So Different Yet So Alike!
Constrained Unsupervised Text Style Transfer

Anonymous ACL submission

Abstract

Transferring text from one domain to the other has seen tremendous progress in the recent past. However, these methods do not aim to explicitly maintain constraints such as similar text length, descriptiveness between the source and the translated text. To this end, we introduce two complementary cooperative losses to the generative adversarial network family. Here, both the generator and the critic reduce the *contrastive* and/or the *classification loss* aiming to satisfy the constraints. These losses allow lexical, syntactic, and domain-specific consistencies to persist across domains. We demonstrate the effectiveness of our method over multiple benchmark datasets, both with single and multi-attribute transfers. The complimentary cooperative losses also improve text quality across datasets as judged by current, automated generation and human evaluation metrics.

1 Introduction

Humans are capable of mapping given inputs from one domain to the other. For example, machine translation converts text between languages, (Vaswani et al., 2017; Artetxe et al., 2018; Lample et al., 2017), or emoji creation maps human faces to emojis (Taigman et al., 2017). Humans do these tasks efficiently, robustly, and without direct supervision. Recently, there has been a surge of interest in similar tasks such as attribute transfer (Jin et al., 2020b) and controlled text generation (Dathathri et al., 2020; Subramanian et al., 2018). These works aim to preserve the semantics of the source sentence ("content"), while changing certain attributes ("style"). Some common works include changing the sentiment (Li et al., 2018), expertise (Cao et al., 2020), formality (Rao and Tetreault, 2018), or multiple attributes (Subramanian et al., 2018).

While most relevant works, offered under the umbrella term *text style transfer*, aim to preserve a vague definition of "content", they do not explicitly enforce any constraints of identity between the source and the translated text. For example, there is no explicit guarantee that the source and the transferred sentence will have a similar length or remain descriptive. Figure 1 shows one such example, where a sentence from the books domain is translated to the movie domain. While the translated sentence "Loved the Movie" is an accepted text style transferred sentence because it retains much of the content and has the target attribute, it does not have the same length, does not have a personal noun ("I") and does not have a domain appropriate proper noun. Comparatively, the higher-fidelity transfer "I absolutely enjoyed Spielberg's direction ", maintains constraints of identity and has the personal pronoun, along with a domain-appropriate proper noun.

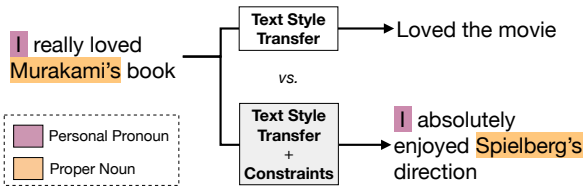


Figure 1: Illustrative example showing transfer of text from books to movies while maintaning constraints of identity.

Enforcing such constraints of identity can help maintain the brand identity when the product descriptions are mapped from one commercial product to another. They can also help in data augmentation for downstream domain adaptation NLP applications. Such constraints of identity are explored extensively in the computer vision task of cross-domain image generation. Taigman et al. (2017) translate human faces to an emoji while maintaining the identity of a face, but these issues are relatively unexplored in NLP.

In this work, we map text between two domains with a focus on maintaining constraints of identity between them. Current methods in text style transfer, aim to maintain the "content" and transfer the "attribute". They neither aim to nor have mechanisms for explicitly enforcing such constraints

between the source and the transferred sentence. To this end, we build upon pioneering text style transfer works and introduce an additional explicit regularization component in the latent space of an Adversarially Regularized Autoencoder (ARAE) through two complementary losses. Unlike the opposing losses that the generator and the critic optimize in ARAE, these losses cooperatively reduce the same objective (Algorithm 1). The first loss is a contrastive loss (Le-Khac et al., 2020) that brings sentences that have similar constraints closer and pushes sentences that are dissimilar farther away. The second loss is a classification loss that maintains the identity constraints from the latent vectors (Odena et al., 2017).

Our approach, while simple and aimed at maintaining constraints, crucially also improves the overall performance of the generation on three datasets, YELP (Zhao et al., 2018b), IMDB (Dai et al., 2019) and POLITICAL (Prabhumoye et al., 2018), with the largest increase of 43.7% compared to works that do not explicitly regularize the latent space (§ 3.4.1). We generate six constraints including lexical, syntactic and domain specific. The introduced cooperative losses satisfy the constraints more effectively compared to strong baselines. Since multiple attributes can change between two domains (Subramanian et al., 2018), we test our method on one such dataset and show that the constraints of identity are maintained more effectively (§ 3.4.2). In summary our contributions are: 1) To the best of our knowledge, the first to introduce cooperative losses in a GAN-like setup for NLP. 2) Maintain constraints of identity for text style transfer while improving overall quality.

2 Method

We consider two corpora: a source $\mathcal{S}$ and a target $\mathcal{T}$. Each of them comprises of a set of sentences with common, known attributes (Jin et al., 2020a). The attributes can range from being sentences with a specific sentiment (positive vs. negative) (Li et al., 2018), political slant (democratic vs republican) or a combination of them (Lample et al., 2019). Let $\{x^1_{src}, x^2_{src} ... x^m_{src}\}$ be the set of sentences in $\mathcal{S}$ and $\{x^1_{trg}, x^2_{trg} ... x^n_{trg}\}$ be the set of sentences in $\mathcal{T}$. Let $\mathcal{C} = \{c_1, c_2, ..., c_{|\mathcal{C}|}\}$ be a set of constraints that should remain invariant between $\mathcal{S}$ and $\mathcal{T}$. We maintain these constraints at various levels including lexical, syntactic and domain specific. (c.f § 3.1). The objective is to transfer a sentence $s_i \in \mathcal{S}$ to an analogous sentence $t_i \in \mathcal{T}$, while maintaining the constraints $\mathcal{C}$.

2.1 Background

Adversarially Regularized Autoencoder(ARAE): To perform unsupervised transfer, we consider *seq–seq* models that can effectively regularize and produce smooth latent spaces, making it easy to sample and generate text with desired properties. Inspired from Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), ARAEs (Zhao et al., 2018b) are one such class of generative latent variable models, that has been widely adopted in unsupervised text generation (Huang et al., 2020), topic modeling (Hu et al., 2020), among others. The general framework consists of a deterministic auto-encoder with an encoder $enc_\theta : \mathcal{X} \rightarrow \mathcal{Z}$ that encodes text $x \in \mathcal{X}$ into a latent representation $z \sim P_z$, and a decoder $dec_\phi : \mathcal{Z} \rightarrow \mathcal{X}$ that decodes (generates) text conditioned on the latent representations. ARAE regularizes the latent space utilizing a GAN-like setup. A sample s is first drawn from a simple prior, such as a Gaussian: $\mathcal{N}(0,1)$, and a generator $g_\psi : \mathcal{N}(0,1) \rightarrow \tilde{\mathcal{Z}}$ maps it to a realistic distribution. A critic $C_\xi : \mathcal{Z} \rightarrow \mathbb{R}$ distinguishes between real and generated samples. The generator is trained to fool the critic while the critic is trained to distinguish the real from the generated text. This results in a min–max optimization which implicitly minimizes the JS-Divergence between the two distributions $P_{\tilde{z}}$ and P_z.

$$\min_{\psi} \max_{\xi} \mathop{\mathbb{E}}_{z \sim P_z} [C_\xi(z)] - \mathop{\mathbb{E}}_{\tilde{z} \sim P_{\tilde{z}}} [C_\xi(\tilde{z})]$$

The training involves (a) reducing the auto-encoder loss – which tries to reconstruct the input and encourages copying behavior and maintain semantics similar to original text (Eq. 1), (b) optimizing the critic to distinguish between real and fake samples (Eq. 2), and (c) training the encoder to fool the critic (Eq. 3).

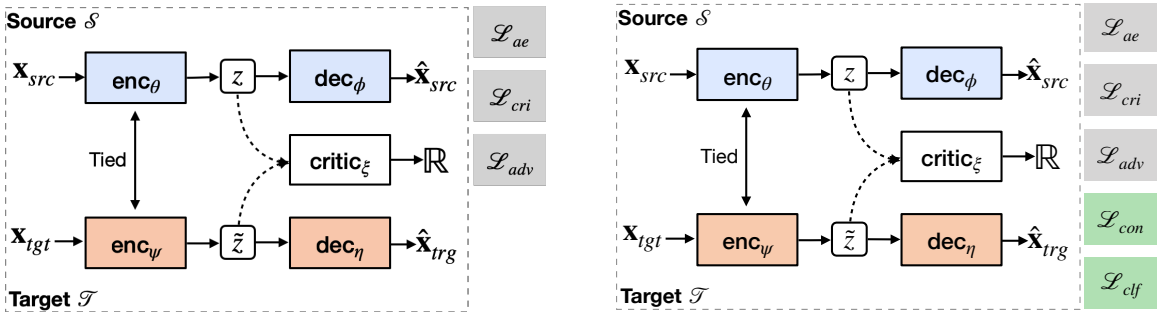
$$\mathcal{L}_{ae}(z;\theta,\phi) = \mathop{\mathbb{E}}_{z \sim P_z} [-log\, p_\phi(x|z)] \quad (1)$$

$$\mathcal{L}_{cri}(z,\tilde{z};\xi) = -\left(\mathop{\mathbb{E}}_{z \sim P_z} [C_\xi(z)] - \mathop{\mathbb{E}}_{\tilde{z} \sim P_{\tilde{z}}} [C_\xi(\tilde{z})] \right) \quad (2)$$

$$\mathcal{L}_{adv}(z,\tilde{z};\psi;\theta) = \mathop{\mathbb{E}}_{z \sim P_z} [C_\xi(z)] - \mathop{\mathbb{E}}_{\tilde{z} \sim P_{\tilde{z}}} [C_\xi(\tilde{z})] \quad (3)$$

2.2 Architecture

Base Model (DCT-ARAE): The main idea of the base model architecture (Figure 2a) is to replace the noise sampling mechanism with an encoder that encodes text from $\mathcal{T}$. Instead of sampling s from a noise distribution like $\mathcal{N}(0,1)$ and passing it through a generator g_ψ, we replace it with an encoder enc_ψ



(a) (b)

Figure 2: a) DCT-ARAE – We replace the generator of ARAE with an encoder that encodes text from $\mathcal{T}$. (b) Adding our proposed cooperative losses to the model.

that encodes text from the target domain $\mathcal{T}$ and a decoder dec_η that decodes text in $\mathcal{T}$. Inspired from Cycle-GAN (Zhu et al., 2017), instead of matching an arbitrary distribution, we match the distribution of $\mathcal{T}$. In addition, we tie the weights of the encoders from the two domains, so that the encoders learn to encode domain agnostic information. Tying encoder weights has also been used by unsupervised machine translation (Artetxe et al., 2018; Lample et al., 2017) and multiple other works (Mai et al., 2020; Huang et al., 2020; Hu et al., 2020; Artetxe et al., 2018)[1]. Any such architecture changes to ARAE can be used as the base model in our work.

Cooperative Contrastive Learning: To maintain the constraints between the two domains $\mathcal{S}$ and $\mathcal{T}$, we introduce the novel step of *introducing a self-supervised learning metric in both encoders, and the critic that controls the distance between instances having similar constraints* (Figure 2b). The idea is to regularize the latent space more by encouraging the encoders to produce representations that bring two sentences that share similar constraints closer together, and force dissimilar ones away.

To this end, we use contrastive representation learning. There are several self-supervised metric losses under the umbrella of contrastive losses (Le-Khac et al., 2020) including Triplet Loss (Hoffer and Ailon, 2015) and NT-Xent loss (Chen et al., 2020). We use one that is amenable to multiple positive instances (Khosla et al., 2020). Given a sentences $s_i \in \mathcal{S}$ in a mini-batch of size B, we mine P positive sentences each from $\mathcal{S}$ and $\mathcal{T}$ that share the same constraints with s_i. This contrastive loss is given by:

$$\mathcal{L}_{con}(z_i,\mathcal{C}_i;\theta,\psi,\xi) = -\frac{1}{|P|} log\left(\sum_{j=1}^{P} \frac{exp(z_i,z_j)}{\sum_{k=1}^{B\setminus\{i\}} exp(z_i,z_k)}\right) \quad (4)$$

[1]We tried with separate encoders and decoders, but encoders with tied-weights work best

where z's are representations obtained from the encoders in $\mathcal{S}$, $\mathcal{T}$ (Line 34 in Algo 1) or representations obtained from the last layer of critic C_ξ. $\mathcal{C}_i$ (Line 14) are a set of constraints for a sentence. Recently, (Kang and Park, 2020) introduced the cooperative loss in the adversarial setup where contrastive losses are added to both the *critic* and *generator* for GANs. Unlike the normal opposing losses of the generator and the critic, both of them cooperatively reduce the contrastive loss. We follow a similar principle and add the loss to both the encoders and the critic. (Lines 17 & 35).

Cooperative Classification: Contrastive learning might be sub-optimal if we do not mine good quality positive and negative samples (Tian et al., 2020). To address this, we propose another way to regularize the latent space. Similar to ACGAN (Odena et al., 2017), we encourage the encoders and the critic to cooperatively reduce a classification loss. We include a classifier $D_\delta : \mathcal{Z} \rightarrow \mathbb{R}^{|\mathcal{C}|}$ that predicts the different constraints $\mathcal{C}$ of the sentences and the binary cross entropy loss is reduced.

$$\mathcal{L}_{clf}(z_i;\theta,\phi,\xi,\delta) = \sum_{c=1}^{|\mathcal{C}|} -\Big[y_c log(\sigma(l_c)) + (1-y_c) log(1-\sigma(l_c))\Big] \quad (5)$$

where $|\mathcal{C}|$ is the number of constraints per sentence, σ is the sigmoid function and l_c are the logits produced by the classifier for z_i. As in contrastive loss, the z_i can be produced by encoders of $\mathcal{S}$, $\mathcal{T}$ (Lines 34–36) or from the hidden layers of the critic (Line 18).

Final Loss: The overall loss is a linear combination of the losses:

$$\mathcal{L} = \mathcal{L}_{ae} + \mathcal{L}_{cri} + \mathcal{L}_{adv} + \lambda_1 \mathcal{L}_{con} + \lambda_2 \mathcal{L}_{clf}. \quad (6)$$

Algorithm 1: Overall training procedure for DCT-ARAE(+ [CLF] + [CONTRA])

Input : # iteration of critic / iteration of enc: n_{dis}, lrs- autoencoder: lr_{ae}, critic: lr_{dis}, adv train: lr_{adv}. Adam params: β_1,β_2.

Initialize (θ,ϕ,ψ,η,ξ,δ)
for *1, . . . # iterations* **do**
  (1) Train the autoencoders
  $\mathbf{x_{src}} \leftarrow \{x^i_{src}\}^m_{i=1} \sim \mathcal{S}$
  $\mathbf{x_{trg}} \leftarrow \{x^j_{trg}\}^n_{j=1} \sim \mathcal{T}$
  $\mathbf{z^s} = enc_\theta(\mathbf{x_{src}})$, $\mathbf{z^t} = enc_\psi(\mathbf{x_{trg}})$
  $\theta, \phi \leftarrow$ Adam($\mathcal{L}_{ae}(\mathbf{z^s};\theta,\phi), \beta_1, \beta_2, lr_{ae}$)
  $\psi, \eta \leftarrow$ Adam($\mathcal{L}_{ae}(\mathbf{z^t};\psi,\eta), \beta_1, \beta_2, lr_{ae}$)
  (2) Train the Critic
  for *1, . . . # n_{dis}* **do**
    $\mathbf{x_{src}} \leftarrow \{x^i_{src}\}^m_{i=1} \sim \mathcal{S}$
    $\mathbf{x_{trg}} \leftarrow \{x^j_{trg}\}^n_{j=1} \sim \mathcal{T}$
    $\mathbf{z^s} = enc_\theta(\mathbf{x_{src}})$, $\mathbf{z^t} = enc_\psi(\mathbf{x_{trg}})$
    // Last layer representations
    $\mathbf{z^s_{cri}} = C^{hid}_\xi(\mathbf{z^s})$, $\mathbf{z^t_{cri}} = C^{hid}_\xi(\mathbf{z^t})$
    $l_{cri} \leftarrow \mathcal{L}_{cri}(\mathbf{z^s}, \mathbf{z^t}; \xi)$
    (2a) Critic Cooperative Training
    $l^{con}_{cri} \leftarrow \mathcal{L}_{con}(\mathbf{z^s_{cri}}, \mathcal{C}; \xi)$ // contrastive loss
    $l^{clf}_{cri} \leftarrow \mathcal{L}_{clf}([\mathbf{z^s_{cri}};\mathbf{z^t_{cri}}], \mathcal{C}_i; \xi, \delta)$ // clf loss
    $\mathcal{L}_{cri} \leftarrow l_{cri} + \lambda_1 l^{con}_{cri} + \lambda_2 l^{clf}_{cri}$
    $\xi \leftarrow$ Adam($\mathcal{L}_{cri}, \beta_1, \beta_2, lr_{dis}$)
  end
  (3) Adversarial Training
  (3a) Adv Training of Target Encoder
  $\mathbf{x_{trg}} \leftarrow \{x^j_{trg}\}^n_{j=1} \sim \mathcal{T}$
  $\mathbf{z^t} = enc_\theta(\mathbf{x_{trg}})$
  $\theta \leftarrow Adam(\underset{\mathbf{z^t}}{\mathbb{E}}[C_\xi(\mathbf{z^t})], \beta_1, \beta_2, lr_{adv})$
  (3b) Adv Training of Source Encoder
  $\mathbf{x_{src}} \leftarrow \{x^i_{src}\}^m_{i=1} \sim \mathcal{S}$
  $\mathbf{z^s} = enc_\theta(\mathbf{x_{src}})$
  $\psi \leftarrow Adam(-\underset{\mathbf{z^s}}{\mathbb{E}}[C_\xi(\mathbf{z_t})], \beta_1, \beta_2, lr_{adv})$
  (4) Encoder Cooperative Training
  $\mathbf{x_{src}} \leftarrow \{x^i_{src}\}^m_{i=1} \sim \mathcal{S}$
  $\mathbf{x_{trg}} \leftarrow \{x^j_{trg}\}^n_{j=1} \sim \mathcal{T}$
  $\mathbf{z^s} = enc_\theta(\mathbf{x_{src}})$, $\mathbf{z^t} = enc_\psi(\mathbf{x_{trg}})$
  $l^{con}_{enc} \leftarrow \mathcal{L}_{con}(\mathbf{z^s}, \mathcal{C}; \phi, \psi)$ // contrastive loss
  $l^{clf}_{enc} \leftarrow \mathcal{L}_{clf}([\mathbf{z^s};\mathbf{z^t}], \mathcal{C}_i; \phi, \psi, \delta)$ // clf loss
  $\mathcal{L}_{coop} = \lambda_1 . l^{con}_{enc} + \lambda_2 . l^{clf}_{enc}$
  $\theta, \psi \leftarrow Adam(\mathcal{L}_{coop}, \beta_1, \beta_2, lr_{ae})$
end

Here, λ_1 and λ_2 control the importance of different losses. For our experiments, e set λ_1, λ_2 in $\{0,1\}$.

3 Experiments

For the encoders, we use a one layer LSTM network with 300 hidden dimensions for YELP, IMDB and 500 hidden dimension for the larger POLITICAL dataset. For the critics and classification loss, we use a 2 layer MLP. Our learning rates and methods to stabilize training are discussed in Appendix C.

3.1 Datasets

We use three datasets to compare our method. Statistics of these datasets are available in Appendix A. Only a single attribute changes between these datasets. We use them to compare our model against others that use similar datasets. However, we also show more pronounced results on datasets where multiple attributes change in § 3.4.2. 1) **Yelp Reviews**; business reviews listed on Yelp, labelled as either a positive or negative sentiment. We use the splits provided by Zhao et al. (2018b). 2) **IMDb Movie Reviews**: consists of movie reviews (Dai et al., 2019). Examples in this dataset are also either positive or negative. 3) **Political Slant**: consists of Facebook posts from the politicians of the United States Senate and the House of Representatives Prabhumoye et al. (2018). Sentences in the dataset are labelled as either democratic or republican.

Generation Constraints: We constrain every sentence along *six* diverse dimensions that we desire to control between the two domains. All of these labels are categorical. **a) Lexical**: Sentence length – The transferred sentence should maintain a similar (binarized) length to the original sentence. Length is binarized to long (10+ words) and short (≤ 10). **b) Syntactic**: Presence of personal pronouns – binarized to indicate the presence of a personal pronoun; number of adjectives – categorical up to 5; number of proper nouns – categorical up to 3; syntactic tree height – categorical up to 10, and **c) Domain specific** – number of domain-specific attributes (Li et al., 2018) – categorical up to 5. We chose the different labels to ensure that at least 90% of the instances are assigned a distinct label. Further, we label the sentence with a constraint-specific, catch-all label if the bounds are beyond what we mention above. Since the distribution of the labels may be different we report the F1 score on our constraints.

3.2 Automatic Evaluation

Krishna et al. (2020) highlighted the shortcomings of traditional measures like BLEU and PPL, where obtaining higher scores on these metrics may not indicate good quality transfer. We adopt their suggestion and calculate a sentence-level aggregated metric of their suggested components: 1) **Semantic Similarity** (SIM): based on a model trained on sub-word units, 2) **Transfer Accuracy** (ACC): binary measure indicating whether the generated sentence has been successfully transferred, and 3) **Fluency** (FL): binary measure indicating the linguistic ac-

	YELP				IMDB				POLITICAL			
p	ACC	FL	SIM	AGG	ACC	FL	SIM	AGG	ACC	FL	SIM	AGG
DRG AND ARAE (IN THAT ORDER)												
0.0	67.4	54.5	**43.6**	16.7	56.5	44.3	**54.1**	14.4	61.3	35.7	**38.7**	8.8
0.0	**93.1**	67.9	31.2	19.8	95.0	76.3	26.4	19.9	63.0	72.1	17.3	11.0
DCT-ARAE												
0.0	88.0	63.1	34.7	19.7	92.5	71.3	38.5	26.7	**96.8**	52.2	26.7	13.3
0.6	88.1	62.8	34.4	19.4	93.9	70.8	38.8	26.9	96.8	51.6	26.5	13.0
0.9	88.2	61.3	34.0	18.8	93.1	66.5	38.4	24.7	96.8	49.6	26.2	12.3
DCT-ARAE + CLF												
0.0	89.0	64.7	33.6	19.8	95.0	**83.2**	34.2	27.5	96.1	51.4	28.1	13.4
0.6	88.9	64.3	33.4	19.6	95.1	83.1	34.7	27.8	96.0	50.6	**28.0**	13.1
0.9	91.2	72.1	30.6	19.9	95.6	82.6	34.0	27.7	96.0	49.1	27.7	12.6
DCT-ARAE + CONTRA												
0.0	91.3	**74.0**	31.1	**20.7**	**96.1**	80.6	36.0	**28.6**	96.6	**61.5**	23.2	13.3
0.6	91.2	73.3	31.1	20.4	95.0	78.2	35.6	27.4	96.5	60.2	23.1	13.0
0.9	91.2	72.1	30.6	19.9	94.5	79.1	34.8	26.5	96.2	57.9	22.8	12.3
DCT-ARAE + CLF + CONTRA												
0.0	89.3	69.2	32.9	20.6	93.8	74.8	36.4	26.4	94.8	53.7	27.7	**13.8**
0.6	89.4	68.6	32.8	20.4	94.3	72.7	35.8	26.0	94.7	53.3	27.6	13.5
0.9	89.5	67.0	32.3	19.7	93.4	72.9	35.9	25.2	94.7	51.1	27.2	12.7

Table 1: Evaluation of DCT-ARAE against ACC (transfer accuracy), FL(fluency) and SIM (semantic similarity), AGG (joint accuracy). Cooperatively reducing the contrastive or the classification loss is better than ARAE. p indicates the probability used in nucleus sampling. p=0 indicates greedy decoding.

ceptability of the transferred sentence. Appendix B has more details.

$$\text{AGG}(\text{ACC},\text{SIM},\text{FL}) = \frac{1}{|S|}\sum_{s\in S}\text{ACC}(s)\cdot\text{SIM}(s)\cdot\text{FL}(s)$$

3.3 Baselines

We compare DCT-ARAE on all the datasets against the following baselines: **a) DRG**: The Delete Retrieve and Generate method that deletes domain specific attributes, retrieves a template and generates the target domain text (Li et al., 2018). We use the stronger, entire system rather than the weaker DELETEONLY and RETRIEVEONLY baselines; **b) ARAE**: Adversarially regularized autoencoders our system is based on (Zhao et al., 2018b); **c) DCT-ARAE**: Our proposed model without the contrastive learning or cooperative classifier; **d) DCT-ARAE + CONTRA**: Our proposed model with the contrastive learning; **e) DCT-ARAE + CLF**: Our proposed model with the cooperative classifier; **f) DCT-ARAE+CONTRA+CLF**: Our proposed model with both the cooperative losses.

3.4 Results

3.4.1 Overall Results

DCT-ARAE + CONTRA and DCT-ARAE + CLF consistently perform better than DRG and ARAE on the AGG score (Table 1). The AGG for YELP is 20.69 (vs 19.8), for IMDB it is 28.57 (vs 19.9) and for POLITICAL 13.76 (vs 11.0). Quantitatively POLITICAL has lower AGG scores compared to the other two, because the dataset has many out-of-vocabulary words and longer sentences compared to the other two datasets. Although, cooperative loss reduction is aimed at satisfying the constraints between two domains, it shows that further regularization of the latent space, not only brings advantages in satisfying the different constraints, but it can improve the performance in general (Lavoie-Marchildon et al., 2020).

Effect of Cooperative Loss Reduction on SIM and FL: Across datasets, reducing cooperative losses improves SIM and FL while producing text with similar ACC to ARAE. Although, DRG produces sentences with high SIM because most of the text from the original sentence is retained after the delete step, there is a large trade-off with ACC resulting in low AGG scores. But, compared to ARAE, adding cooperative losses significantly increases the SIM, with the highest percentage increase of 63.6% observed for POLITICAL (28.13 vs 17.3). The reasons for this could be two fold: 1) since we mine positive sentences from a corpus that is grounded in real world events, most lexically similar sentences can also be semantically similar (Guu et al., 2018), and 2) since we tie the encoders from the source and target domain, we extract domain agnostic information before generation which retains content.

The FL also improves on 2 of 3 datasets with

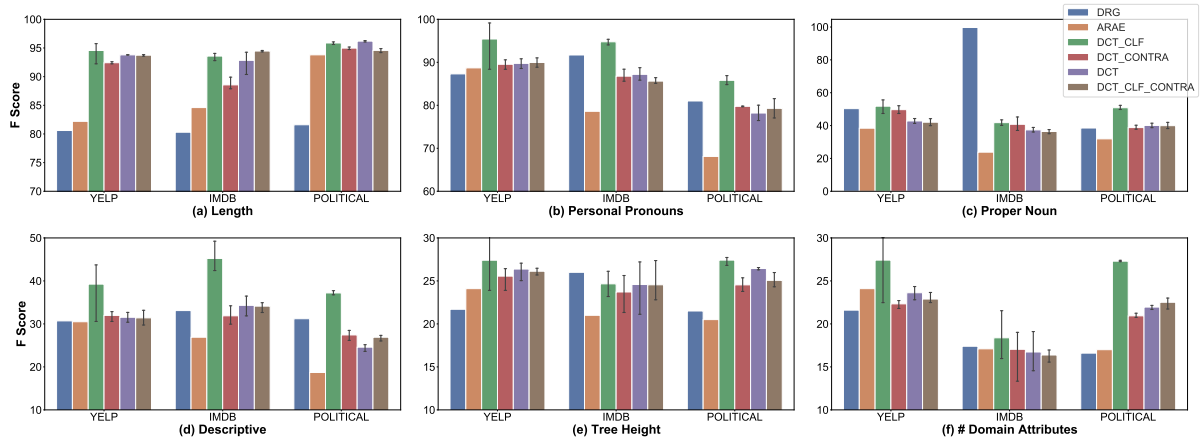


Figure 3: F-scores of different constraints. Adding cooperative losses helps in better maintaining the constraints. The error bars show the variance of generating text using greedy decoding and nucleus sampling with $p = \{0.6, 0.9\}$.

73.93 (vs 67.9) for YELP, and 83.2 (vs 76.3) for IMDB. We hypothesize that reducing cooperative losses regularizes the latent space bringing fluent sentences closer together, enabling the decoder to produce semantically similar and linguistically acceptable sentences. For POLITICAL, based on qualitative analysis, we found that source sentences in themselves are less fluent and contain many US political acronyms and our system produces many out-of-vocab words affecting the overall fluency.

Nucleus Sampling: We find that our system achieves the highest AGG score with greedy decoding. We also experiment with nucleus sampling (Holtzman et al., 2019) with different p values as shown in Table 1. Nucleus sampling produces diverse sentences and increases ACC as expected. However we find that, with higher values of p, there is a trade-off with SIM resulting in a lower AGG score overall.

Effect of the Number of Positives: The number of positive and negative samples used for contrastive learning (Eq. 4) have a significant effect on the overall performance (Khosla et al., 2020; Chen et al., 2020; Henaff, 2020). Table 2 (rows p{1,2,5,10}) shows the AGG scores on IMDB on which the contrastive losses produce the highest gain in AGG (43.7%), for different number of positives. We find that AGG is the highest with 2 positives per sample as also used by Khosla et al. (2020). Although increasing the number of negatives is beneficial for contrastive learning, when more than one positive examples are available, making use of them brings further improvements (Khosla et al., 2020).

Cooperative Losses are Important on Both the Generator and Critic: Table 2 shows results on IMDB where we remove the cooperative losses on the generator and critic. First, we see that adding the cooperative losses on both the generator and the critic is crucial for the overall performance. While adding the cooperative contrastive loss to both the generator and critic increases FL and ACC while maintaining similar levels of SIM, adding the cooperative classification loss improves SIM which shows the complementary nature of the losses.

Model	ACC	FL	SIM	AGG
DCT-ARAE + CLF	95.0	83.2	34.2	27.5
– generator	96.2	87.2	31.3	26.7
– critic	94.9	84.4	30.8	25.5
DCT-ARAE + CONTRA	96.1	80.6	36	28.6
– generator	93.5	78.8	34.0	26.0
– critic	90.1	67.8	39.5	24.9
p1	92.4	75.5	36.6	26.2
p2	96.1	80.6	36.0	28.6
p5	96.0	84.0	31.4	26.0
p10	95.5	83.3	31.8	26.0

Table 2: Ablation study showing for cooperative losses not added to the generator (–generator) and the critic (–critic) and with different # of positives on IMDB.

Human Evaluation: For our human evaluation, we randomly sample 100 samples from each of the three datasets and hire 3 researchers to rate every sentence for FL, SIM and ACC on a 3 point scale. We average the results and present it in Table 3. DRG produces marginally better semantically similar sentences. Compared to ARAE our model performs well except for in YELP. This may be because, here we use nucleus sampling with 0.9 which optimizes for diversity rather than similarity. On other metrics we perform on-par or better than our competing systems. We provide more details in Appendix F.

Qualitative Examples: Examples are in Table 4.

Dataset	Model	ACC	FL	SIM
YELP	DRG	2.3	2.1	2.1
	ARAE	**2.8**	**2.4**	**2.1**
	OURS	**2.8**	**2.4**	2.0
IMDB	DRG	1.9	2.0	**2.2**
	ARAE	2.5	2.1	1.4
	OURS	**2.6**	**2.2**	2.1
POLITICAL	DRG	2.3	2.2	2.1
	ARAE	2.1	2.1	1.5
	OURS	**2.5**	**2.4**	**2.2**

Table 3: Human Evaluation of generated sentences.

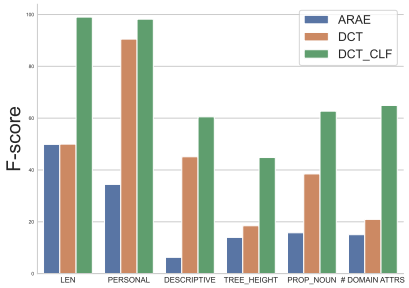


Figure 4: Comparison of ARAE, DCT-ARAE and DCT-ARAE + CLF for different constraints.

More examples and negative examples are in Appendix D. Mistakes made by the model can be attributed to not understanding the semantics, lack of diversity and not producing attribute specific words.

3.4.2 Maintaining Constraints

Figure 3 shows that introducing the cooperative losses significantly outperform DRG and ARAE in maintaining constraints. Specifically the DCT-ARAE + CLF model performs better than DCT-ARAE+ CONTRA. One reason could be that, finding the appropriate positives and strong negatives can be problematic for contrastive learning. On the other hand, the classifier's objective is simpler and forces the encoder to produce representations that satisfy the different constraints effectively.

A seemingly easy to maintain constraint is the length of the sentence. However, *seq-seq* systems have a difficulty of maintaining appropriate lengths (Murray and Chiang, 2018). With no additional regularization ARAE does not maintain the length as well as DCT-ARAE + CLF. On the other hand, compared to the lexical constraints, syntactic attributes like descriptiveness, tree height and domain specific constraints present challenges, with significantly lower F-scores. DCT-ARAE + CLF produces significantly better results in maintaining them. This shows that obtaining improvements on the overall AGG, does not necessarily translate to producing outputs that satisfy constraints and there is a room for further improvement. DRG maintains the proper noun for IMDB very effectively, because it contains a wide variety actor and movie names. They are retained verbatim after the delete operation.

Multiple Attribute Datasets: In order to test whether our model can satisfy constraints across domains where multiple attributes change, we use the multi-attribute dataset released by (Lample et al., 2019). We chose the *Asian* and *Mexican* as two domains. Each of these domains can have multiple attributes like positive and negative sentiment text, different gender attributions to sentences etc. We compare our DCT-ARAE + CLF model with the DCT-ARAE and ARAE in Figure 4. The results are more pronounced in this case with DCT-ARAE + CLF having clear advantage over DCT-ARAE. This shows that even with multiple attributes changing between domains, cooperatively reducing losses can satisfy different constraints more effectively.

Qualitative Examples: Table 5 shows examples of our model maintaining constraints compared to ARAE. Sometimes, ARAE hallucinates and adds personal pronouns like "my" to the text even when there are no personal pronouns (Row 1) and in other cases, it fails to ensure that the personal pronoun is retained (Row 2). Also, our model produces sentences where the number of proper nouns are retained (Chris Klein vs Robert De Niro) whereas ARAE does not.

4 Discussion

Cycle Consistency Loss: a) *In Latent Spaces* - Cycle consistency has been shown to improve cross lingual dictionary construction (Mohiuddin and Joty, 2019), topic modeling (Hu et al., 2020) etc which are word level tasks. A recent work from (Huang et al., 2020) claims to improve unsupervised text style transfer. In our experiments, it did not result in any noticeable performance improvement [2]. Cycle consistency might be too restrictive for sentence level tasks. b) *Using Back-Translation*- Back-translation is another alternative to ensure semantic consistency between source and the target sentence (Prabhumoye et al., 2018; Artetxe et al., 2018; Lample et al., 2017). However, in our case, since we are training an ARAE, it would involve an additional inference and autoencoder training step which is more expensive.

Using Transformers: We also replace our LSTM-autoencoders with pre-trained transformer (Rothe et al., 2020) and randomly initialized transformer encoder-decoders. Although we found an increase

[2]Repeated attempts to obtain source codes failed

Dataset	Input	Output (Ours)	Output (ARAE)
YELP	they close earlier than posted hours	they're open late night	they keep me getting better
YELP	i will not go back to this hotel again	i will definitely go back again and again	i will definitely go back to return and again
IMDB	this movie is a very poor attempt to make money using a classical theme.	this movie is a very good example of a film that will never be forgotten.	this is a film that has been a lot of times and it's really good.
IMDB	it was wooden, totally unrealistic and had no plot or meaning to the story	it was also very funny, and i was pleasantly surprised by the story	it was also a great cast , and the characters are so hard to find.
POLITICAL	what are you doing about the border ?	what are what you are doing about gun violence?	so why are you and the republican party ?
POLITICAL	i wish u would bring change	and i wish you would help bring democracy	and i 'm not sure mr.trump.

Table 4: Example outputs generated by the best system according to AGG score.

Constraint			Explanations
Personal Pronoun	*Source Sentence* (IMDB)	jean seberg had not one iota of acting talent.	ARAE hallucinates and introduces my because it reflects the training distribution
	Ours	michael keaton was also great in his role.	
	ARAE	john abraham had one of my favorite roles .	
Personal Pronoun	*Source Sentence* (IMDB)	oh , i forgot,there was one redeeming feature - the scenery was nice	Our model reproduces the word i even in the target sentence while ARAE misses it.
	Ours	overall, i was pleasantly surprised, this was one of the best animated films	
	ARAE	although this was n't one, it was nice, the first hour of an episode.	
Proper Noun	*Source Sentence* (IMDB)	chris klein's character was unlikable from the start and never made an improvement	Our model retains the number of proper nouns in the sentence unlike ARAE
	Ours	robert de niro was very good as the man and she 's never been	
	ARAE	both of his character was made and had a huge smile on me	

Table 5: Table showing constraints satisfied by our system compared to ARAE.

in the AGG, it was mostly because of very high SIM and very low ACC. Reducing the number of layers, attention heads would still result in a large model that is still prone to copying text. This reveals the potential challenges of training transformers with unpaired mappings and is an important future work.

5 Related Work

Unsupervised text attribute transfer has been tackled by disentangling *attributes* from *content* in the latent dimension. which a decoder uses to generate a sentence (Jin et al., 2020a). Adversarial methods inspired by GANs are prevalent (Zhao et al., 2018a) for achieving such disentanglement. Much of these works transfer a single attribute like, sentiment (Zhao et al., 2018a), expertise (Cao et al., 2020) etc. Adversarial methods involve critics that identify attributes. Although applicable to multiple-attribute transfer become infeasible with an increasing number of attributes. Hence, recent works on multiple attribute transfer (Subramanian et al., 2018) use techniques from unsupervised machine translation which do not involve critics (Artetxe et al., 2018; Lample et al., 2017). Nonetheless, these works aim to preserve content and ignore other desirable constraints of identity. Compared to previous text style transfer works, we aim to maintain these constraints of identity and find that it improves the overall performance. In a similar vein, Lavoie-Marchildon et al. (2020) argue that preserving domain invariant semantic features and identity improves unsupervised image translation. Also, recent works in CV introduced contrastive learning to GANs (Kang and Park, 2020; Sinha et al., 2021) To the best of our knowledge, we are the first to introduce cooperative losses in the GAN setup for NLP. We discuss other closely related approaches in Appendix E.

6 Conclusion

In this work, we highlight that text style transfer focuses on retaining "content" and changing the "style" of sentences, but it does not maintain other desirable constraints. To that end, we introduce two cooperative losses to the GAN inspired Adversarially Regularized Autoencoder that further regularizes the latent space. While satisfying the constraints effectively irrespective of multiple attributes changing, it brings surprising improvements to the overall score by as much as 47.6%. While we focused on simple constraints at the sentence and word level as a first step, future work can add phrase level and more fine-grained constraints like maintaining the syntactic tree structure. Further, potential future work includes using Reinforcement Learning losses to directly optimize the constraints (Liu et al., 2021), produce adversarial examples from different domains for a given NLP task. We hope that the future style transfer works consider satisfying constraints while focusing on improving other metrics.

References

Martín Arjovsky and Léon Bottou. 2017. Towards principled methods for training generative adversarial networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation.

Yixin Cao, Ruihao Shui, Liangming Pan, Min-Yen Kan, Zhiyuan Liu, and Tat-Seng Chua. 2020. Expertise style transfer: A new task towards better communication between experts and laymen. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1061–1071, Online. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

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.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative adversarial networks. *CoRR*, abs/1406.2661.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. 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 1321–1330. PMLR.

Kelvin Guu, Tatsunori B. Hashimoto, Yonatan Oren, and Percy Liang. 2018. Generating sentences by editing prototypes. *Transactions of the Association for Computational Linguistics*, 6:437–450.

Behnam Hedayatnia, Karthik Gopalakrishnan, Seokhwan Kim, Yang Liu, Mihail Eric, and Dilek Hakkani-Tur. 2020. Policy-driven neural response generation for knowledge-grounded dialog systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 412–421, Dublin, Ireland. Association for Computational Linguistics.

Olivier Henaff. 2020. Data-efficient image recognition with contrastive predictive coding. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4182–4192. PMLR.

Elad Hoffer and Nir Ailon. 2015. Deep metric learning using triplet network. In *Similarity-Based Pattern Recognition*, pages 84–92, Cham. Springer International Publishing.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *CoRR*, abs/1904.09751.

Xuemeng Hu, Rui Wang, Deyu Zhou, and Yuxuan Xiong. 2020. Neural topic modeling with cycle-consistent adversarial training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9018–9030, Online. Association for Computational Linguistics.

Yufang Huang, Wentao Zhu, Deyi Xiong, Yiye Zhang, Changjian Hu, and Feiyu Xu. 2020. Cycle-consistent adversarial autoencoders for unsupervised text style transfer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2213–2223, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2020a. Deep learning for text style transfer: A survey. *CoRR*, abs/2011.00416.

Di Jin, Zhijing Jin, and Rada Mihalcea. 2020b. Deep learning for text attribute transfer: A survey.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.

Minguk Kang and Jaesik Park. 2020. Contragan: Contrastive learning for conditional image generation. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *CoRR*, abs/1711.00043.

Guillaume Lample, Sandeep Subramanian, Eric Michael Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Samuel Lavoie-Marchildon, Faruk Ahmed, and Aaron C. Courville. 2020. Integrating categorical semantics into unsupervised domain translation. *CoRR*, abs/2010.01262.

Phuc H. Le-Khac, Graham Healy, and Alan F. Smeaton. 2020. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934.

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, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4262–4273. Association for Computational Linguistics.

Florian Mai, Nikolaos Pappas, Ivan Montero, Noah A. Smith, and James Henderson. 2020. Plug and play autoencoders for conditional text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6076–6092. Association for Computational Linguistics.

Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating style transfer for text. 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 495–504, Minneapolis, Minnesota. Association for Computational Linguistics.

Tasnim Mohiuddin and Shafiq Joty. 2019. Revisiting adversarial autoencoder for unsupervised word translation with cycle consistency and improved training. 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 3857–3867, Minneapolis, Minnesota. Association for Computational Linguistics.

Kenton Murray and David Chiang. 2018. Correcting length bias in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 212–223, Brussels, Belgium. Association for Computational Linguistics.

Augustus Odena, Christopher Olah, and Jonathon Shlens. 2017. Conditional image synthesis with auxiliary classifier gans. 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 2642–2651. PMLR.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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, Melbourne, Australia. Association for Computational Linguistics.

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

Sudha Rao and Joel R. 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, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 129–140. Association for Computational Linguistics.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Abhishek Sinha, Kumar Ayush, Jiaming Song, Burak Uzkent, Hongxia Jin, and Stefano Ermon. 2021. Negative data augmentation. *CoRR*, abs/2102.05113.

Sandeep Subramanian, Guillaume Lample, Eric Michael Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2018. Multiple-attribute text style transfer. *CoRR*, abs/1811.00552.

Yaniv Taigman, Adam Polyak, and Lior Wolf. 2017. Unsupervised cross-domain image generation. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. 2020. What makes for good views for contrastive learning? In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. 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 5998–6008.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU:training neural machine translation with semantic similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355, Florence, Italy. Association for Computational Linguistics.

Junbo Zhao, Yoon Kim, Kelly Zhang, Alexander Rush, and Yann LeCun. 2018a. Adversarially regularized autoencoders. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5902–5911, Stockholmsmässan, Stockholm Sweden. PMLR.

Junbo Jake Zhao, Yoon Kim, Kelly Zhang, Alexander M. Rush, and Yann LeCun. 2018b. Adversarially regularized autoencoders. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 5897–5906. PMLR.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2242–2251. IEEE Computer Society.

A Dataset Statistics

Dataset Statistics: We provide a summary of the dataset statistics in Table 6. We include datasets of varied length and complexity. Apart from having different topics, the IMDB dataset is more formal compared to the more colloquial YELP. We fix the maximum vocabulary size for YELP, IMDB and POLITICAL at 30K which is also the default maximum vocab size used in (Zhao et al., 2018b).

Dataset	Attributes	Train	Dev	Test	Avg len.	Vocab
YELP	Positive	266,041	25,278	50,278	8.9	10K
	Negative	177,218	38,205	76,392		
IMDB	Positive	178,869	2K	1K	18.5	30K
	Negative	187,597	2K	1K		
POLITICAL	Democratic	270,000	2K	28K	16	30K
	Republican	270,000	2K	28K		

Table 6: Dataset splits for YELP, IMDB and POLITICAL.

B Evaluation Measures

Semantic Similarity (SIM) A transferred sentence should be similar in meaning to the source sentence. Earlier works have mainly adopted *n-gram* overlap metrics like *BLEU* (Papineni et al., 2002). However, they have weak correlations with human judgment of semantic similarity. Instead, Krishna et al. (2020) proposed to use encoders that consider subwords (Wieting et al., 2019), which perform well in measuring textual semantic similarity. We directly use this model obtained from their repository.[3]

Transfer Accuracy (ACC) To measure how well the generated sentences have been transferred from the source domain, a popular way of measuring is to report accuracy of a classifier trained to distinguish between the source and the target sentences. We build these classifiers using *fastText* (Joulin et al., 2017) for every dataset. We achieve accuracies of 97.9 for YELP, 96.9 for IMDB and 97.1 for POLITICAL. Unlike Krishna et al. (2020) who use ROBERTa-large, we achieve good accuracies with a simple and fast classifier trained using fastText.

Fluency (FL) A transferred sentence should be grammatically correct. Previous studies rely on trained language models and use perplexity as a measure of fluency. Training a language model for evaluation is cumbersome and language model perplexity does not correlate with human judgments of fluency (Mir et al., 2019). Similar to Krishna et al. (2020), we fine-tune a ROBERTa-large on the corpus of linguistic acceptability that measures and use it to measure whether a sentence is linguistically acceptable.

Aggregate score (AGG) : Calculates sentence level aggregate score of SIM, ACC and FL. This measure strongly penalizes the system if the accuracy of the transferred sentence is 0 and if the fluency score is 0, both of which are discrete. We urge future works to use this metric to enable a fair comparison.

$$\text{AGG}(\text{ACC}, \text{SIM}, \text{FL}) = \frac{1}{|S|}\sum_{s \in S} \text{ACC}\,(s) \cdot \text{SIM}\,(s) \cdot \text{FL}\,(s)$$

C Hyper-parameter Details

Training : For all our experiments we set the learning rate of the auto-encoder (lr_{ae}) to 1e-3 and (lr_{disc}) to 1e-4. The number of discriminator steps (n_{dis}) is set to 5. The Adam optimizer parameters β_1=0.5 and β_2=0.9, which ensures a more conservative optimization and is known to improve stability. We also add a gradient penalty to the loss function of the discriminator that stabilizes training. All the suggestions for stabilizing training are mostly obtained from (Arjovsky and Bottou, 2017).

Inference We used nucleus sampling with $p \in [0.6, 0.9]$. We tried different temperatures of scaling the softmax (Guo et al., 2017) - 0.4, 0.5, 0.6, 0.7 and chose the one that produced the best result on the dev set.

D Transfer Results

More transfer results are mention in Table 8. Examples where our system fails with plausible explanation are given in Table 9. Examples of translation from the multi-attribute dataset is shown in Table 10.

E More Related Work

While our method is closely related to adversarial learning based methods, various other approaches to improve retaining "content" have been developed. Unsupervised machine translation and back-translation approaches have inspired a flurry of works (Prabhumoye et al., 2018; Subramanian et al., 2018; Krishna et al., 2020). Prabhumoye et al. (2018) show that a back translated sentence retrains the content and removes *attribute* related content. A similar approach is followed by Krishna et al. (2020) who observe that paraphrasing has a similar effect. Another group of approach are inspired by Cycle consistency used in CycleGAN (Zhu et al.,

[3]

2017) which is closely related to back translation. The idea is to ensure that a representation/image when translated to another domain, should map to a similar representation/image when mapped back. It is known to improve generating synthesized images in computer vision, and has improved topic modeling in NLP (Hu et al., 2020) and unsupervised cross lingual dictionary induction (Mohiuddin and Joty, 2019). However, cycle consistency loss for sentence level tasks has only been recently tackled by Huang et al. (2020) and yet to prove effective.

Transformer Based Methods - Since pre-trained transformers are known to generate fluent sentences (Radford et al., 2019), newer style transfer works employ them in their pipeline. Style-Transformer (Dai et al., 2019) train a transformer model from scratch for style transfer. But, pretrained transformers with an adversarial classifier have also proven effective (Dathathri et al., 2020). Recently (Krishna et al., 2020) also fine tune a GPT-2 model on pseudo-parallel dataset that is formed by passing sentences through a paraphrase model. However, their use in an adversarially regularized auto-encoder framework is not explored and our initial exploration found that training such systems with transformers is ineffective.

F More details on Human Evaluation

For FL, 0 indicates not fluent at all, 1 indicates somewhat fluent and 2 is a completely fluent sentence. We explicitly ask the annotators to consider semantic similarity for SIM, irrespective of whether the target sentence shares some phrases with the source sentence, with 1 indicating no semantic similarity and 3 indicating complete semantic similarity. For ACC, 1 indicates that the target sentence has only the source sentence style while 2 indicates good transfer to the target style.

Dataset	Metric	α
YELP	ACC	0.69
	FL	0.33
	SIM	0.49
IMDB	ACC	0.60
	FL	0.38
	SIM	0.48
POLITICAL	ACC	0.76
	FL	0.71
	SIM	0.71

Table 7: Krippendorff's alpha showing inter annotator agreement for three datasets YELP, IMDB and POLITICAL

We calculate the Krippendorff's alpha to assess the inter annotator agreement. Table 7 shows the inter-annotator agreement. An α of 0.4 is considered good agreeement (Hedayatnia et al., 2020). We have moderate to good agreements on all the datasets for different measures. On more inspection we found that the disagreements in fluency mostly arrives for small phrases like "my fav" although is an accepted phrase in social media text is considered 2 by one annotator and 3 by another. We also further note that, smaller sentences were easier to judge and had better agreement rates on SIM compared to longer sentences.

Dataset	Source	Target
YELP	consistently slow.	consistently good.
YELP	so nasty.	so delicious!
YELP	i hate mayonnaise.	i love chipotle!
YELP	i 'm so disappointed!	i 'm so impressed!
YELP	but service was horrible both times.	but service was really good & fast.
YELP	now the service i experienced was bad.	now i have the best service.
YELP	the chicken tenders did n't taste like chicken	wtf?,the chicken marsala , really good tomato , love!
YELP	the food was nothing special and the service was slow.	the food was amazing , the service is good.
YELP	that's why i think its shady .	that's why i think its finest.
YELP	that stuff was awful.	that's delicious!
YELP	disgusting all around.	great , all around.
YELP	the rice was dry.	the rice was delicious.
YELP	the sweet and sour chicken is hit and miss.	the sweet and sour chicken is a winner here.
IMDB	the dialog is poorly written	the writing and direction are so precise, and he captures the spirit.
IMDB	i'm a sucker for a good pirate movie, but this ain't it.	i'm a huge fan of the genre , but this movie is definitely worth it.
IMDB	don't see this movie.	don't miss this movie.
IMDB	terrible movie made on zero budget.	absolutely amazing movie on tv.
IMDB	maybe the worse movie i have ever see.	maybe the best movie i have ever seen.
IMDB	never would i recommend this movie to my worst enemy, yet anybody i actually like.	i would recommend this movie to anyone who enjoys good wholesome, clean fun.
IMDB	tedious, not hilarious.	real, great.
IMDB	this movie is truly one of the worst movies i 've ever seen.	this movie is one of the best movies i 've ever seen.
IMDB	it was one of the shortest movies i 've ever seen, and thank god!	it was one of the most original films i've ever seen, and i'm glad.
IMDB	do not watch this movie sober.	do not miss this movie.
IMDB	wesley snipes is a far more accomplished actor than to be in this.	rob roy is a great actor in his own right to date.
IMDB	this film is a real yawner.	this film is a true delight.
IMDB	my rating : 2/10.	my vote : 9/10.
IMDB	some competent acting talent was squandered.	an excellent performance by everyone.
POLITICAL	support you, rand.	support you, elizabeth.
POLITICAL	borders first.	equal rights
POLITICAL	keep telling yourself that	ted.,keep telling that truth, keith.
POLITICAL	just love the constitution.	just love the dnc.
POLITICAL	for supporting clemson and for working for a balance budget .	for supporting student loans for a working and fair job.
POLITICAL	for you service trey !	for you service kamala!
POLITICAL	save america!	save us elizabeth
POLITICAL	stand with your constituents and vote to defund obama care.	stand with your constituents and vote for bernie sanders' bill!!
POLITICAL	poliquin has been a strong voice for the people of northern maine he has my vote	carol has been doing a great job for the people of this state of ohio.

Table 8: More examples from our best performing model for YELP, IMDB, POLITICAL

Dataset	Source	Target	Explanation
YELP	completely outdated, old hotel.	completely charming and old school.	The model produces mixed sentiments without understanding that “old school” has negative connotations
YELP	bad service, bad food.	great food, amazing food.	Lack of diversity in the generation and the model does not produce outputs with respect to service
IMDB	music is boring, and starts to annoy after 15-20 minutes.	its an epic and very moving film, without being preachy.	The model fails to produce semantically similar sentence. Probably because music is not a frequent topic in the dataset
IMDB	brad pitt overacts appallingly.	john woo does it.	Although the the model reproduces a name, it does not produce a fluent sentence
POLITICAL	obamacare, no one wants it!!	al, no one cares it.	Does not understand that "Obamacare" is an entity and hallucinates and uses "care" as a verb
POLITICAL	are clearly not republican anymore!	are not enough sen booker.	Hallucinates Sen Booker which appears frequently in the dataset

Table 9: Mistakes that DCT-ARAE makes and plausible explanations

Source	Target
by far, the best spot for ramen. simple menu. fast service. silky, creamy chicken broth.	by far the best breakfast tacos in the area. friendly staff. great food. ask for the spicy chicken, and they have a great selection.
try sushi boat. it's totally amazing. they offer good food and high quality. good sake is ready. thank you for good place.	love it. good food. they have good margaritas and good food. good prices. there's a good amount of food for you.
best thai in austin. we love the atmosphere, the service and obviously the food. they did a great job last time we were there since our party had specific requirements like <unk> free and <unk>.	best mexican food in the area. the service was great and the food was so good. we had a party of 10 and they were very accommodating to our group of us. we were there last night and the food was good
fabulous, delicious, authentic. at lunch on a saturday the place was packed! 20 minute wait for a table. i was one of two customers who was not chinese. i 'll be back frequently.	awesome mexican food, a little on the corner of a <unk>. i was here on a saturday night. they were busy, but we were able to get a table. i will definitely be back!
this place is great! i grew up going to china inn in chamblee plaza and it's the same owner! lunch service is fast and delicious! give it a shot, you won't be disappointed !	this place is awesome!! i've been coming to this location for years and it's always clean and the service is fast and friendly. it's a great mexican restaurant, you can't go wrong with the food!
awful. i'm writing this as i eat it now. worst poke bowl i've ever had. the smallest portion of poke possible, <unk> overcooked rice, and barely got any ponzu. most standard toppings cost extra too.	awful! i've never had a bad meal here. i only ordered two of them. the only thing i didn't like was the <unk>. it's not much flavor, but the meat is dry.
worst chinese food experience i ever had. told the manager about my allergies and that all i wanted was vegetable fried rice no soy sauce they couldn't even handle that!!! amateur hour here don't waste your time. go to china blossom	worst experience ever. i ordered the <unk> and they were all wrong with that i couldn't eat the food. that's how i don't care about how they charge you for the fajitas. no one ever came to eat here.
the food was terrible. it definitely was not fresh. the broccoli was over cooked on my beef broccoli. my chicken chow mean fried rice just looked and tasted like last weeks rice. there was one chunk of chicken and <unk> pieces of egg in	the food was just ok. the chicken was dry. it was very dry. i ordered the chicken chimichanga and it was just plain gross. the only thing that was <unk> was the chicken burrito. there was only one other person in the <unk>

Table 10: Examples for multiple-attribute dataset