A Supplemental Material

A.1 Derivation

$$\mathbb{E}_{\mathbf{x}}[\mathrm{KL}(q(\mathbf{z}|\mathbf{x})|p(\mathbf{z}))] = \quad (1)$$
$$\mathbb{E}_{q(\mathbf{z}|\mathbf{x})p(\mathbf{x})}[\log(q(\mathbf{z}|\mathbf{x})) - \log(p(\mathbf{z}))]$$
$$= -H(\mathbf{z}|\mathbf{x}) - \mathbb{E}_{q(\mathbf{z})}[\log(p(\mathbf{z}))] \quad (2)$$
$$= -H(\mathbf{z}|\mathbf{x}) + H(\mathbf{z}) + \mathrm{KL}(q(\mathbf{z})\|p(\mathbf{z})) \quad (3)$$
$$= I(\mathbf{z}, \mathbf{x}) + \mathrm{KL}(q(\mathbf{z})\|p(\mathbf{z})) \quad (4)$$

where $q(\mathbf{z}) = \mathbb{E}_{\mathbf{x}}[q(\mathbf{z}|\mathbf{x})]$ and $I(\mathbf{z}, \mathbf{x}) = H(\mathbf{z}) - H(\mathbf{z}|\mathbf{x})$ is mutual information between $\mathbf{z}$ and $\mathbf{x}$ by definition.

A.2 Interpolating Latent Space

Bowman et al., (2015) have shown that one can transform between two sentences by interpolating in the latent space of continuous VAEs. We found that our DI-VAE enjoys the same property. Specifically, two sentences $\mathbf{x}_1$ and $\mathbf{x}_2$ are sampled and their latent code are $q_{\mathcal{R}}(\mathbf{z}_1|\mathbf{x}_1)$ and $q_{\mathcal{R}}(\mathbf{z}_2|\mathbf{x}_2)$. We can then interpolate by flipping each latent code from $\mathbf{z}_1^m$ to $\mathbf{z}_2^m$, $m \in [1, M]$. For models with M latent variables, one sentence can transform to another one in at most M steps. Table 1 shows an example.

So you can keep record of all the checks you write.
So you can get all kinds of information and credit cards.
So you can keep track of all the credit cards.
So you kind of look at the credit union.
So you know of all the credit cards.
Yeah because you know of all the credit cards.
Right you know at least a lot of times.
Right you know a lot of times.

Table 1: Interpolating from the source sentence (top) to a target sentence (bottom) by sequentially setting the source latent code to the target code.

A.3 Data Details

The details of the three dialog datasets are shown in Table 2.

	SMD	DD	SW
Type	Task	Chat	Chat
Vocab Size	1,835	17,705	24,503
# of Dialogs	3,031	13,118	2,400
Avg Dialog Len	6.36	9.84	59.2
Avg Utterance Len	12.1	16.3	22.1

Table 2: Statistics of the dialog datasets. Vocabulary in DD and SW are capped to the most frequent 10K word types.

A.4 Training Details

All RNNs use GRUs (Chung et al., 2014). The GRU-RNNs for DI-VAE and DI-VST have hidden size 512. The utterance encoder in LAED has hidden size 256 for one direction and the context encoder and response decoder have hidden size 512. The word embedding is shared everywhere with embedding size 200. The temperature of Gumbel Softmax is set to 1. We train with Adams (Kingma and Ba, 2014) with initial learning rate 0.001.

References

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. 2015. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349* .

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* .

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