So Different Yet So Alike!
Constrained Unsupervised Text Style Transfer

Anonymous ACL submission

Abstract

Automatic transfer of text between domains has become popular in recent times. One of its aims is to preserve the semantic content while adapting to the target domain. However, it does not explicitly maintain other attributes between the source and translated text: e.g., text length and descriptiveness. Maintaining constraints in transfer has several downstream applications, including data augmentation and debiasing. We introduce a method for such constrained unsupervised text style transfer by introducing two complementary losses to the generative adversarial network (GAN) family of models. Unlike the competing losses used in GANs, we introduce cooperative losses where the discriminator and the generator cooperate and reduce the same loss. The first is a *contrastive* loss and the second is a *classification* loss — aiming to regularize the latent space further and bring similar sentences across domains closer together. We demonstrate that such training retains lexical, syntactic, and domain-specific constraints between domains for multiple benchmark datasets, including ones where more than one attribute change. We show that the complementary cooperative losses improve text quality, according to both automated and human evaluation measures.

1 Introduction

Modern neural networks methods are capable of mapping data from one domain to another. Prominent examples include translation of text between languages (Vaswani et al., 2017; Artetxe et al., 2018; Lample et al., 2017), emoji creation from human faces (Taigman et al., 2017), and stylistic transfer of speech (Yuan et al., 2021). In Natural Language Processing (NLP), the umbrella term *attribute transfer* (Jin et al., 2020b) (or *domain transfer*) refers to similar methods[1]. The aim is to maximally preserve the semantics of the source sentence ("content") but change other properties ("attributes"), such as sentiment (Jin et al., 2020b), expertise (Cao et al., 2020), formality (Rao and Tetreault, 2018) or a combination of them (Subramanian et al., 2018).

[1]While the literature primary utilizes the term *style transfer*, we adopt the more general term *attribute* as suggested by Jin et al. (2020a).

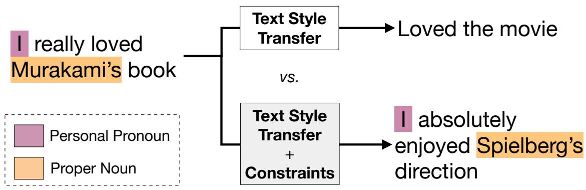


Figure 1: Illustrative example showing transfer of text from books to movies while maintaining constraints of identity.

Text style transfer, a popular form of attribute transfer, regards "style" as any attribute that changes between datasets (Jin et al., 2020a). Building on the progress of supervised transfer models, recent works have focused on *unsupervised style transfer* that avoids costly annotation of parallel sentences. However, models built using unsupervised methods perform poorly when compared to supervised (parallel) training (Artetxe et al., 2020). These methods, while capable of achieving the target domain characteristics, often fail to maintain the invariant content. Figure 1 illustrates one such example, where a sentence from the BOOKS domain is translated to the MOVIE domain. While the translated sentence "*Loved the movie*" has correctly transferred the attribute (style), it does not have the same length, does not retain the personal noun ("I"), nor use a domain-appropriate proper noun. Comparatively, the higher-fidelity transfer "I absolutely enjoyed Spielberg's direction", maintains such *constraints of identity*, in addition to being an aptly transferred sentence.

This problem setting is an important application of text transfer, as enforcing 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 (§ 5). Constraints of identity are explored extensively in the computer vision task of cross-domain image generation. (Taigman et al., 2017), but these issues are unexplored in NLP.

In this paper, we improve unsupervised attribute transfer by enforcing invariances via explicit constraints. Current methods in text attribute transfer lack mechanisms to explicitly enforce such constraints between the source and the transferred sentence. In this work, we map text between two domains with a focus on maintaining constraints of identity between them. To this end, we build upon unsupervised text style transfer work by introducing an additional explicit regularization component in the latent space of a GAN-based *seq2seq* network through two complementary losses. Unlike the adversarial losses in the GAN framework, our proposed losses cooperatively reduce the same objective. 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 helps maintain the sentence identity via constraints from the latent vectors (Odena et al., 2017).

Our approach, while simple and aimed at maintaining constraints, improves the overall performance of the generation. We demonstrate these gains over three datasets: YELP (Zhao et al., 2018b), IMDB (Dai et al., 2019) and POLITICAL (Prabhumoye et al., 2018), generating six constraints including lexical, syntactic and domain-specific. The introduced cooperative losses satisfy the constraints more effectively compared against 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 (§ 4.4.2). To the best of our knowledge, our approach is the first to introduce cooperative losses in a GAN-like setup for NLG.

2 Preliminaries

Task Setup: We consider two sets of sentences (or corpora) $\mathcal{S}= \{x^1_{src}, x^2_{src}, \dots x^m_{src}\}$ and $\mathcal{T}= \{x^1_{trg}, x^2_{trg}, \dots x^n_{trg}\}$, as the *source* and *target* domains, respectively. Each corpus — which we interpret as domains — contain discernable attributes, ranging from sentiment (e.g., positive vs. negative), topics, political slant (e.g., democratic vs. republican), or some combination (Li et al., 2018; Lample et al., 2019). The overall task is to rewrite a piece of text $s_i \in \mathcal{S}$ to $t_i \in \mathcal{T}$, such that the translation changes the attributes varying across the two domains but retains the remaining content. While content retention is not explicitly defined in the literature, we design this new task of constrained unsupervised attribute transfer that assigns explicit constraints $\mathcal{C} = \{c_1, c_2, \dots, c_{|\mathcal{C}|}\}$, to be retained. These constraints can be defined at various levels of a sentence: lexical, syntactic and domain-specific.

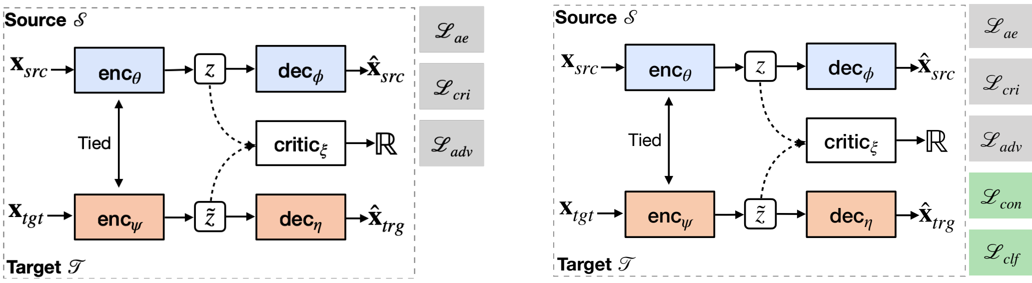
Adversarially Regularized Autoencoder (ARAE): To perform unsupervised attribute transfer, we consider *seq2seq* models that encode source sentences to a latent space and then decodes them to the target sentences. ARAEs (Zhao et al., 2018b) are the auto-encoder variants of the Generative Adversarial Network (GAN) (Goodfellow et al., 2014) framework. They learn smooth latent spaces (by imposing implicit priors) to ease the sampling of latent sentences. ARAEs have been widely adopted in tasks like unsupervised text generation (Huang et al., 2020), topic modeling (Hu et al., 2020), among others, and form the backbone of our proposed model.

ARAE consists of an auto-encoder with a deterministic encoder $enc_\theta : \mathcal{X} \rightarrow \mathcal{Z}$ that encodes sentences into a latent space; i.e., $\mathbf{z} = enc_\theta(\mathbf{x}) \sim P_z$, and a conditional decoder $p_\phi(\mathbf{x}|\mathbf{z})$ that generates a sentence given a latent code. ARAE regularizes this latent space utilizing a GAN-like setup that includes an implicit prior obtained from a parameterized generator network $enc_\psi : \mathcal{N}(0, I) \rightarrow \mathcal{Z}$. Here, enc_ψ maps a noise sample $s \sim \mathcal{N}(0, I)$ to the corresponding prior latent code $\bar{\mathbf{z}} = enc_\psi(s) \sim P_{\bar{z}}$.

A critic $crc_\xi : \mathcal{Z} \rightarrow \mathbb{R}$ then learns to distinguish between real and generated samples, whereas both enc_θ and enc_ψ are adversarially trained to fool the critic. This results in a minimax optimization which implicitly minimizes the JS-Divergence between the two distributions P_z and $P_{\bar{z}}$:

$$\min_{\psi} \max_{\xi} \quad \mathop{\mathbb{E}}_{\mathbf{z}\sim P_z} [crc_\xi(\mathbf{z})] - \mathop{\mathbb{E}}_{\bar{\mathbf{z}}\sim P_{\bar{z}}} [crc_\xi(\bar{\mathbf{z}})] \quad (1)$$

The training involves three optimizations: i) reducing the auto-encoder loss $\mathcal{L}_{ae}$, which tries to reconstruct the input and encourages copying behavior and maintain semantics similar to original text (Eq. 2); ii) optimizing the critic's loss $\mathcal{L}_{cri}$ to distinguish between real and fake samples (Eq. 3); and iii) training the encoder and generator loss



(a) (b)

Figure 2: a) ARAE$_{seq2seq}$ – We replace the generator of ARAE with an encoder that encodes text from $\mathcal{T}$. (b) Adding our proposed cooperative losses to the model.

$\mathcal{L}_{adv}$ to fool the critic (Eq. 4):

$$\mathcal{L}_{ae}(\theta,\phi) = \mathop{\mathbb{E}}_{\mathbf{z}\sim P_z}\left[-\log p_\phi(\mathbf{x}|\mathbf{z})\right], \quad (2)$$

$$\mathcal{L}_{crc}(\xi) = -\mathop{\mathbb{E}}_{\mathbf{z}\sim P_z}\left[crc_\xi(\mathbf{z})\right] + \mathop{\mathbb{E}}_{\bar{\mathbf{z}}\sim P_{\bar{z}}}\left[crc_\xi(\bar{\mathbf{z}})\right], \quad (3)$$

$$\mathcal{L}_{adv}(\theta,\psi) = \mathop{\mathbb{E}}_{\mathbf{z}\sim P_z}\left[crc_\xi(\mathbf{z})\right] - \mathop{\mathbb{E}}_{\bar{\mathbf{z}}\sim P_{\bar{z}}}\left[crc_\xi(\bar{\mathbf{z}})\right]. \quad (4)$$

3 Proposed Method

3.1 Base Model (ARAE$_{seq2seq}$)

While ARAE is an auto-encoder that recreates input $\mathbf{x} \rightarrow \hat{\mathbf{x}}$, our requirement is to translate sentences from one domain to another. Given this, we modify the ARAE to a *seq2seq* variant such that we can translate two input sentences from both source and target domains; i.e., $\mathbf{x}_{src} \rightarrow \hat{\mathbf{x}}_{tgt}$ and $\mathbf{x}_{tgt} \rightarrow \hat{\mathbf{x}}_{src}$.

To achieve this, we utilize enc_θ to encode $\mathbf{x}_{src}$ and repurpose enc_ψ to encode $\mathbf{x}_{tgt}$. We obtain their latent codes $(\mathbf{z},\bar{\mathbf{z}})$ which we name as $(\mathbf{z}^s,\mathbf{z}^t)$, i.e., $\mathbf{z}^s = enc_\theta(\mathbf{x}_{src})$ and $\mathbf{z}^t = enc_\psi(\mathbf{x}_{tgt})$.

Next, to generate sentences, we consider two decoders $\hat{\mathbf{x}}_{src} \sim p_\phi(\mathbf{x}|\mathbf{z})$ and $\hat{\mathbf{x}}_{tgt} \sim p_\eta(\mathbf{x}|\mathbf{z})$. Here, $\mathbf{z}$ can be either $\mathbf{z}^s$ or $\mathbf{z}^t$ based on whether we auto-encode (e.g., $p_\phi\left(\mathbf{x}|\mathbf{z}^s = enc_\theta(\mathbf{x}_{src})\right)$) or translate (e.g., $p_\phi\left(\mathbf{x}|\mathbf{z}^t = enc_\psi(\mathbf{x}_{tgt})\right)$). Unlike ARAE's single decoder, we incorporate two decoders to enable bi-directional translation.

In the above process, instead of sampling s from a noise distribution like $\mathcal{N}(0,I)$ and passing it through a generator enc_ψ, we feed it text from the target domain $\mathcal{T}$ and a decoder dec_η that decodes text in $\mathcal{T}$. This is inspired from Cycle-GAN (Zhu et al., 2017), where instead of matching the noise distribution $\mathcal{N}$, we match the distribution of $\mathcal{T}$.

In addition, we tie the weights of the encoders from both 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)[2].

Algorithm 1: ARAE$_{seq2seq}$ + CLF + CONTRA

for *each training iteration* **do**
- 1) Train the Auto-encoders:
  - Sample $\mathbf{x}_{src} \sim \mathcal{S}$, $\mathbf{x}_{trg} \sim \mathcal{T}$
  - $\mathbf{z}^s = enc_\theta(\mathbf{x}_{src})$, $\mathbf{z}^t = enc_\psi(\mathbf{x}_{trg})$
  - Backprop loss, $\mathcal{L}_{ae}(\theta,\phi)$, $\mathcal{L}_{ae}(\psi,\eta)$
- 2) Train the Critic:
  - Sample $\mathbf{x}_{src} \sim \mathcal{S}$, $\mathbf{x}_{trg} \sim \mathcal{T}$
  - $\mathbf{z}^s = enc_\theta(\mathbf{x}_{src})$, $\mathbf{z}^t = enc_\psi(\mathbf{x}_{trg})$
  - $\mathbf{z}^s_{crc} = crc^{hid}_\xi(\mathbf{z}^s)$, $\mathbf{z}^t_{crc} = crc^{hid}_\xi(\mathbf{z}^t)$
  - $l_{crc} \leftarrow \mathcal{L}_{crc}(\xi)$
- 2a) Critic Co-op Training:
  - Backprop loss, $l_{crc} + \lambda_1\mathcal{L}_{con}(\xi) + \lambda_2\mathcal{L}_{clf}(\xi,\delta)$
- 3) Adversarial Training:
  - Sample $\mathbf{x}_{src} \sim \mathcal{S}$, $\mathbf{x}_{trg} \sim \mathcal{T}$
  - $\mathbf{z}^s = enc_\theta(\mathbf{x}_{src})$, $\mathbf{z}^t = enc_\psi(\mathbf{x}_{trg})$
  - Backprop loss, $\mathcal{L}_{adv}(\theta,\psi)$
- 3a) Encoder Co-op Training:
  - Backprop loss, $\lambda_1\mathcal{L}_{con}(\theta,\phi) + \lambda_2\mathcal{L}_{clf}(\theta,\phi,\delta)$

3.2 Adding Constraints via Co-op Training

While the latent space in ARAE$_{seq2seq}$ learns to match $\mathcal{S}$ and $\mathcal{T}$ sentences, there is no guarantee on translations maintaining the "content". This issue is particularly pronounced in unsupervised attribute transfer due to lack of parallel sentences between $\mathcal{S}$ and $\mathcal{T}$.

To alleviate the issue, we propose to learn a structured latent space which embodies notions of our constraints in its embedded latent codes. This ensure that instances with similar constraints are closer in the latent space. In particular, we propose

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

two types of optimization — self-supervised and discriminative — to maintain the constraints better.

3.2.1 Cooperative Contrastive Learning

We use contrastive representation learning to regularize the latent space, such that encoders bring two sentences sharing similar constraints closer together (positive pairs), and force dissimilar ones away (negative pairs). For example, sentences of similar lengths (irrespective of their domains) should be closer together.

Among many self-supervised metric losses such as 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 sentence $s_i \in \mathcal{S}$ 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:

$$\mathcal{L}_{con}(\theta,\psi,\xi) = -\frac{1}{|P|}\log\left(\sum_{j=1}^{P}\frac{e^{(\mathbf{z}_i\cdot\mathbf{z}_j)}}{\sum_{k=1}^{B\setminus\{i\}}e^{(\mathbf{z}_i\cdot\mathbf{z}_k)}}\right), \tag{5}$$

where $\mathbf{z}$'s are representations obtained from the encoders in $\mathcal{S}, \mathcal{T}$ or representations obtained from the last layer of critic crc_ξ. $\mathcal{C}_i$ 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 18).

3.2.2 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 $D_\delta : \mathcal{Z} \to \mathbb{R}^{|\mathcal{C}|}$ that predicts the different constraints $\mathcal{C}$ of the sentences and the binary cross entropy loss is reduced.

$$\mathcal{L}_{clf}(\theta,\phi,\xi,\delta) = -\sum_{c=1}^{|\mathcal{C}|}\log\Big(\sigma(l_c)^{y_c}(1-\sigma(l_c))^{1-y_c}\Big), \tag{6}$$

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 $\mathcal{S}$, $\mathcal{T}$ or from the hidden layers of the critic.

The overall training process is highlighted in Algorithm 1 where $\mathcal{L}_{con}$ and $\mathcal{L}_{clf}$ are weighted by λ_1 and λ_2. We choose $\lambda_1, \lambda_2 \in \{0,1\}$.

4 Experiments

Datasets. We use three datasets with single attribute changes: i) **Yelp Reviews**: business reviews listed on Yelp, labeled as either a positive or negative sentiment. ii) **IMDb Movie Reviews**: consists of movie reviews (Dai et al., 2019) also labelled as positive or negative. iii) **Political Slant**: consists of Facebook posts from the politicians of the United States Senate and the House of Representatives (Prabhumoye et al., 2018), labeled with either democratic/republican slant. See Appendix A for dataset statistics.

Constraints: We constrain every sentence along *six* diverse dimensions that we desire to control between the two domains: i) **Lexical**: *Sentence length* – The transferred sentence should maintain a length similar to the original sentence (binarized to long sentences with 10 or or more words or short otherwise). ii) **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). iii) **Domain specific** – number of *domain-specific attributes* (Li et al., 2018) (categorical up to 5). 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.

4.1 Model Details

For the encoders, we use a one-layer LSTM network with 300 hidden dimensions for all the datasets. For the critics and classification loss, we use a two-layer multilayer perceptron with 100 hidden units. Our learning rates and methods to stabilize training are discussed in Appendix B.

4.2 Evaluation Setup

Automatic Evaluation: Our automatic evaluation considers the following three prominent criteria: i) **Semantic Similarity (SIM):** Measured between source and translated target sentences using encoders (Wieting et al., 2019), instead of *n-gram* metrics like BLEU (Papineni et al., 2002) which

Model	Sampling	YELP				IMDB				POLITICAL			
		ACC	FL	SIM	AGG	ACC	FL	SIM	AGG	ACC	FL	SIM	AGG
DRG	greedy	67.4	54.5	**43.6**	16.7	56.5	44.3	**54.1**	14.4	61.3	35.7	38.7	8.8
ARAE	greedy	**93.1**	67.9	31.2	19.8	95.0	76.3	26.4	19.9	63.0	72.1	17.3	11.0
ARAE$_{seq2seq}$	greedy	88.3	66.0	34.4	20.2	95.4	70.5	**36.4**	26.0	95.80	53.1	28.5	14.1
	nucleus($p=0.6$)	86.7	63.9	35.3	19.9	95.1	69.8	**36.4**	25.6	95.8	52.2	28.4	13.9
ARAE$_{seq2seq}$ + CLF	greedy	85.7	63.4	36.7	20.2	96.0	73.6	35.4	26.2	98.6	55.0	**44.4**	**25.5**
	nucleus($p=0.6$)	85.6	63.0	36.6	20.0	95.8	72.8	35.3	25.7	98.6	54.4	44.2	25.1
ARAE$_{seq2seq}$ + CONTRA	greedy	89.6	**69.7**	32.0	20.1	97.6	82.9	32.5	27.0	99.0	56.5	40.8	24.2
	nucleus($p=0.6$)	89.7	69.2	31.9	20.0	97.7	83.2	32.2	26.7	99.0	55.9	40.7	23.9
ARAE$_{seq2seq}$ + CLF + CONTRA	greedy	89.3	69.2	32.9	**20.6**	**97.8**	**84.0**	33.5	**28.1**	**99.0**	**56.8**	41.8	24.9
	nucleus($p=0.6$)	89.4	68.6	32.8	20.4	97.1	82.6	33.6	27.4	99.0	56.0	41.6	24.4

Table 1: Evaluation of ARAE$_{seq2seq}$ 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. We report the mean of five runs for our experiments. The bolded measures are the best results

have weak correlations with human judgments. *ii*) **Transfer Accuracy (ACC):** The transferred sentence should belong to the target domain and a classifier is trained to distinguish between the source and the target sentence. We use *fastText* classifiers (Joulin et al., 2017) for every dataset. We achieve accuracy of 97.9 for YELP, 96.9 for IMDB and 97.1 for POLITICAL. *iii*) **Fluency (FL):** A transferred sentence should be grammatically correct. We fine-tune a RoBERTa-large model on the COLA (Warstadt et al., 2018) dataset to indicate whether a sentence is linguistically acceptable. Finally, we combine the three scores into an aggregate, following the criteria suggested by Krishna et al. (2020):

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

Human Evaluation: We also perform an indicative human evaluation where we randomly sample 100 samples from each of the three datasets and hire three researchers to rate every sentence for FL, SIM and ACC on a 3-point scale (Krishna et al., 2020).

4.3 Baselines

We compare ARAE$_{seq2seq}$ with the following baselines: **a) DRG**: The Delete, Retrieve, 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) ARAE$_{seq2seq}$**: Our model without the contrastive learning or cooperative classifier; **d) ARAE$_{seq2seq}$ + CONTRA**: Our model with the contrastive learning; **e) ARAE$_{seq2seq}$ + CLF**: Our model with the cooperative classifier; **f) ARAE$_{seq2seq}$+CLF+CONTRA**: Our model with both the cooperative losses. The closest model to ours is from (Huang et al., 2020). However, we were not able to reproduce the results.[3]

4.4 Results

4.4.1 Overall Results

ARAE$_{seq2seq}$ + CONTRA and ARAE$_{seq2seq}$ + CLF consistently perform better than DRG and ARAE on the AGG score (Table 1). The AGG for YELP is 20.6 (vs. 19.8), for IMDB it is 28.1 (vs. 19.9) and for POLITICAL 25.5 (vs. 11.0). Although cooperative loss reduction aims to satisfy the constraints between two domains, our results show that further regularization of the latent space not only brings advantages in satisfying the constraints but also improves performance (Lavoie-Marchildon et al., 2020).

Effect of Cooperative Loss Reduction on ACC and FL and SIM: Across datasets, reducing cooperative losses improves ACC and FL and SIM to ARAE. Although DRG produces sentences with high SIM as 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. Also, compared to ARAE, adding cooperative losses significantly increases the SIM, with the highest increase observed for POLITICAL. The reasons for this could be two-fold: *i*) since we mine positive sentences from a corpus that is grounded in real world events, most lexically-similar sentences may also be semantically similar (Guu et al., 2018), and *ii*) since we tie the encoders from the source and target domain, we extract domain-agnostic information before generation, which retains content.

Fluency (FL) also improves over all datasets. We hypothesize that reducing cooperative losses reg-

[3]Repeated attempts to obtain the original source code failed.

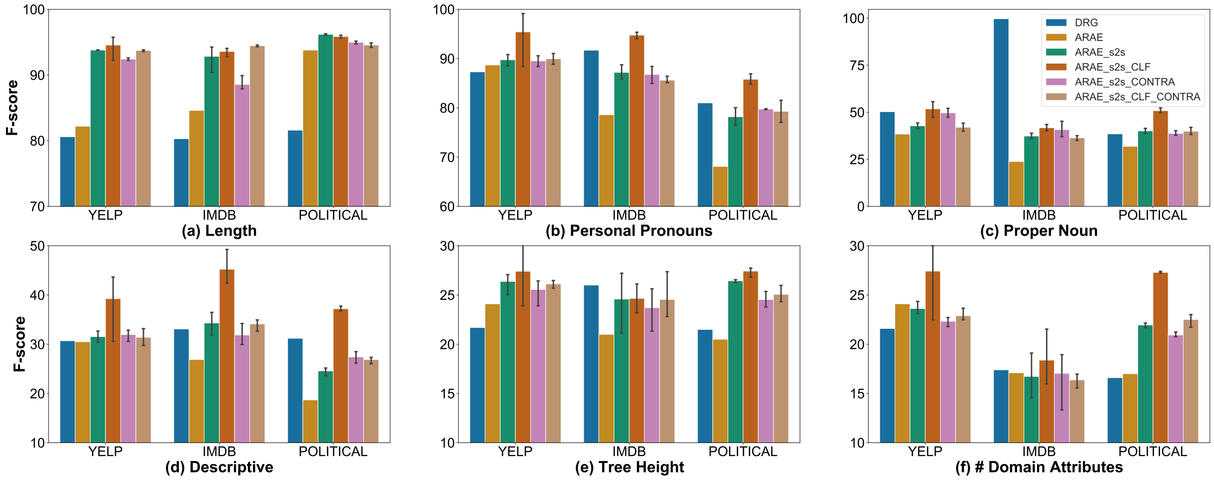


Figure 3: F-scores of different constraints. Adding cooperative losses helps in better maintaining the constraints. The error bars show the variance of generating text using greedy decoding and nucleus sampling with $p = \{0.6, 0.9\}$.

ularizes the latent space bringing fluent sentences closer together, enabling the decoder to produce semantically similar and linguistically acceptable sentences. The improvement for POLITICAL is less; we find these source sentences themselves are less fluent and contain many U.S. political acronyms, and that our system produces many out-of-vocabulary words affecting fluency.

Nucleus Sampling: 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 in Table 1, which does produce more diversity, increasing 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 — similar to Krishna et al. (2020).

Effect of the Number of Positives: The number of positive and negative samples used for contrastive learning (Eq. 5) have a significant effect on the overall performance (Khosla et al., 2020; Chen et al., 2020; Henaff, 2020). Table 2 (*rows* $|P| \in \{1, 2, 5, 10\}$) shows the AGG scores on IMDB (for one of the runs), 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 example is 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 the importance of adding 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.

Model	ACC	FL	SIM	AGG
$\text{ARAE}_{seq2seq}$ + 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
$\text{ARAE}_{seq2seq}$ + 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
$\|P\| = 1$	92.4	75.5	36.6	26.2
$\|P\| = 2$	96.1	80.6	36.0	28.6
$\|P\| = 5$	96.0	84.0	31.4	26.0
$\|P\| = 10$	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.

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.

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.

Human Evaluation: 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 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. (See Appendix D)

Dataset	Input	Output (Ours)	Output (ARAE)
YELP	they close earlier than posted hours	they're open late night	they keep me getting better
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.
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		
Personal Pronoun	*Source* (IMDB)	jean seberg had not one iota of acting talent.
	Ours	michael keaton was also great in his role.
	ARAE	john abraham had one of my favorite roles .
Proper Pronoun	*Source* (IMDB)	chris klein's character was unlikable from the start and never made an improvement
	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. Our method maintains constraints like number of proper nouns between sentences.

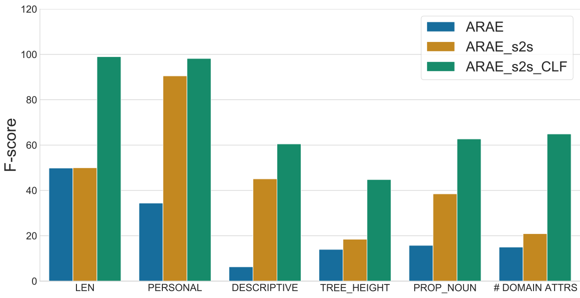


Figure 4: Comparison of ARAE, ARAE$_{seq2seq}$ and ARAE$_{seq2seq}$ + CLF for different constraints.

Qualitative Examples: Table 4 shows examples of the quality of transferred examples (see Appendix C for more). Mistakes made by the model can be attributed to poor understanding of the original semantics, lack of diversity, and not producing attribute-specific words.

4.4.2 Maintaining Constraints

Figure 3 shows that introducing the cooperative losses significantly outperform DRG and ARAE in maintaining constraints. Specifically the ARAE$_{seq2seq}$ + CLF model performs better than ARAE$_{seq2seq}$+ 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, *seq2seq* 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 ARAE$_{seq2seq}$ + 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. ARAE$_{seq2seq}$ + 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. DRG maintains the proper noun for IMDB effectively, because it contains a wide variety of actor and movie names. They are retained verbatim after the delete operation.

Multiple Attribute Datasets: 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 ARAE$_{seq2seq}$ + CLF model with the ARAE$_{seq2seq}$ and ARAE in Figure 4. The results are more pronounced in this case with ARAE$_{seq2seq}$ + CLF having clear advantage over ARAE$_{seq2seq}$. 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.

5 Discussion

Cycle Consistency Loss: a) *In Latent Spaces* - Cycle consistency in latent spaces has been shown to improve word level tasks, such as cross lingual dictionary construction (Mohiuddin and Joty, 2019) and topic modeling (Hu et al., 2020). A recent work from (Huang et al., 2020) claims to improve unsupervised style transfer using such losses. In our experiments, however, it did not result in any noticeable performance improvement [4]. Given this, we hypothesize that 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 auto-encoder training step which is expensive and we defer exploring this.

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

Transferred sentences as Adversarial Examples: We demonstrate an important application of our proposed constrained transfer by considering them as adversarial examples for domain adaptation. Domain Adversarial Neural Network (DANN) (Ganin et al., 2017) is an unsupervised domain adaptation method that improves performance of an end-task (e.g, sentiment analysis) on a target domain considering only supervised data from source domain. We train DANN for sentiment analysis on amazon reviews dataset (He and McAuley, 2016) with DVD as source and ELECTRONICS as the target domain – achieving an accuracy of 83.75% on ELECTRONICS.

Next, we train the best variant of ARAE$_{seq2seq}$ to transfer a separate set DVD reviews to ELECTRONICS reviews and use them as adversarial examples to test the DANN model [5]. We find that the accuracy of DANN on the ELECTRONICS domain reduces by $\sim$3 points. This shows the potential application of domain transferred sentences as adversarial examples. Similar ideas have been tried for image style transfer (Xu et al., 2020), but needs more investigation in text attribute transfer.

[4] Repeated attempts to obtain source codes failed.

[5] Since each of DVD and ELECTRONICS contain positive and negative reviews, we test whether transferred sentences maintain the appropriate sentiment and find the accuracy to be 79%.

6 Related Work

Text attribute transfer has a vast literature (Jin et al., 2020a) with deep learning methods becoming popular. The methods are either supervised – requiring parallel data and unsupervised. Supervised methods repurpose Sequence to Sequence models used in machine translation to achieve the goals (Rao and Tetreault, 2018). However, obtaining parallel data is cumbersome and thus unsupervised methods that consider pseudo-parallel data have become popular.

Disentanglement approaches are the prevalent approach to tackle unsupervised attribute transfer: *attributes* and *content* are separated in latent dimension. To disentangle the attributes adversarial methods maximize the loss of a pretrained attribute classifier (Li et al., 2020; Fu et al., 2018; Zhao et al., 2018a; John et al., 2019). However, the literature has paid little attention in defining and preserving content. Cycle consistency losses – imposing that reconstruction from the target style sentence should resemble the source sentence – is the most prevalent (Prabhumoye et al., 2018; Logeswaran et al., 2018; Dai et al., 2019; Huang et al., 2020; Yi et al., 2020). However, this is expensive, non differentiable requiring reinforcement learning techniques to enforce it. Our work defines the different constraints that should be preserved and adds simple differentiable contrastive learning losses to preserve them.

In recent times, text style transfer models are moving away from disentanglement approaches (Subramanian et al., 2018). Recent works that use transformers for style transfer also have adopted this (Dai et al., 2019; Krishna et al., 2020). However, these methods do not explicitly maintain the constraints between the two styles which is the main aim of our work.

7 Conclusion

Text style transfer works focuses on retaining content and changing the style of sentences but does not maintain other desirable constraints. We address this by introducing two cooperative losses to the GAN-inspired Adversarially Regularized Autoencoder (ARAE) that further regularizes the latent space. While satisfying the constraints our methods brings significant improvements in overall score. While we focus on simple constraints at the sentence- and word-level, future work can add phrase-level and more fine-grained constraints. Potential future work may explore reinforcement learning losses to directly optimize the constraints.

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.

Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. 2020. A call for more rigor in unsupervised cross-lingual learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7375–7388, Online. Association for Computational Linguistics.

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.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 663–670. AAAI Press.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. 2017. Domain-adversarial training of neural networks. In Gabriela Csurka, editor, *Domain Adaptation in Computer Vision Applications*, Advances in Computer Vision and Pattern Recognition, pages 189–209. Springer.

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.

Ruining He and Julian J. McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 507–517. ACM.

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.

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

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.

Yuan Li, Chunyuan Li, Yizhe Zhang, Xiujun Li, Guoqing Zheng, Lawrence Carin, and Jianfeng Gao. 2020. Complementary auxiliary classifiers for label-conditional text generation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8303–8310. AAAI Press.

Lajanugen Logeswaran, Honglak Lee, and Samy Bengio. 2018. Content preserving text generation with attribute controls. *CoRR*, abs/1811.01135.

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.

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.

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.

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.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2018. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.

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.

Qiuling Xu, Guanhong Tao, Siyuan Cheng, Lin Tan, and Xiangyu Zhang. 2020. Towards feature space adversarial attack. *CoRR*, abs/2004.12385.

Xiaoyuan Yi, Zhenghao Liu, Wenhao Li, and Maosong Sun. 2020. Text style transfer via learning style instance supported latent space. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3801–3807. ijcai.org.

Siyang Yuan, Pengyu Cheng, Ruiyi Zhang, Weituo Hao, Zhe Gan, and Lawrence Carin. 2021. Improving zero-shot voice style transfer via disentangled representation learning. *arXiv preprint arXiv:2103.09420*.

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 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.

C 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.

D 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.

Information about participants: We hire three graduate researchers in NLP (average age 25) for the annotation task who are well versed in English. We obtained permission for their participation and compensated them appropriately according to hourly wages in the country. The specific instruction given to them for the evaluation are as follows.
Consider two sentences

- **Source sentence**: Sentence from the source domain

- **Target sentence**: The transferred sentence produced by one of the systems

For every target sentence you will be asked to rate it according to three measures described below.

Fluency: Indicate how fluent the target sentence is (regardless of whether the sentence is appropriately transferred to the target sentence)

1 - Not fluent at all - Does not look like an English sentence.

2 - Fluent but with some mistakes - Fluent but with some grammatical errors

3 - Entirely fluent. - A good English Sentence

Similarity: Indicate how semantically similar the target sentence is.

1 - Does not share any words/phrases with the source sentence and/or is not semantically similar (does not share high level topics of the sentence)

2 - Shares some words/phrases with the source sentence and/or has moderate level of semantic similarity (talks about similar high level topics)

3 - Shares appropriate words/phrases with the source sentence and is highly semantically similar

Accuracy: Indicate whether the target sentence is accurately transferred to the target domain

Sentiment Transfer

1 - The target sentiment is not evident in the target sentence at all. Has words expressing opposite sentiment

2 - Neutral Sentiment. Choose this option, if it has both positive and negative sentiment

3 - The target sentiment is evident in the target sentiment. Has appropriate sentiment bearing words.

If the sentence itself has no sentiment then chose 2

Political Orientation

1 - Talks about topics with the other orientation. For example, if the target style is democratic and the target sentence talks about conservative issues like abortion, gun control

2 - Neutral.

3 - Talks about topics with the correct orientation. For example, if the target style is democratic and talks about progressive issues like liberty, free speech, Elizabeth Warren, Joe Biden, gay rights etc.

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 ARAE$_{seq2seq}$ 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