Insufficient Data Can Also Rock! Learning to Converse Using Smaller Data with Augmentation

Juntao Li,[1,2,*] **Lisong Qiu,**[1,2,*] **Bo Tang,**[3] **Dongmin Chen,**[1] **Dongyan Zhao,**[1,2] **Rui Yan**[1,2†]
[1]Center for Data Science, Academy for
Advanced Interdisciplinary Studies, Peking University, Beijing, China
[2]Institute of Computer Science and Technology, Peking University, Beijing, China
[3]Department of Computer Science and Engineering, Southern University of Science and Technology
{lijuntao,qiuls,dongminchen,zhaody,ruiyan}@pku.edu.cn
tangb3@sustc.edu.cn

Abstract

Recent successes of open-domain dialogue generation mainly rely on the advances of deep neural networks. The effectiveness of deep neural network models depends on the amount of training data. As it is laboursome and expensive to acquire a huge amount of data in most scenarios, how to effectively utilize existing data is the crux of this issue. In this paper, we use data augmentation techniques to improve the performance of neural dialogue models on the condition of insufficient data. Specifically, we propose a novel generative model to augment existing data, where the conditional variational autoencoder (CVAE) is employed as the generator to output more training data with diversified expressions. To improve the correlation of each augmented training pair, we design a discriminator with adversarial training to supervise the augmentation process. Moreover, we thoroughly investigate various data augmentation schemes for neural dialogue system with generative models, both GAN and CVAE. Experimental results on two open corpora, Weibo and Twitter, demonstrate the superiority of our proposed data augmentation model.

Introduction

Open-domain dialogue generation is becoming a research hotspot in the community of natural language processing due to their penitential applications. Along with the flourishing development of neural networks, plenty systems have been proposed to improve the quality of generated dialogues from many aspects such as diversity (Li et al. 2015; Zhao, Zhao, and Eskenazi 2017), topic (Xing et al. 2017), persona modeling (Zhang et al. 2018) and emotion controlling (Zhou et al. 2017). Generally, in the paradigm of deep neural networks, a large quantity of training data is required for facilitating the convergence of these models. For instance, Li et al. (2016) collected over 24 millions of query-response training pairs for building personalized dialogue generation systems. Hence, the crucial point of building dialogue systems depends on whether exists proper dataset with sufficient training pairs. In other words, current dialogue generation models mainly focused on using effective model frameworks, e.g. CVAE (Shen et al. 2017), and model training, e.g. adversarial and reinforcement learning (Li et al. 2017) to achieve improvement on an existing dataset.

As progress on dialogue generation is restricted by the size of training datasets, we investigate to move the frontier of dialogue generation forward from a different angle, i.e. training dialogue system on smaller data through using data augmentation techniques. To augment existing dialogue training data, there are two main issues to be addressed. The first one is how to get sufficient alternative expressions of the raw query response pairs since small data will result in the over-fitting problem and make conversation system vulnerable to unseen data (Hou et al. 2018). In real-world chit-chats, there are many proper responses for one specific query and there are also alternative responses with different syntactical structure but same semantic information. Another challenge is how to control the quality of augmented query response training pairs. It is expected that the data augmentation method can generate more valid training data that can supplement the expressions of original data and further improve the performance of trained dialogue systems. To obtain more valid training data and limit negative noise, an effective data augmentation model is required to generate each query-response pair with high relevance, and meanwhile reduce repetitions between augmented query-response pairs.

While important, there are only a few studies presented for addressing the aforementioned challenges in dialogue data augmentation. To improve the performance of task-oriented dialogue system, Kurata et al. (2016) introduced an auto-encoder with random noise to produce more different utterances. Beyond that, another data augmentation approach for dialogue language understanding was presented through combing sequence-to-sequence and diversity rank to generate more diverse utterances (Hou et al. 2018). Although these methods have achieved performance improvement on specific tasks, there is still noticeable gap between these models and the aforementioned requirements. For one thing, they suffer from data sparsity as there are a few proper transformations only considering the lexical-level alternative expressions. For another, integrating random noise to generate more data without supervision and feedback about the feasibility of the augmented utterances is risky.

To address the aforementioned challenges, we propose a novel data augmentation model for open-domain dialogue generation. Specifically, we use a conditional variational au-

[*] Equal contribution

[†] Corresponding author: Rui Yan (ruiyan@pku.edu.cn)

toencoder (CVAE) model for generating sufficient alternative expressions with diversified words by introducing a latent variable. To improve the relevance of each augmented query-response pair, we combine the CVAE model with a discriminator. The whole model is trained in an adversarial fashion to feed loss information from the discriminator to train the CVAE part, which allows generating more valid training pairs. Moreover, we introduce a distillation strategy to filter repetition query-response pairs in augmented data.

For evaluating the performance of our proposed data augmentation model, we conduct experiments on two open datasets, Weibo and Twitter. Automatic metrics and human evaluations indicate that our model can generate more valid training pairs with diversified expressions and good relevance within each augmented pair. Furthermore, we utilize sequence to sequence (Sutskever, Vinyals, and Le 2014; Klein et al. 2017) as the conversation generation model to evaluate whether the augmented dataset can achieve improvement over the original one. Both quantitative and qualitative analysis confirm that after augmentation, the sequence-to-sequence model can generate better responses.

Problem Formulation

In this paper, our goal is to enrich training data under low-source condition. To formulate this task, the dialogue generation and data augmentation processes are described with necessary notations, show as follows.

Following previous work (Li et al. 2015), open-domain dialogue generation involves automatically generating a response $R = (r_0, \ldots, r_j, \ldots, r_m)$ for a user-issued query $Q = (q_0, \ldots, q_k, \ldots, q_{m'})$, where r_j refers to the embedding representation of i-th word in a response and q_k denotes the k-th word's embedding of query. m and $m^{'}$ are the length of a response and a query, respectively. The entire dialogue system is trained under D, i.e. maximizing the $P(R_i \,|\, Q_i)$, where $D = \{(Q_i, R_i)\}_{i=0}^{N}$ is the dataset and N refers to the number of training query-reply pairs. For the data augmentation task, the original dataset D is increased to $D^{'} = \{(Q_i, R_i)\}_{i=0}^{nN}$, where n is the magnification of augmentation. Correspondingly, the response generation changes from $argmax_R P(R \,|\, Q, D)$ to $argmax_R P(R \,|\, Q, D^{'})$.

Recall that dialogues are diversified, e.g. one specific query can match many proper responses and vice versa. Inspired by this, we design three different schemes for enhancing D to $D^{'}$, which is shown in Figure 1. The first one is to construct n queries conditioned on each single query-response pair, as a result of which each query refers to n proper responses, named one-to-many (1-n). Similarly, we refer the second one to generating n queries based on each single query-response pair, defined as many-to-one (n-1). Another one is named many-to-many (n-n) which represents mapping each query-response pair to n pairs, where the generation process is accomplished in an paired fashion, i.e. both queries and replies are different with the original one.

The Approach

In this section, we will elaborate the details of data augmentation model, augmenting process and dialogue modeling.

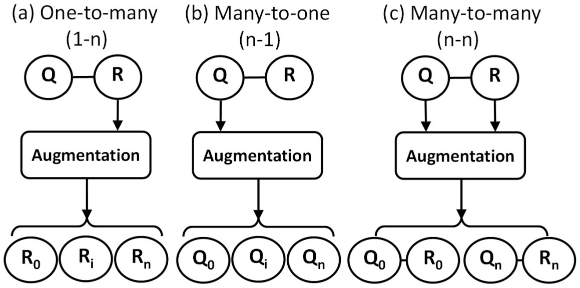

Figure 1: Three different paradigms of data augmentation for open-domain dialogue generation.

As demonstrated in Figure 2, our proposed data augmentation model comprises two parts: CVAE and a discriminator.

CVAE

To generate diversified query-response pairs, we employ CVAE as the core of our data augmentation model. On the whole, the CVAE part consists of an encoder and a decoder. The encoder is a bidirectional RNN (Schuster and Paliwal 1997) with gated recurrent units (GRU) (Sutskever, Vinyals, and Le 2014), where input x is mapped to a latent variable z under given condition c, i.e. $(x, c) \mapsto z$. In details, the encoder computes a posterior distribution $q_\theta(z \,|\, x, c)$ given the input x and condition c. Note that input x can be either query or response depending on which augmentation scheme mentioned above is used, so does c. For instance, x refers to response while c refers to query for the one-to-many setting. Hereinafter, we take the one-to-many setting as example to describe the augmentation process of our proposed model while other settings will be elaborated in the following subsection. During encoding, a query and a response are represented by the concatenation of forward and backward vectors, i.e. $h_q = [\overrightarrow{h}_q, \overleftarrow{h}_q]$ and $h_r = [\overrightarrow{h}_r, \overleftarrow{h}_r]$.

The decoder is a one-layer RNN with GRU cell, which takes $[z, c]$ as input to construct the input x, i.e. $(z, c) \mapsto x$. Specifically, the decoding process is formulated as computing $p_\theta(x \,|\, z, c)$, where z follows a prior distribution $p_\theta(z \,|\, c)$, e.g, a standard Gaussian distribution. θ denotes the parameters of both encoder and decoder. Since the integral of the marginal likelihood $p_\theta(x \,|\, c)$ is intractable for large datasets (Kingma and Welling 2014), the true posterior is replaced by its variational approximation $q_\phi(z \,|\, x, c)$. ϕ refers to parameters of q.

For training the CVAE part, the objective is to maximize the log-likelihood $p_\theta(x \,|\, c)$ over input x conditioned on c. Through pushing up the variational lower bound of the objective, written by

$$\begin{aligned}\mathbb{L}(\theta, \phi; x, c) = &- \mathrm{KL}(q_\phi(z|x, c) \parallel p_\theta(z|c)) \\ &+ \mathbf{E}_{q_\phi(z|x,c)}[\log p_\theta(x|z, c)]\end{aligned} \tag{1}$$

the whole model is optimized, where KL($\cdot$) is the KL-divergence term which is used as the regularization for encouraging the approximated posterior $q_\phi(z|x, c)$ to approach the prior $p_\theta(z \,|\, c)$. E[$\cdot$] is the reconstruction term to evaluate how well the decoding process goes conditioned on

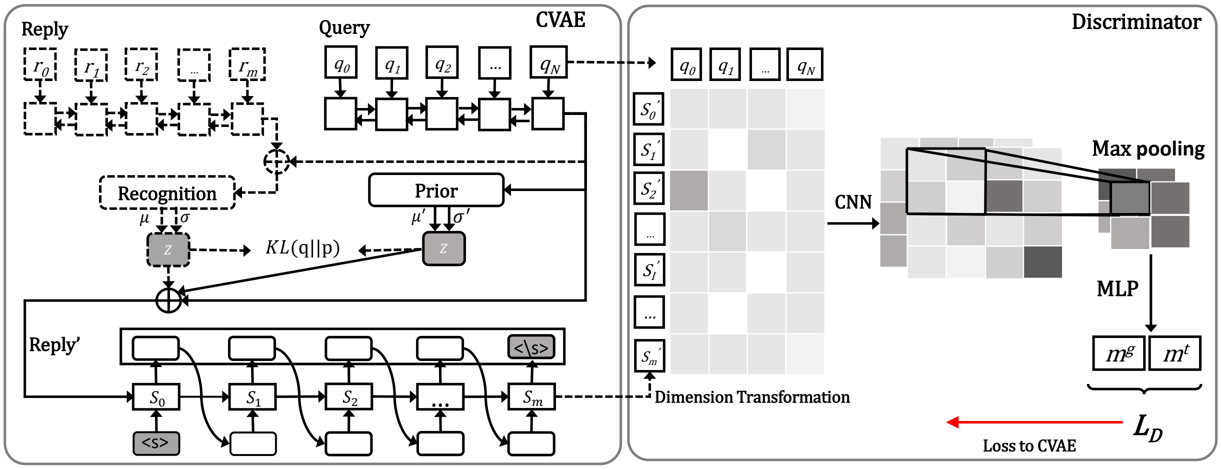

Figure 2: The architecture of our proposed data augmentation model, where the left is the CVAE and the right depicts the discriminator. The operations labeled by red color refers to the loss signal from the discriminator. Notice that the discriminator and objects with dashed lines are only used during training while other parts are used during both training and testing process.

the approximated posterior $q_\phi(z|x,c)$. Following previous work, (Kingma and Welling 2014; Zhao, Zhao, and Eskenazi 2017), we hypothesize the approximated posterior follows a multivariate Gaussian $\mathcal{N}$, i.e $q_\phi(z|x,c) = \mathcal{N}(\mu, \sigma^2 \boldsymbol{I})$, where μ and σ^2 represents the mean and variance of $\mathcal{N}$ and I has a diagonal structure. Thus modeling the apprximiated posterior is converted to learn μ and σ, computed by

$$\begin{bmatrix}\mu \\ \log\left(\sigma^2\right)\end{bmatrix} = W_q \begin{bmatrix}x \\ c\end{bmatrix} + b_q \tag{2}$$

where W_q and b_q are trainable parameters. In the same vein, the prior $p_\theta(z \,|\, c)$ is another multivariate Gaussian $\mathcal{N}(\mu', \sigma'^2 \boldsymbol{I})$. The key parameters μ' and σ'^2 are learned by a single-layer fully-connected network (MLP) with the $tanh(\cdot)$ activation function, formulated as:

$$\begin{bmatrix}\mu' \\ \log\left(\sigma'^2\right)\end{bmatrix} = \mathrm{MLP}_p\left(c\right) \tag{3}$$

Both the encoder and decoder are used during the training process while only part with solid lines in Figure 2 are used during prediction. It is worth noticing that the hidden states of the decoder, $(s_0, \ldots, s_i, \ldots, s_m)$, and the word embedding representation of a query, $(q_0, \ldots, q_k, \ldots, q_{m'})$, will be passed to the discriminator as inputs.

The Discriminator

The discriminator in our model is to evaluate whether the generated responses are feasible for a given query. The loss signal from the discriminator is back-propagated to the CVAE part for enhancing its training process. In this study, we utilize the implementation that contains two steps. We first calculate the interaction (or matching) matrix between the generated response R' and the query Q, where R' refers to the constructed result of the original R. We then use a convolutional neural network (CNN) to extract features from the interaction matrix and output a score for calibrating the matching degree between R' and the given query. As the training is conducted in an adversarial fashion, we treat R' as the negative instance while R as positive one.

Specifically, R' is represented by $(s_0, \ldots, s_i, \ldots, s_m)$ and Q is $(q_0, \ldots, q_k, \ldots, q_{m'})$, which are from the CVAE part. A dimension transformation is first conducted on R' to align it with Q, formulated as:

$$s_i' = ReLU(W_d s_i + b_d) \tag{4}$$

where $ReLU$ is activation function and W_d and b_d are trainable parameters. In this way, the dimension of s_i' is same as word embedding. With generated response $R_d' = (s_0', \ldots, s_i', \ldots, s_m')$, the interaction matrix between the reconstructed response and query Q is formulated as below:

$$M^g = R_d' \cdot Q \tag{5}$$

where $M^g \in \mathbb{R}^{(m+1)\times(m'+1)}$; " $\cdot$ " refers to the matrix multiplication operation.

In the second step, a feature matrix is learned by $\mathcal{F} = CNN(M^g)$, which corresponds to extracting features from the interaction matrix using a CNN network. Once this is done, a max-overtime pooling strategy (Collobert et al. 2011) is used for filtering features in $\mathcal{F}$. Then, a one-layer fully-connected neural network with a $sigmoid$ activation function is utilized to flatten the feature matrix, resulting in the final matching score $m^g \in (0,1)$. Apart from m^g, the matching score m^t between the positive instance R and query Q is also calculated as the above-mentioned procedure, except the dimension transformation operation. Note that R is represented by $(r_0, \ldots, r_j, \ldots, r_m)$, where r_j has the same dimension with word embedding.

In the paradigm of generative adversarial networks (GAN) (Goodfellow et al. 2014), the training objective of the discriminator is to maximize the matching score of positive instances and meanwhile minimize the matching score of negative ones, formulated as:

$$\mathcal{L}_D = \log(m^t) + \log(1 - m^g) \tag{6}$$

Finally, the overall objective of our proposed data augmentation model can be formulated as the following one:

$$\mathcal{L}_{\text{CVAE-D}} = \mathcal{L}_{CVAE} - \lambda \mathcal{L}_D \tag{7}$$

which is maximized for updating the parameters of the CVAE, where $\mathcal{L}_{CVAE}$ is the objective of CVAE, i.e. $\mathcal{L}_{CVAE} = \mathbb{L}(\theta, \phi; x, c)$. For training the discriminator, this objective is minimized such that $\mathcal{L}_D$ is maximized, which corresponds to encourage better distinction between positive and negative instances. λ is a balancing parameter. The CVAE and the discriminator is trained alternatively in a two-step adversarial fashion (Zhang, Barzilay, and Jaakkola 2017). This training process is repeated until the whole objective $\mathcal{L}_{\text{CVAE-D}}$ is converged.

Data Augmentation

Recall that we investigate to utilize three different augmentation schemes for generating query-response pairs, i.e., one-to-many, many-to-one, many-to-many. For the one-to-many (1-n) scheme, we are targeting at generating n corresponding responses according to each query-response pair, where n is the magaification. In this setting, the aforementioned condition c of CVAE refers to the query and x refers to the response. In the test, the trained CVAE-GAN model generates n responses conditioned on c and x, i.e. the original query-response pair. As to many-to-one (n-1), the augmentation process is conduced in a similar fashion, where the reply is presented as c while the query refers to x. The setting of many-to-many (n-n) refers to iteratively using the trained models in one-to-many and many-to-one settings, i.e., generating a response conditioned a query and then outputting a query according to the generated response at each iteration. After n iterations, a query-response pair is enhanced to n pairs. As there could exist alike or irrelevant instances, we propose a distillation strategy to filter the augmented results. Specifically, we use the Jaccard distance to depict the word-level semantic similarity between utterances, where highly similar utterances will be removed.

Dialogue Generation

As a most popular dialogue generation model, the sequence to sequence model with attention (S2S) is adopted as the benchmark to verify the efficiency of our proposed approach. In S2S, given a query $Q = (q_1, ..., q_k, ..., q_{m'})$ and a reply $R = (r_0, \ldots, r_j, \ldots, r_m)$, Q is encoded by applying $h_t^e = RNN_{enc}(x_t|h_{t-1}^e)$ and the final hidden state h_k is fed into the decoder RNN_{dec} as the initial state h_0^d. At each timestep of decoding, the hidden state of decoder is computed as $s_i = RNN_{dec}(r_{i-1}, s_{i-1}, c_i)$, where c_i is calculated through the attention mechanism (Luong, Pham, and Manning 2015) and the probability distribution of candidate words is computed through softmax. The objective function is to minimize the negative log probability:

$$\mathcal{L}_{\text{S2S}} = \sum_{j=1}^{N} -logP(R_j|Q_j) \tag{8}$$

where $P(R_j|Q_j)$ is calculated by the chain rule.

Rela.	Does the query-response pair correlate well?
Divt.	Does the utterance narrate with diversified words?
Red.	Is the utterance grammatically formed?
Ovr.	The average score of the above three criteria.

Table 1: Criteria of human evaluation.

Experiment Setup

Datasets

To evaluate the effect of data augmentation, we conduct experiments on two open dialogue corpora in different languages, the Weibo (Wang et al. 2013) and Twitter (Ritter, Cherry, and Dolan 2010). Concretely, the Weibo dataset consists of short-text online chit-chat dialogues in Chinese, which is crawled from Sina Weibo [1]. The Twitter dataset is in English, which is collected from the microblogging service, Twitter [2]. Totally, there are 0.6 million query-response pairs in the Weibo corpus and 1.3 million pairs in the Twitter dataset. In our experiments, we randomly extract [20k, 50k, 100k, 200k, 300k, 500k] utterance pairs from both datasets for training to simulate the scenario of data augmentation. In addition to the training sets, we also collect 10k and 20k pairs for validation and testing, respectively. For preprocessing, we follow the conventional settings (Ritter, Cherry, and Dolan 2010; Wang et al. 2013).

Baselines

Our model is compared with several highly related and strong baselines, including:

Sequence to sequence (S2S), the vanilla RNN-based sequence-to-sequence dialogue system (Klein et al. 2017).

Noising Autoencoder (NAE), the conventional RNNs encoder-decoder with random perturbations added to the encoded vectors (Kurata, Xiang, and Zhou 2016).

DAGAN, a basic implementation of DAGAN (Antoniou, Storkey, and Edwards 2017) for text generation task, which is the combination of our introduced discriminator and the NVE model.

CVAE, the conventional CVAE model (Zhao, Zhao, and Eskenazi 2017; Shen et al. 2017), which is used for investigating the performance of CVAE for data augmentation.

Model Settings

All models are trained with the following hyper-parameters: both encoder and decoder are set to one layer with GRU cells, where the hidden state size of GRU is 500; the utterance length is limited to 50, and the vocabulary size is

[1] https://www.weibo.com/

[2] https://twitter.com/

40,000; word embedding size is 500; all trainable parameters are initialized from a uniform distribution [-0.08, 0.08]; we employ the Adam (Kingma and Ba 2014) for optimization with a mini-batch size 64; the gradient clipping strategy is utilized to avoid gradient explosion, where the gradient clipping value is set to be 5. We stop training the dialogue model S2S if the perplexity keeps increasing in two successive epochs. During decoding, we use the beam search strategy, with the beam size set to 5. In addition to the shared hyper-parameters, we have the following settings for CVAE, DAGAN, and CVAE-GAN. The dimension of the latent variable z in CVAE and CVAE-GAN is set to 300 and the layer size of MLP_p is 400. For the discriminator in DAGAN and CVAE-GAN, the kernel size of CNN is (5, 5) with the stride size k set to 2. Following the conventional setting (Hu et al. 2017), we set the balancing parameter λ to 0.1. For data augmentation, the magnification n is 10. For each original query-response pair, its corresponding augmented utterances that have Jaccard distances larger than 0.8 with others are filtered during the distillation process, where the threshold value is set empirically.

Evaluation Metrics

To comprehensively evaluate the quality of augmented query-response pairs and their influence to dialogue generation system, we utilize the following metrics:

BLEU: In dialogue generation, BLEU is widely used in previous studies (Zhao, Zhao, and Eskenazi 2017; Lei et al. 2018). We follow their settings in this paper.

Distinctness: To distinguish safe and common responses, the distinctness score (Li et al. 2015) is designed to measure word diversity by counting the distinctive [1,4]-grams with the final distinctness values normalized to [0, 100].

Human Evaluation: For assessing the quality of augmented query-response pairs, we utilize the criteria in previous work (Tao et al. 2018; Shang et al. 2018), i.e. relevance (Rela.), diversity (Divt.), Readability (Red.). Details of the criteria are given in Table 1, where each criterion is scored from $\{1, 2, 3\}$, denoting bad, normal, good, respectively. To evaluate the performance of dialogue generation model, we conduct human preference judges (Fan, Lewis, and Dauphin 2018) on generated utterances, i.e. choosing the better one from two competing utterances. For both settings, we randomly sample 200 generated utterances for each model, where each utterance is ranked by 5 well-educated annotators. The whole evaluation is conducted in a blind fashion.

Results and Analysis

Recall that we conduct experiments on different raw data sizes to investigate the effects of data augmentation on dialogue generation. Table 2 presents the results of human judges and Table 3 illustrates the results of automatic evaluation. We analyze these results from the following aspects.

Ablation Study

This section is to investigate how each part of the data augmentation model affects the generated query-response pairs and the resulting performance of dialogue generation. There are two groups of ablation observations, i.e. the effects of the CVAE and the influence of the introduced discriminator.

Datasets	**Models**	**Rela.**	**Divt.**	**Red.**	**Ovr.**
Weibo	**NAE**	1.87	2.02	2.22	2.04
	DAGAN	1.96	1.97	2.27	2.07
	CVAE	2.20	**2.41**	2.35	2.32
	CVAE-GAN	**2.38**	2.35	**2.45**	**2.39**
Twitter	**NAE**	1.67	1.89	1.98	1.85
	DAGAN	1.79	1.92	2.04	1.92
	CVAE	1.84	2.21	2.10	2.05
	CVAE-GAN	**2.04**	**2.23**	**2.22**	**2.16**

Table 2: Results of human evaluations on augmented query-response pairs (p<0.01), where **Rela.**, **Divt.**, **Red.**, **Ovr.** represent relevance, diversity, readability, overall, respectively.

To investigate whether using a latent variable and variational inference can improve the diversity of augmented query-response pairs, we summarize two groups comparison of human evaluation results in Table 2, i.e, NAE $v.s.$ CVAE, DAGAN $v.s.$ CVAE-GAN. The results show that CVAE outperforms all baselines in terms of diversity score on two datasets, which means the CVAE part can generate query-response pairs with diversified expressions. With diversified expressions in the augmented data, CVAE also achieves a better user experience, which is confirmed by the overall score in human evaluation. These results confirm that CVAE is effective for supplementing the original training data with proper query-response pairs in diversified words. Though effective for improving the diversity, the CVAE model yields substandard correlations between augmented queries and their corresponded responses. As the variational latent variable introduces randomness into the augmentation process, query-response pairs from CVAE model have weaker correlation than those from CVAE-GAN, which is proven by the relevance score in Table 2.

To study the influence of augmented training data from CVAE, we analyze the results of dialogue generation in Table 3. It can be seen that the dialogue model enhanced with augmented data from CVAE outputs responses with better distinct scores than other baseline settings, which confirms that dataset augmented by CVAE can substantially improve the diversity of dialogue generation. We also observe that training dialogue systems with the augmented data from CVAE improves the performance in terms of BLEU scores.

The discriminator is introduced for supervising data augmentation process so as to generate query-response pairs correlated well with each other. We conduct two groups of comparisons to study the effectiveness of the discriminator, i.e., DAGAN $v.s.$ NAE, CVAE $v.s.$ CVAE-GAN. We observe that the discriminator is effective for improving the correlations between generated query-response pairs. As presented by the relevance score in Table 2, the CVAE and NAE model can generate correlated query-response pairs with the enhancement of the discriminator. Meanwhile, the introduced discriminator does not result in inferior performance in readability and diversity.

As for its influence to the performance of S2S dialogue generation model, the results show that the introduced dis-

Metrics	S2S	NAE			DAGAN			CVAE			CVAE-GAN		
		1-n	**n-1**	**n-n**	**1-n**	**n-1**	**n-n**	**1-n**	**n-1**	**n-n**	**1-n**	**n-1**	**n-n**
BLEU-1	9.42	9.79	9.78	9.54	10.1	9.46	9.15	10.1	9.19	9.22	**10.9**	9.18	9.71
BLEU-2	1.44	1.28	1.46	1.40	1.30	1.47	1.37	1.51	1.50	1.55	1.62	1.60	**1.72**
BLEU-3	0.58	0.47	0.74	0.69	0.47	0.73	0.71	0.62	0.79	0.79	0.57	0.85	**0.95**
BLEU-4	0.34	0.29	0.58	0.52	0.28	0.58	0.54	0.39	0.64	0.60	0.32	0.68	**0.76**
Dist-1	5.00	4.94	5.50	7.49	4.16	7.34	6.77	8.29	8.41	8.20	**8.87**	8.68	8.62
Dist-2	25.9	24.1	24.2	31.1	20.2	32.4	31.0	36.6	35.5	34.8	35.8	**37.5**	36.2
Dist-3	51.0	51.6	47.7	56.3	43.7	57.5	56.4	**66.5**	61.3	60.9	63.7	63.0	63.2
Dist-4	69.7	73.9	66.9	73.9	65.0	74.7	73.4	**84.4**	77.5	77.5	82.0	68.8	80.0

Table 3: Results of automatic evaluation for dialogue systems with distillation under the setting of 100k raw Weibo training data. **BLEU-n** refers to BLEU scores on n-gram; **Dist-n** denotes the distinctness of n-gram, with $n = 1$ to 4; **1-n**, **n-1**, **n-n** represent augmenting responses (one-to-many), queries (many-to-one), query-response pairs (many-to-many), respectively.

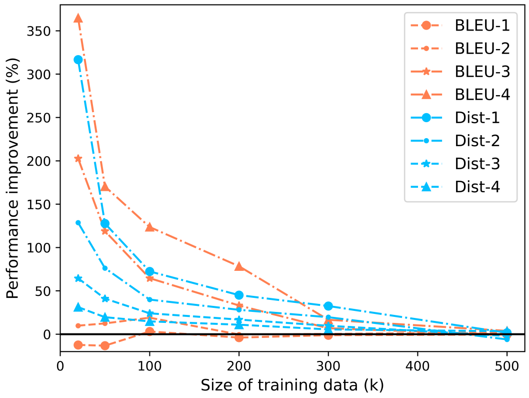

Figure 3: The performance improvements with different raw training data sizes, where augmentation is done by the CVAE-GAN model under the many-to-many (n-n) setting.

criminator boosts the dialogue generation model with a substantial improvement in terms of BLEU scores, which represents more overlappings with the expected conversations. Moreover, the discriminator does not introduce negative effects to the S2S dialogue generation model, which is confirmed in Table 3 that CVAE-GAN and DAGAN achieves comparable results from other aspects, e.g. diversity.

For our CVAE-GAN model, it outperforms all baselines in human evaluations for dialogue data augmentation. As illustrated in Table 2, training pairs augmented from CVAE-GAN are better than other baselines from three aspects, i.e. relevance, diversity, and readability. We also observe that the CVAE-GAN model further improves the quality of dialogue generation model trained on the augmented data pairs, which is confirmed by the significant improvement of BLEU and distinct scores. These results indicate that the CVAE-GAN data augmentation model is effective for enhancing the S2S dialogue generation model, especially improving diversity.

The Effect of Data Augmentation

As illustrated in Table 3, each data augmentation models have three different schemes for enhancing the raw data, i.e. one-to-many (1-n), many-to-one (n-1), and many-to-many (n-n). Through analyzing these results, we have the following three main observations: under the setting of 1-n, the augmented data set achieves the best results of diversity while obtaining relatively low BLEU scores; for the n-1 setting, the augmented data set yields higher BLEU scores and shows competitive diversity; as for n-n setting, it achieves a balance between the above-mentioned two settings. To show the superiority of our proposed data augmentation scheme, we also conduct human preference judges on the results of S2S dialogue system. Table 4 shows the results of training the S2S model with augmented data from CVAE-GAN under the many-to-many setting (n-n). The results indicate that the CVAE-GAN data augmentation model substantially improves user experience of the final trained dialogue system.

Model	S2S	S2S+DA
Weibo	38.4%	61.6%
Twitter	43.2%	56.8%

Table 4: Results of human perference judges on the dialogue systems, where **DA** refers to data augmentation.

In addition to different data augmentation schemes, we also conduct experiments to investigate whether the distillation strategy can boost the performance of dialogue generation. The results show that the distillation strategy can significantly increase the BLEU scores while sometimes decreasing the performance of diversity. Moreover, we conduct data augmentation for datasets with different scales. These results are sketched in Figure 3. It can be easily seen that with data augmentation, the quality of generated conversations in terms of diversity and BLEU-3, 4 increases significantly while the improvement drops as the size of raw training data grows. We attribute this phenomenon to that when raw data is sufficient to train the dialogue model the gains from augmented data are not able to overpass the noise incurred by augmentation at the same time, which may result in the less improvement or even deterioration.

Scheme	**Query**	Have a look at these various mistakes.
	Response	The one in the kitchen is so funny, hahaha.
1-n	**Response1**	How can the last expression be so funny!
	Response2	The last one is really funny. So cute!
	Response3	Sure enough, it's hilariously funny, hahaha.
	Response4	Ah, impressive! This is so funny.
n-1	**Query1**	What do you think of it?
	Query2	Have a look at these brilliant replies.
	Query3	Find the highlight, especially in those comments.
	Query4	What's going on? Can you share it?
n-n	**Query1**	This is the right way to be funny.
	Response1	o o hahaha, it's so cute. We laugh our heads off.
	Query2	There are various funny mistakes. Have a look!
	Response2	Wow, the kid is so funny, hahaha.
	Query3	Take a look at these various mistakes as you wish.
	Response3	LoL, it's awesome, hahaha.
	Query4	It cracks me up.
	Response4	This is really cute, hahaha.

Figure 4: Example pairs augmented by CVAE-GAN model.

Case Study

In addition to quantitative results, we also launch case study to illustrate the superiority of our proposed data augmentation approach. Figure 4 presents some examples produced by the CVAE-GAN model on the Weibo corpus where the Chinese utterances are translated into English. We can observe that in all three settings, our data augmentation model can generate query-pairs correlated well with the original ones while limiting repetitions. Figure 5 gives some cases which illustrate the influence of augmented data to dialogue generation, where case 1, 2 are obtained from the dialogue model trained on Weibo dataset while the others refer to the Twitter corpus. These results confirm that dialogue system trained with augmented data can produce utterances with diverse expressions that are proper and feasiable.

Related Work

Deep Generative Models

This work can be seen as the extension of deep generative models (Bengio et al. 2014) in natural language generation. Conventionally, generative models, inluding VAEs (Kingma and Welling 2014) and GANs (Goodfellow et al. 2014), are mainly used for image generation (Sohn, Yan, and Lee 2015; Yan et al. 2016). In the field of natural language processing (NLP), owing to the discrete nature of text (Yu et al. 2017), previous success for text generation with generative models concentrates on using VAE (Bowman et al. 2016) and CVAE (Serban et al. 2017; Shen et al. 2017; Zhao, Zhao, and Eskenazi 2017). As the GAN framework facilitates training the generator, Larsen et.al. (2016) propose to integrate VAE and GAN, where VAE is treated as the generator. Hu et.al. (2017) first combine VAE and GAN for text generation. To the best of our knowledge, we are the first to combine CVAE-GAN (Li et al. 2018) for data augmentation.

Data Augmentation

In the paradigm of deep learning, data augmentation is an effective way to boost the performance of neural models. Previous success of data augmentation is mainly observed in computer vision while there are only a few works designed for tasks in natural language processing. A denosiong autoencoder is utilized for generating more utterances through introducing noise to the decoding process (Kurata, Xiang, and Zhou 2016). Besides, Hou et.al. (2018) combined sequence to sequence model with a diversity rank to produce alternatives of utterances for language understanding.

Case 1	**Query**	The new photography of the twenty-four solar terms is so overwhelming.
	Response	I love this so much that I can't be thin to settle.
	S2S	I am grieved to watch the game.
	S2S+DA	I like it so much, and I've already collected it.
Case 2	**Query**	In Tibet, you can see such starry sky as you look up. Do you like it?
	Response	I wanna to go to Tibet.
	S2S	I just want to ask where you are.
	S2S+DA	I wanna go to the place of Tibet.
Case 3	**Query**	It was nice. I had my nieces 2nd birthday get together and it was nice hanging with children.
	Response	Awwww, was she all happy ?
	S2S	It was nice. I thought it was fun today.
	S2S+DA	Glad she had a wonderful birthday !

Figure 5: Case study of utterances generated by dialogue systems. **DA** denotes augmentation with the CVAE-GAN model under the many-to-many (n-n) scheme.

Conclusion

In this paper, we propose an effective model that combines CVAE and GAN for augmenting dialogue data. Concretely, we utilized CVAE to improve the diversity of augmented query-response pairs. A discriminator is used with adversarial training for enhancing the relevance of generated training pairs. Moreover, we designed three data augmentation schemes for query-response pair generation, i.e. one-to-many, many-to-one and many-to-many. Experimental results on two open corpora, Weibo and Twitter, indicate that through combing CVAE with the discriminator, notable improvement has been achieved on the quality of augmented training data over those generated by existing models. With the augmented training data, the dialogue generation model also gains a substantial performance improvement.

Acknowledgments

We would like to thank the anonymous reviewers for their constructive comments. This work was supported by the National Key Research and Development Program of China (No. 2017YFC0804001), the National Science Foundation of China (NSFC No. 61672058; NSFC No. 61876196). Rui Yan was sponsored by CCF-Tencent Open Research Fund, Alibaba Innovative Research (AIR) fund, and Microsoft Research Asia (MSRA) Collaborative Research Program. Bo Tang was supported by Guangdong Natural Science Foundation (Grant No. 2018A030310129) and the National Science Foundation of China (NSFC No. 61802163).

References

Antoniou, A.; Storkey, A.; and Edwards, H. 2017. Data Augmentation Generative Adversarial Networks. *arXiv preprint arXiv:1711.04340.*

Bengio, Y.; Thibodeau-Laufer, E.; Alain, G.; and Yosinski, J. 2014. Deep Generative Stochastic Networks Trainable by Backprop. In *ICML*, 226–234.

Bowman, S. R.; Vilnis, L.; Vinyals, O.; Dai, A.; Jozefowicz, R.; and Bengio, S. 2016. Generating Sentences from a Continuous Space. In *SIGNLL*, 10–21.

Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; and Kuksa, P. 2011. Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research* 12(Aug):2493–2537.

Fan, A.; Lewis, M.; and Dauphin, Y. 2018. Hierarchical Neural Story Generation. *arXiv preprint arXiv:1805.04833*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In *NIPS*, 2672–2680.

Hou, Y.; Liu, Y.; Che, W.; and Liu, T. 2018. Sequence-to-Sequence Data Augmentation for Dialogue Language Understanding. *arXiv preprint arXiv:1807.01554*.

Hu, Z.; Yang, Z.; Liang, X.; Salakhutdinov, R.; and Xing, E. P. 2017. Toward controlled generation of text. *arXiv preprint arXiv:1703.00955*.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P., and Welling, M. 2014. Auto-Encoding Variational Bayes. *stat* 1050:10.

Klein, G.; Kim, Y.; Deng, Y.; Senellart, J.; and Rush, A. M. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*.

Kurata, G.; Xiang, B.; and Zhou, B. 2016. Labeled Data Generation with Encoder-Decoder LSTM for Semantic Slot Filling. *INTERSPEECH*.

Larsen, A. B. L.; Sønderby, S. K.; Larochelle, H.; and Winther, O. 2016. Autoencoding Beyond Pixels Using a Learned Similarity Metric. In *ICML*, 1558–1566.

Lei, W.; Jin, X.; Kan, M.-Y.; Ren, Z.; He, X.; and Yin, D. 2018. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In *ACL*, volume 1, 1437–1447.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Li, J.; Galley, M.; Brockett, C.; Spithourakis, G.; Gao, J.; and Dolan, B. 2016. A persona-based neural conversation model. In *ACL*, volume 1, 994–1003.

Li, J.; Monroe, W.; Shi, T.; Jean, S.; Ritter, A.; and Jurafsky, D. 2017. Adversarial learning for neural dialogue generation. In *EMNLP*, 2157–2169.

Li, J.; Song, Y.; Zhang, H.; Chen, D.; Shi, S.; Zhao, D.; and Yan, R. 2018. Generating classical chinese poems via conditional variational autoencoder and adversarial training. In *EMNLP*, 3890–3900.

Luong, M.-T.; Pham, H.; and Manning, C. D. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Ritter, A.; Cherry, C.; and Dolan, B. 2010. Unsupervised modeling of twitter conversations. In *NAACL*, 172–180.

Schuster, M., and Paliwal, K. K. 1997. Bidirectional Recurrent Neural Networks. *IEEE Transactions on Signal Processing* 45(11):2673–2681.

Serban, I. V.; Sordoni, A.; Lowe, R.; Charlin, L.; Pineau, J.; Courville, A. C.; and Bengio, Y. 2017. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. In *AAAI*, 3295–3301.

Shang, M.; Fu, Z.; Peng, N.; Feng, Y.; Zhao, D.; and Yan, R. 2018. Learning to converse with noisy data: Generation with calibration. In *IJCAI*, 4338–4344.

Shen, X.; Su, H.; Li, Y.; Li, W.; Niu, S.; Zhao, Y.; Aizawa, A.; and Long, G. 2017. A Conditional Variational Framework for Dialog Generation. In *ACL*, volume 2, 504–509.

Sohn, K.; Yan, X.; and Lee, H. 2015. Learning Structured Output Representation using Deep Conditional Generative Models. In *NIPS*, 3483–3491.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *NIPS*, 3104–3112.

Tao, C.; Gao, S.; Shang, M.; Wu, W.; Zhao, D.; and Yan, R. 2018. Get the point of my utterance! learning towards effective responses with multi-head attention mechanism. In *IJCAI*, 4418–4424.

Wang, H.; Lu, Z.; Li, H.; and Chen, E. 2013. A dataset for research on short-text conversations. In *EMNLP*, 935–945.

Xing, C.; Wu, W.; Wu, Y.; Liu, J.; Huang, Y.; Zhou, M.; and Ma, W.-Y. 2017. Topic Aware Neural Response Generation. In *AAAI*, volume 17, 3351–3357.

Yan, X.; Yang, J.; Sohn, K.; and Lee, H. 2016. Attribute2image: Conditional Image Generation from Visual Attributes. In *ECCV*, 776–791. Amsterdam, Netherlands: Springer.

Yu, L.; Zhang, W.; Wang, J.; and Yu, Y. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. *AAAI*.

Zhang, Y.; Barzilay, R.; and Jaakkola, T. 2017. Aspect-augmented Adversarial Networks for Domain Adaptation. *TACL* 5(1):515–528.

Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; and Weston, J. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too? *ACL*.

Zhao, T.; Zhao, R.; and Eskenazi, M. 2017. Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders. In *ACL*, volume 1, 654–664.

Zhou, H.; Huang, M.; Zhang, T.; Zhu, X.; and Liu, B. 2017. Emotional chatting machine: Emotional conversation generation with internal and external memory. *arXiv preprint arXiv:1704.01074*.