Enhancing the Open-Domain Dialogue Evaluation in Latent Space

Zhangming Chan^{1,3,*}, Lemao Liu², Juntao Li⁴, Haisong Zhang², Dongyan Zhao^{1,3}, Shuming Shi² and Rui Yan⁵
¹ Wangxuan Institute of Computer Technology, Peking University
² Tencent AI Lab ³ Center for Data Science, AAIS, Peking University
⁴ School of Computer Science and Technology, Soochow University
⁵ Gaoling School of Artificial Intelligence, Renmin University of China {zhangming.chan, zhaody}@pku.edu.cn ljt@suda.edu.cn, ruiyan@ruc.edu.cn {redmondliu, hansonzhang, shumingshi}@tencent.com

Abstract

The notorious one-to-many nature of opendomain dialogues poses huge challenges for automatic evaluation methods. Recent studies attempt to mitigate this issue by considering the similarity of the generated response with the conversational context and design discriminative models to learn from multiple positive responses. Despite the promising results, they can not be applied to general scenarios where training data with multiple responses is unavailable. To this end, in this paper, we propose a self-supervised setting to obtain a smooth latent space that can both capture discourse-level context information and implicitly model more references in latent space. Specifically, we present EMS, an Enhanced dialogue evaluation Metric in latent Space. Experimental results on two real-world dialogue datasets confirm the superiority of our method for open-domain dialogue evaluation, where both Pearson and Spearman correlations with human judgments outperform all baselines.

1 Introduction

With the surge of deep learning techniques, generation-based open-domain dialogue systems have witnessed significant improvement in recent years. Plenty of novel and effective models (Sutskever et al., 2014; Serban et al., 2016; Li et al., 2015; Serban et al., 2016; Zhao et al., 2017; Gu et al., 2018; Qiu et al., 2019; Chan et al., 2019b; Serban et al., 2017; Wolf et al., 2019; Hu et al., 2019; Chen et al., 2020) are proposed and have greatly promoted the development of the opendomain dialogue generation. Unlike the endless emergence of novel methods, however, there is still no meaningful and widely accepted automatic evaluation metric for dialogue generation yet. As we

know, automatic evaluation allows quick and effective comparison between different systems and is crucial for the development of natural language generation (NLG) tasks (Dathathri et al., 2019; Gu et al., 2019; Gao et al., 2019; Chan et al., 2019a, 2020). The lack of meaningful automatic evaluation metrics has become a significant impediment for open-domain dialog generation research.

Over the past decade, many automatic evaluation metrics are proposed to evaluate the opendomain dialogue systems. Among them, the word overlap-based automatic evaluation metrics from NLG tasks, such as BLEU (Papineni et al., 2002) in machine translation and ROUGE (Lin, 2004) in text summarization, are popular. In addition, Embedding Metrics (Mitchell and Lapata, 2008; Forgues et al., 2014; Rus and Lintean, 2012) have been utilized to evaluate the open-domain dialogue systems (Gu et al., 2018; Chan et al., 2019b; Shen et al., 2018). Recently, with the fantastic development of the large-scale pre-training model (Devlin et al., 2018; Liu et al., 2019; Radford et al., 2019), researchers proposed to enhance the embedding metrics by converting the dialogue sentences to hidden space via pre-training model (Zhang et al., 2019; Sellam et al., 2020; Zhao et al., 2019; Xiang et al., 2021). The common idea behind these metrics is that they measure the semantic similarity between a reference response and a generated response, independent on the conversational context. However, due to the notorious one-to-many nature (Li et al., 2015; Zhao et al., 2017; Qiu et al., 2019; Gu et al., 2018) of open-domain dialogue, a good response should be related well to its context yet may be largely different from a reference response in semantics.

Some other works (Tao et al., 2018; Ghazarian et al., 2019; Sinha et al., 2020) thereby proposed to build automatic dialogue evaluation metrics by considering the similarity of the generated

^{*} This work was done while Z. Chan was an intern at Tencent AI Lab. Corresponding Author: Rui Yan.

responses with the conversational context. Specifically, these works design discriminative models which can judge whether the generated responses match the conversational context well, which learn from {conversational context, response reference, *negative sample*} pairs in unsupervised learning manner. Zhao et al. (2020) further proposed to enhance such discriminative evaluation metrics by finetuning on a few human-annotated data to improve the robustness. These discriminative metrics trained using a single relevant response and multiple negative samples. However, Sai et al. (2020) argued that such discriminative metrics should be trained on multiple relevant responses (i.e., positive samples) and multiple negative samples, to favor the one-to-many nature in open-domain dialogues. Therefore, they collected a new dataset which contains multiple relevant and irrelevant responses for any given conversational context to train their discriminative evaluation model and the model trained by multiple relevant responses shows impressive performance. However, there are no organized relevant multiple responses in most existing datasets. Collecting a new dataset is expensive and timeconsuming. Thus, we aim to learn multiple reference information with limited data.

Inspired by the impressive effectiveness of the Variational Auto-encoder (VAEs) and Conditional Variational Auto-encoder (CVAEs) on the representation learning and dialogue modeling, we propose to learn the dialogue representations via VAEs/CVAEs for better evaluation. Equip with such dialogue representations, we obtain an Enhanced dialogue evaluation Metric in latent Space (EMS). EMS is a self-supervised evaluation metric with a two-stage training procedure. It represents dialogue sentences in a smooth latent space to both capture discourse-level context information and model more feasible latent references. Specifically, in the first stage, we build a VAE based model to map the dialogue sentences into a latent (or semantic) space. Li et al. (2019) showed that VAEs can be viewed as a regularized version of the auto-encoder and learn a smooth latent space through the regularization from the Gaussian prior. Then, we train our model by optimizing CVAEs' objective which forces the prior distribution to capture the feasible latent references information (details in Section 3.3). In the second stage, we combine the dialogue representations and the captured feasible latent reference information to

train a discriminative model. Meanwhile, we give a potential explanation of our motivation about why using feasible latent reference information can lead to a better evaluation (details in Section 3.1). Experimental results on two real-world dialogue datasets confirm the superiority of our method for opendomain dialogue evaluation, where both Pearson and Spearman correlations with human judgments outperform all baseline methods.

In a nutshell, our contributions can be summarized as follows:

• We proposed a novel automatic evaluation metric, i.e., EMS, for open-domain dialogue systems;

• We proposed a pre-training variational model to capture the feasible latent references;

• Experiments performed on two large datasets demonstrate the effectiveness of our proposed model and outperform all baseline methods.

2 Related Work

Word overlap-based Metrics. Several word overlap-based automatic evaluation metrics, such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and ROUGE (Lin, 2004), have been widely used to evaluate the quality of generated responses. These word overlap-based metrics measure how many words overlap in a given generated response when compared to a reference response. Liu et al. (2016); Lowe et al. (2017); Tao et al. (2018) argued that these word overlap-based metric scores are weakly correlated to human judgment due to ignoring the notorious one-to-many nature of the open-domain dialogues. Therefore, Yuma et al. (2020) proposed the improved BLEU, which compares the generated response with multiply diverse references.

Embedding-based Metrics. Unlike word overlap-based metrics comparing two raw sentences, Embedding Metrics (Mitchell and Lapata, 2008; Forgues et al., 2014; Rus and Lintean, 2012) map sentences to a high dimensional space, and calculate similarity based on the high-dimensional representations. Embedding Metrics are recently popular for evaluating the generation tasks, such as text summarization (Gao et al., 2020; Chen et al., 2021), question answer (Gao et al., 2019) and text generation (Hashimoto et al., 2019; Chan et al., 2020). Meanwhile, several works (Qiu et al., 2019; Chen et al., 2020; Gao et al., 2021) have shown their effectiveness in the open-domain dialogue systems. With the development of the largescale pre-training model (Devlin et al., 2018; Liu et al., 2019; Radford et al., 2019), some studies, e.g., BERTScore (Zhang et al., 2019) and Mover-Score (Zhao et al., 2019), further enhance the quality of representations via a large-scale pre-training model. However, these embedding-based metrics perform a better comparison compared to word overlap-based metrics but still ignore the one-to-many nature of open-domain dialogues.

Learning-based Metrics. Recent studies (Tao et al., 2018; Sinha et al., 2020) attempt to mitigate the one-to-many issue by considering the similarity of the generated response with the conversational contexts. The similarity is calculated by a designed discriminative model which learns to evaluate whether a response matches the conversational context well. The discriminative model is learned from tuples of data, {conversational con*text*, *response reference*, *negative sample*}, in an unsupervised learning manner. However, these learning-based metrics rely on a sophisticated sampling technique. Lan et al. (2020) proposed a sampling strategy to collect the valuable negative samples for the discriminative training. Bak and Oh (2020) conduct speaker sensitive response evaluation by conducting negative sampling from several levels. To further improve the robustness, Zhao et al. (2020) proposed to enhance the discriminative model by finetuning on a few human-annotated data. Sai et al. (2020) argued that these discriminative metrics should be trained on multiple relevant responses and multiple irrelevant samples for any given context. Therefore, they collected such a dataset and improved the evaluation performance greatly. However, collecting a new dataset is expensive and time-consuming. In this work, we propose a method to improve the effectiveness of the discriminative metrics based on the VAEs/CVAEs.

3 Methodology

In this paper, we propose an Enhanced dialogue evaluation Metric in latent Space (EMS), which contains two training stages (illustrated in Fig. 2). In this section, we first conduct some discussions about our motivations in Section 3.1. Then, we introduce the overall architecture in Section 3.2. The two training stages are described in Section 3.3 and Section 3.4, respectively. Finally, we describe the inference process in Section 3.5.

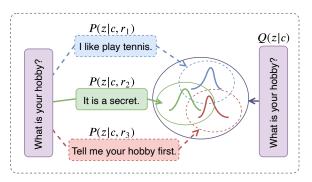


Figure 1: Distributions in latent space. Each circle represents a Gaussian distribution while three small circles refer to a special Gaussian distribution for each corresponding response. Naturally, the biggest circle indicates the prior Gaussian distribution. We use the prior distribution to approximate all the response conditional distribution. Dotted lines indicate the latent response.

3.1 Discussion about Motivation

We conduct a discussion about our motivation from information theory. Let r_i denotes a feasible response coming from $\{r_k\}_{k=1}^N$ which represents N feasible latent references. Assume a binary label $l \in \{0, 1\}$ indicates whether a response matches its context well. Existing works (Tao et al., 2018; Ghazarian et al., 2019; Sinha et al., 2020; Zhao et al., 2020) training with single relevant response are actually maximizing $I(l; c, r_i)$. Recently, Sai et al. (2020) proposed to training with multiple relevant responses, which actually maximizes $I(l; c, \{r_k\}_{k=1}^N)$. An intuitive explanation for the surprising improvement in Sai et al. (2020) is that $I(l; c, \{r_k\}_{k=1}^N) \ge I(l; c, r_i)^1$.

However, there are no organized relevant multiple responses in existing datasets and collecting a new dataset is expensive and time-consuming. Therefore, we aim to capture the feasible latent reference information with limited data. Inspired by previous works which model multiple responses for dialogue (Zhao et al., 2017; Qiu et al., 2019; Chan et al., 2019b), we utilize CVAEs (details in Section 3.2) which build a prior distribution P(z|c)to capture the feasible latent reference information in the latent space. Specifically, when training CVAEs, P(z|c) is forced to be close to the posterior distribution $Q(z|c, r_i)$ for any reference response r_i as illustrated in Fig. 1. In this sense, if z is sampled from P(z|c), z may contain some information of any r_i in some extent, and z can be used as a surrogate of $\{r_k\}_{k=1}^K$. Therefore, we can expect

¹There is a brief proof in Apendix A.

 $I(l; c, \{r_k\}_{k=1}^N) \ge I(l; c, r_i, z) \ge I(l; c, r_i).$

3.2 Overall Architecture

Previous works (Li et al., 2019; Gururangan et al., 2019; Li et al., 2020) concluded that VAEs can learn a smooth latent space through the regularization from the gaussian prior. Inspired by Li et al. (2020), we propose a novel architecture which can be regarded as a large-scale pretrained language model (PLM) based on VAEs/CVAEs.

Encoder. Li et al. (2019) argue that the VAEs might benefit from initialization with a noncollapsed encoder, because the encoder provides useful information from the beginning of training. We use the Masked PLMs (Devlin et al., 2018; Liu et al., 2019) as the text encoder because of their impressive effectiveness in natural language understanding tasks. We describe the encoding process as following,

$$h_q = \mathsf{PLM}([c; r]),$$

$$h_p = \mathsf{PLM}(c)$$
(1)

where c, r indicate conversational context and response reference, respectively.

Latent Variable Modeling. For modeling the latent variable, we hypothesize that the approximated variational prior and posterior follows an isotropic multivariate Gaussian distribution $\mathcal{N}(\mu, \sigma^2 I)$, where I represents the diagonal covariance. We use a recognition network $q_{\phi}(z|h_q)$ and a prior network $p_{\theta}(z|h_p)$ to approximate the posterior Q(z|c, r) and the prior P(z|c), respectively.

Decoder. The reconstruction process² forces the latent variable to contain the useful posterior information, which is a crucial step in the variational training. We use another PLM as the decoder to reconstruct the original input texts. For transporting the latent variable to the PLM decoder, we use the memory mechanism mentioned in Li et al. (2020) where the latent variable plays the role of an additional memory vector for the PLM decoder to attend. Specifically, the latent variable z is converted through a Multilayer Perceptron (MLP) and separated into several vectors, each of which is transported to the PLM decoder via attention mechanism.

3.3 Stage 1: Representation in Latent Space

Our first stage is to learn the latent representation of the dialogues and capturing the feasible latent reference information. Specifically, we first optimize our model via VAEs' objective to model a smooth latent space. Then, we train our model by CVAEs' objective to capture the feasible latent reference information. We describe the details as following.

A smooth latent space. Following Li et al. (2020), we first train the posterior module by optimizing the VAEs' objective. Li et al. (2019) showed that VAEs can be viewed as a regularized version of the autoencoder and can learn a smooth latent space. Based on this, we convert sentences in a universal smooth latent space. In a smooth latent space, latent representation of similar sentences should be close to each other and vice versa (Li et al., 2019). Therefore, it is a great outset for our model. To train this model, the log-likelihood objective is maximized through pushing up its variational lower bound:

$$-\mathbf{KL}(p_{\theta}(z|h_q)||q(z)) +\mathbf{E}_{p_{\theta}(z|h_q)}[\log p_{\theta}([c;r]|z,h_q)],$$
(2)

where $\mathbf{KL}(\cdot)$ represents the KL-divergence term, which serves as the regularization that encourages $p_{\theta}(z|h_q)$ to approach to q(z), i.e., a standard Gaussian distribution; $\mathbf{E}[\cdot]$ is the term of reconstruction loss, reflecting how well the decoder performs.

Feasible latent reference information. The oneto-many nature of the open-domain dialogues poses that there can be a lot of reasonable responses for the same conversational context. Therefore, we handle this one-to-many nature by CVAE as previous works (Zhao et al., 2017; Gu et al., 2018; Chan et al., 2019b) to capture the feasible latent reference information. As shown in Fig. 1, CVAE builds Gaussian posterior distributions for each feasible reference and forces the prior distribution to approach the posterior distributions. Ideally, a welllearned prior distribution will cover all the feasible latent reference information. We train our model by optimizing the following variational lower bound:

$$- \mathbf{KL}(q_{\phi}(z|h_{q})||p_{\theta}(z|h_{p})) + \mathbf{E}_{q_{\phi}(z|h_{q})}[\log p_{\theta}([c;r]|z,h_{q})]$$
(3)

where $\mathbf{KL}(\cdot)$ represents the KL-divergence term, which serves as the regularization that encourages the prior $p_{\phi}(z|h_p)$ to approach the approximated

²Note that the reconstruction process doesn't only indicate the autoregressive generation but also the masked causal generation.

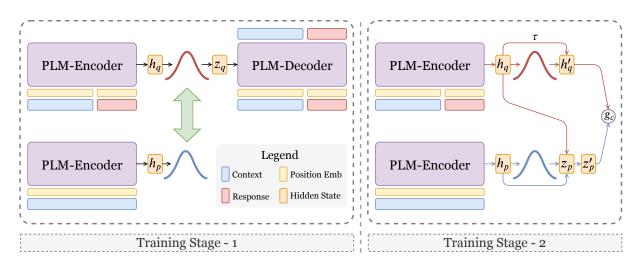


Figure 2: The detailed architecture of our proposed EMS and there are two stages in the training process. The first and second training stage is described in Section 3.3 and Section 3.4, respectively. In the second training stage, the response input is replaced by the generated responses which need be evaluated. The *Position Emb* indicates the Position Embedding.

posterior $q_{\theta}(z|h_q)$, i.e. a conditional Gaussian distribution.

Training procedures. Previous works (Bowman et al., 2015; Zhao et al., 2017) mentioned that VAE and CVAE training is challenging due to the KL vanishing issue, where the decoder ignores the conditional information and all the resulting posteriors almost collapse to a same Gaussian prior. To mitigate this issue, first, we initialize our model with Optimus (Li et al., 2020), a large-scale VAE-based PLM model, while optimizing Eq. 2.

To mitigate the same issue while optimizing Eq. 3, we use the cyclical KL annealing schedule (Fu et al., 2019). Specifically, we add a hyperparameter α to control the weight of the KLdivergence in Eq. 3. We set α close to zero in the first half of cyclic schedule, linearly anneal α to 1 in the next one-fourth of cyclic schedule and kept $\alpha = 1$ in the remaining cyclic schedule.

Moreover, the Free Bits (Bowman et al., 2015) is also crucial for the training. It replaces the KLdivergence in Eq. 3 by a hinge loss

$$\max(\gamma, \mathbf{KL}(q_{\phi}(z|h_q)||p_{\theta}(z|h_p)))$$
(4)

where γ is a hyperparameter which controls the information space for the each dimension of the latent variable. Finally, an extra bag-of-word loss (Zhao et al., 2017) is also used during the training.

3.4 Stage 2: Matching Training

In the second stage, we learn to judge the similarity between the conversational context and the response using the learned representations. Li et al. (2020) argue that the KL regularization applied on z has a large impact on the preceding layer feature, thus, the preceding layer feature also contains the information of z. Therefore, we consider combining h_q and z into the final representation,

$$h'_q = (1 - \tau) \cdot h_q + \tau \cdot z_q, \tag{5}$$

where τ is a hyperparameter and z_q indicates the latent representation from the posterior network. Meanwhile, we use the feasible latent reference information, captured by our prior network, to enhance the matching. We combine these two representations as following,

$$e = \sigma(\mathbf{W}_g \cdot [h_p; h_q; z_p] + \mathbf{b}_g)$$

$$z'_p = e \cdot z_p + (1 - e) \cdot h_p$$
(6)

where W_g and b_g are trainable parameters, and e is learned by the gate mechanism that controls the fusion of z_p and h_p . Note that z_p indicates the latent representation from the prior network. The activation function σ is sigmoid. Finally, we infer matching score between the conversational context and generated reference as follows,

$$g_c = \sigma(\mathbf{W}_s \cdot ([h'_q; z'_p]) + \mathbf{b}_s) \tag{7}$$

where W_s and b_s are trainable parameters and the activation function σ is sigmoid. Finally, we optimize our model with positive sampling and negative sampling (Lan et al., 2020) based on the discriminative training scheme.

3.5 Inference

In the inference process, we input the conversational context and response candidate as the c and r in Eq. 1, and conduct the operation as Eq. 5, Eq. 6 and Eq. 7 to obtain the score g_c . We use the g_c as matching degree of the response candidate.

4 Experiments

4.1 Dataset

To evaluate the effectiveness of our proposed automatic evaluation metric EMS, we conduct experiments on the following two open-access datasets.

Persona-Chat Dataset. The persona-chat dataset (Zhang et al., 2018) is a large persona-conditioned chit-chat style dialogue dataset which consists of 10,907 multi-turn dialogue sessions³.

DailyDialog Dataset. The dailydialog dataset (Li et al., 2017) is an another widely-used large collection of human-human dialogues which consists of 13,118 multi-turn dialogue sessions⁴.

Human-annotated Dataset. We collect the human-annotated datasets from Amazon Mechanical Turk and obtain two human-annotated datasets which consist of 750 context-response pairs in the persona-chat dataset and 800 ones in the dailydialog dataset, respectively. Following Zhao et al. (2020), the generated references come from several classical dialogue models, i,e., Seq2Seq (Sutskever et al., 2014), Seq2Seq with Attention (Serban et al., 2016), HRED (Serban et al., 2017), VHRED (Serban et al., 2016), GPT-2 (Wolf et al., 2019).

4.2 Baselines

We compare our proposed method with the following highly related and strong baselines.

BLEU. We utilize BLEU score (Papineni et al., 2002) to measure n-grams overlaps between response reference and generated response. Specifically, we follow the conventional setting in Sinha et al. (2020) and use the multi-bleu⁵.

ROUGE. ROUGE (Lin, 2004) is the most popular metric in the text summarization task. There, we report the f1 score of Rouge- $\{1,2,3,L\}$ and Rouge-L identifies the longest common subsequence between the generated and reference to better account

for sentence-level structure when computing word overlap.

METEOR. The METEOR (Banerjee and Lavie, 2005) is designed as an improvement on BLEU using a harmonic mean of precision and recall, as well as stemming and synonyms.

Embedding Metrics. Embedding Metrics compute the similarity between the embeddings representations of generated results and reference. The used embeddings come from glove⁶. In particular, we calculate three metrics: 1) *Average*, cosine similarity between the averaged word embeddings in the two sentences (Mitchell and Lapata, 2008); 2) *Extrema*, cosine similarity between the largest extreme values among the word embeddings in the two sentences (Forgues et al., 2014); 3) *Greedy*, i.e., greedily matching words in two sentences based on the cosine similarities, and the total scores are then averaged across all words (Rus and Lintean, 2012).

BERTScore. BERTScore (Zhang et al., 2019) uses a strong PLM model to greedily match each word in a reference response with one word in the generated response. By doing so, it computes the recall of the generated sequence. BERTScore was shown to have strong system-level and segment-level correlation with human judgment on several machine translation tasks.

BLEURT. BLEURT (Sellam et al., 2020) is based on the BERTScore and finetuned on human judgments after pretraining on large-scale synthetic data with multiple automatic metrics as supervision signals. BLEURT has shown its strong correlation with human judgment on machine translation tasks.

RUBER. RUBER (Tao et al., 2018) is an unsupervised automatic evaluation metric that considering the similarity of the generated response with conversational context and response reference.

MAUDE. MAUDE (Sinha et al., 2020) proposed an unreferenced automated evaluation metric that uses large-scale PLMs to extract hidden representations of dialogue sentences, and leverages the temporal transitions that exist between them.

4.3 Settings

The dimension of latent variable z is set to 768 to improve the information bottleneck. As we mentioned before, the encoder and the decoder in our model are BERT (Devlin et al., 2018) and GPT-2 (Radford et al., 2019), respectively. We use the

³http://parl.ai

⁴http://yanran.li/dailydialog

⁵https://github.com/OpenNMT/

OpenNMT-py/blob/master/tools/multi-bleu.
perl

⁶http://nlp.stanford.edu/data/glove. 840B.300d.zip

Tokenizer of BERT to conduct the texts for BERT and tokenizer of GPT-2 for GPT-2 in all experiments. We use the Optimus (Li et al., 2020) to save the time consumption of the VAE training. The recognition network and prior network are consist of 3 layers MLP with Dropout layer and GELU function. When training, we set the mini-batch size to 16. The AdamW optimizer is used to train the VAEs module with the initial learning rate 5e-5, and the learning rate warmup and decay strategy is employed. The value of τ in Eq. 5 is set to 0.01. When we conduct the matching training, we change the initial learning rate to 3e-6, and the learning rate warmup and decay strategy is also employed.

5 Results and Analysis

5.1 Overall Performance

We examine the performance of our model compared with baselines on two open-access datasets and the results.

The word-overlap metrics based on n-gram perform worst. As shown in Table 1 and Table 2, the word-overlap evaluation metrics, i.e., BLEU, ROUGE, and METEOR, obtain the worst performance in the dialogue evaluation on both two datasets. Among them, BLEU (hybrid) scores on two datasets are both less than 0.1, though it is the most widely used metric in machine translation. Intuitively, the information from n-gram is more accurate with a larger n (the most accurate information comes from the whole sentence). However, as the results shown in the Table 1 and Table 2, the correlation score decays when n increases. The same phenomenon is observed when using ROUGE. It seems using n-grams as the representation of the dialogue sentence is not a good choice.

PLM is an effective representation extractor for dialogue sentences. From Table 1 and Table 2, we can see that most embedding-based metrics, i.e., Average, Extrema, Greedy, BERTScore and BLEURT, using pretrained embedding to represent the sentences, perform better than wordoverlap metrics which uses n-grams as representation. Furthermore, traditional embedding-based metrics with Glove-based embedding, i.e., Average, Extrema, Greedy, perform worse than the embedding-based metrics with PLM-based embedding, i.e., BERTScore and BLEUET. Thus, we can know that using PLM to represent the dialogue sentence is more effective for the evaluation.

Learning-based discriminative metrics outper-

Table 1: Performance of our EMS methods and all baselines evaluated on the Persona-Chat Dataset. The BERTScore and BLEURT are built based on the base and large version of the Roberta (Liu et al., 2019).

Metrics	Pearson's ρ	Spearmans's r		
BLEU				
Hybrid	0.0602 (0.100)	0.0623 (0.088)		
1-gram	0.2486 (<0.001)	0.2201 (<0.001)		
2-gram	0.1991 (<0.001)	0.1891 (<0.001)		
3-gram	0.0998 (<0.001)	0.1138 (<0.001)		
4-gram	0.0558 (0.127)	0.0622 (0.002)		
ROUGE				
1-gram	0.2663 (<0.001)	0.2553 (<0.001)		
2-gram	0.1706 (<0.001)	0.1715 (<0.001)		
3-gram	0.1062 (0.004)	0.1157 (0.001)		
L	0.2777 (<0.001)	$0.2619 \ (< 0.001)$		
METEOR	0.2507 (0.033)	0.2297 (0.020)		
Embedding	Metrics			
Average	0.1330 (<0.001)	0.1452 (<0.001)		
Extrema	0.2354 (<0.001)	0.2272 (<0.001)		
Greedy	0.2585 (<0.001)	0.2502 (<0.001)		
BERTScore				
Base	0.2943 (<0.001)	0.2781 (<0.001)		
Large	0.2782 (<0.001)	$0.2619 \ (< 0.001)$		
BLEURT				
Base	0.3560 (<0.001)	0.3482 (<0.001)		
Large	0.3045 (<0.001)	0.2857 (<0.001)		
RUBER	0.4581 (<0.001)	0.4570 (<0.001)		
MAUDE	0.5619 (<0.001)	0.5631 (<0.001)		
EMS	0.5856 (<0.001)	0.5921 (<0.001)		

form training-free metrics. From Table 1 and Table 2, we can observe that all the learning-based discriminative metrics (i.e., RUBER, MAUDE, and our EMS) outperform all other training-free metrics (i.e., BLEU, ROUGE, METEOR, Average, Extrema, Greedy, BERTScore, and BLEURT) in both the Pearson and Spearman correlation with human judgment. Therefore, we conclude the learning-based discriminative metrics are very powerful to evaluate the open-domain dialogue systems.

Our proposed EMS metric performs the best. Our EMS metric achieves the best performance with 0.5856, 0.5921 in Pearson and Spearman correlation with the human judgment on the personachat dataset, respectively. Meanwhile, on the dailydialog dataset, EMS obtains 0.5331, 0.5253 in Pearson and Spearman score. These experimental

Table 2: Performance of our EMS methods and all baselines evaluated on the DailyDialog Dataset. The BERTScore and BLEURT are built based on the base and large version of the Roberta (Liu et al., 2019).

Metrics	Pearson's ρ	Spearmans's r			
BLEU					
Hybrid	0.0240 (0.498)	0.0211 (0.551)			
1-gram	0.2285 (<0.001)	0.2050 (<0.001)			
2-gram	0.1115 (<0.001)	0.0812 (<0.001)			
3-gram	0.0364 (0.304)	0.0401 (0.257)			
4-gram	0.0261 (0.461)	0.0211 (0.551)			
ROUGE					
1-gram	0.1424 (<0.001)	0.0921 (0.009)			
2-gram	0.1226 (<0.001)	0.1104 (0.002)			
3-gram	0.0534 (0.131)	0.0467 (0.187)			
L	0.1401 (<0.001)	0.0901 (0.011)			
METEOR	0.2095 (<0.001)	0.1809 (<0.001)			
Embedding	Metrics				
Average	0.0272 (0.443)	0.0440 (0.214)			
Extrama	0.1763 (<0.001)	0.1652 (<0.001)			
Greedy	0.1311 (<0.001)	0.1186 (<0.001)			
BERTScore	BERTScore				
Base	0.1139 (0.001)	0.1132 (0.001)			
Large	0.1273 (<0.001)	0.1251 (<0.001)			
BLEURT					
Base	0.3478 (<0.001)	0.3356 (<0.001)			
Large	0.2260 (<0.001)	0.1916 (<0.001)			
RUBER	0.4328 (<0.001)	0.4310 (<0.001)			
MAUDE	0.5162 (<0.001)	0.5343 (<0.001)			
EMS	0.5331 (<0.001)	0.5253 (<0.001)			

results show our method outperforms all existing baselines, indicating the superiority of our method.

5.2 Analysis

Our model aims to enhance the dialogue evaluation via variational training. Hence, in this subsection, we examine whether variational training can improve the performance by ablation study.

First, we replace the hidden representation ("w/o q") in Eq. 5 by the one from pure BERT, i.e., CLS. From the performance in Table 3, the KL regularization enhances the performance of EMS metric on both two datasets which proves a smooth latent space (via VAE training) is important. Second, as shown in Table 3, without z_p in Eq. 6 ("w/o p") which captures the feasible latent reference information, EMS gains a performance drop. Therefore,

Table 3: Performance of the ablation study. Note that ρ and r indicate the Pearson's ρ and Spearmans's r.

Metrics	Persona-Chat Dataset		DailyDialog Dataset	
	ρ	r	ρ	r
EMS	0.5856	0.5921	0.5331	0.5253
w/o q w/o p	0.5732 0.5621	0.5741 0.5618	0.5298 0.5193	0.5221 0.5218

Table 4: Two evalution cases from DailyDialog Dataset. _ eos_ indicates the end of the current sentence.

Context:	did you get your bus pass? _ eos_ i have n't gone to get it yet eos_ why have n't you got it? _ eos_ i do n't know where to go to get one eos_ i know where to get them from.			
Reference:	where do i get it?			
Generated:	what kind of buses are they on?			
Human (1-5)	BLEU (0-1)	MAUDE (1-5)	EMS (1-5)	
3.00	0.00	4.99	3.32	
Context:	what a beautiful home! _ eos_ you'll notice that the window treatments, carpeting, and drapes are all new eos_ i like the way the blinds give you privacy from the street eos_ follow me into the kitchen. you will love it eos_ i love that they put a wine storage area in the kitchen.			
Reference:	the best part is the bedroom and attached bathroom.			
Generated:	i'm sure you will.			
Human (1-5)	BLEU (0-1)	MAUDE (1-5)	EMS (1-5)	
3.50	0.00	1.17	2.29	

it proves the effectiveness of the feasible latent reference information.

5.3 Case Study

To explain more intuitively, we show two cases of our experiments in Table 4. In the first case, we can observe the golden score from the human is 3.00, however, MAUDE predicts the score as 4.99. We find the MAUDE gives such a high score because there is a keyword ("what kind of <u>buses</u> are they on ?") in the generated response which also exists in the conversational context. In the second case, MAUDE gives an extremely low score, i.e., 1.17, since no repeated words in the generated response and the context. However, our EMS gives scores similar to the Human Score, 3.32 and 2.29 in the first and second case, respectively. It proves that our EMS is more similar to human evaluation.

6 Conclusion and Future Work

In this study, we propose a two-stage automatic evaluation metric, i.e., EMS, which can obtain a smooth latent space that can both capture discourselevel context information and model more feasible latent references for evaluating the open-domain dialogues. Experimental results on two dialogue datasets confirm the superiority of our method for open-domain dialogue evaluation, where both Pearson and Spearman correlations with human judgments outperform all baseline methods.

Owing to the promising performance of the variational training, we pursue to design the training procedures for better representation in latent space. Besides, we will explore more efficient methods to obtain more useful feasible reference information.

Acknowledgement

We would like to thank the anonymous reviewers for their constructive comments. This work is supported by the National Key Research and Development Program of China (No. 2020YFB1406702), the National Science Foundation of China (NSFC No. 61876196) and Beijing Outstanding Young Scientist Program No.BJJWZYJH012019100020098. Rui Yan is sponsored by Tencent Collaborative Research Fund. Rui Yan is the corresponding author, and is supported as a Young Fellow of Beijing Institute of Artificial Intelligence (BAAI).

Ethics Impact

In this paper, we propose a two-stage automatic evaluation metric EMS for open-domain dialogue systems. The positive impact lies in that it allows quick and effective comparison between different dialogue systems and is crucial for the development of open-domain dialogue tasks. The negative impact may be that in some extreme cases, the system may give high scores to the rude and dirty responses. Hence, in such situation, the training dataset is crucial and should be examined before employed in practice.

References

- JinYeong Bak and Alice Oh. 2020. Speaker sensitive response evaluation model. *arXiv preprint arXiv:2006.07015*.
- Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings*

of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, pages 65–72.

- 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.
- Zhangming Chan, Xiuying Chen, Yongliang Wang, Juntao Li, Zhiqiang Zhang, Kun Gai, Dongyan Zhao, and Rui Yan. 2019a. Stick to the facts: Learning towards a fidelity-oriented e-commerce product description generation. In *Proceedings of the* 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4960–4969.
- Zhangming Chan, Juntao Li, Xiaopeng Yang, Xiuying Chen, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019b. Modeling personalization in continuous space for response generation via augmented wasserstein autoencoders. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)*, pages 1931–1940.
- Zhangming Chan, Yuchi Zhang, Xiuying Chen, Shen Gao, Zhiqiang Zhang, Dongyan Zhao, and Rui Yan. 2020. Selection and generation: Learning towards multi-product advertisement post generation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 3818–3829.
- Wang Chen, Piji Li, and Irwin King. 2021. A trainingfree and reference-free summarization evaluation metric via centrality-weighted relevance and selfreferenced redundancy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics.*
- Xiuying Chen, Zhi Cui, Jiayi Zhang, Chen Wei, Jianwei Cui, Bin Wang, Dongyan Zhao, and Rui Yan. 2020. Reasoning in dialog: Improving response generation by context reading comprehension. *arXiv preprint arXiv:2012.07410*.
- Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
- Gabriel Forgues, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay. 2014. Bootstrapping dialog systems with word embeddings. In *Nips, workshop*, volume 2.

- Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. 2019. Cyclical annealing schedule: A simple approach to mitigating kl vanishing. arXiv preprint arXiv:1903.10145.
- Jun Gao, Wei Bi, Ruifeng Xu, and Shuming Shi. 2021. Ream: An enhancement approach to referencebased evaluation metrics for open-domain dialog generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics: Findings*.
- Shen Gao, Zhaochun Ren, Yihong Zhao, Dongyan Zhao, Dawei Yin, and Rui Yan. 2019. Product-aware answer generation in e-commerce question-answering. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, pages 429–437.
- Yang Gao, Wei Zhao, and Steffen Eger. 2020. Supert: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. *arXiv preprint arXiv:2005.03724*.
- Sarik Ghazarian, Johnny Tian-Zheng Wei, Aram Galstyan, and Nanyun Peng. 2019. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. *arXiv preprint arXiv:1904.10635*.
- Jiatao Gu, Changhan Wang, and Jake Zhao. 2019. Levenshtein transformer. *arXiv preprint arXiv:1905.11006*.
- Xiaodong Gu, Kyunghyun Cho, Jung-Woo Ha, and Sunghun Kim. 2018. Dialogwae: Multimodal response generation with conditional wasserstein autoencoder. *arXiv preprint arXiv:1805.12352*.
- Suchin Gururangan, Tam Dang, Dallas Card, and Noah A Smith. 2019. Variational pretraining for semi-supervised text classification. *arXiv preprint arXiv:1906.02242*.
- Tatsunori B Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. *arXiv preprint arXiv:1904.02792*.
- Wenpeng Hu, Zhangming Chan, Bing Liu, Dongyan Zhao, Jinwen Ma, and Rui Yan. 2019. Gsn: A graphstructured network for multi-party dialogues. *arXiv* preprint arXiv:1905.13637.
- Tian Lan, Xian-Ling Mao, Wei Wei, Xiaoyan Gao, and Heyan Huang. 2020. Pone: A novel automatic evaluation metric for open-domain generative dialogue systems. *ACM Transactions on Information Systems* (*TOIS*), 39(1):1–37.
- Bohan Li, Junxian He, Graham Neubig, Taylor Berg-Kirkpatrick, and Yiming Yang. 2019. A surprisingly effective fix for deep latent variable modeling of text. *arXiv preprint arXiv:1909.00868*.

- Chunyuan Li, Xiang Gao, Yuan Li, Xiujun Li, Baolin Peng, Yizhe Zhang, and Jianfeng Gao. 2020. Optimus: Organizing sentences via pre-trained modeling of a latent space. *arXiv preprint arXiv:2004.04092*.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.
- Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. arXiv preprint arXiv:1603.08023.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.
- Ryan Lowe, Michael Noseworthy, Iulian V Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic turing test: Learning to evaluate dialogue responses. *arXiv preprint arXiv:1708.07149*.
- Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. *NAACL-HLT*, pages 236–244.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311– 318. ACL.
- Lisong Qiu, Juntao Li, Wei Bi, Dongyan Zhao, and Rui Yan. 2019. Are training samples correlated? learning to generate dialogue responses with multiple references. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3826–3835.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Vasile Rus and Mihai Lintean. 2012. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 157– 162. ACL.

- Ananya B Sai, Akash Kumar Mohankumar, Siddhartha Arora, and Mitesh M Khapra. 2020. Improving dialog evaluation with a multi-reference adversarial dataset and large scale pretraining. *Transactions* of the Association for Computational Linguistics, 8:810–827.
- Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *ACL*.
- Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.
- Iulian Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.
- Xiaoyu Shen, Hui Su, Wenjie Li, and Dietrich Klakow. 2018. Nexus network: Connecting the preceding and the following in dialogue generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4316– 4327.
- Koustuv Sinha, Prasanna Parthasarathi, Jasmine Wang, Ryan Lowe, William L Hamilton, and Joelle Pineau. 2020. Learning an unreferenced metric for online dialogue evaluation. *arXiv preprint arXiv:2005.00583*.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*.
- Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.
- Jiannan Xiang, Yahui Liu, Deng Cai, Huayang Li, Defu Lian, and Lemao Liu. 2021. Assessing dialogue systems with distribution distances. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics: Findings*.
- Tsuta Yuma, Naoki Yoshinaga, and Masashi Toyoda. 2020. ubleu: Uncertainty-aware automatic evaluation method for open-domain dialogue systems. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 199–206.

- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ICLR*.
- Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. *ACL*.
- Tianyu Zhao, Divesh Lala, and Tatsuya Kawahara. 2020. Designing precise and robust dialogue response evaluators. *arXiv preprint arXiv:2004.04908*.
- Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1909.02622*.

A Proof

As well we know, the the mutual information (MI) of X and Y is defined as

$$I(X;Y) = H(X) - H(X | Y) = H(Y) - H(Y | X)$$
(8)

where $H(\cdot)$ denotes the entropy.

Obviously, we can get

$$H(l \mid c, \{r_k\}_{k=1}^N) = H(l \mid c, r_i, \{r_k\}_{k=1}^N \setminus r_i) \leq H(l \mid c, r_i).$$
(9)

Then, we can compare the MI of using the feasible latent reference information and not as follows

$$I(l; c, \{r_k\}_{k=1}^N) = H(l) - H(l \mid c, \{r_k\}_{k=1}^N)$$

$$\geq H(l) - H(l \mid c, r_i)$$

$$= I(l; c, r_i).$$
(10)

where we can observe the MI is enhanced.