Abstract

Large-scale auto-regressive models have achieved great success in dialogue response generation, with the help of Transformer layers. However, these models do not learn a representative latent space of the sentence distribution, making it hard to control the generation. Recent works have tried to learn sentence representations using Transformer-based framework, but do not model the context-response relationship embedded in the dialogue datasets. In this work, we aim to construct a robust sentence representation learning model, that is specifically designed for dialogue response generation, with Transformer-based encoder-decoder structure. An utterance-level contrastive learning is proposed, encoding predictive information in each context representation for its corresponding response. Extensive experiments are conducted to verify the robustness of the proposed representation learning mechanism. By using both reference-based and reference-free evaluation metrics, we provide detailed analysis on the generated sentences, demonstrating the effectiveness of our proposed model.

1 Introduction

Large-scale pretrained language models with Transformer-based architecture (Radford et al., 2019; Raffel et al., 2019; Zhang et al., 2019) have achieved state-of-the-art performance across a variety of natural language processing (NLP) tasks, including open-domain dialogue response generation. The majority of these models are auto-regressive (AR) language models, creating a matching mechanism between the input context and the generated response. However, such an approach often ignores the semantic meaning shared among sentences from an abstract level (Li et al., 2020), generating responses that have matching components with the context, but are not logically sound. Figure 1 demonstrates this potential problem by using some generated samples, where the response generation model generates similar patterns to the context but fail to predict semantically reasonable responses to continue the conversation.

In comparison, variational autoencoders (VAEs) (Kingma and Welling, 2013; Zhao et al., 2017; Gu et al., 2018) are able to explicitly model sentences in a representation space, providing control on the generated responses. While most VAE-based language models adopt shallow network such as LSTMs, it is straightforward to construct such a latent representation learning with a Transformer framework. Recent works (Li et al., 2020) aim to build a pretrained language model that uses a sentence-level VAE objective with a BERT (Devlin et al., 2018) encoder and a GPT-2 (Radford et al., 2019) decoder. Nevertheless, being designed for multiple language understanding and generation tasks, they do not utilize the enriched context-response relationship in the dialogue response generation task.

Figure 1: Faulty generated sentences based on given context, via AR-based models. The provided dialogue is based on speakers ‘A’ and ‘B’. Three model generated samples are given. Each pair of matching words/components are marked in the same color. The Ground Truth Response is provided for comparison.
relationship between context and response latent representations and hence constrains the response generation quality.}

In this work, we aim to design a dialogue response understanding and generation framework, that combines a higher-level sentence representation learning with a Transformer-based encoder-decoder architecture, yielding controllable response generation from an abstract level, while achieving top-tier generation quality. We adopt the approach that uses a BERT encoder and a GPT-2 decoder as the model backbone, and enrich the underlying information in the learned representation by performing utterance-level contrastive learning. Specifically, the latent representation of the context input (one or multiple utterances) is used to predict that of the response input (a single utterance). Note that our proposed contrastive loss is built based upon utterances, while the entire context input is encoded by a single context encoder.

To benefit the contrastive learning of the latent representation, we propose to use the hard negative sampling mechanism. Such an approach has been used in the computer vision domain (Robinson et al., 2020), guiding a learning method to correct its mistakes more quickly. In our case, we select the negative samples by using a pretrained context-response matching model (Cai et al., 2020). Given a context input, the responses with the top matching scores would be considered as the negative samples, and used for the contrastive objective.

In terms of model architecture, we select conditional variational autoencoder (CVAE) (Zhao et al., 2017). This provides an encoder-decoder structure, in which the BERT-GPT backbone can be applied.

2 Preliminary

2.1 Contrastive Predictive Coding

The key idea of contrastive predictive coding (CPC) (Oord et al., 2018) is to learn representations that encode the underlying shared information between different parts of the high-dimensional signal, with the help of next step prediction. Given a sequence of inputs \( \{x_1, x_2, \ldots, x_t, \ldots, x_T\} \) (e.g., speech signal input), we obtain the latent representations \( h_t = g_{\text{enc}}(x_t) \) for each \( x_t \). A context representation \( z_t = g_{\text{AR}}(h_{\leq t}) \) summarizes all \( \{h_1, h_2, \ldots, h_t\} \) using an auto-regressive model \( g_{\text{AR}} \). To predict the future observations \( x_{t+k} \), the following positive real score is considered.

\[
f_k(x_{t+k}, z_t) = \exp(h_{t+k}^T W_k z_t),
\]

where \( W_k \) is linear transformation for step \( k \). To train the predictive model, a negative contrastive estimation (NCE) objective is used.

\[
L_{\text{CPC}} = -\mathbb{E}_X[\log \frac{f_k(x_{t+k}, z_t)}{\sum_{x_j \in X} f_k(x_j, z_t)}],
\]

where \( X = \{x_1, \ldots, x_N\} \) is a set of \( N \) random samples containing the positive sample \( x_{t+k} \) and \( N-1 \) negative samples from some distribution \( p(x_{t+k}) \). The negative samples aim to provide guiding on learning predictive information in the latent representations of the context inputs.

2.2 Conditional VAE

The conditional VAE (CVAE) (Zhao et al., 2017) provides a way to utilize the context feature as a conditional input to the VAE framework. Originally, VAE contains an encoder \( q_\phi(z|x) \) and a decoder \( p_\theta(x|z) \), where \( x \) is the input data and \( z \) is the latent representation. It assumes a Gaussian prior distribution \( p(z) \), and by using the following objective, it approximates the expected log-likelihood.

\[
L_{\text{CVAE}}(x; \theta, \phi) = \mathbb{E}_{z \sim q_\phi(z|x)} [\log p_\theta(x|z)] - \mathbb{D}_{KL}(q_\phi(z|x) \parallel p(z)),
\]

where the conditional likelihood term (first) and the KL term (second) characterize reconstruction and generalization capabilities, respectively. In practice, reparameterization trick is used, in order to achieve Gaussian posterior distribution and compute the KL term in (3). Specifically, the encoder yields mean \( \mu \) and standard deviation \( \sigma \), and we can sample from such Gaussian distribution:

\[
z = \mu + \sigma \circ \epsilon,
\]

where \( \epsilon \sim \mathcal{N}(0, I) \). The CVAE, in comparison, has an additional condition input \( c \). In the dialogue response problem setup, the condition \( c \) corresponds to the context input, while \( x \) is the response input. Both the encoder and the decoder take an extra input \( c \), denoted as \( q_\phi(z|x, c) \) and \( p_\theta(x|z, c) \), respectively. For the prior, a prior network \( p_\psi(z|c) \) is required, resulting in the updated objective as follow.

\[
L_{\text{CVAE}}(x; \theta, \phi, \psi) = \mathbb{E}_{z \sim q_\phi(z|x,c)} [\log p_\theta(x|z, c)] - \mathbb{D}_{KL}(q_\phi(z|x, c) \parallel p_\psi(z|c)).
\]

During inference, we sample \( z \sim p_\psi(z|c) \) and generate using the decoder \( p_\theta(x|z, c) \), while the
3 Method

We include detailed model design in the following section. By introducing the utterance-level contrastive learning, our proposed model learns sentence representations that are capable of predicting future utterances. CVAE is considered as our main model architecture, where we use a BERT encoder and a GPT-2 decoder.

3.1 Representation Learning with a Transformer-Based Framework

We first introduce our model backbone, which is basically a CVAE model with a BERT encoder and a GPT-2 decoder. When encoding, we consider $h_{[CLS]}$, the output of the special token [CLS] of the last-layer, as the sentence-level representation, and obtain the latent representation by a linear mapping $e(\cdot)$ parameterized as $z = W_\epsilon h_{[CLS]}$. In the CVAE framework, since we have the posterior and prior networks, we denote the former as the utterance encoder $f_u(\cdot)$, and define the latter as a combination of two encoders $f_c(f_u(\cdot))$, both of which are BERT encoders with outputs $h_x$ and $h_c$, respectively. Correspondingly, we have $e_r(\cdot)$ and $e_c(\cdot)$ as recognition encoder and prior encoder (i.e., linear mappings), that take $(h_x, h_c)$ and $h_c$ as input, respectively. The outputs of the two encoders are $z_{post} \sim q_\theta(z|x, c)$ and $z_{prior} \sim p_\psi(z|c)$, respectively.

When decoding, we follow the Memory approach in Optimus (Li et al., 2020). Specifically, we first compute $h_M = W_M z$, where $W_M \in \mathbb{R}^{LH}$ is a weight matrix. The output $h_M \in \mathbb{R}^{LH}$ is separated into $L$ vectors of length $H$, each of which is attended by one layer of the GPT-2 decoder $g(\cdot)$. Figure 2 shows the CVAE framework for both training and inference. Note that the outputs of the recognition encoder and the prior encoder are $\mu$ and $\sigma$, as the mean and standard deviation of the learnt $q_\theta(z|x, c)$ and $p_\psi(z|c)$, respectively. As discussed in Sec 2.2, such a reparameterization trick is often used in VAE framework, where $z = \mu + \sigma \odot \epsilon$ for some $\epsilon \sim \mathcal{N}(0, I)$.

3.2 Utterance Level Contrastive Learning

With latent representation learnt, it is straightforward to adapt the contrastive learning in our model backbone. Different from speech signal data on which the original CPC objective is applied, dialogue datasets contain multi-turn utterances between two speakers for each conversation input. In this case, we consider encoding $c = \{u_1, u_2, \ldots, u_{k-1}\}$ as a concatenation of all utterances in the context history. The output $h_c$ is then sent to the latent encoder $e(\cdot)$, with an output $z = e(h_c)$ that is used to predict the sentence representation $h_x$ of the next time step utterance $x = u_k$. 
We further improve our contrastive learning (CL) word permutation. The traditional CL usually uniformly selects negative or positive. When training our proposed model, we defined the final objective of our proposed model,

\[
\mathcal{L}_{\text{CL}} = -\mathbb{E}_{x_0 \sim p(x)} \log \frac{\exp(h_x W_k z_c)}{\sum_{x_j \in X} \exp(h_j W_k z_c)},
\]

where \(W_k\) is the weight matrix. Note that the proposed contrastive learning procedure can be directly applied to the CVAE model, in which \(z_c\) is basically the representation learnt by the prior encoder \(e_c(\cdot)\). By combining the CVAE objective and the contrastive loss, we defined the final objective of the proposed model,

\[
\mathcal{L} = \mathcal{L}_{\text{CVAE}} - \lambda \mathbb{E}_{x_0 \sim p(x)} \log \frac{\exp(h_x W_k z)}{\sum_{x_j \in X} \exp(h_j W_k z)}
\]

where \(\lambda\) is the hyper-parameter.

### 3.3 Improved CL with Hard Negative Sampling

We further improve our contrastive learning (CL) procedure by introducing hard negative sampling. The traditional CL usually uniformly selects negative samples from the data distribution. In the dialogue setup, this would be sampling random response samples \(x_j\) for a given context input \(c\). However, it is known in the computer vision domain that using negative samples, which look similar to the positive sample, would help correct the mistakes made by the encoder and learn useful information much more quickly (Robinson et al., 2020). For image datasets, one may consider adding noisy pixels, or rotating the positive images to obtain the negative ones. Similar approaches can be done in the natural language domain by using word permutation.

In our case, the multi-turn dialogue data setup allows us to further utilize the context-response relationship, and conduct hard negative sampling by using context-response matching models. Following (Cai et al., 2020), we consider training a Multi-hop Selector Network (MSN) (Yuan et al., 2019) which provides matching scores between the context and response inputs. Specifically, we construct a dialogue dataset, in which each context input \(c\) is paired with one positive response sample \(x\), and multiple randomly sample distractor response samples \(x_j\). The model is trained to predict whether a given context-response pair is positive or negative. When training our proposed CVAE model, the hard negative samples are selected based on higher matching scores, provided by the MSN model.

### 4 Related Work

**Transformer-based representation learning for text generation.** Transformer layers have been applied in various language models and achieve engaged in such a framework, it aims to generate novel ideas and great results (Devlin et al., 2018; Yang et al., 2019; Keskar et al., 2019). Many are pre-trained models and can be fine-tuned to adapt to various downstream tasks (Le and Mikolov, 2014; Dai and Le, 2015; Kiros et al., 2015). Recent works try to learn an interpretable representation space for the sentence encoding, using Transformer-based encoders and decoders. OPTIMUS (Li et al., 2020) provides a pretrained language model with a BERT encoder and a GPT-2 decoder. With variational inference sentences by learning a universal latent space through the pretraining. PLATO (Bao et al., 2019) also adopts BERT framework. It is specifically designed for dialogue generation while learning a discrete latent variable for action recognition, upon which the generation is based.

**Contrastive learning in NLP.** Contrastive learning has been widely used with applications on various computer vision (Chopra et al., 2005; Schroff et al., 2015) and NLP problems (Mikolov et al., 2013; Logeswaran and Lee, 2018). It aims to learn a latent representation by contrasting positive and negative pairs in different setups. In the NLP domain, contrastive learning is applied in Word2Vec (Mikolov et al., 2013), a word embedding model, in which neighbouring words are predicted from context. Other than that, it is also used for sentence representation, where the positive and negative pairs are sampled as two contiguous sentences and sentences from other documents, respectively (Logeswaran and Lee, 2018). CLAPS (Lee et al., 2020) is proposed recently for conditional text generation, sampling positive and negative sentences by adding perturbations to the input sequence. Being seq2seq model, CLAPS focuses on machine translation and text summarization, while our model tackles dialogue response generation, contrasting each utterance against other matching responses from the dialogue dataset.

### 5 Experiments

In this section, we present the experiment results of the proposed model, with comparison to sev-
<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>BOW</th>
<th>Intra</th>
<th>Inter</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>P</td>
<td>F1</td>
<td>A</td>
</tr>
<tr>
<td>GPT-2</td>
<td>0.467</td>
<td>0.14</td>
<td>0.215</td>
<td>0.891</td>
</tr>
<tr>
<td>PLATO</td>
<td>0.331</td>
<td>0.174</td>
<td>0.228</td>
<td>0.929</td>
</tr>
<tr>
<td>iVAE</td>
<td>0.425</td>
<td>0.248</td>
<td>0.313</td>
<td>0.923</td>
</tr>
<tr>
<td>CVAE</td>
<td>0.489</td>
<td>0.226</td>
<td>0.309</td>
<td>0.928</td>
</tr>
<tr>
<td>CVAE + CL</td>
<td><strong>0.498</strong></td>
<td>0.239</td>
<td><strong>0.323</strong></td>
<td>0.937</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>BOW</th>
<th>Intra</th>
<th>Inter</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>P</td>
<td>F1</td>
<td>A</td>
</tr>
<tr>
<td>GPT-2</td>
<td>0.372</td>
<td>0.203</td>
<td>0.263</td>
<td>0.924</td>
</tr>
<tr>
<td>PLATO</td>
<td>0.382</td>
<td>0.237</td>
<td>0.293</td>
<td>0.934</td>
</tr>
<tr>
<td>GLC</td>
<td>0.433</td>
<td>0.243</td>
<td>0.311</td>
<td>0.928</td>
</tr>
<tr>
<td>CVAE</td>
<td>0.415</td>
<td>0.256</td>
<td>0.317</td>
<td>0.949</td>
</tr>
<tr>
<td>CVAE + CL</td>
<td>0.419</td>
<td>0.267</td>
<td><strong>0.326</strong></td>
<td>0.95</td>
</tr>
</tbody>
</table>

Table 2: Reference-based Results on PersonaChat.

eral baseline models across different dialogue response datasets. Detailed experimental setups are included, with analysis on both reference-based and reference-free evaluation.

5.1 Implementation Details
We implement our proposed model based on the CVAE model backbone. As discussed in Section 3.1, we apply the contrastive learning procedure (Section 3.2) on the CVAE framework, in which the context latent representation $z_c$ is used to predict the response hidden state $h_x$, with negative samples selected by the pretrained MSN model. Our model uses BERT encoder (12 transformer layers with hidden size 768 and self-attention heads 12), and GPT-2 decoder (12 transformer layers with hidden size 768 and self-attention heads 12).

5.2 Datasets
We mainly evaluate our model on two benchmark datasets, namely the DailyDialog dataset (Li et al., 2017) and the PersonaChat dataset (Zhang et al., 2018). The former contains 13K daily conversations for an English learner, with an average number of turns in a conversation as 8.5. The latter is a knowledge grounded conversation dataset where two participants chat naturally and try to get to know each other. A total of 11K dialogues are contained in PersonaChat with an average number of turns per conversation as 15. For both datasets, we process each utterance as the response of the previous context utterances from both speakers. Both datasets are separated into train, validation, and test subsets, with a separation ratio of 10:1:1 for DailyDialog, and 9:1:1 for PersonaChat.

5.3 Baseline Models
We select State-Of-The-Art (SOTA) models from different model categories as our baseline models. For auto-regressive model, we consider DialogGPT (Zhang et al., 2019), which is a GPT-2 based model that is specifically designed for dialogue response generation. For dialogue response generation model that uses contrastive learning, we include group-wise contrastive learning (GCL) (Cai et al., 2020), which conduct CL between target dialogue model and a pretrained reference model. PLATO (Bao et al., 2019) is another model that uses transformer-based model architecture while including a discrete latent variable to tackle the one-to-many mapping problem. Last but not the least, we compare our model to the SOTA VAE model on text generation, the implicit VAE (iVAE) (Fang et al., 2019), in which the original Gaussian assumption is replaced by an implicit distribution, learnt in an adversarial fashion.

5.4 Evaluation Metrics
In our work, we consider the traditional reference-based, as well as several newly proposed reference-
Table 3: Reference-free Results on Dailydialog (H vs R: Human vs Random; H vs M: Human vs Machine). Higher scores indicate better generation quality in the corresponding categories.

<table>
<thead>
<tr>
<th></th>
<th>DialogRPT</th>
<th>RoBERTa Eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ground Truth</td>
<td></td>
<td>2.812</td>
</tr>
<tr>
<td>GPT-2</td>
<td>0.311</td>
<td>2.739</td>
</tr>
<tr>
<td>PLATO</td>
<td>0.314</td>
<td>2.132</td>
</tr>
<tr>
<td>CVAE</td>
<td>0.392</td>
<td>1.931</td>
</tr>
<tr>
<td>CVAE + CL</td>
<td><strong>0.407</strong></td>
<td><strong>2.106</strong></td>
</tr>
</tbody>
</table>

Table 4: Reference-Free Results on Personachat.

<table>
<thead>
<tr>
<th></th>
<th>DialogRPT</th>
<th>RoBERTa Eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT-2</td>
<td>0.339</td>
<td>1.383</td>
</tr>
<tr>
<td>PLATO</td>
<td>0.358</td>
<td>1.582</td>
</tr>
<tr>
<td>GLC</td>
<td>0.361</td>
<td>1.607</td>
</tr>
<tr>
<td>CVAE</td>
<td>0.367</td>
<td>1.678</td>
</tr>
<tr>
<td>CVAE + CL</td>
<td><strong>0.372</strong></td>
<td><strong>1.702</strong></td>
</tr>
</tbody>
</table>

free evaluation metrics. The former basically measures the relevance between the reference response data (ground truth) and the generated response samples. In comparison, the reference-free evaluators do not need the reference responses, while computing the score based on a trained evaluation model.

We mainly consider three reference-based evaluation metrics: 1) BLEU score measures how many n-gram ($n = 4$ in our experiments) from generated response overlaps with the references. We sample 10 responses; BLEU-precision and BLEU-recall are defined as average and maximum scores. 2) Cosine similarity of bag-of-words (BOW) embedding between the generated response and the reference is considered (e.g., 3 types of embedding including greedy, average, and extreme). 3) Distinct evaluates the diversity of the generation; dist-$n$ is the ratio of unique $n$-grams ($n = 1, 2$) over all $n$-grams in the generated responses. We evaluate both within each sampled response and among all responses as intra-dist and inter-dist, respectively.

For reference-free evaluators, we consider the following two metrics. The first one is DialogRPT (Gao et al., 2020), which provides ranking scores of responses based on human evaluation on $i$) Width: the number of dialogues after the current turn ($c, x$); $ii$) Depth: the maximum length of the dialogue after the current turn; and $iii$) Up-down: the number of up votes minus the number of down votes (by human raters). Another one is the RoBERTa Evaluator (Zhao et al., 2020), which is first trained on next-sentence-prediction task in an unsupervised manner, and then trained on predicting human annotated data in a semi-supervised setup. The model is trained using the RoBERTa encoder (Liu et al., 2019), on the concatenation of the input context $c$ and the generated response $\tilde{x}$, with output rescaled to the human rating scale.

5.5 Results

Table 1 and 2 show the reference-based results on DailyDialog and PersonaChat datasets, respectively. In terms of BLEU score, our proposed CVAE/CVAE+CL achieves better performance than the other Transformer-based models. By adding the proposed utterance level contrastive learning (CL) to CVAE, the BLEU score is improved from 0.309 to 0.318, outperforming GPT-2 and PLATO, while being comparable to iVAE. Meanwhile, CVAE-based models provide much higher generation diversity in both inter and intra-dist. Note that adding contrastive learning does slightly drop the generation diversity of CVAE, which is likely due to the additional predictive information learnt in the representation of context inputs. Overall, our proposed model CVAE+CL achieves top-tier performance in the majority of the evaluation categories, across both datasets.

Table 3 and 4 present the reference-free results on DailyDialog and PersonaChat, respectively. For
Table 5: Ablation Study on DailyDialog via Reference-based Evaluation Metrics.

<table>
<thead>
<tr>
<th>Models</th>
<th>Recall</th>
<th>Precision</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed Model</td>
<td>0.498</td>
<td>0.239</td>
<td>0.323</td>
</tr>
<tr>
<td>- HNS</td>
<td>0.49</td>
<td>0.235</td>
<td>0.317</td>
</tr>
<tr>
<td>- CL</td>
<td>0.489</td>
<td>0.226</td>
<td>0.309</td>
</tr>
<tr>
<td>- TF</td>
<td>0.265</td>
<td>0.222</td>
<td>0.242</td>
</tr>
</tbody>
</table>

Table 6: Sensitiveness of Contrastive Learning Objective against Hyper-parameter $\lambda$. Results on DailyDialog via Reference-based Evaluation Metrics.

<table>
<thead>
<tr>
<th>Objective</th>
<th>$\lambda$</th>
<th>Recall</th>
<th>Precision</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{L}<em>\text{CVAE} + \lambda \mathcal{L}</em>\text{CL}$</td>
<td>0.0</td>
<td>0.489</td>
<td>0.226</td>
<td>0.309</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.489</td>
<td>0.232</td>
<td>0.315</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0.493</td>
<td>0.232</td>
<td>0.316</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.498</td>
<td>0.239</td>
<td>0.323</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>0.496</td>
<td>0.24</td>
<td>0.323</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>0.492</td>
<td>0.235</td>
<td>0.318</td>
</tr>
</tbody>
</table>

DailyDialog, CVAE provides better overall DialogRPT scores than the other baseline models. However, since responses are generated based on the context latent representations (instead of direct sentence inputs), CVAE does not provide matching words across context and response, reducing the RoBERTa score for the DailyDialog dataset. For PersonaChat, CVAE outperforms the other baseline models in both evaluation metrics. Overall, we see the effectiveness of adding CL to the CVAE framework, improving both DialogRPT and RoBERTa scores across the two datasets.

Ablation study is conducted based on each components in the proposed CVAE. Table 5 shows the BLEU score (i.e., recall, precision, and F1) of our model on the DailyDialog dataset, in which we gradually remove one component at a time to demonstrate the effectiveness from each of them. We first remove hard negative sampling (HNS) and use random negative samples (from the same training mini batch) instead. The contrastive learning (CL) objective is then removed, where we notice a drop of 0.014 (from 0.323 to 0.309) in F1. Finally, the model performance drops drastically when we replace the transformer layers (TF) in the CVAE with LSTM (i.e., original CVAE framework); F1 decreases from 0.309 to 0.242.

We further show the sensitiveness of our proposed utterance-based contrastive loss against different values of hyper-parameter $\lambda$. By tuning $\lambda$ from 0.0 to 2.0, we report the BLEU scores of each training objectives. As shown in Table 6, the model with $\lambda = 0.5$ and $\lambda = 1.0$ achieves the best overall performance, while the models with the other values of $\lambda$ also provide competitive results. This demonstrates that our model is relatively insensitive to the hyper-parameter for contrastive learning objective. Meanwhile, we can see that adding the contrastive loss is improving the performance (regardless of different values of $\lambda$) when comparing the baseline CVAE (i.e., $\lambda = 0.0$).

Figure 4 provides examples of hard negative sentences selected by MSN. Each example includes context sentence(s), a ground truth response sentence (reference), as well as 3 selected hard negative samples. As can be seen, the negative sentences share relevant words and contents with the input context, which provides a more efficient training for our model.

### 6 Conclusion
In this work, we propose a dialogue response generation model that combines the controllability of representation learning models and SOTA performance of AR-based frameworks. An efficient utterance level contrastive learning is introduced to enrich the information learned in the latent representations (i.e., predictive information of utterance in the next time step). Our proposed model achieves competitive performance comparing to the SOTA models in different benchmark datasets.
References


Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In International conference on machine learning, pages 1188–1196. PMLR.


