Improving Disentangled Text Representation Learning with Information-Theoretic Guidance

Pengyu Cheng[1], **Martin Renqiang Min**[2], **Dinghan Shen**[4], **Christopher Malon**[2], **Yizhe Zhang**[3], **Yitong Li**[1], **Lawrence Carin**[1]
[1]Duke University [2]NEC Labs America [3]Microsoft Research
[4]Microsoft Dynamics 365 AI
pengyu.cheng@duke.edu

Abstract

Learning disentangled semantic representations of natural language is essential for many NLP tasks, *e.g.*, conditional text generation, style transfer, personalized dialogue systems, *etc*. Similar problems have been studied extensively for other forms of data, such as images and videos. However, the discrete nature of natural language makes the disentangling of textual representations more challenging (*e.g.*, the manipulation over the data space cannot be easily achieved). Inspired by information theory, we propose a novel method that effectively manifests disentangled representations of text, without any supervision on semantics. A new mutual information upper bound is derived and leveraged to measure dependence between style and content. By minimizing this upper bound, the proposed method induces style and content embeddings into two independent low-dimensional spaces. Experiments on both conditional text generation and text-style transfer demonstrate the high quality of our disentangled representation in terms of content and style preservation.

1 Introduction

Disentangled representation learning (DRL), which maps different aspects of data into distinct and independent low-dimensional latent vector spaces, has attracted considerable attention for making deep learning models more interpretable. Through a series of operations such as selecting, combining, and switching, the learned disentangled representations can be utilized for downstream tasks, such as domain adaptation (Liu et al., 2018), style transfer (Lee et al., 2018), conditional generation (Denton et al., 2017; Burgess et al., 2018), and few-shot learning (Kumar Verma et al., 2018). Although widely used in various domains, such as images (Tran et al., 2017; Lee et al., 2018), videos (Yingzhen and Mandt, 2018; Hsieh et al., 2018), and speech (chieh Chou et al., 2018; Zhou et al., 2019), many challenges in DRL have received limited exploration in natural language processing (John et al., 2019).

To disentangle various attributes of text, two distinct types of embeddings are typically considered: the *style embedding* and the *content embedding* (John et al., 2019). The content embedding is designed to encapsulate the semantic meaning of a sentence. In contrast, the style embedding should represent desired attributes, such as the sentiment of a review, or the personality associated with a post. Ideally, a disentangled-text-representation model should learn representative embeddings for both style and content.

To accomplish this, several strategies have been introduced. Shen et al. (2017) proposed to learn a semantically-meaningful content embedding space by matching the content embedding from two different style domains. However, their method requires predefined style domains, and thus cannot automatically infer style information from unlabeled text. Hu et al. (2017) and Lample et al. (2019) utilized one-hot vectors as style-related features (instead of inferring the style embeddings from the original data). These models are not applicable when new data comes from an unseen style class. John et al. (2019) proposed an encoder-decoder model in combination with an adversarial training objective to infer both style and content embeddings from the original data. However, their adversarial training framework requires manually-processed supervised information for content embeddings (*e.g.*, reconstructing sentences with manually-chosen sentiment-related words removed). Further, there is no theoretical guarantee for the quality of disentanglement.

In this paper, we introduce a novel **I**nformation-theoretic **D**isentangled **E**mbedding **L**earning method (IDEL) for text, based on guidance from

information theory. Inspired by Variation of Information (VI), we introduce a novel information-theoretic objective to measure how well the learned representations are disentangled. Specifically, our IDEL reduces the dependency between style and content embeddings by minimizing a sample-based mutual information upper bound. Furthermore, the mutual information between latent embeddings and the input data is also maximized to ensure the representativeness of the latent embeddings (*i.e.*, style and content embeddings). The contributions of this paper are summarized as follows:

A principled framework is introduced to learn disentangled representations of natural language. By minimizing a novel VI-based DRL objective, our model not only explicitly reduces the correlation between style and content embeddings, but also simultaneously preserves the sentence information in the latent spaces.

A general sample-based mutual information upper bound is derived to facilitate the minimization of our VI-based objective. With this new upper bound, the dependency of style and content embeddings can be decreased effectively and stably.

The proposed model is evaluated empirically relative to other disentangled representation learning methods. Our model exhibits competitive results in several real-world applications.

2 Preliminary

2.1 Mutual Information Variational Bounds

Mutual information (MI) is a key concept in information theory, for measuring the dependence between two random variables. Given two random variables x and y, their MI is defined as

$$\mathrm{I}(x;y) = \mathrm{E}_{p(\boldsymbol{x};\boldsymbol{y})}[\log \frac{p(x;y)}{p(x)p(y)}]: \qquad (1)$$

where $p(x;y)$ is the joint distribution of the random variables, with $p(x)$ and $p(y)$ representing the respective marginal distributions.

In disentangled representation learning, a common goal is to minimize the MI between different types of embeddings (Poole et al., 2019). However, the exact MI value is difficult to calculate in practice, because in most cases the integral in Eq. (1) is intractable. To address this problem, various MI estimation methods have been introduced (Chen et al., 2016; Belghazi et al., 2018; Poole et al., 2019). One of the commonly used estimation approaches is the Barber-Agakov lower bound (Barber and Agakov, 2003). By introducing a variational distribution $q(xjy)$, one may derive

$$\mathrm{I}(x;y) \quad \mathrm{H}(x) + \mathrm{E}_{p(\boldsymbol{x};\boldsymbol{y})}[\log q(xjy)]; \qquad (2)$$

where $\mathrm{H}(x) = \mathrm{E}_{p(\boldsymbol{x})}[\ \log p(x)]$ is the entropy of variable x.

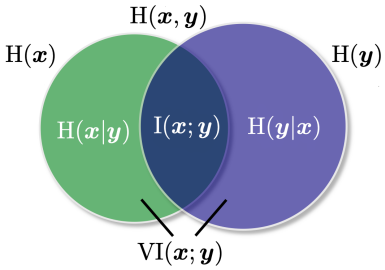

Figure 1: The green and purple circles represent the entropy of x and y, respectively. The intersection (blue region) is the mutual information between x and y. The symmetric difference of the two circles (green and purple regions) is $\mathrm{VI}(x;y)$.

2.2 Variation of Information

In information theory, Variation of Information (VI, also called Shared Information Distance) is a measure of independence between two random variables. The mathematical definition of VI between random variables x and y is

$$\mathrm{VI}(x;y) = \mathrm{H}(x) + \mathrm{H}(y) \quad 2\mathrm{I}(x;y); \qquad (3)$$

where $\mathrm{H}(x)$ and $\mathrm{H}(y)$ are entropies of x and y, respectively (shown in Figure 1). Kraskov et al. (2005) show that VI is a well-defined metric, which satisfies the triangle inequality:

$$\mathrm{VI}(y;x) + \mathrm{VI}(x;z) \quad \mathrm{VI}(y;z); \qquad (4)$$

for any random variables x, y and z. Additionally, $\mathrm{VI}(x;y) = 0$ indicates x and y are the same variable (Meilă, 2007). From Eq. (3), the VI distance has a close relation to mutual information: if the mutual information is a measure of "dependence" between two variables, then the VI distance is a measure of "independence" between them.

3 Method

Consider data $f(x_i;y_i)g_{i=1}^N$, where each x_i is a sentence drawn from a distribution $p(x)$, and y_i

is the label indicating the style of x_i. The goal is to encode each sentence x_i into its corresponding style embedding s_i and content embedding c_i with an encoder $q(s, c|x)$:

$$s_i, c_i \sim q(s, c|x_i) \quad (5)$$

The collection of style embeddings $\{s_i\}_{i=1}^N$ can be regarded as samples drawn from a variable s in the style embedding space, while the collection of content embeddings $\{c_i\}_{i=1}^N$ are samples from a variable c in the content embedding space. In practice, the dimension of the content embedding is typically higher than that of the style embedding, considering that the content usually contains more information than the style (John et al., 2019).

We first give an intuitive introduction to our proposed VI-based objective, then in Section 3.1 we provide the theoretical justification for it. To disentangle the style and content embedding, we try to minimize the mutual information $I(s; c)$ between s and c. Meanwhile, we maximize $I(c; x)$ to ensure that the content embedding s sufficiently encapsulates information from the sentence x. The embedding s is expected to contain rich style information. Therefore, the mutual information $I(s; y)$ should be maximized. Thus, our overall disentangled representation learning objective is: $\mathcal{L}_{\text{Dis}} = I(s; c) - I(c; x) - I(s; y)$:

3.1 Theoretical Justification of the Objective

The objective $\mathcal{L}_{\text{Dis}}$ has a strong connection with the independence measurement in information theory. As described in Section 2.2, Variation of Information (VI) is a well-defined metric of independence between variables. Applying the triangle inequality from Eq. (4) to s, c and x, we have $VI(s; x) + VI(x; c) \geq VI(s; c)$: Equality occurs if and only if the information from variable x is totally separated into two independent variable s and c, which is an ideal scenario for disentangling sentence x into its corresponding style embedding s and content embedding c.

Therefore, the difference between $VI(s; x) + VI(x; c)$ and $VI(s; c)$ represents the degree of disentanglement. Hence we introduce a measurement:

$$D(x; s; c) = VI(s; x) + VI(x; c) - VI(c; s):$$

From Eq. (4), we know that $D(x; y; z)$ is always non-negative. By the definition of VI in Eq. (3), $D(x; s; c)$ can be simplified as:

$$\begin{aligned} &VI(c; x) + VI(x; s) - VI(s; c) \\ =&2H(x) + 2[I(s; c) - I(x; c) - I(x; s)]: \end{aligned}$$

Since $H(x)$ is a constant associated with the data, we only need to focus on $I(s; c) - I(x; c) - I(x; s)$.

The measurement $D(x; s; c)$ is symmetric to style s and content c, giving rise to the problem that without any inductive bias in supervision, the disentangled representation could be meaningless (as observed by Locatello et al. (2019)). Therefore, we add inductive biases by utilizing the style label y as supervised information for style embedding s. Noting that $s \to x \to y$ is a Markov Chain, we have $I(s; x) \geq I(s; y)$ based on the MI data-processing inequality (Cover and Thomas, 2012). Then we convert the minimization of $I(s; c) - I(x; c) - I(x; s)$ into the minimization of the upper bound $I(s; c) - I(x; c) - I(y; s)$, which further leads to our objective $\mathcal{L}_{\text{Dis}}$.

However, minimizing the exact value of mutual information in the objective $\mathcal{L}_{\text{Dis}}$ causes numerical instabilities, especially when the dimension of the latent embeddings is large (Chen et al., 2016). Therefore, we provide several MI estimations to the objective terms $I(s; c)$, $I(x; c)$ and $I(s; y)$ in the following two sections.

3.2 MI Variation Lower Bound

To maximize $I(x; c)$ and $I(s; y)$, we derive two variational lower bounds. For $I(x; c)$, we introduce a variational decoder $q(x|c)$ to reconstruct the sentence x by the content embedding c. Leveraging the MI variational lower bound from Eq. (2), we have $I(x; c) \geq H(x) + \mathbb{E}_{p(x,c)}[\log q(x|c)]$: Similarly, for $I(s; y)$, another variational lower bound can be obtained as: $I(s; y) \geq H(y) + \mathbb{E}_{p(y,s)}[\log q(y|s)]$, where $q(y|s)$ is a classifier mapping the style embedding s to its corresponding style label y. Based on these two lower bounds, $\mathcal{L}_{\text{Dis}}$ has an upper bound:

$$\begin{aligned} \mathcal{L}_{\text{Dis}} \leq\ & I(s; c) - [H(x) + \mathbb{E}_{p(x,c)}[\log q(x|c)]] \\ & - [H(y) + \mathbb{E}_{p(y,s)}[\log q(y|s)]]: \end{aligned} \quad (6)$$

Noting that both $H(x)$ and $H(y)$ are constants from the data, we only need to minimize:

$$\begin{aligned} \mathcal{L}_{\text{Dis}} = I(s; c) &- \mathbb{E}_{p(x,c)}[\log q(x|c)] \\ &- \mathbb{E}_{p(y,s)}[\log q(y|s)]: \end{aligned} \quad (7)$$

As an intuitive explanation of $\mathcal{L}_{\text{Dis}}$, the style embedding s and content embedding c are expected to be independent by minimizing mutual information $I(s; c)$, while they also need to be representative: the style embedding s is encouraged to give

Algorithm 1: Disentangling s and c

Input: Data $\{x_j\}_{j=1}^M$, encoder $q(s, c|x)$, approximation network $p(s|c)$.
for *each training iteration* **do**
  Sample $\{s_j, c_j\}_{j=1}^M$ from $q(s, c|x)$;
  $\mathcal{L}(\sigma) = \frac{1}{M}\sum_{j=1}^M \log p(s_j|c_j)$;
  Update $p(s|c)$ by maximize $\mathcal{L}(\sigma)$;
  for $j = 1$ **to** M **do**
    Sample k' uniformly from $\{1, 2, \ldots, M\}$;
    $\hat{R}_j = \log p(s_j|c_j) - \log p(s_j|c_{k'})$;
  end
  Update $q(s, c|x)$ by minimize $\frac{1}{M}\sum_{j=1}^M \hat{R}_j$;
end

a better prediction of style label y by maximizing $\mathbb{E}_{p(y;s)}[\log q(y|s)]$; the content embedding should maximize the log-likelihood $\mathbb{E}_{p(x;c)}[\log q(x|c)]$ to contain sufficient information from sentence x.

3.3 MI Sample-based Upper Bound

To estimate $I(s; c)$, we propose a novel sample-based upper bound. Assume we have M latent embedding pairs $\{(s_j; c_j)\}_{j=1}^M$ drawn from $p(s; c)$. As shown in Theorem 3.1, we derive an upper bound of mutual information based on the samples. A detailed proof is provided in the Supplementary Material.

Theorem 3.1. *If* $\{(s_j; c_j)\}_{j=1}^M \sim p(s; c)$, *then*

$$I(s; c) \le \mathbb{E}\left[\frac{1}{M}\sum_{j=1}^M R_j\right] =: \hat{I}(s; c), \quad (8)$$

where $R_j = \log p(s_j|c_j) - \frac{1}{M}\sum_{k=1}^M \log p(s_j|c_k)$.

Based on Theorem 3.1, given embedding samples $\{(s_j; c_j)\}_{j=1}^M$, we can minimize $\frac{1}{M}\sum_{j=1}^M R_j$ as an unbiased estimation of the upper bound $\hat{I}(s; c)$. The calculation of R_j requires the conditional distribution $p(s|c)$, whose closed form is unknown. Therefore, we use a variational network $p(s|c)$ to approximate $p(s|c)$ with embedding samples.

To implement the upper bound in Eq. (8), we first feed M sentences $\{x_j\}$ into encoder $q(s; c|x)$ to obtain embedding pairs $\{(s_j; c_j)\}$. Then, we train the variational distribution $p(c|x)$ by maximizing the log-likelihood $L(\theta) = \frac{1}{M}\sum_{j=1}^M \log p(s_j|c_j)$. After the training of $p(s|c)$ is finished, we calculate R_j for each embedding pair $(s_j; c_j)$. Finally, the gradient for $\frac{1}{M}\sum_{j=1}^M R_j$ is calculated and back-propagated to encoder $q(s; c|x)$. We apply the re-parameterization trick (Kingma and Welling, 2013) to ensure the gradient back-propagates through the sampled embeddings $(s_j; c_j)$. When the encoder weights are updated, the distribution $q(s; c|x)$ changes, which leads to the changing of conditional distribution $p(s|c)$. Therefore, we need to update the approximation network $p(s|c)$ again. Consequently, the encoder network $q(s; c|x)$ and the approximation network $p(s|c)$ are updated alternately during training.

In each training step, the above algorithm requires M pairs of embedding samples $\{(s_j; c_j)\}_{j=1}^M$ and the calculation of all conditional distributions $p(s_j|c_k)$. This leads to $O(M^2)$ computational complexity. To accelerate the training, we further approximate term $\frac{1}{M}\sum_{k=1}^M \log p(s_j|c_k)$ in R_j by $\log p(s_j|c_{k'})$, where k' is selected uniformly from indices $\{1; 2; \ldots; M\}$. This stochastic sampling not only leads to an unbiased estimation $\hat{R}_j$ to R_j, but also improves the model robustness (as shown in Algorithm 1).

Symmetrically, we can also derive an MI upper bound based on the conditional distribution $p(c|s)$. However, the dimension of c is much higher than the dimension of s, which indicates that the neural approximation to $p(c|s)$ would have worse performance compared with the approximation to $p(s|c)$. Alternatively, the lower-dimensional distribution $p(s|c)$ used in our model is relatively easy to approximate with neural networks.

3.4 Encoder-Decoder Framework

One important downstream task for disentangled representation learning (DRL) is conditional generation. Our MI-based text DRL method can be also embedded into an Encoder-Decoder generative model and trained end-to-end.

Since the proposed DRL encoder $q(s; c|x)$ is a stochastic neural network, a natural extension is to add a decoder to build a variational autoencoder (VAE) (Kingma and Welling, 2013). Therefore, we introduce another decoder network $p(x|s; c)$ that generates a new sentence based on the given style s and content c. A prior distribution $p(s; c) = p(s)p(c)$, as the product of two multivariate unit-variance Gaussians, is used to regularize the posterior distribution $q(s; c|x)$ by KL-divergence minimization. Meanwhile, the log-likelihood term for text reconstruction should be maximized. The objective for VAE is:

$$\mathcal{L}_{\text{VAE}} = \text{KL}(q(s; c|x) \| p(s; c)) - \mathbb{E}_{q(s;c|x)}[\log p(x|s; c)].$$

We combine the VAE objective and our MI-based disentanglement term to form an end-to-end learning framework (as shown in Figure 2). The total

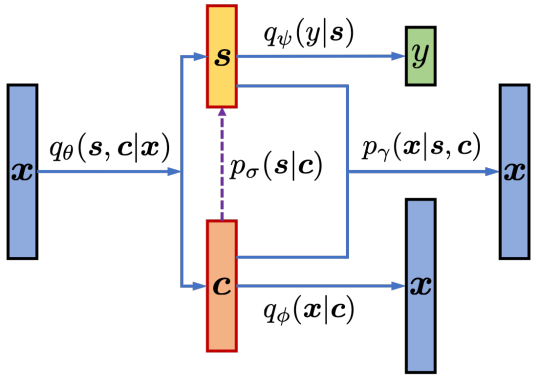

Figure 2: Proposed framework: Each sentence x is encoded into style embedding s and content embedding c. The style embedding s goes through a classifier $q_\psi(y|s)$ to predict the style label y; the content embedding c is used to reconstruct x. An auxiliary network $p_\sigma(s|c)$ helps disentangle the style and content embeddings. The decoder $p_\gamma(x|s; c)$ generates sentences based on the combination of s and c.

loss function is $L_{\text{total}} = \quad L_{\text{Dis}} + L_{\text{VAE}}$, where L_{Dis} replaces $I(s; c)$ in L_{Dis} (Eq. (7)) with our MI upper bound $\hat{I}(s; c)$ from Eq. (8); > 0 is a hyper-parameter re-weighting DRL and VAE terms. We call this final framework Information-theoretical Disentangled text Embedding Learning (IDEL).

4 Related Work

4.1 Disentangled Representation Learning

Disentangled representation learning (DRL) can be classified into two categories: unsupervised disentangling and supervised disentangling. Unsupervised disentangling methods focus on adding constraints on the embedding space to enforce that each dimension of the space be as independent as possible (Burgess et al., 2018; Chen et al., 2018). However, Locatello et al. (2019) challenge the effectiveness of unsupervised disentangling without any induced bias from data or supervision. For supervised disentangling, supervision is always provided on different parts of disentangled representations. However, for text representation learning, supervised information can typically be provided only for the style embeddings (*e.g.* sentiment or personality labels), making the task much more challenging. John et al. (2019) tried to alleviate this issue by manually removing sentiment-related words from a sentence. In contrast, our model is trained in an end-to-end manner without manually adding any supervision on the content embeddings.

4.2 Mutual Information Estimation

Mutual information (MI) is a fundamental measurement of the dependence between two random variables. MI has been applied to a wide range of tasks in machine learning, including generative modeling (Chen et al., 2016), the information bottleneck (Tishby et al., 2000), and domain adaptation (Gholami et al., 2020). In our proposed method, we utilize MI to measure the dependence between content and style embedding. By minimizing the MI, the learned content and style representations are explicitly disentangled.

However, the exact value of MI is hard to calculate, especially for high-dimensional embedding vectors (Poole et al., 2019). To approximate MI, most previous work focuses on lower-bound estimations (Chen et al., 2016; Belghazi et al., 2018; Poole et al., 2019), which are not applicable to MI minimization tasks. Poole et al. (2019) propose a leave-one-out upper bound of MI; however it is not numerically stable in practice. Inspired by these observations, we introduce a novel variational MI upper bound for disentangled representation learning, which stably minimizes the correlation between content and style embedding in a principled manner.

5 Experiments

5.1 Datasets

We conduct experiments to evaluate our models on the following real-world datasets:

Yelp Reviews: The Yelp dataset contains online service reviews with associated rating scores. We follow the pre-processing from Shen et al. (2017) for a fair comparison. The resulting dataset includes 250,000 positive review sentences and 350,000 negative review sentences.

Personality Captioning: Personality Captioning dataset (Shuster et al., 2019) collects captions of images which are written according to 215 different personality traits. These traits can be divided into three categories: *positive*, *neutral*, and *negative*. We select sentences from *positive* and *negative* classes for evaluation.

5.2 Experimental Setup

We build the sentence encoder $q(s; c|x)$ with a one-layer bi-directional LSTM plus a multi-head attention mechanism. The style classifier $q(y|s)$ is parameterized by a single fully-connected network

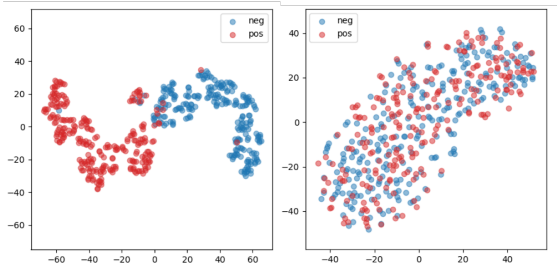

Figure 3: Latent spaces t-SNE plots of IDEL on Yelp.

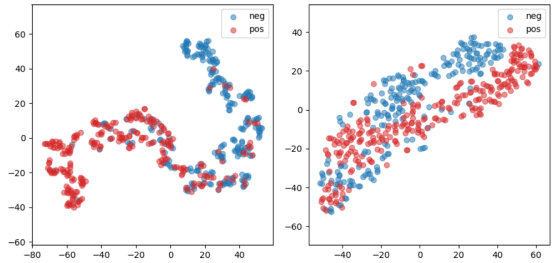

Figure 4: t-SNE plots of IDEL$^-$ without $\hat{I}(s;c)$.

with the softmax activation. The content-based decoder $q(x|c)$ is a one-layer uni-directional LSTM appended with a linear layer with vocabulary size output, outputting the predicted probability of the next words. The conditional distribution approximation $p(s|c)$ is represented by a two-layer fully-connected network with ReLU activation. The generator $p(x|s;c)$ is built by a two-layer uni-directional LSTM plus a linear projection with output dimension equal to the vocabulary size, providing the next-word prediction based on previous sentence information and the current word.

We initialize and fix our word embeddings by the 300-dimensional pre-trained GloVe vectors (Pennington et al., 2014). The style embedding dimension is set to 32 and the content embedding dimension is 512. We use a standard multivariate normal distribution as the prior of the latent spaces. We train the model with the Adam optimizer (Kingma and Ba, 2014) with initial learning rate of $5 \quad 10^{-5}$. The batch size is equal to 128.

5.3 Embedding Disentanglement Quality

We first examine the disentangling quality of learned latent embeddings, primarily studying the latent spaces of IDEL on the Yelp dataset.

Latent Space Visualization: We randomly select 1,000 sentences from the Yelp testing set and visualize their latent embeddings in Figure 3, via t-SNE plots (van der Maaten and Hinton, 2008). The blue and red points respectively represent the positive and negative sentences. The left side of the figure shows the style embedding space, which is well separated into two parts with different colors. It supports the claim that our model learns a semantically meaningful style embedding space. The right side of the figure is the content embedding space, which cannot be distinguished by the style labels (different colors). The lack of difference in the pattern of content embedding also provides evidence that our content embeddings have little correlation with the style labels.

For an ablation study, we train another IDEL model under the same setup, while removing our MI upper bound $\hat{I}(s;c)$. We call this model IDEL in the following experiments. We encode the same sentences used in Figure 3, and display the corresponding embeddings in Figure 4. Compared with results from the original IDEL, the style embedding space (left in Figure 4) is not separated in a clean manner. On the other hand, the positive and negative embeddings become distinguishable in the content embedding space. The difference between Figures 3 and 4 indicates the disentangling effectiveness of our MI upper bound $\hat{I}(s;c)$.

Label-Embedding Correlation: Besides visualization, we also numerically analyze the correlation between latent embeddings and style labels. Inspired by the statistical two-sample test (Gretton et al., 2012), we use the sample-based divergence between the positive embedding distribution $p(c|y = 1)$ and the negative embedding distribution $p(c|y = 0)$ as a measurement of label-embedding correlation. We consider four divergences: Mean Absolute Deviation (MAD) (Geary, 1935), Energy Distance (ED) (Sejdinovic et al., 2013), Maximum Mean Discrepancy (MMD) (Gretton et al., 2012), and Wasserstein distance (WD) (Ramdas et al., 2017). For a fair comparison, we re-implement previous text embedding methods and set their content embedding dimension to 512 and the style embedding dimension to 32 (if applicable). Details about the divergences and embedding processing are shown in the Supplementary Material.

From Table 2, the proposed IDEL achieves the lowest divergences between positive and negative *content* embeddings compared with Ctrl-Gen (Hu et al., 2017), CAAE (Shen et al., 2017), ARAE (Zhao et al., 2018), BackTranslation (BT) (Lample et al., 2019), and DRLST (John et al., 2019), indicating our model better disentangles the

	Yelp Dataset							Personality Captioning Dataset						
	Conditional Generation			Style Transfer				Conditional Generation			Style Transfer			
	ACC	**BLEU**	**GM**	**ACC**	**BLEU**	**S-BLEU**	**GM**	**ACC**	**BLEU**	**GM**	**ACC**	**BLEU**	**S-BLEU**	**GM**
CtrlGen	82.5	20.8	41.4	83.4	19.4	31.4	37.0	73.6	18.9	37.0	73.3	18.9	30.0	34.6
CAAE	78.9	19.7	39.4	79.3	18.5	28.2	34.6	72.2	19.5	37.5	72.1	18.3	27.4	33.1
ARAE	78.3	**23.1**	42.4	78.5	21.3	32.5	37.9	72.8	**22.5**	40.4	71.5	20.4	31.6	35.8
BT	81.4	20.2	40.5	**86.3**	24.1	**35.6**	**41.9**	74.1	21.0	39.4	**75.9**	23.1	34.2	39.1
DRLST	83.7	22.8	43.7	85.0	23.9	34.9	41.4	74.9	22.0	40.5	75.7	21.9	33.8	38.3
IDEL	78.1	20.3	39.8	79.1	20.1	27.5	35.1	72.0	19.7	37.7	72.4	19.7	27.1	33.8
IDEL	**83.9**	23.0	**43.9**	85.7	**24.3**	35.2	**41.9**	**75.1**	22.3	**40.9**	75.6	**23.3**	**34.6**	**39.4**

Table 1: Performance comparison of text DRL models. For conditional generation, the GM scores are calculated over ACC and BLEU. For style transfer, the GMs are calculated over ACC, BLEU, S-BLEU(self-BLEU).

Method	**MAD**	**ED**	**WD**	**MMD**
CtrlGen	0.261	0.105	0.311	0.063
CAAE	0.285	0.112	0.306	0.078
ARAE	0.194	0.050	0.248	0.042
BT	0.211	0.053	0.269	0.049
DRLST	0.181	0.048	0.215	0.031
IDEL	0.217	0.077	0.293	0.051
IDEL	**0.063**	**0.015**	**0.084**	**0.010**

Table 2: Sample divergences between positive and negative *content* embeddings.

Method	**MAD**	**ED**	**WD**	**MMD**
DRLST	1.024	0.503	1.375	0.286
IDEL	0.996	0.489	1.124	0.251
IDEL	**1.167**	**0.583**	**1.392**	**0.302**

Table 3: Sample divergences between positive and negative *style* embeddings.

content embeddings from the style labels. For *style* embeddings, we compare IDEL with DRLST, the only prior method that infers the text style embeddings. Table 3 shows a larger distribution gap between positive and negative style embeddings with IDEL than with DRLST, which demonstrates the proposed IDEL has better style information expression in the style embedding space. The comparison between IDEL and IDEL supports the effectiveness of our MI upper bound minimization.

5.4 Embedding Representation Quality

To show the representation ability of IDEL, we conduct experiments on two text-generation tasks: style transfer and conditional generation.

For style transfer, we encode two sentences into a disentangled representation, and then combine the style embedding from one sentence and the content embedding from another to generate a new sentence via the generator $p(x|s,c)$. For conditional generation, we set one of the style or content embeddings to be fixed and sample the other part from the latent prior distribution, and then use the combination to generate text. Since most previous work only embedded the content information, for fair comparison, we mainly focus on fixing style and sampling context embeddings under the conditional generation setup.

To measure generation quality for both tasks, we test the following metrics (more specific description is provided in the Supplementary Material).

Style Preservation: Following previous work (Hu et al., 2017; Shen et al., 2017; John et al., 2019), we pre-train a style classifier and use it to test whether a generated sentence can be categorized into the correct target style class.

Content Preservation: For style transfer, we measure whether a generation preserves the content information from the original sentence by the self-BLEU score (Lample et al., 2019). The self-BLEU is calculated between one original sentence and its style-transferred sentence.

Generation Quality: To measure the generation quality, we calculate the corpus-level BLEU score (Papineni et al., 2002) between a generated sentence and the testing data corpus.

Geometric Mean: We use the geometric mean (GM) (John et al., 2019) of the above metrics to obtain an overall evaluation metric of representiveness of DRL models.

We compare our IDEL with previous state-of-the-art methods on Yelp and Personality Captioning datasets, as shown in Table 1. The references to the other models are mentioned in Section 5.3. Note that the original BackTranslation (BT) method (Lample et al., 2019) is a Auto-Encoder framework, that is not able to do conditional generation. To compare with BT fairly, we add a standard Gaussian prior in its latent space to make it a variational auto-encoder model.

From the results in Table 1, ARAE performs well on the conditional generation. Compared to

Content Source	Style Source	Transferred Result
I **enjoy** it thoroughly!	never before had a **bad** experience at the habit until tonight.	I **dislike** it thoroughly.
quality is **just so so**.		quality is so **bad**.
I am so **grateful**.		I am so **disgusted**.
never before had a **bad** experience at the habit until tonight.	I am so **grateful**.	never had a service that was **enjoyable** experience tonight.
	quality is **just so so**.	never had a **unimpressed** experience until tonight.
	quality of food is **fantastic**.	never had **awesome** routine until tonight.
I am so **disappointed** with palm today.	we were both so **impressed**.	I am so **impressed** with palm again.
	quality of food is **fantastic** .	I am **good** with palm today.
	never before had a **bad** experience at the habit until tonight.	I am so **disgusted** with palm today.

Table 4: Examples of text style transfer on Yelp dataset. The style-related words are bold.

	SA	CP	SF	GM
CtrlGen	71.2 (3.56)	3.25	3.12	3.30
CAAE	63.1 (3.16)	2.83	3.06	3.01
ARAE	68.0 (3.40)	**3.44**	3.09	3.31
IDEL	**73.7 (3.69)**	3.39	**3.21**	**3.42**

Table 5: Manual evaluation for style transfer on Yelp. The style accuracy (SA) scores are scaled in range $[0; 5]$ for compatible calculation of geometric mean (GM).

	ACC	BLEU	S-BLEU	GM
L_{VAE}	52.1	**24.7**	20.8	29.9
$L_{\text{VAE}} + \text{I}(s; y)$	**86.1**	23.3	16.4	32.0
$L_{\text{VAE}} + \text{I}(x; c)$	50.2	24.0	**36.3**	34.7
IDEL	79.1	20.1	27.5	35.1
IDEL	85.5	24.0	35.0	41.5
IDEL	85.7	24.3	35.2	**41.9**

Table 6: Ablation tests for style transfer on Yelp.

ARAE, our model performance is slightly lower on content preservation (BLEU). In contrast, the style classification score of IDEL has a large margin above that of ARAE. The BackTranslation (BT) has a better performance on style transfer tasks, especially on the Yelp dataset. Our IDEL has a lower style classification accuracy (ACC) than BT on the style transfer task. However, IDEL achieves high BLEU on style transfer, which leads to a high overall GM score on the Personality-Captioning dataset. On the Yelp dataset, IDEL also has a competitive GM score compared with BT. The experiments show a clear trade-off between style preservation and content preservation, in which our IDEL learns more representative disentangled representation and leads to a better balance.

Besides the automatic evaluation metrics mentioned above, we further test our disentangled representation effectiveness by human evaluation. Due to the limitation of manual effort, we only evaluate the style transfer performance on Yelp datasets. The generated sentences are manually evaluated on style accuracy (SA), content preservation (CP), and sentence fluency (SF). The CP and SF scores are between 0 to 5. Details are provided in the Supplementary Material. Our method achieves better style and content preservation, with a little performance sacrifice on sentence fluency.

Table 4 shows three style transfer examples from IDEL on the Yelp dataset. The first example shows three sentences transferred with the style from a given sentence. The other two examples transfer each given sentence based on the styles of three different sentences. Our IDEL not only transfers sentences into target sentiment classes, but also renders the sentence with more detailed style information (*e.g.*, the degree of the sentiment).

In addition, we conduct an ablation study to test the influence of different objective terms in our model. We re-train the model with different training loss combinations while keeping all other setups the same. In Table 1, IDEL surpasses IDEL (without MI upper bound minimization) with a large gap, demonstrating the effectiveness of our proposed MI upper bound. The vanilla VAE has the best generation quality. However, its transfer style accuracy is slightly better than a random guess. When adding $\text{I}(s; y)$, the ACC score significantly improves, but the content preservation (S-BLEU) becomes worse. When adding $\text{I}(c; x)$, the content information is well preserved, while the ACC even decreases. By gradually adding MI terms, the model performance becomes more balanced on all the metrics, with the overall GM monotonically increasing. Additionally, we test the influence of the stochastic calculation of R_j in Algorithm 1 (IDEL) with the closed form from Theorem 3.1 (IDEL). The stochastic IDEL not only accelerates the training but also gains a performance improvement relative to IDEL .

6 Conclusions

We have proposed a novel information-theoretic disentangled text representation learning framework. Following the theoretical guidance from information theory, our method separates the textual

information into independent spaces, constituting style and content representations. Concurrently, the original text information is well preserved by maximizing the mutual information between the input sentence and the latent representation. A general sample-based mutual information upper bound is derived to assist the training process. This upper bound could be further applied to various deep learning tasks, such as domain adaptation and zero-shot learning. In our experiments, we introduce several two-sample test statistics to measure label-embedding correlation. The proposed model achieves competitive performance compared with previous methods on both conditional generation and style transfer. For future work, our model can be extended to disentangled representation learning with non-categorical style labels, and applied to zero-shot style transfer with newly-coming unseen styles.

References

David Barber and Felix V Agakov. 2003. The im algorithm: a variational approach to information maximization. In *Advances in neural information processing systems*, page None.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Devon Hjelm, and Aaron Courville. 2018. Mutual information neural estimation. In *International Conference on Machine Learning*, pages 530–539.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. 2018. Understanding disentangling in beta-vae. *arXiv preprint arXiv:1804.03599*.

Tian Qi Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. 2018. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, pages 2610–2620.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180.

Ju chieh Chou, Cheng chieh Yeh, Hung yi Lee, and Lin shan Lee. 2018. Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations. In *Proc. Interspeech 2018*, pages 501–505.

Thomas M Cover and Joy A Thomas. 2012. *Elements of information theory*. John Wiley & Sons.

Emily L Denton et al. 2017. Unsupervised learning of disentangled representations from video. In *Advances in neural information processing systems*, pages 4414–4423.

Roy C Geary. 1935. The ratio of the mean deviation to the standard deviation as a test of normality. *Biometrika*, 27(3/4):310–332.

Behnam Gholami, Pritish Sahu, Ognjen Rudovic, Konstantinos Bousmalis, and Vladimir Pavlovic. 2020. Unsupervised multi-target domain adaptation: An information theoretic approach. *IEEE Transactions on Image Processing*, 29:3993–4002.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773.

Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li F Fei-Fei, and Juan Carlos Niebles. 2018. Learning to decompose and disentangle representations for video prediction. In *Advances in Neural Information Processing Systems*, pages 517–526.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1587–1596. JMLR. org.

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

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv:1412.6980v9*.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*.

Alexander Kraskov, Harald Stögbauer, Ralph G Andrzejak, and Peter Grassberger. 2005. Hierarchical clustering using mutual information. *EPL (Europhysics Letters)*, 70(2):278.

Vinay Kumar Verma, Gundeep Arora, Ashish Mishra, and Piyush Rai. 2018. Generalized zero-shot learning via synthesized examples. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4281–4289.

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

Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. 2018. Diverse

image-to-image translation via disentangled representations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–51.

Yen-Cheng Liu, Yu-Ying Yeh, Tzu-Chien Fu, Sheng-De Wang, Wei-Chen Chiu, and Yu-Chiang Frank Wang. 2018. Detach and adapt: Learning cross-domain disentangled deep representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8867–8876.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, pages 4114–4124.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing high-dimensional data using t-SNE. *JMLR*.

Marina Meilă. 2007. Comparing clusterings—an information based distance. *Journal of multivariate analysis*, 98(5):873–895.

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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. 2019. On variational bounds of mutual information. In *International Conference on Machine Learning*, pages 5171–5180.

Aaditya Ramdas, Nicolás Trillos, and Marco Cuturi. 2017. On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, Kenji Fukumizu, et al. 2013. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in neural information processing systems*, pages 6830–6841.

Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, and Jason Weston. 2019. Engaging image captioning via personality. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12516–12526.

Naftali Tishby, Fernando C Pereira, and William Bialek. 2000. The information bottleneck method. *arXiv preprint physics/0004057*.

Luan Tran, Xi Yin, and Xiaoming Liu. 2017. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1415–1424.

Li Yingzhen and Stephan Mandt. 2018. Disentangled sequential autoencoder. In *International Conference on Machine Learning*, pages 5656–5665.

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

Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. 2019. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9299–9306.

A Proofs of Theorems

Proof of Theorem 3.1. First, we show that

$$\mathrm{E}_{p(\boldsymbol{s},\boldsymbol{c})}[\log p(s|c)] \quad \mathrm{E}_{p(\boldsymbol{s})p(\boldsymbol{c})}[\log p(s|c)] \quad \mathrm{I}(s;c). \tag{9}$$

Calculate the gap between the left-hand side and right-hand side of Eq. (9):

$$\begin{aligned}
&= \mathrm{E}_{p(\boldsymbol{s},\boldsymbol{c})}[\log p(s|c)] \quad \mathrm{E}_{p(\boldsymbol{s})p(\boldsymbol{c})}[\log p(s|c)] \quad \mathrm{I}(s;c) \\
&= \mathrm{E}_{p(\boldsymbol{s},\boldsymbol{c})}[\log p(s|c)] \quad \mathrm{E}_{p(\boldsymbol{s})p(\boldsymbol{c})}[\log p(s|c)] \quad \mathrm{E}_{p(\boldsymbol{s},\boldsymbol{c})}[\log p(s|c) \quad \log p(s)] \\
&= \mathrm{E}_{p(\boldsymbol{s},\boldsymbol{c})}[\log p(s)] \quad \mathrm{E}_{p(\boldsymbol{s})}\mathrm{E}_{p(\boldsymbol{c})}[\log p(s|c)] \\
&= \mathrm{E}_{p(\boldsymbol{s})} \quad \log p(s) \quad \mathrm{E}_{p(\boldsymbol{c})}[\log(p(s|c)] \\
&= \mathrm{E}_{p(\boldsymbol{s})} \quad \log \quad \mathrm{E}_{p(\boldsymbol{c})}[p(s|c)] \quad \mathrm{E}_{p(\boldsymbol{c})}[\log p(s|c)] \quad 0. \qquad \text{(Jensen's Inequality)}
\end{aligned}$$

Therefore, the inequality in Eq. (9) holds.

Given sample pairs $\{(s_j, c_j)\}_{j=1}^{M} \quad p(s,c)$, the left-hand side of Eq. (9) has an unbiased estimation:

$$\begin{aligned}
&\frac{1}{M}\sum_{j=1}^{M} \mathrm{E}_{(\boldsymbol{s}_j,\boldsymbol{c}_j) \quad p(\boldsymbol{s},\boldsymbol{c})}[\log p(s_j|c_j)] \quad \frac{1}{M^2}\sum_{j=1}^{M}\sum_{k=1}^{M} \mathrm{E}_{\boldsymbol{s}_j \quad p(\boldsymbol{s})}\mathrm{E}_{\boldsymbol{c}_k \quad p(\boldsymbol{c})}[\log p(s_j|c_k)] \\
&= \mathrm{E}\left[\frac{1}{M}\sum_{j=1}^{M} \log p(s_j|c_j) \quad \frac{1}{M}\sum_{k=1}^{M} p(s_j|c_k)\right] = \mathrm{E}\left[\frac{1}{M}\sum_{j=1}^{M} R_j\right],
\end{aligned}$$

which is what we claim in Theorem 3.1. □

Proof of Lower Bounds in Eq. (6).

$$\begin{aligned}
\mathrm{I}(c;x) &= \mathrm{E}_{p(\boldsymbol{x},\boldsymbol{c})}[\log p(x|c) \quad \log p(x)] = \mathrm{H}(x) + \mathrm{E}_{p(\boldsymbol{x},\boldsymbol{c})}[\log p(x|c)] \\
&= \mathrm{H}(x) + \mathrm{E}_{p(\boldsymbol{x},\boldsymbol{c})}[\log p(x|c) \quad \log q \ (x|c) + \log q \ (x|c)] \\
&= \mathrm{H}(x) + \mathrm{E}_{p(\boldsymbol{x},\boldsymbol{c})}[\log p(x|c) \quad \log q \ (x|c)] + \mathrm{E}_{p(\boldsymbol{x},\boldsymbol{c})}[\log q \ (x|c)] \\
&= \mathrm{H}(x) + \mathrm{KL}(p(x|c)\|q \ (x|c)) + \mathrm{E}_{p(\boldsymbol{x},\boldsymbol{c})}[\log q \ (x|c)] \\
&\quad H(x) + \mathrm{E}_{p(\boldsymbol{x},\boldsymbol{c})}[\log q(x|c)].
\end{aligned}$$

The inequality is based on the fact that the KL-divergence is always non-negative. The lower bound for $\mathrm{I}(s;y)$ can be also derived in the similar way. □

B Sample-based Embedding Divergences

In this section we introduce the implementation details of the calculation about label-embedding correlation. As mentioned in Section 5.4 , the distribution divergence between $p(c|y=0)$ and $p(c|y=1)$ measures the correlation between content embeddings and style labels. Assume $c_1^{(0)}, c_2^{(0)}, \ldots, c_{N_0}^{(0)} \quad p(c|y=0)$, and $c_1^{(1)}, c_2^{(1)}, \ldots, c_{N_1}^{(1)} \quad p(c|y=1)$, then the four metrics MAD, ED, WD, MMD are calculated based on the two groups of samples. With a ground distance $d(\ ,\ )$, the implementaion of the above four metrics are demonstrated in following:

$$D_{\mathrm{MAD}} = d\left(\frac{1}{N_0}\sum_{i=1}^{N_0} c_i^{(0)}, \frac{1}{N_1}\sum_{j=1}^{N_1} c_j^{(1)}\right). \tag{10}$$

$$D_{\mathrm{ED}} = \frac{2}{N_0 N_1}\sum_{i=1}^{N_0}\sum_{j=1}^{N_1} d(c_i^{(0)}, c_j^{(1)}) \quad \frac{1}{N_0^2}\sum_{i=1}^{N_0}\sum_{j=1}^{N_0} d(c_i^{(0)}, c_j^{(0)}) \quad \frac{1}{N_1^2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_1} d(c_i^{(1)}, c_j^{(1)}) \tag{11}$$

$$D_{\mathrm{WD}} = \min_{p_{ij}} \sum_{i=1}^{N_0} \sum_{j=1}^{N_1} p_{ij}\, d(c_i^{(0)}, c_j^{(1)}) \quad s.t. \sum_{i=1}^{N_0} p_{ij} = \frac{1}{N_1}, \quad \sum_{j=1}^{N_1} p_{ij} = \frac{1}{N_0}. \tag{12}$$

$$D_{\mathrm{MMD}} = \frac{1}{N_0^2} \sum_{i=1}^{N_0} \sum_{j=1}^{N_0} K(c_i^{(0)}, c_j^{(0)}) + \frac{1}{N_1^2} \sum_{i=1}^{N_1} \sum_{j=1}^{N_1} K(c_i^{(1)}, c_j^{(1)}) - \frac{2}{N_0 N_1} \sum_{i=1}^{N_0} \sum_{j=1}^{N_1} K(c_i^{(0)}, c_j^{(1)}), \tag{13}$$

where $K(\cdot, \cdot)$ is a kernel function. Here we choose $K(\cdot, \cdot)$ from RBF kernel family with bandwidth $w = 1$.

For style embedding, the calculation formats are the same as in above equations. The style embeddings and content embeddings have different dimensions, which leads to the ground metric $d(\cdot, \cdot)$ inconsistent. Therefore, instead of using Euclidean distance, we use the cosine distance as the ground metric.

C Details in Representation Quality Evaluation

For style preservation, we pretrain a style classifier on each dataset. The style classifier is built by a one-layer LSTM appended with a multi-head attention layer. The number of the attention head is set to 6. The classifiers reach 95% prediction accuracy on Yelp and 93% prediction accuracy on Personality-Captioning. We input transferred sentences into the classifier and test whether the predicted style label is the same as the target style label.

For human evaluation, we transferred 1000 sentences with randomly selected style labels. After the transferring, we ask 10 human annotators to justify the style label, content preservation and content fluency. The style label is 0 or 1 representing the positive or negative sentiment of the given sentence. The content preservation and the content fluency is scored between 0 to 5. To make the style accuracy compatible with the other two scores, we scale it into range [0,5]. If the scores from the two annotators have a difference larger than 2, the scores will not be recorded. In this way, we ensure the evaluation criteria of annotators are similar.