Unsupervised Controllable Text Generation with Global Variation Discovery and Disentanglement

Peng Xu¹, Yanshuai Cao¹, Jackie Chi Kit Cheung¹,²,³
¹Borealis AI
{peng.z.xu,yanshuai.cao}@borealisai.com
²McGill University
jcheung@cs.mcgill.ca
³Canada CIFAR Chair, Mila

Abstract

Existing controllable text generation systems rely on annotated attributes, which greatly limits their capabilities and applications. In this work, we make the first successful attempt to use VAEs to achieve controllable text generation without supervision. We do so by decomposing the latent space of the VAE into two parts: one incorporates structural constraints to capture dominant global variations implicitly present in the data, e.g., sentiment or topic; the other is unstructured and is used for the reconstruction of the source sentences. With the enforced structural constraint, the underlying global variations will be discovered and disentangled during the training of the VAE. The structural constraint also provides a natural recipe for mitigating posterior collapse for the structured part, which cannot be fully resolved by the existing techniques. On the task of text style transfer, our unsupervised approach achieves significantly better performance than previous supervised approaches. By showcasing generation with finer-grained control including Cards-Against-Humanity-style topic transitions within a sentence, we demonstrate that our model can perform controlled text generation in a more flexible way than existing methods.

1 Introduction

Controllable text generation aims at generating realistic text with control over various attributes including sentiment, topic and other high-level properties. Such controllable text generation models help in a wide range of application, e.g., dialogues systems [17]. Backed by the recent success of deep generative models, such as Variational Autoencoders (VAEs) [10], Generative Adversarial Nets (GANs) [5], and autoregressive models [12], existing models have made progress towards controllable text generation [15][8][9][18]. However, all these models rely on learning from annotated attributes to generate the text in a controllable fashion. The high cost of labeling large training corpora with attributes of interest limits the usage of these models, as pre-existing annotation often do not align with some downstream goal. Even if cheap labels are available, for example, review scores as proxy for sentiment, the control over text generation for these models are limited to the variation defined by the attributes.

Recently, large-scale pre-training with generative models on language [13][4][14] have achieved stunning success on various tasks including generating coherent text. By conditioning on certain observed text, these pre-trained models can also exert some degree of control on text generation. However, these models are essentially black boxes. The control over the generated text using seed sentences is weak, unpredictable and often uninterpretable.

¹A popular board game, https://en.wikipedia.org/wiki/Cards_Against_Humanity

Preprint. Under review.
While unsupervised disentanglement, controllable generation and manipulation has been successfully done for images [3,7], it is unclear whether similar capabilities are feasible on text considering its discrete nature and inherent intricacy as human construction. In this paper, we provide positive evidence by making the first successful attempt to use VAEs as the backbone for controllable and interpretable generation of text without supervision. Instead of depending on annotated attributes, we want the underlying high-level global variations to be discovered and disentangled automatically during the training. The resulting latent variables allow flexible and interpretable controlled generation. To the best of our knowledge, the proposed method is the first to simultaneously achieve unsupervised attribute discovery, and controllable generation in the form of attribute transfer and attribute transition.

Achieving this goal is challenging for several reasons. First, information about topics, semantics and syntax are highly intertwined in language. Second, jointly discovering such variation while generating texts that exhibit them requires learning about “re-entangle” the factors in a natural way after disentangling them, which poses another level of difficulty.

In this paper, we propose an approach based on VAEs to do unsupervised controllable text generation while addressing the above challenges effectively. We decompose the latent space of the VAE into two parts: a structured latent space to capture the dominant global variations in the dataset, whether it be topic, sentiment, or other unknown factors; and a second unstructured latent space to capture any other useful information to help the decoder reconstruct the source sentences. In order to discover and disentangle the underlying global variation, we warp the Gaussian posterior distribution into a learned probability simplex, whose vertices are the unsupervisedly learned basis vectors capturing global variations. A reconstruction loss of the latent sample and a regularization term to encourage orthogonality of the learned basis vectors provide additional structured constraint for the posterior beyond the usual isotropic Gaussian prior regularization from the ELBO. Facilitated by the large-scale pre-trained models as a feature extractor, this constraint makes the basis reflect the dominant variation in the data. Enforcing this structured constraint also has the side effect of preventing posterior collapse for the structured part when optimizing VAEs, which cannot be fully resolved with the existing techniques.

Experimental results show that our unsupervised approach outperforms previous supervised approaches significantly on the task of sentiment manipulation, and achieves comparable results on the task of topic modelling as compared to strong unsupervised baselines. On three domains of text, we show that our model achieves the following capabilities: a) discovering finer-grained text attributes that go beyond manually annotated gold labels; b) generating realistic sentences or performing style transfer with a specified global variation discovered; c) controlling generation with natural transitions among different discovered global variations. Table 1 gives a preview of one of the capabilities, where our model generates natural sentences with controlled topic transitions happening in the middle of generation, in a similar fashion to the game Cards Against Humanity.

Table 1: Cards-Against-Humanity-style topic transition during generation: three pairs of samples generated without and with topic transition. The first sentence in the pair is generated with a topic fixed throughout the generation; while the second sentence is generated with topic transition, the generated outputs after switching are marked as bold.

| World throughout | A federal judge on Friday ordered a federal appeals court to overturn a federal appeals court ruling that the Visa and MasterCard credit card associations violated federal antitrust law by barring the names of the state. |
| World to Sci/Tech | A federal judge on Friday ordered a federal appeals court to overturn a decision by the Supreme Court to overturn a decision by the Federal Communications Commission to block the company’s antitrust case against Microsoft Corp. |
| Sports throughout | NEW YORK (Reuters) - Roger Federer, the world’s No. 1 player, will miss the rest of the season because of a sore quadriceps. |
| Sports to Business | NEW YORK (Reuters) - Roger Federer, the world’s No. 1 player, will miss the rest of the year because of a bid-rigging scandal. |
| Sci/Tech throughout | NewsFactor - IBM (NYSE: IBM) has unveiled the latest version of its Windows XP operating system, which is designed to make it easier for administrators to create and deploy honeypots. |
| Sci/Tech to Sports | NewsFactor - IBM (NYSE: IBM) has unveiled the latest version of its Windows XP operating system, which is designed to make it easier for the Olympics. |
2 Related Work

In order to perform controllable text generation, previous methods assume either annotated attributes or multiple text datasets with different known styles [15, 8, 9, 18]. The requirement of labelled data largely restricts the capabilities and the applications of these models. Instead, all our framework needs is raw text without any annotated attribute. The underlying dominant global variations in the given corpus will be discovered and disentangled automatically, which can in turn be used for controlled generation.

Topic modelling is a long-studied line of research which aims to discover underlying topics that occur in a collection of documents unsupervisedly. Recently, efforts have been committed to leverage deep neural networks for topic modelling [6, 16]. [6] use a neural attention model to model topics by exploiting the distribution of word co-occurrences through the use of pre-trained word embeddings. Since this model aims at extracting aspects in documents, it cannot perform text generation. [16] propose a model which can capture both the global semantic meaning and the local word-ordering structure in a document by combining a neural topic model with a Mixture-of-Experts language model. Though this model can perform text generation within a specified topic, it will drop all the information from the source document except the topic. As a result, it cannot perform style transfer and other finer-grained controls over the generation.

3 Method

We now present the details of our proposed framework, outlined in Figure 1.

![Figure 1: Overview of our proposed framework and how the examples shown in Table I are generated.](image)

3.1 Latent Decomposition

The backbone of our framework is a variational autoencoder (VAE) in which the text is generated conditioned on some latent code, $z$. Given observed text $x$, the VAE is trained to optimize a tractable variational lower bound of $\log p_\theta(x)$:

$$
\mathcal{L}_{VAE}(x; \theta, \phi) = \mathbb{E}_{z \sim q_\phi(x|z)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p(z)),
$$

(1)
where \( q_\phi(z|x) \) is a variational distribution parameterized by an inference network with parameters \( \phi \), and \( p_\theta(z|x) \) denotes the generator network with parameters \( \theta \). This lower bound tries to minimize the reconstruction error of the observed text, and at the same time regularize \( q_\phi(z|x) \) towards the prior \( p(z) \). In this paper, \( p(z) \) is chosen as \( N(0, I) \).

Due to the inherent difficulties as discussed in Sec. 1, limited success has been achieved in distilling high-level semantics from text by an unstructured latent code \( z \) alone. In order to make this process easier, we attempt to separately model global variations and local information by splitting \( z \) into two parts, \( z^{(1)} \) and \( z^{(2)} \). The first part captures the global variations that are dominant in the data without an inductive bias from external signals, while the second part learns to encode the remaining local information that is useful for reconstructing the source sentences. As a result, \( q_\phi(z|x) \) is decomposed into \( q_\phi(z^{(1)}|x)q_\phi(z^{(2)}|x) \) where \( \phi = \phi_1 \cup \phi_2 \). With diagonal covariances, the KL divergence term in Eq. 4 splits into two separate KL terms.

### 3.2 Global Variation Discovery and Disentanglement

We now describe how to discover and disentangle the underlying global variations by enforcing a structural constraint on the latent space of \( z^{(1)} \in \mathbb{R}^N \). We would like to constraint \( z^{(1)} \) to have a structure as follows:

\[
z^{(1)} = \sum_{i=1}^{K} p_i e_i, \quad \sum_{i=1}^{K} p_i = 1, \quad \langle e_i, e_j \rangle = 0, i \neq j, \quad K \leq N
\]

where \( e_i \) are vectors representing the global variations, \( p_i \) is the proportion of \( i \)-th global variation encoded in \( z^{(1)} \) and \( K \) is a hyperparameter indicating the number of global variations to discover. In other words, the latent space of \( z^{(1)} \) is constrained to be a \( K \)-dimension probability simplex in \( \mathbb{R}^N \) whose vertices are represented by the orthogonal basis vector \( e_i, i = 1, \ldots, K \).

With the structural constraint, all the observed text during training are encoded into latent codes which are convex combinations of the learned basis vectors. Due to the fact that the learned basis vectors and the latent codes are located in the same simplex, it is easier for the model to generate natural outputs given the learned basis vectors as the latent code. As verified in Sec. 4, it enables the control in a more flexible way.

Initially, we draw a raw latent code \( \tilde{z}^{(1)} \sim q_{\phi_1}(z^{(1)}|x) \) from the Gaussian variational posterior. We enforce the structural constraint by reconstructing \( \tilde{z}^{(1)} \) by \( z^{(1)} = Ep \), where \( E = [e_1, \ldots, e_K] \) is a learnable embedding matrix representing the bases, and \( p = [p_1, \ldots, p_K]^T \) can be obtained by:

\[
p = \text{softmax}(W\tilde{z}^{(1)} + b),
\]

where \( W \), the weight matrix, and \( b \), the bias vector, are learnable parameters. Similar to an autoencoder, we want to minimize the reconstruction error of the structured latent code. For each input sentence, we randomly sample \( m \) sentences from the training data as negative samples. With the same encoding and reconstructing process, we get the latent code \( u_i \) for each negative sample. Our goal is to make the raw latent code \( \tilde{z}^{(1)} \) similar to the restructured latent code \( z^{(1)} \) while different from latent code \( u_i \) of negative samples. Following [6], we formulate the structured reconstruction loss as a margin loss:

\[
L_{S-REC}(x; \phi_1, \lambda) = \mathbb{E}_{z^{(1)} \sim q_{\phi_1}(z^{(1)}|x)} \left[ \frac{1}{m} \sum_{i=1}^{m} \max(0, 1 - z^{(1)} \cdot \tilde{z}^{(1)} + z^{(1)} \cdot u_i) \right],
\]

where \( \lambda = \{ E, W, b \} \).

If the labelled attributes are available, one could also assign each basis vector a class and train with a discriminator similar to [8]. However, this supervised alternative would introduce a strong inductive bias about what attributes to learn. For example, we often think of certain aspects of variation in opposition to one another, like positive v.s. negative in sentiment, but the way these aspects manifest is still not clear. Inevitably, the model can also be effected by the potential label noise in the data. By contrast, our unsupervised approach can avoid the above problems completely and show potential to outperform the supervised methods, as seen in Sec. 4.

With the structured reconstruction loss as defined with Eq. 4 alone, we cannot achieve the orthogonality of \( e_i \) defined in Eq. 2. Our basis vectors \( E \) may have high correlations, leading to incomplete
disentanglement of the underlying global variations. To encourage the orthogonality of the basis vectors, a regularization term is added to the objective function:

$$L_{\text{REG}}(\mathbf{x}; \lambda) = \| \mathbf{E}^\top \mathbf{E} - I \|,$$

where $I$ is the identity matrix. Our final objective function is defined as follows:

$$L(\mathbf{x}; \theta, \phi, \lambda) = L_{\text{VAE}} + L_{\text{S-REC}} + L_{\text{REG}}.$$  

As a generative model, (uncontrolled) sampling from our model works as follows: in the structured branch, we first draw $\tilde{\mathbf{z}}^{(1)}$ from the isotropic Gaussian prior, then project the sample into the convex hall as described above, which is then concatenated with unstructured prior sample $\mathbf{z}^{(2)}$ before passed into the decoder as in the typical sequence VAE. For controlled generation, one can choose a vertex or any desired point in the probability simplex.

### 3.3 Preventing Posterior Collapse for $\mathbf{z}^{(1)}$

Posterior collapse [2] is a phenomenon where the model ignores the latent code $\mathbf{z}$ during the training of VAEs. It is exacerbated when the generator $p_{\theta}(\mathbf{x}|\mathbf{z})$ is parametrized with a strong autoregressive neural network, which is often the case for text generation. Further complicating matters is the fact that there is an abundance of signals for predicting the next token in the text, but the signals indicating high-level semantics are quite sparse. It is thus unlikely that the VAEs can capture useful global information from raw text without collapse, and at the same time filter out irrelevant noisy signals.

Recent successes of large-scale pre-training on various tasks on language showcase the capabilities of these pre-trained models to capture high-level semantic information. We thus use BERT [4] as a sentence-level feature extractor $f(\cdot)$ to parametrize $\mathbf{z}^{(1)}$, which is a Gaussian distribution:

$$\mathbf{\mu} = W_\mu f(\mathbf{x}) + b_\mu, \quad \log \sigma^2 = W_\sigma f(\mathbf{x}) + b_\sigma, \quad \mathbf{z}^{(1)} \sim \mathcal{N}(\mathbf{\mu}, \text{diag}(\sigma^2)),$$

where $W_\mu, b_\mu, W_\sigma, b_\sigma$ are all learnable parameters.

Prior works tried to mitigate posterior collapse in various ways, which can be applied to resolve the issue of posterior collapse for the unstructured part, $\mathbf{z}^{(2)}$. However, due to differences between the encoders, existing methods [2, 7] are not enough to fully resolve the posterior collapse for the structured part, $\mathbf{z}^{(1)}$, as we can see in Sec. 4. Here, we show that the structure constraint, mentioned in Sec. 3.2 introduces a natural recipe to prevent posterior collapse for $\mathbf{z}^{(1)}$. Note that the KL divergence between $q_\phi(\mathbf{z}^{(1)}|\mathbf{x})$ and $p(\mathbf{z}^{(1)})$ is

$$D_{\text{KL}}(q_\phi(\mathbf{z}^{(1)}|\mathbf{x})||p(\mathbf{z}^{(1)})) = \frac{1}{2} \mu^2 + \frac{1}{2} \left( \sigma^2 - \log \sigma^2 - 1 \right).$$

If we only apply the structural constraint on $\mu$, with orthogonality, the first term in the above equation can be factorized into

$$\mu^2 = (\sum_i p_i \mathbf{e}_i)^2 = \sum_i p_i^2 \mathbf{e}_i^2.$$

With Eq. [5] we have $\mathbf{e}_i^2 = 1$, so that $\mu^2 = \sum_i p_i^2$ reaching its minimum $\frac{1}{K}$ when $\mathbf{p}$ is a uniform distribution. Due to this term, we can see that the KL term will never fully collapse with the structural constraint. In practice, due to the structured reconstruction loss $L_{\text{S-REC}}$, $\mathbf{p}$ will be pushed away from a uniform distribution and $\mu^2$ will be much larger than $\frac{1}{K}$. Another hyperparameter $\alpha$ can be added into Eq. [5] to further enhance the effect by making $\mathbf{e}_i^2 = \alpha$ and $\mu^2 = \alpha \sum_i p_i^2$:

$$L_{\text{REG}}(\mathbf{x}; \lambda) = \| \mathbf{E}^\top \mathbf{E} - \alpha I \|.$$
4 Experiments

To demonstrate the effectiveness of our proposed approach, we experiment on three different domains of text: reviews, news and questions. Two tasks are used for quantitative analysis: sentiment manipulation and topic modelling. We use a single-layer LSTM for the decoder \( p_{θ}(z|x) \) and a single-layer bi-directional LSTM for the encoder of \( z^{(2)} \), \( q_{φ}(z^{(2)}|x) \). To avoid posterior collapse, we use KL annealing \[2\] and \( β-VAE \[7\] for the different datasets. For decoding, we use beam search with a beam size of 5. Detailed configurations can be found in the supplement.

4.1 Text Style Transfer

**Experimental setup:** We use the truncated Yelp restaurant reviews dataset without reviews exceeding 15 words and the same data split as \[15\]. However, we did not use the sentiment labels, which is different from \[15\] and other supervised approaches. To decide which basis vector corresponds to which sentiment (positive or negative), we feed a sentence with strong sentiment (“awesome!” for positive; “terrible!” for negative) to the encoder and choose the basis vector with the highest \( p_i \) according to Eq.\[3\] yielding \( e_{\text{pos}} \) and \( e_{\text{neg}} \). If \( e_{\text{pos}} \) and \( e_{\text{neg}} \) are chosen to be the same vector, we choose the index with the second highest \( p_i \) for \( e_{\text{pos}} \). To perform style transfer, we fix \( z^{(1)} \) to be the chosen basis vector, that is, \( e_{\text{pos}} \) or \( e_{\text{neg}} \). For evaluation, we follow previous work \[15\] in measuring whether transferred sentences have the correct sentiment according to a pre-trained CNN-based sentiment classifier. In addition, we measure the BLEU score of the transferred sentences against their original sentences following \[9, 18\], since we want the transferred sentence to preserve the original content as much as possible except for the sentiment.

**Quantitative results:** We choose two strong supervised models, Cross-Aligned Autoencoder \[15\] and Adversarial Regularized Autoencoder \[9\] as our baselines. As we can see from Table 2, our proposed unsupervised approach outperforms these supervised baselines on both metrics by a noticeable margin, demonstrating that our unsupervised approach not only discovers the underlying global variations but also helps better disentangle them than previous approaches.

Table 2: Results for sentiment transfer.

| Model                   | Acc(%) | BLEU   | KL  | MI  | Collision |
|-------------------------|--------|--------|-----|-----|-----------|
| Cross-Aligned AE        | 83.63  | 23.68  |     |     |           |
| Adversarial Regularized AE | 83.49  | 25.64  |     |     |           |
| Ours (\( α = 1 \))     | 89.54  | 33.12  |     |     |           |
| Ours (\( α = 10 \))    | 93.78  | 31.62  |     |     |           |

Table 3: Ablation study. The tick mark in the last column means that \( e_{\text{pos}} \) and \( e_{\text{neg}} \) collide and the index with the second highest \( p_i \) is chosen for \( e_{\text{pos}} \).

| Model                        | Acc(%) | BLEU | KL  | MI  | Collision |
|------------------------------|--------|------|-----|-----|-----------|
| Ours w/o \( L_{S-REC} \)     | 81.93  | 37.03| 2.96| 0.93|           |
| Ours w/o \( L_{REG} \)       | 15.03  | 35.73| 0.36| 0.11|           |
| Ours w/o \( f(\cdot) \)      | 16.52  | 23.68| 5.64| 0.84|           |
| Ours setting \( N = K \)     | 69.66  | 35.24| 4.27| 1.80|           |
| Ours with fixed \( E \)      | 16.05  | 43.69| 0.77| 0.24|           |
| Ours (\( α = 1 \))          | 89.54  | 33.12| 3.56| 1.09|           |
| Ours (\( α = 10 \))         | 93.78  | 31.62| 4.44| 1.41|           |

**Ablation study:** We conduct an ablation study by removing \( L_{S-REC} \), \( L_{REG} \), by replacing \( f(\cdot) \) with a LSTM encoder, set \( N \) as \( K \) which is 3 in this setting or fix \( E \) with random orthogonal bases. We also report \( D_{KL}(q_{φ}(z^{(1)}|x)||p(z)) \) (KL), mutual information \( I_q = D_{KL}(q_{φ}(z^{(1)}|x)||p(z)) - D_{KL}(q_{φ}(z^{(1)})||p(z)) \) (MI) and whether \( e_{\text{pos}} \) and \( e_{\text{neg}} \) collide to check the behaviours of our approach in each case. As shown in Table 3, we can see that all three components are crucial to our approach. Without \( L_{S-REC} \), the accuracy drops by a large portion while the BLEU score rises which means that the model gives higher priority to preserve the original content instead of performing style transfer. Without \( L_{REG} \), both KL and MI are close to 0 indicating posterior collapse for \( z^{(1)} \). As a result, the model loses its ability to transfer sentences and fails with respect to accuracy. If we replace \( f(\cdot) \) with a LSTM encoder, we can see that KL does not collapse for \( z^{(1)} \). However, the model collapses to one basis vector and seems unable to acquire the ability to perform style transfer by raw text alone without the help of \( f(\cdot) \). To check whether necessary to have a low-rank instead of full-rank, we force
Table 4: Sentiment manipulation results.

| Sentiment | Review                                                                 |
|-----------|------------------------------------------------------------------------|
| Positive  | Caesar salad and prosciutto melon were wonderful.                      |
| Neutral   | Caesar salad and salad was completely disgusting.                     |
| Negative  | Caesar salad and artichoke dip was meh.                               |
| Positive  | Service was good, food was fun for kids.                              |
| Neutral   | Service was okay, but nothing special for this place.                 |

Table 5: Topic transfer results.

| Science                | Question                                                                 |
|------------------------|--------------------------------------------------------------------------|
| ⇒ Music                | How many times do you think it will take you to go on the moon if you had a million dollars? |
| ⇒ Politics             | How many people voted for Bush in the 2004 presidential election before he leaves office before he leaves office? |
| Music                  | Who has seen "An American Haunting" and how was it?                      |
| ⇒ Science              | What is the wavelength of electromagnetic wave and how is it measured?   |
| ⇒ Politics             | Who thinks George W. Bush is going to be an American President?          |
| Politics               | Who was really behind the JFK assassination?                             |
| ⇒ Science              | What is the speed of light emitted?                                     |
| ⇒ Music                | Who was your favorite character of all time?                            |

$N = K$. We can see that small $N$ will restrict the amount of information encoded in $z^{(1)}$, leading to significant decrease in the transfer ability. With fixed $E$, our model collapses to one basis vector and barely learns useful information for $z^{(1)}$. At the same time, it achieves the highest BLEU score indicating that our model just ignores $z^{(1)}$ and tries to preserve as much as content from the source sentences. In addition, we vary the value of $\alpha$ and find that larger $\alpha$ can force the model to give more attention to $z^{(1)}$ and help prevent posterior collapse for $z^{(1)}$, leading to better performance.

Automatic discovery of finer-grained sentiment: After examining the generated outputs of our model, we found another basis vector which can generate outputs with neutral sentiment. In Table 10, we show two sets of sentiment manipulation results generated by our model. We can see that our model can transfer reviews not only to traditional binarized sentiment but also to the discovered neutral sentiment, which is beyond the capabilities of existing supervised methods for text transfer that are trained on binary datasets. It also points to one possible reason why our unsupervised approach can outperform the supervised approaches. Since text data is inherently noisy and there exist some reviews with neutral or mixed sentiment, supervised approaches may be misled by those noisy examples. By contrast, our approach is able to effectively discover and disentangle these variations without supervision, avoiding the influence of label noise. However, the discovered neutral sentiment is still far from perfect and often fails to generate the desired output. More samples including different kinds of failure cases can be found in the supplement.

Topic transfer: Following [9], we also apply style transfer on the more challenging Yahoo QA dataset [20]. We use the questions only and test on transfer to three topic classes: SCIENCE & MATH, ENTERTAINMENT & MUSIC and POLITICS & GOVERNMENT, for this experiment. Randomly selected generations are shown in Table 5. See the supplement for additional examples.

4.2 Topic Discovery and Generation

Experimental setup: We use the AG news dataset for this task constructed by [20]. It contains four topic categories which are World, Sports, Business and Sci/Tech, with the title and description fields. In this paper, we drop the title and just use the description field. We compare our approach to two standard baselines for unsupervised topic modelling: (1) LDA [1], a standard implementation of LDA is used for this baseline; (2) k-means. To show the power of our approach beyond the pre-trained feature extractor $f(\cdot)$, we perform $k$-means clustering directly on the sentence embedding obtained from $f(\cdot)$. Following [11], we assign each inferred topic to one of the gold-standard topics with the optimal mapping and report the precision (a.k.a. purity), recall (a.k.a. collocation) and $F_1$.

\footnote{https://radimrehurek.com/gensim/}
score. The number of topics is chosen to be 10. The results reported for the baselines and our model are the average over 10 runs.

**Quantitative results:** The results are shown in Table 6. We can see that our approach achieves comparable results to LDA while significantly outperforming \textit{k-means} in all four categories, indicating that our approach can go beyond just clustering on pre-trained feature representations.

**Generation with topic transition:** To further explore the potential of our proposed model, we conduct the following exploratory experiments. We first choose a basis vector discovered by our model and generate a few tokens. Then, we switch the basis vector and continue the generation until the end-of-seq token is generated. Generated samples are shown in Table 1. We see that our model learns to transition the sentence from one topic to another in a natural and fluent way. Several observations can be made based on these samples: (1) it is good at detecting name entities and replacing them with the name entities related to the chosen topic; (2) it is able to preserve the general syntactic structure of the original sentence, demonstrating that useful syntactic information has been learnt by the unstructured latent code \( z^{(2)} \); (3) there is no hard restriction on when to switch the topic; the model will determine an appropriate way to do the transition by itself. In Table 7, we show a typical failure case for topic transition, which is no transition at all. We find that it is the dominating cases for failure made and indicates that our model will perform the transition only if the transition can be conducted in a natural way to avoid awkwardness. Due to space limitations, we put more samples in the supplement.

**Generation with sentiment transition:** We also run the same experiment on sentiment transition. Since reviews in the truncated Yelp dataset are too short to perform the transition, we use the Yelp dataset without truncation following [19]. Randomly selected generations are shown in Table 8. See the supplement for additional examples.

### Table 6: Results for topic identification.

| Topic          | Model   | Precision | Recall | \( F_1 \) |
|----------------|---------|-----------|--------|-----------|
| World          | LDA     | 69.73     | 75.32  | 72.14     |
|                | \textit{k-means} | 67.64     | 47.63  | 55.90     |
|                | Ours    | 80.83     | 70.55  | 74.59     |
| Sports         | LDA     | 79.17     | 82.50  | 80.22     |
|                | \textit{k-means} | 47.66     | 89.50  | 62.04     |
|                | Ours    | 81.14     | 78.88  | 79.49     |
| Business       | LDA     | 72.10     | 66.45  | 68.46     |
|                | \textit{k-means} | 53.06     | 53.16  | 53.11     |
|                | Ours    | 64.04     | 64.53  | 63.97     |
| Sci/Tech       | LDA     | 66.55     | 59.77  | 61.60     |
|                | \textit{k-means} | 81.32     | 31.59  | 44.67     |
|                | Ours    | 65.20     | 71.74  | 66.77     |

### Table 7: One failure case for topic transition.

| Sci/Tech throughout | Reuters - Microsoft Corp. on Monday said it will offer a new version of its desktop search tool for mobile devices. |
|---------------------|-------------------------------------------------------------------------------------------------------------|
| Sci/Tech to Sports  | Reuters - Microsoft Corp. on Monday said it will offer a new version of its desktop search tool for mobile phones. |

### Table 8: Sentiment transition results.

| Sentiment transition | Positive throughout                                                                        | Negative throughout                                                                 |
|----------------------|---------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|
|                      | this is a great place to go for lunch or dinner . the food is good , and the staff is friendly . the food is good , and the staff is friendly . the food is good , and the staff is friendly . | this is the worst indian food i have ever had in my life . i have been here several times and have never been disappointed . this is the worst indian food i have ever had . |
|                      | Positive to Negative                                                                         | Negative to Positive                                                                  |
|                      | this is a great place to go for lunch or dinner . the food is good , and the staff is friendly . the food is good , but the service is horrible . the food is mediocre at best and the service is terrible . i will never go back . | this is the worst indian food i have ever had in my life . i have been to this location several times , and i 've never had a bad experience here . the food is consistently good , and the staff is friendly and helpful . |

### 5 Conclusion

In this work, we propose an unsupervised framework to perform controllable text generation leveraging VAEs and large-scale pre-training on language. With a latent decomposition and a structured latent space, our framework can discover and disentangle the dominant global variations in a dataset without supervision, leading to more flexible control over text generation than previous approaches.
References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[2] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.

[3] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[6] R. He, W. S. Lee, H. T. Ng, and D. Dahlmeier. An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 388–397, 2017.

[7] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, volume 3, 2017.

[8] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1587–1596. JMLR. org, 2017.

[9] Y. Kim, K. Zhang, A. M. Rush, Y. LeCun, et al. Adversarially regularized autoencoders. *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[10] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[11] C. Manning, P. Raghavan, and H. Schütze. Introduction to information retrieval. *Natural Language Engineering*, 16(1):100–103, 2010.

[12] A. v. d. Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016.

[13] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[14] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1:8, 2019.

[15] T. Shen, T. Lei, R. Barzilay, and T. Jaakkola. Style transfer from non-parallel text by cross-alignment. *Advances in neural information processing systems*, pages 6830–6841, 2017.

[16] W. Wang, Z. Gan, W. Wang, D. Shen, J. Huang, W. Ping, S. Satheesh, and L. Carin. Topic compositional neural language model. *arXiv preprint arXiv:1712.09783*, 2017.

[17] T.-H. Wen, D. Vandyke, N. Mrksic, M. Gasic, L. M. Rojas-Barahona, P.-H. Su, S. Ultes, and S. Young. A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562*, 2016.

[18] Z. Yang, Z. Hu, C. Dyer, E. P. Xing, and T. Berg-Kirkpatrick. Unsupervised text style transfer using language models as discriminators. In *Advances in Neural Information Processing Systems*, pages 7287–7298, 2018.

[19] Z. Yang, Z. Hu, R. Salakhutdinov, and T. Berg-Kirkpatrick. Improved variational autoencoders for text modeling using dilated convolutions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3881–3890. JMLR. org, 2017.

[20] X. Zhang, J. Zhao, and Y. LeCun. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657, 2015.
A Experimental Details

All the experiments are implemented with PyTorch\(^3\). Different settings are used for each dataset:

- **Yelp-truncated**: short reviews for sentiment manipulation.
- **Yahoo-QA**: questions for topic transfer.
- **AG-news**: news for topic transition.
- **Yelp-full**: moderately long reviews for sentiment transition.

A.1 Feature Extractor

We use the open-sourced pre-trained models of BERT\(^4\) and extract the feature vector of the [CLS] token in the last layer as \(f(\cdot)\), whose dimension is 1024. For **Yelp-truncated** and **Yelp-full**, we use BERT-Large Cased; for **Yahoo-QA** and **AG-news**, BERT-Large Uncased is used.

A.2 Optimization Details

Across all the datasets, we use Adam with learning rate 0.001 to update the parameters for the encoding \(\phi \cup \lambda\), while SGD with learning rate 1.0 to update the parameters for the decoding \(\theta\). The batch size is chosen to be 32. Dropouts with drop probability 0.5 are applied after the embedding layer and the LSTM layer in the decoder. We train the model until the reconstruction loss \(\mathbb{E}_{z \sim q_{\phi}(z|x)} [\log p_{\theta}(x|z)]\) stops decreasing.

A.3 Mitigating Posterior Collapse

For the structured part \(z^{(1)}\), we use \(\beta\)-VAE setting \(\beta\) as 0.2 across all the datasets. For the unstructured part \(z^{(2)}\), different strategies are employed for each dataset:

- **Yelp-truncated**: \(\beta\)-VAE setting \(\beta\) as 0.4.
- **Yahoo-QA**: \(\beta\)-VAE setting \(\beta\) as 0.6.
- **AG-news**: KL annealing, from 0.1 to 1.0 in 10 epochs.
- **Yelp-full**: KL annealing, from 0.1 to 1.0 in 10 epochs.

A.4 Hyperparameter Settings

The hyperparameters are chosen by checking \(L_{\text{VAE}}\), KL, MI and the generated outputs on the development set for **Yelp-truncated** and **AG-News**. Except \(K\), **Yahoo-QA** follows the same setting as **Yelp-truncated** and **Yelp-full** follows **AG-News**.

Table 9: Hyperparameter settings.

|                         | Yelp-truncated | Yahoo-QA | AG-News | Yelp-full |
|-------------------------|----------------|----------|---------|-----------|
| Number of variations \(K\) | 3              | 5        | 10      | 5         |
| Input dimension for LSTM encoder | 128            | 128      | 512     | 512       |
| Hidden dimension for LSTM encoder | 1024           | 1024     | 1024    | 1024      |
| Dimension for \(z^{(2)}\) | 32             | 32       | 96      | 96        |
| Dimension for \(z^{(1)}\) | 16             | 16       | 32      | 32        |
| Input dimension for LSTM decoder | 128            | 128      | 512     | 512       |
| Hidden dimension for LSTM decoder | 1024           | 1024     | 1024    | 1024      |

\(^3\)https://pytorch.org/
\(^4\)https://github.com/google-research/bert
### B Text Transfer Examples

#### B.1 Sentiment manipulation on Yelp restaurant reviews (truncated)

Table 10: Sentiment manipulation results.

| Positive                                      | ⇒ Negative                                      |
|-----------------------------------------------|-------------------------------------------------|
| great customer service and all around!        | horrible customer service and all around!       |
| ⇒ Neutral                                     | great customer service and all around!          |
| Positive                                      | so grateful to god for leading my family and i here .|
| ⇒ Negative                                    | so run to this hotel and will never return on them .|
| ⇒ Neutral                                     | so kudos to dr. cohen for my wife and i know .   |
| Positive                                      | food , service , and prices were perfect!       |
| ⇒ Negative                                    | food , service , and service were not!          |
| ⇒ Neutral                                     | food , service , and drinks were $ _num_ !       |
| Positive                                      | each dish comes with a great helping of hot rice which i also love .|
| ⇒ Negative                                    | each time with a side of cheese no bread i have never tasted .|
| ⇒ Neutral                                     | my boyfriend ‘s $ _num_ with a side of chicken they have n’t ordered .|
| Negative                                      | overall this place just needs to go away .      |
| ⇒ Positive                                    | overall this place has to go to it .            |
| ⇒ Neutral                                     | overall this place does n’t take it home .      |
| Negative                                      | i can only say , ripped off .                   |
| ⇒ Positive                                    | i can only say , my childhood .                  |
| ⇒ Neutral                                     | i can only say , my buddy .                     |
| Negative                                      | i was thankful i did n’t get sick .              |
| ⇒ Positive                                    | i was impressed with my first visit .            |
| ⇒ Neutral                                     | i was like i did n’t get sick .                  |
| Negative                                      | not good .                                      |
| ⇒ Positive                                    | highly recommended .                            |
| ⇒ Neutral                                     | not good .                                      |
### B.2 Topic transfer on Yahoo QA

Table 11: Topic transfer examples.

| Science | What is pi ? |
|---------|--------------|
| ⇒ Music | What is your favorite song ? |
| ⇒ Politics | What is this ? |
| Science | what role does <unk> play in the cell wall of a bacteria ? |
| ⇒ Music | what are your views on the sex offender registry for a band ? |
| ⇒ Politics | what role did communism play in the united states of a dictatorship ? |
| Science | Are skunks part of the rodent family ? |
| ⇒ Music | Are you looking forward to the actor ? |
| ⇒ Politics | Are there any filipinos siding with the Jews ? |
| Music | why does some good actors doesn't get chances in movies ? |
| ⇒ Science | why are there so many metals in other metals ? |
| ⇒ Politics | why are there so many black people in politics ? |
| Music | what kinda dances do the pussycat dolls dance ? |
| ⇒ Science | what makes the red hot chili peppers ? |
| ⇒ Politics | what ever happened with the downing street ? |
| Music | who sings this song ? |
| ⇒ Science | what is the antiderivative of -1 ? |
| ⇒ Politics | who sings this song ? |
| Politics | What was so bad about the British response to the IRA in Northern Ireland ? |
| ⇒ Science | What will happen to the atoms in the periodic table of elements ? |
| ⇒ Music | What was your favorite movie from the 80 's and 80 's ? |
| Politics | Are corporations covered by the 5th amendment ’s privilege against self <unk> so , who holds it ? |
| ⇒ Science | Are there any chemical properties that determine the amount of moles of water , and why ? |
| ⇒ Music | Are you satisfied with the fact that you are going to vote for your birthday party ? |
| Politics | how long can a will be in probate before the government steps in ? |
| ⇒ Science | how long does it take for a substance in the human body ? |
| ⇒ Music | how long do you think you are in the military for a good job ? |
### C Text Transition Examples

#### C.1 Topic transition on AG news

Table 12: Topic transition examples.

| World throughout | BAGHDAD (Reuters) - Iraq’s interim prime minister, Iyad Allawi, said on Monday that the United States had no intention of withdrawing from the country to end the violence in Iraq. |
| World to Sports   | BAGHDAD (Reuters) - Iraq’s interim prime minister, Iyad Allawi, said on Monday that the United States had no intention of withdrawing its troops from the country to the end of the year. |
| World to Business | BAGHDAD (Reuters) - Iraq’s interim prime minister, Iyad Allawi, said on Monday that the United States had no intention of withdrawing its troops from the country to the country. |
| World to Sci/Tech | BAGHDAD (Reuters) - Iraq’s interim prime minister, Iyad Allawi, said on Monday that the United States had no intention of withdrawing its uranium enrichment program to the United States. |
| Sports throughout | For the first time in four years, the US men’s basketball team won the gold medal in the men’s 400-meter medley relay. |
| Sports to World   | For the first time in four years, the US men’s basketball team won the gold medal at the Athens Olympics in Athens, where the United States and the United States have agreed to a peace deal. |
| Sports to Business| For the first time in four years, the US men’s basketball team won the gold medal at the Athens Olympics on Wednesday, with a surge in crude oil prices. |
| Sports to Sci/Tech| For the first time in four years, the US men’s basketball team won the gold medal in the men’s Olympic basketball tournament in Beijing on Tuesday. |
| Business throughout | NEW YORK (Reuters) - U.S. stocks opened higher on Friday, as oil prices climbed above $48 a barrel and the Federal Reserve raised interest rates by a quarter percentage point. |
| Business to World | NEW YORK (Reuters) - U.S. stocks opened higher on Friday, as oil prices climbed above $48 a barrel and the Federal Reserve raised interest rates by a quarter percentage point. |
| Business to Sports | NEW YORK (Reuters) - U.S. stocks opened higher on Friday, as oil prices climbed above $48 a barrel and the Federal Reserve raised interest rates by a quarter percentage point. |
| Business to Sci/Tech | NEW YORK (Reuters) - U.S. stocks opened higher on Friday, as oil prices climbed above $48 a barrel and the Federal Communications Commission said it would allow the companies to use mobile phones. |
| Sci/Tech throughout | SINGAPORE (Reuters) - South Korea’s Hynix Semiconductor Inc. said on Tuesday it had developed a prototype micro fuel cell recharger for a range of security vulnerabilities in India. |
| Sci/Tech to World | SINGAPORE (Reuters) - South Korea’s Hynix Semiconductor Inc. said on Tuesday it had developed a prototype micro fuel cell aimed at ending a standoff with North Korea. |
| Sci/Tech to Sports | SINGAPORE (Reuters) - South Korea’s Hynix Semiconductor Inc. said on Tuesday it had developed a prototype micro fuel cell aimed at protecting the world’s biggest gold medal. |
| Sci/Tech to Business | SINGAPORE (Reuters) - South Korea’s Hynix Semiconductor Inc. said on Tuesday it had developed a prototype micro fuel cell aimed at protecting the world’s largest oil producer. |
C.2 Sentiment transition on Yelp restaurant reviews (full)

Table 13: Sentiment transition examples (I): first row with **Positive** throughout; second row with transition from **Positive** to **Negative**

| Positive throughout | Positive to Negative |
|---------------------|----------------------|
| i am a fan of this place. i’ve been here a few times and have never had a bad experience. the staff is always friendly and helpful. the staff is always friendly and helpful. the staff is always friendly and helpful. the staff is always friendly and helpful. the staff is always friendly and helpful. i would recommend this place to anyone looking for a great experience. |
| i am a fan of this place. i’ve been here a few times and have never had a bad experience. the last time i was there i had to wait for my car to get my car washed. i had to wait a long time to get my car washed. i will never go back. |

| Negative throughout | Negative to Positive |
|---------------------|----------------------|
| i’ve been here a few times and have never been disappointed. the food is good, and the service is good. the food is good, and the service is good. the food is good, and the service is good. the food is good, and the service is good. |
| this is the worst sandwich i have ever had in my life. i ordered a turkey sandwich and was told it would be ready to be ready. i asked for a refund and was told that they were out of stock. i will never go back. |

Table 14: Sentiment transition examples (II): first row with **Negative** throughout; second row with transition from **Negative** to **Positive**

| Negative throughout | Negative to Positive |
|---------------------|----------------------|
| this is the worst sandwich i have ever had in my life. i ordered a turkey sandwich and it was very good. i also had a side of fries with a side of fries and a side of fries. the service was friendly and prompt. i will definitely be back. |
| this is the worst sandwich i have ever had in my life. i ordered a turkey sandwich and it was very good. i also had a side of fries with a side of fries and a side of fries. the service was friendly and prompt. i will definitely be back. |

| Negative throughout | Negative to Positive |
|---------------------|----------------------|
| this place is horrible! i went in for a pedicure and they didn’t have any polish on my nails. i asked for a gel manicure and they didn’t even give me a discount. i will never go back. |
| this place is horrible! i went in to get my nails done and they didn’t even tell me it would be ready. i was very pleased with the service and the staff was very friendly. i will definitely be back! |

| Negative throughout | Negative to Positive |
|---------------------|----------------------|
| this is the worst airline i’ve ever been to. i’ve been here twice and the last time i was there i had to wait for a taxi to get to the airport. the driver was rude and did n’t even apologize. i will never stay here again. |
| this is the worst airline i’ve ever been to. i’ve been here twice and both times i’ve been disappointed. i’ve been here a few times, and i’ve never had a bad experience here. it’s a great place to stay if you’re in the area. |