instead of using OT-based reward, we add another auxiliary task in which the model is trained to reconstruct the POS tag of the input sentence. Note that, for this auxiliary task, the model is trained to re-convert the generated sentence D' back to D; (5) **Graph-based**: Instead of directly encoding the interaction between the syntax and semantics via OT-based distance, in this baseline we encode the syntax and the semantics together using a Graph Convolution Network (GCN) (Kipf and Welling, 2017). Specifically, before generating the words of D', the representations H obtained from the GPT-2 model are further abstracted using a two-layer GCN that takes the dependency tree of the input sentence as the input graph. We evaluate the models on the development set of E&M domain for formal and informal style transfer (i.e., both direction)[3].

Table 9 shows the results. This table shows that removing both syntax and semantics scores from OT will hurt the performance. However, syntactic information has more importance as removing them results in more performance loss. Moreover, it is clear from this table that reconstructing syntactic features is not as effective as OT-based reward. This inferiority is better evident from the loss in BLEU and BLEURT scores. Our hypothesis for better performance of the OT-based reward is that OT can encode the interaction between the syntax and semantics while reconstruction makes these two tasks separate. Finally, this table shows that the graph-based model has poor performance compared to OT4. Our hypothesis for this observation is that while the GCN model can encode the syntactic structure of both sentences, it cannot encode the

[3]Note that the same pattern is observed in other settings and domains too

ID	Sentence	Important Alignments (source → generated)
Source #1	Oh, I'm literally dismal in my writing skills but a hero in singing	(literally → very), (dismal → poor), (skills → abilities)
Generated #1	My writing abilities are very poor but my singing is excellent.	(hero → excellent), (reading → reading)
Source #2	though she is not his gf but he hangs out with her a lot!	(though → However), (gf → girlfriend), (hangs → spend),
Generated #2	However she is not his girlfriend, he spends too much time with her.	(hangs → time), (lot → much), (her → her)

Table 8: Case Study for Informal to Formal conversion. The important alignments are the alignments with the highest probability predicted by solving the Optimal Transport problem for the given two sentences.

Model	BLEURT	BLEU	ACC	HM
No Semantics	0.079	0.545	0.921	0.684
No Syntax	0.064	0.521	0.912	0.663
No OT	0.050	0.509	0.889	0.647
Reconstruct	0.059	0.560	0.929	0.698
Graph-based	0.063	0.555	0.921	0.692
OT4 (Full)	**0.114**	**0.621**	**0.940**	**0.747**

Table 9: Ablation study on the development set of E&M domain for formal ↔ informal conversion

Original Input	It actually turned out to be pretty decent as far as B-list horror/suspense films go
RACoLN	It is a terrible movie for a category of this genre.
OT4 (Ours)	It seems to be an unsatisfactory movie for the genre of horror.

Table 10: Generated text by the proposed model and the prior SOTA model for an IMDB sample text.

alignment between the words of the input sentence and the generated sentence. Hence, hindering the content preservation computation.

4.3 Case Study

To provide more insight into the performance of the proposed model, in this section we conduct a qualitative analysis. Specifically, we study how the OT-based model is able to find the perfect alignment between the words of the input sentence and the generated sentence. Note that in case of a successful style conversion, there should be a small Wasserstein distance between the two sentences, hence, the semantically related words will be aligned with each other. Table 8 shows two informal sentences along with their converted formal counterparts generated by OT4. To study the role of Optimal Transport, we report the alignments with the highest probability which are obtained by solving the OT problem for the given two sentences. This table shows that there is a high semantic similarity between the aligned words. More importantly, the aligned words have the same syntactic connections with the other words in their sentence. For instance, in the first example, the word "*dismal*" and its child in the dependency tree, i.e., "*skills*", are aligned with the word "*poor*" and its child in the dependency tree, i.e., "*abilities*". This example shows that the OT alignment considers both semantic and syntactic relations between aligned words. However, OT is not restricted to semantic or syntactic structures and it can relax the alignments whenever it is needed. For instance, in the second example, we observe that the word "*time*" and "*hangs*" are aligned with each other while serving different syntactic roles in the sentence. It shows that when the semantic relation is more important, OT can break the syntactic constraints to find a perfect alignment and the lowest Wasserstein distance. This example shows that the prior work for reconstructing the syntactic roles regardless of their semantic importance will be inferior to our proposed OT-based approach.

Finally, to qualitatively study the improvement obtained by the proposed model on the unsupervised setting, we present the generated text for a sample text from the IMDB dataset. Specifically, we compare our model output with the prior SOTA model, i.e., RACoLN. Table 10 shows the samples. It is clear from this table that the proposed model can retain more content from the input text. In particular, while RACoNL omits the genre of the movie, OT4 successfully keeps this information in the generated text. We hypothesis that the similar distance of the word "*horror*" to the opinion words in the input and the generated text, i.e., "*decent*" and "*unsatisfactory*", are helping to keep this information in OT4 output.

5 Conclusion

We propose a new model for encouraging content preservation in text style transfer. We demonstrate that both syntax and semantics of the input sentence and the generated sentence should be taken into account for content preservation. More importantly, we empirically show that the interaction between syntax and semantics of the input and target sentences is necessary for TST. We conducted extensive experiments on benchmark datasets in supervised and unsupervised settings, achieving state-of-the-art performance on multiple datasets.

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, Online. Association for Computational Linguistics.

Liqun Chen, Shuyang Dai, Chenyang Tao, Haichao Zhang, Zhe Gan, Dinghan Shen, Yizhe Zhang, Guoyin Wang, Ruiyi Zhang, and Lawrence Carin. 2018. Adversarial text generation via feature-mover's distance. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 4671–4682.

Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. 2019. Style transformer: Unpaired text style transfer without disentangled latent representation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5997–6007, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A probabilistic formulation of unsupervised text style transfer. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1587–1596. PMLR.

Fei Huang, Zikai Chen, Chen Henry Wu, Qihan Guo, Xiaoyan Zhu, and Minlie Huang. 2021. NAST: A non-autoregressive generator with word alignment for unsupervised text style transfer. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1577–1590, Online. Association for Computational Linguistics.

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, Doha, Qatar. Association for Computational Linguistics.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Nicholas I. Kolkin, Jason Salavon, and Gregory Shakhnarovich. 2019. Style transfer by relaxed optimal transport and self-similarity. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 10051–10060. Computer Vision Foundation / IEEE.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762, Online. Association for Computational Linguistics.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2021. Thank you BART! rewarding pre-trained models improves formality style transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 484–494, Online. Association for Computational Linguistics.

Dongkyu Lee, Zhiliang Tian, Lanqing Xue, and Nevin L. Zhang. 2021a. Enhancing content preservation in text style transfer using reverse attention and conditional layer normalization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 93–102, Online. Association for Computational Linguistics.

Dongkyu Lee, Zhiliang Tian, Lanqing Xue, and Nevin L. Zhang. 2021b. Enhancing content preservation in text style transfer using reverse attention and conditional layer normalization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 93–102, Online. Association for Computational Linguistics.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874, New Orleans, Louisiana. Association for Computational Linguistics.

Yixin Liu, Graham Neubig, and John Wieting. 2021. On learning text style transfer with direct rewards. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4262–4273, Online. Association for Computational Linguistics.

Yun Ma, Yangbin Chen, Xudong Mao, and Qing Li. 2019. A syntax-aware approach for unsupervised text style transfer. In *OpenReview*.

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, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Gabriel Peyre and Marco Cuturi. 2019. Computational optimal transport: With applications to data science. In *Foundations and Trends in Machine Learning*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Eric Risser. 2020. Optimal textures: Fast and robust texture synthesis and style transfer through optimal transport. volume abs/2010.14702.

Abhilasha Sancheti, Kundan Krishna, Balaji Vasan Srinivasan, and Anandhavelu Natarajan. 2020. Reinforced rewards framework for text style transfer. In *Advances in Information Retrieval*, volume 12035, page 545. Nature Publishing Group.

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, Hong Kong, China. Association for Computational Linguistics.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6830–6841.

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, Hong Kong, China. Association for Computational Linguistics.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. In *Kluwer Academic*.

Jingjing Xu, Hao Zhou, Chun Gan, Zaixiang Zheng, and Lei Li. 2021. Vocabulary learning via optimal transport for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7361–7373, Online. Association for Computational Linguistics.

Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *Proceedings of COLING 2012*, pages 2899–2914, Mumbai, India. The COLING 2012 Organizing Committee.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.