Educating Text Autoencoders: Latent Representation Guidance via Denoising Supplementary Material

A. Wasserstein Distance

The AAE objective can be connected to a relaxed form of the Wasserstein distance between model and data distributions (Tolstikhin et al., 2017). Specifically, for cost function $c(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and deterministic decoder mapping $G : \mathcal{Z} \to \mathcal{X}$, it holds that:

$$\begin{aligned} &\inf_{\Gamma \in \mathcal{P}(x\sim p_{\text{data}}, y\sim p_G)} \mathbb{E}_{(x,y)\sim\Gamma}[c(x,y)] \\ = &\inf_{q(z|x):q(z)=p(z)} \mathbb{E}_{p_{\text{data}}(x)}\mathbb{E}_{q(z|x)}[c(x,G(z))] \end{aligned} \tag{7}$$

where the minimization over couplings Γ with marginals p_{data} and p_G can be replaced with minimization over conditional distributions $q(z|x)$ whose marginal $q(z) = \mathbb{E}_{p_{\text{data}}(x)}[q(z|x)]$ matches the latent prior distribution $p(z)$. Relaxing this marginal constraint via a divergence penalty $D(q(z)\|p(z))$ estimated by adversarial training, one recovers the AAE objective (Eq. 1). In particular, AAE on discrete x with the cross-entropy loss is minimizing an upper bound of the total variation distance between p_{data} and p_G, with c chosen as the indicator cost function (Zhao et al., 2018).

Our model is optimizing over conditional distributions $q(z|x)$ of the form (6), a subset of all possible conditional distributions. Thus, after introducing input perturbations, our method is still minimizing an upper bound of the Wasserstein distance between p_{data} and p_G described in (7).

B. Toy Experiments With Latent Dimension 5

Here we repeat our toy experiment with clustered data, this time using a larger latent space with 5 dimensions.

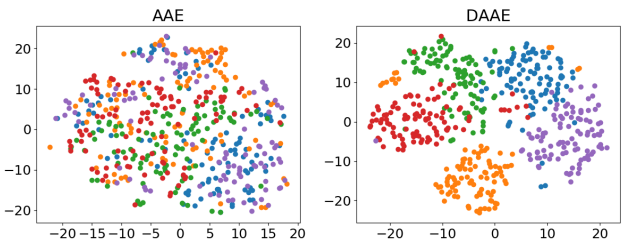

Figure B.1. t-SNE visualization of 5-D latent representations learned by AAE and DAAE when mapping clustered sequences in $\mathcal{X} = \{0,1\}^{50}$ to $\mathcal{Z} = \mathbb{R}^5$. The training data stem from 5 underlying clusters, with 100 sequences sampled from each (colored accordingly by cluster identity).

C. Proof of Theorem 1

Theorem 1. *For any one-to-one encoder mapping E from $\{x_1, \cdots, x_n\}$ to $\{z_1, \cdots, z_n\}$, the optimal value of objective $\max_{G\in\mathcal{G}_L} \frac{1}{n}\sum_{i=1}^n \log p_G(x_i|E(x_i))$ is the same.*

Proof. Consider two encoder matchings x_i to $z_{\alpha(i)}$ and x_i to $z_{\beta(i)}$, where both α and β are permutations of the indices $\{1,\ldots,n\}$. Suppose G_α is the optimal decoder model for the first matching (with permutations α). This implies

$$p_{G_\alpha} = \arg\max_{G\in\mathcal{G}_L} \sum_{i=1}^n \log p_G(x_i|z_{\alpha(i)})$$

Now let $p_{G_\beta}(x_i|z_j) = p_{G_\alpha}(x_{\beta\alpha^{-1}(i)}|z_j), \forall i,j$. Then G_β can achieve exactly the same log-likelihood objective value for matching β as G_α for matching α, while still respecting the Lipschitz constraint. □

D. Proof of Theorem 2

Theorem 2. *Let d be a distance metric over $\mathcal{X}$. Suppose x_1, x_2, x_3, x_4 satisfy that with some $\epsilon > 0$: $d(x_1,x_2) < \epsilon$, $d(x_3,x_4) < \epsilon$, and $d(x_i,x_j) > \epsilon$ for all other (x_i,x_j) pairs. In addition, z_1, z_2, z_3, z_4 satisfy that with some $0 < \delta < \zeta$: $\|z_1 - z_2\| < \delta$, $\|z_3 - z_4\| < \delta$, and $\|z_i - z_j\| > \zeta$ for all other (z_i,z_j) pairs. Suppose our perturbation process C reflects local $\mathcal{X}$ geometry with: $p_C(x_i|x_j) = 1/2$ if $d(x_i,x_j) < \epsilon$ and $= 0$ otherwise. For $\delta < 1/L \cdot (2\log(\sigma(L\zeta)) + \log 2)$ and $\zeta > 1/L \cdot \log\left(1/(\sqrt{2}-1)\right)$, the denoising objective $\max_{G\in\mathcal{G}_L} \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^n p_C(x_j|x_i)\log p_G(x_i|E(x_j))$ (where $n = 4$) achieves the largest value when encoder E maps close pairs of x to close pairs of z.*

Proof. Let $[n]$ denote $\{1,\ldots,n\}$, and assume without loss of generality that the encoder E maps each x_i to z_i. We also define $A = \{1,2\}, B = \{3,4\}$ as the two x-pairs that lie close together. For our choice of $C(x)$, the training objective to be maximized is:

$$\begin{aligned} &\sum_{i,j\in A} \log p_G(x_i|E(x_j)) + \sum_{k,\ell\in B} \log p_G(x_k|E(x_\ell)) \\ = &\sum_{i,j\in A} \log p_G(x_i|z_j) + \sum_{k,\ell\in B} \log p_G(x_k|z_\ell) \end{aligned} \tag{8}$$

The remainder of our proof is split into two cases:

Case 1. $||z_j - z_\ell|| > \zeta$ for $j \in A, \ell \in B$

Case 2. $||z_j - z_\ell|| < \delta$ for $j \in A, \ell \in B$

Under Case 1, x points that lie far apart also have z encodings that remain far apart. Under Case 2, x points that lie far apart have z encodings that lie close together. We complete the proof by showing that the achievable objective value in Case 2 is strictly worse than in Case 1, and thus an optimal encoder/decoder pair would avoid the x, z matching that leads to Case 2.

In Case 1 where $||z_j - z_\ell|| > \zeta$ for all $j \in A, \ell \in B$, we can lower bound the training objective (8) by choosing:

$$p_G(x_i|z_j) = \begin{cases} (1-\gamma)/2 & \text{if } i,j \in A \text{ or } i,j \in B \\ \gamma/2 & \text{otherwise} \end{cases} \tag{9}$$

with $\gamma = \sigma(-L\zeta) \in (0, \frac{1}{2})$, where $\sigma(\cdot)$ denotes the sigmoid function. Note that this ensures $\sum_{i\in[4]} p_G(x_i|z_j) = 1$ for each $j \in [4]$, and does not violate the Lipschitz condition from Assumption 2 since:

$$|\log p_G(x_i|z_j) - \log p_G(x_i|z_\ell)|$$

$$\begin{cases} = 0 & \text{if } j,\ell \in A \text{ or } j,\ell \in B \\ \leq \log\left((1-\gamma)/\gamma\right) & \text{otherwise} \end{cases}$$

and thus remains $\leq L||z_j - z_\ell||$ when $\gamma = \sigma(-L\zeta) \geq \sigma(-L||z_j - z_\ell||) = 1/[1 + \exp(L||z_j - z_\ell||)]$.

Plugging the $p_G(x|z)$ assignment from (9) into (8), we see that an optimal decoder can obtain training objective value $\geq 8 \log[\sigma(L\zeta)/2]$ in Case 1 where $||z_j - z_\ell|| > \zeta$, $\forall j \in A, \ell \in B$.

Next, we consider the alternative case where $||z_j - z_\ell|| < \delta$ for $j \in A, \ell \in B$.

For $i, j \in A$ and for all $\ell \in B$, we have:

$$\begin{aligned} \log p_G(x_i|z_j) &\leq \log p_G(x_i|z_\ell) + L||z_j - z_\ell|| \\ &\qquad \text{(by Assumption 2)} \\ &\leq \log p_G(x_i|z_\ell) + L\delta \\ &\leq L\delta + \log\left[1 - \sum_{k\in B} p_G(x_k|z_\ell)\right] \\ &\qquad \text{(since } \textstyle\sum_k p_G(x_k|z_\ell) \leq 1\text{)} \end{aligned}$$

Continuing from (8), the overall training objective in this case is thus:

$$\begin{aligned} &\sum_{i,j\in A} \log p_G(x_i|z_j) + \sum_{k,\ell\in B} \log p_G(x_k|z_\ell) \\ &\leq 4L\delta + \sum_{i,j\in A} \min_{\ell\in B} \log\left[1 - \sum_{k\in B} p_G(x_k|z_\ell)\right] \\ &\quad + \sum_{k,\ell\in B} \log p_G(x_k|z_\ell) \\ &\leq 4L\delta + \sum_{\ell\in B}\left[2\log\left(1 - \sum_{k\in B} p_G(x_k|z_\ell)\right)\right. \\ &\quad \left. + \sum_{k\in B} \log p_G(x_k|z_\ell)\right] \\ &\leq 4L\delta - 12\log 2 \end{aligned}$$

using the fact that the optimal decoder for the bound in this case is: $p_G(x_k|z_\ell) = 1/4$ for all $k, \ell \in B$.

Finally, plugging our range for δ stated in the Theorem 2, it shows that the best achievable objective value in Case 2 is strictly worse than the objective value achievable in Case 1. Thus, the optimal encoder/decoder pair under the AAE with perturbed x will always prefer the matching between $\{x_1, \ldots, x_4\}$ and $\{z_1, \ldots, z_4\}$ that ensures nearby x_i are encoded to nearby z_i (corresponding to Case 1). □

E. Proof of Theorem 3

Theorem 3. *Suppose $x_1, \cdots, x_n$ are divided into n/K clusters of equal size K, with S_i denoting the cluster index of x_i. Let the perturbation process C be uniform within clusters, i.e. $p_C(x_i|x_j) = 1/K$ if $S_i = S_j$ and $= 0$ otherwise. With a one-to-one encoder mapping E from $\{x_1, \cdots, x_n\}$ to $\{z_1, \cdots, z_n\}$, the denoising objective $\max_{G\in\mathcal{G}_L} \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^n p_C(x_j|x_i) \log p_G(x_i|E(x_j))$ has an upper bound: $\frac{1}{n^2}\sum_{i,j:S_i\neq S_j} \log\sigma(L\|E(x_i) - E(x_j)\|) - \log K$.*

Proof. Without loss of generality, let $E(x_i) = z_i$ for notational convenience. We consider what is the optimal decoder probability assignment $p_G(x_i|z_j)$ under the Lipschitz constraint 2.

The objective of the AAE with perturbed x is to maximize:

$$\begin{aligned} &\frac{1}{n}\sum_i \sum_j p_C(x_j|x_i) \log p_G(x_i|E(x_j)) \\ &= \frac{1}{nK}\sum_j \sum_{i:S_i=S_j} \log p_G(x_i|z_j) \end{aligned}$$

We first show that the optimal $p_G(\cdot|\cdot)$ will satisfy that the same probability is assigned within a cluster, i.e. $p(x_i|z_j) =$

$p(x_k|z_j)$ for all i, k s.t. $S_i = S_k$. If not, let $P_{sj} = \sum_{i:S_i=s} p_G(x_i|z_j)$, and we reassign $p_{G'}(x_i|z_j) = P_{S_i j}/K$. Then G' still conforms to the Lipschitz constraint if G meets it, and G' will have a larger target value than G.

Now let us define $P_j = \sum_{i:S_i=S_j} p_G(x_i|z_j) = K \cdot p_G(x_j|z_j)$ $(0 \le P_j \le 1)$. The objective becomes:

$$\begin{aligned}
&\max_{p_G} \frac{1}{nK} \sum_j \sum_{i:S_i=S_j} \log p_G(x_i|z_j) \\
=\ &\max_{p_G} \frac{1}{n} \sum_j \log p_G(x_j|z_j) \\
=\ &\max_{p_G} \frac{1}{n} \sum_j \log P_j - \log K \\
=\ &\max_{p_G} \frac{1}{2n^2} \sum_i \sum_j (\log P_i + \log P_j) - \log K \\
\le\ &\frac{1}{2n^2} \sum_i \sum_j \max_{p_G} (\log P_i + \log P_j) - \log K
\end{aligned}$$

Consider each term $\max_{p_G}(\log P_i + \log P_j)$: when $S_i = S_j$, this term can achieve the maximum value 0 by assigning $P_i = P_j = 1$; when $S_i \neq S_j$, the Lipschitz constraint ensures that:

$$\begin{aligned}
\log(1 - P_i) &\ge \log P_j - L\|z_i - z_j\| \\
\log(1 - P_j) &\ge \log P_i - L\|z_i - z_j\|
\end{aligned}$$

Therefore:

$$\log P_i + \log P_j \le 2 \log \sigma(L\|z_i - z_j\|)$$

Overall, we thus have:

$$\begin{aligned}
&\max_{p_G} \frac{1}{nK} \sum_j \sum_{i:S_i=S_j} \log p_G(x_i|z_j) \\
\le\ &\frac{1}{n^2} \sum_{i,j:S_i \neq S_j} \log \sigma(L\|z_i - z_j\|) - \log K
\end{aligned}$$

□

F. Experimental Details

We use the same architecture to implement all models with different objectives. The encoder E, generator G, and the language model used to compute Forward/Reverse PPL are one-layer LSTMs with hidden dimension 1024 and word embedding dimension 512. The last hidden state of the encoder is projected into 128/256 dimensions to produce the latent code z for Yelp/Yahoo datasets respectively, which is then projected and added with input word embeddings fed to the generator. The discriminator D is an MLP with one hidden layer of size 512. λ of AAE based models is set to 10 to ensure the latent codes are indistinguishable from the prior. All models are trained via the Adam optimizer (Kingma & Ba, 2014) with learning rate 0.0005, $\beta_1 = 0.5$, $\beta_2 = 0.999$. At test time, encoder-side perturbations are disabled, and we use greedy decoding to generate x from z.

G. Human Evaluation

For the tense transfer experiment, the human annotator is presented with a source sentence and two outputs (one from each approach, presented in random order) and asked to judge which one successfully changes the tense while being faithful to the source, or whether both are good/bad, or if the input is not suitable to have its tense inverted. We collect labels from two human annotators and if they disagree, we further solicit a label from the third annotator.

H. Neighborhood Preservation

Here we include non-generative models AE, DAE and repeat the neighborhood preservation analysis in Section 5.1. We find that an untrained RNN encoder from random initialization has a good recall rate, and we suspect that SGD training of vanilla AE towards only the reconstruction loss will not overturn this initial bias. Note that denoising still improves neighborhood preservation in this case. Also note that DAAE has the highest recall rate among all generative models that have a latent prior imposed.

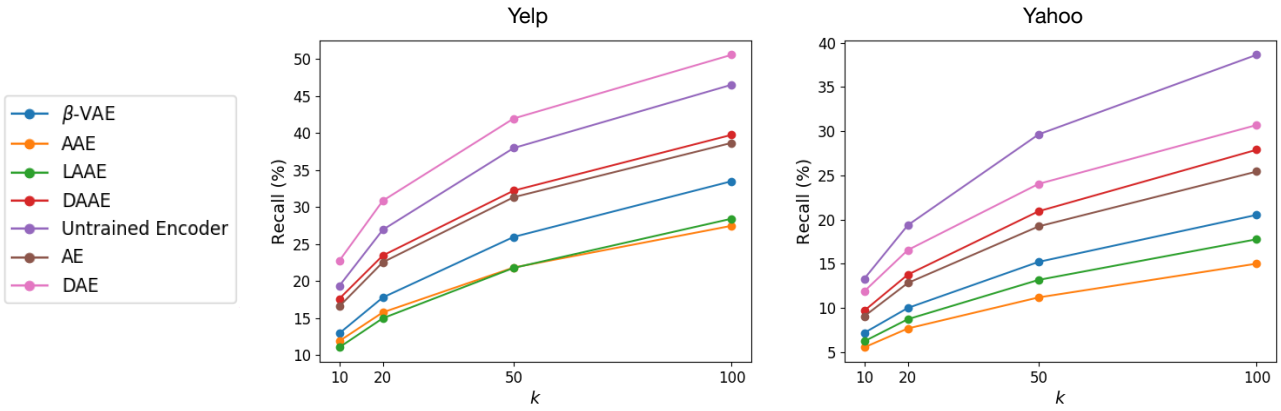

Figure H.2. Recall rate of 10 nearest neighbors in the sentence space retrieved by k nearest neighbors in the latent space of different autoencoders on the Yelp and Yahoo datasets.

Source	**how many gospels are there that were n't included in the bible ?**
5-NN by AAE	there are no other gospels that were n't included in the bible . how many permutations are there for the letters in the word _UNK ' ? anyone else picked up any of the _UNK in the film ? what 's the significance of the number 40 in the bible ? how many pieces of ribbon were used in the _UNK act ?
5-NN by DAAE	there are no other gospels that were n't included in the bible . how many litres of water is there in the sea ? how many _UNK gods are there in the classroom ? how many pieces of ribbon were used in the _UNK act ? how many times have you been grounded in the last year ?
Source	**how do i change colors in new yahoo mail beta ?**
5-NN by AAE	how should you present yourself at a _UNK speaking exam ? how can i learn to be a hip hop producer ? how can i create a _UNK web on the internet ? how can i change my _UNK for female not male ? what should you look for in buying your first cello ?
5-NN by DAAE	how do i change that back to english ? is it possible to _UNK a yahoo account ? how do i change my yahoo toolbar options ? what should you look for in buying your first cello ? who do you think should go number one in the baseball fantasy draft , pujols or _UNK ?

Table H.1. Examples of nearest neighbors in the latent Euclidean space of AAE and DAAE on Yahoo dataset.

I. Generation-Reconstruction Results on the Yahoo Dataset

In this section, we repeat the autoencoder generation-reconstruction analysis in Section 5.2 on the Yahoo dataset. As on the Yelp dataset, our DAAE model provides the best trade-off.

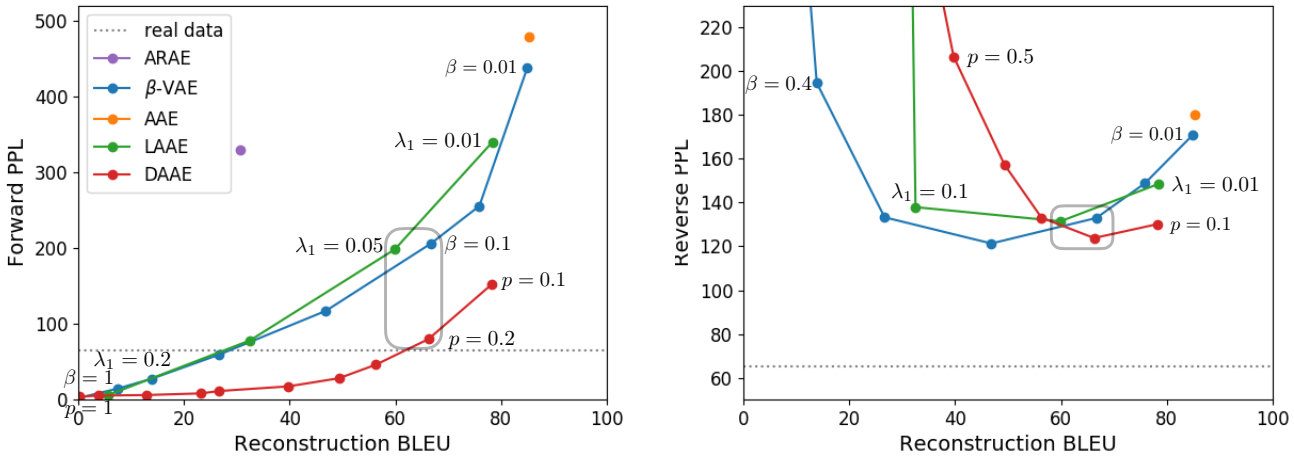

Figure I.3. Generation-reconstruction trade-off of various text autoencoders on the Yahoo dataset. The "real data" line marks the PPL of a language model trained and evaluated on real data. We strive to approach the lower right corner with both high BLEU and low PPL. The grey box identifies hyperparameters we finalize for respective models. Points of severe collapse (Reverse PPL > 300) are removed from the right panel.

J. Additional Examples of Generated Text

This section presents more examples of text decoded after various geometric manipulations of the latent space representations.

Input	**the staff is rude and the dr. does not spend time with you .**	**slow service , the food tasted like last night 's leftovers .**
ARAE	the staff is rude and the dr. does not worth two with you .	slow service , the food tasted like last night 's leftovers .
β-VAE	the staff was rude and the dr. did not spend time with your attitude .	slow service , the food tastes like last place serves .
AAE	the staff was rude and the dr. does not spend time with you .	slow service , the food tasted like last night 's leftovers .
LAAE	the staff was rude and the dr. is even for another of her entertained .	slow service , the food , on this burger spot !
DAAE	the staff was rude and the dr. did not make time with you .	slow service , the food tastes like last night ... .
Input	**they are the worst credit union in arizona .**	**i reported this twice and nothing was done .**
ARAE	they are the worst bank credit in arizona .	i swear this twice and nothing was done .
β-VAE	they were the worst credit union in my book .	i 've gone here and nothing too .
AAE	they are the worst credit union in arizona .	i reported this twice and nothing was done .
LAAE	they were the worst credit union in my heart .	i dislike this twice so pleasant guy .
DAAE	they were the worst credit union in arizona ever .	i hate this pizza and nothing done .

Table J.2. Additional examples of vector arithmetic for tense inversion.

	AAE	DAAE
Input	**this woman was extremely rude to me .**	**this woman was extremely rude to me .**
$+v$	this woman was extremely rude to me .	this woman was extremely nice .
$+1.5v$	this woman was extremely rude to baby .	this staff was amazing .
$+2v$	this woman was extremely rude to muffins .	this staff is amazing .
Input	**my boyfriend said his pizza was basic and bland also .**	**my boyfriend said his pizza was basic and bland also .**
$+v$	my boyfriend said his pizza was basic and tasty also .	my boyfriend said his pizza is also excellent .
$+1.5v$	my shared said friday pizza was basic and tasty also .	my boyfriend and pizza is excellent also .
$+2v$	my shared got pizza pasta was basic and tasty also .	my smoked pizza is excellent and also exceptional .
Input	**the stew is quite inexpensive and very tasty .**	**the stew is quite inexpensive and very tasty .**
$-v$	the stew is quite inexpensive and very tasty .	the stew is quite an inexpensive and very large .
$-1.5v$	the stew is quite inexpensive and very very tasteless .	the stew is quite a bit overpriced and very fairly brown .
$-2v$	the – was being slow - very very tasteless .	the hostess was quite impossible in an expensive and very few customers .
Input	**the patrons all looked happy and relaxed .**	**the patrons all looked happy and relaxed .**
$-v$	the patrons all looked happy and relaxed .	the patrons all helped us were happy and relaxed .
$-1.5v$	the patrons all just happy and smelled .	the patrons that all seemed around and left very stressed .
$-2v$	the patrons all just happy and smelled .	the patrons actually kept us all looked long and was annoyed .

Table J.3. Additional examples of vector arithmetic for sentiment transfer.

Input 1	**what language should i learn to be more competitive in today 's global culture ?**
Input 2	**what languages do you speak ?**
AAE	what language should i learn to be more competitive in today 's global culture ?
	what language should i learn to be more competitive in today 's global culture ?
	what language should you speak ?
	what languages do you speak ?
	what languages do you speak ?
DAAE	what language should i learn to be more competitive in today 's global culture ?
	what language should i learn to be competitive today in arabic 's culture ?
	what languages do you learn to be english culture ?
	what languages do you learn ?
	what languages do you speak ?
Input 1	**i believe angels exist .**
Input 2	**if you were a character from a movie , who would it be and why ?**
AAE	i believe angels exist .
	i believe angels - there was the exist exist .
	i believe in tsunami romeo or <unk> i think would it exist as the world population .
	if you were a character from me in this , would we it be (why !
	if you were a character from a movie , who would it be and why ?
DAAE	i believe angels exist .
	i believe angels exist in the evolution .
	what did <unk> worship by in <unk> universe ?
	if you were your character from a bible , it will be why ?
	if you were a character from a movie , who would it be and why ?

Table J.4. Interpolations between two input sentences generated by AAE and our model on the Yahoo dataset.