RADE: Reference-Assisted Dialogue Evaluation for Open-Domain Dialogue

Zhengliang Shi[1], **Weiwei Sun**[1], **Shuo Zhang**[2], **Zhen Zhang**[1], **Pengjie Ren**[1], **Zhaochun Ren**[1*]
[1]Shandong University, Qingdao, China
[2]Bloomberg, London, United Kingdom
shizhl@mail.sdu.edu.cn {sunnweiwei, zhen.zhang.sdu}@gmail.com
zhaochun.ren@sdu.edu.cn szhang611@bloomberg.net jay.ren@outlook.com

Abstract

Evaluating open-domain dialogue systems is challenging for reasons such as the one-to-many problem, i.e., many appropriate responses other than just the golden response. As of now, automatic evaluation methods need better consistency with humans, while reliable human evaluation can be time- and cost-intensive. To this end, we propose the **R**eference-**A**ssisted **D**ialogue **E**valuation (RADE) approach under the multi-task learning framework, which leverages the pre-created utterance as reference other than the gold response to relief the one-to-many problem. Specifically, RADE explicitly compares reference and the candidate response to predict their overall scores. Moreover, an auxiliary response generation task enhances prediction via a shared encoder. To support RADE, we extend three datasets with additional rated responses other than just a golden response by human annotation. Experiments on our three datasets and two existing benchmarks demonstrate the effectiveness of our method, where Pearson, Spearman, and Kendall correlations with human evaluation outperform state-of-the-art baselines.

1 Introduction

Open-domain dialogue system, which focuses on non-goal-oriented chitchat, may converse on a broad range of arbitrary topics. Recent years have witnessed rapid advances in natural language generation (Zhang et al., 2019b; Roller et al., 2021; Zhao et al., 2023), boosting the development of open-domain dialogue systems. Conversations with such systems resemble human-human interactions as various responses might fit the context, given that users often do not have a specific goal beyond enjoying the conversation. Evaluating these conversations is thus challenging because of the so-called one-to-many problem (Chan et al., 2021; Ji et al., 2022); see Figure 1 where three candidate responses with different semantics fit the context while there is only one golden response.

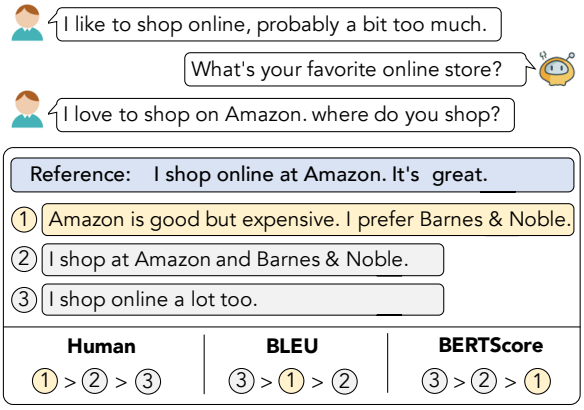

Figure 1: An example to explain the one-to-many nature of open-domain dialogues.

The most common practice of dialogue evaluation is done with reference-based metrics, which compare the generated response with a pre-created response, commonly referred to as the golden standard (Ji et al., 2022). The reference-based metrics calculate the similarity between the generated and gold responses at either lexical level (e.g., ROUGE (Lin, 2004), BLEU (Papineni et al., 2002)) or semantic level (e.g., BERTScore (Zhang et al., 2019a), ADEM (Lowe et al., 2017)). However, these metrics ignore the one-to-many nature of open-domain dialogues. As illustrated at the bottom of Figure 1, the generated response "*Amazon is good but expensive ...*" expresses the opposite semantics to the golden response "*I shop online...*" and is therefore considered a non-good response by the reference-based metrics. Therefore, these metrics may need a higher consistency with humans. Recently, *multi-reference methods* and *reference-free methods* are proposed to address the drawback of reference-based metrics. The former explicitly annotates multiple references for dialogue (Eric et al., 2021), whereas the latter discards the golden response in the evaluation and achieves high cor-

* Corresponding author.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 12856–12875

relations with human judgments (Mehri and Eskenazi, 2020c; Huang et al., 2020). However, drawback still exists in these two classes of methods. Multi-reference methods are costly and hard to generalize to different datasets, while reference-free methods are often unstable and vulnerable to data-induced biases[1].

To overcome the weakness of existing evaluation methods and further resolve the one-to-many problem, we propose a new technique, namely **R**eference-**A**ssisted **D**ialogue **E**valuation (RADE). RADE considers the pre-created response as a reference instead of the golden standard.

To support RADE, we design a new human annotation task to extend existing datasets, which includes metric decompose and pairwise annotation, where a pre-scored golden response is paired with generated responses for rating following a unified rating score. The final scores are arrived at by aggregating ratings with a weighted sum from different sub-metrics. The human annotation collects labels for three high-quality datasets with 10,112 dialogues, which correspond to three downstream open-domain dialogue system tasks, i.e., chitchat, empathetic dialogue, and personal chat. These multi-domain datasets make RADE more robust when generalizing to cross-domain evaluation scenarios while having a better task-specific performance.

We propose a RADE model under the multi-task learning framework for automatic evaluation based on the newly collected datasets. Specifically, RADE first explicitly encodes the relation between dialogue context and generated response with reference assistance. Then RADE discriminates whether the reference or response fits the context better and predicts the scores for each utterance. To relieve the one-to-many problem, we augment RADE with a joint response generation task where RADE learns to generate the reference responses to better perceive the range of candidate responses.

Extensive experiments on our three benchmarks demonstrate that RADE achieves the best correlations with human judgment. We also examine two existing USR benchmark (Mehri and Eskenazi, 2020c) where RADE outperforms the state-of-the-art methods, e.g., pushing the Pearson correlation coefficient to 48% (6.8% absolute improvement) and Spearman correlation coefficient to 46.6% (4.3% absolute improvement). Experiments also verify the generalizability of our proposed method.

Our contributions can be summarized as follows: (1) We propose the reference-assisted evaluation method, i.e., RADE, for open-domain dialogue evaluation; (2) We design a new human annotation task and collect three new dialogue evaluation datasets; (3) Experiments on our benchmarks and two existing benchmarks verify the effectiveness and robustness of the proposed methods; (4) We release three new benchmarks and the pre-trained evaluation model to facilitate future research on dialogue evaluation.

2 Related work

2.1 Reference-based dialogue evaluation

Previous reference-based methods compare the generated response with the pre-created response at the lexical or semantic level. Lexical-level metrics, e.g., ROUGE (Lin, 2004), BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005), count the n-gram overlap between the candidate response and the reference response. These methods usually correlate poorly with human evaluation results due to the lexical mismatch problem (Liu et al., 2016). Semantic-level metrics evaluate address lexical mismatch problem by calculating similarity with high-dimension embeddings. For example, Sharma et al. (2017) measures the embedding distance between golden and generated response. Ghazarian et al. (2019) and Zhang et al. (2019a) enhance the text representation using the large pre-train model, which has shown exemplary performance in capturing semantic similarity. However, they suffer from the one-to-many problem when evaluating open-domain dialogues since responses with various semantics may fit the dialogue context.

Recent works tend to relieve this drawback by annotating multiple references for dialogue, commonly referred to as multi-reference methods (Li et al., 2017; Sai et al., 2020), which are costly and hard to generalize to agnostic scenarios. The proposed RADE aims to consider the pre-created response as a candidate instead of the golden standard to address the one-to-many problem of dialogue evaluation.

[1]The data-induced biases included two aspects: (1) Noise collected in data/annotations, (2) The reference-free models tend to favor the underlying models' outputs and those from similar models or trained with similar datasets. (Khalid and Lee, 2022; Deutsch et al., 2022)

2.2 Reference-free dialogue evaluation

The reference-free methods are gaining more attention as they correlate more with human judgment only with the dialogue context and response. For example, MAUDE predicts the score of dialogue using pre-trained language models, GRADE (Huang et al., 2020) evaluates the coherence of dialogues with the augmentation of the commonsense graph, EMS (Chan et al., 2021) enhances the dialogue evaluation by capturing the representation of the context and response in latent space. Some methods further decompose the evaluation of responses into multiple perspectives (Mehri and Eskenazi, 2020a,c; Phy et al., 2020), such as relevance, fluency, and engagingness, then aggregate the overall score from different sub-metrics with a weighted average. However, some recent studies (Khalid and Lee, 2022; Deutsch et al., 2022) reveal that the reference-free methods are vulnerable to data-induced biases and inherently biased toward models which are more similar to their own. In contrast, this paper proposes a reference-assisted approach, which enhances the robustness of the model using reference responses as a benchmark.

3 Task Formulation

In this work, we propose two tasks: (1) extending the existing datasets by human annotation, and (2) leveraging the rated references collected in (1) to enhance automatic evaluation.

Human annotation Human annotation aims to extend existing datasets with multiple rated responses to facilitate automatic evaluation. Given a dialogue context c, which is always paired with a golden response (denoted as reference) r_h, we employ the generation models, e.g., BlenderBot (Roller et al., 2021), to generate one more response r_a. We then assign a fixed overall score or derive from existing datasets to the reference as s_h. The annotators are instructed to rate r_a as s_a, following the same scale while taking the reference as a benchmark. The annotators are also asked to revise the reference score s_h if s_h is inappropriate.

Automatic evaluation Given a dialogue context c, the proposed RADE learns to evaluate the response r_a with the assistance of reference r_h under the multi-task learning framework. The first task explicitly models the relation between reference and response and discriminates which fits the context better. The scores of reference and response are predicted simultaneously. And the second task enhances the score prediction task by implicitly estimating the distribution of candidate responses.

Relevance †**:** *Whether the response matches dialogue context semantically.*
Engagingness†**:** *Whether the response is engaging or interesting rather than rigid template.*
Fluency†**:** *Whether the response is fluent and natural throughout the conversation.*
Understandability‡**:** *Is there any external knowledge contained in the response.*
Emotional-awareness‡**:** *Whether the agent capture the emotion of user and support empathic support.*
Personality-awareness‡**:** *Whether the response conforms to given personality.*

Table 1: **Criteria in human annotation.** Metrics with † are general metrics for all dialogue tasks, while metrics ‡ are metrics for specific dialogue tasks (e.g., understandability for chitchat, emotion-awareness for emotional dialogue and personal-awareness for personal chat).

4 Human Annotation

Our human annotation task aims to rate the candidate responses following a pre-scored reference as a benchmark. Since there are multiple perspectives to assess the response, we simplify by sorting the possible aspects into two categories: the general view and the task-specific view. As listed in Table 1, the former contains relevance, engagingness, and fluency, which are suitable for all dialogue agents. And task-specific criteria consist of understandability, emotional awareness, and personality awareness, which correspond to chitchat dialogue, emotional dialogue, and persona dialogue. We annotate rates on each metric and calculate the overall rating score by weighting these sub-metrics. Specifically, the weights are obtained based on the preference of users (see section A.1.3 for more details).

4.1 Data preparation

We consider three datasets to extend: • *DSTC–ChitChat (ChitChat)* (Hori and Hori, 2017), a chitchat dataset collected from Twitter, each example derived from the conversation between a customer and an agent. • *Empathetic Dialogues (EmpaDial)* (Rashkin et al., 2019), which consists of 25k dialogues grounded in emotional situations.

Domain	ChitChat	EmpaDial	PersonaChat
# Dialogues	2,090	4,022	4,000
Kappa	0.540	0.554	0.533
Distribution of the score			
Rating 1	0.5%	1.2%	3.7%
Rating 2	15.6%	12.5%	12.6%
Rating 3	48.3%	42.0%	50.5%
Rating 4	29.5%	32.0%	23.9%
Rating 5	5.1%	12.3%	9.4%

Table 2: **The statistics of the collected datasets.** For each example, the overall score of the response is mean of all sub-metrics.

• *PersonaChat* (Zhang et al., 2018), a real-world dataset consisting of 10k dialogues where each participant plays the part of an assigned persona.

Then, we collect model-generated responses using the following seven well-performing dialogue models on these datasets: BlenderBot (Roller et al., 2021), DialoGPT (Zhang et al., 2019b), KEMP (Li et al., 2020b), MoEL (Lin et al., 2019), MIME (Majumder et al., 2020), EmpDG (Li et al., 2020a), PersonaGPT (Tang et al., 2021). The train-dev-test of collected datasets are split as Chitchat (1490/300/300, 5/1/1), Empathetic Dialogue (3022/500/500, 6/1/1), and Persona Chat (3000/500/500, 6/1/1). More details of these models are available in Appendix A.1.1.

4.2 Human annotation detalis

We hire 40 annotators for data annotation. Following a five-scale standard, they are asked to label sub-metrics as listed in Table 1. The five-scale allows the annotators to factor in their subjective interpretation of the extent of success or failure of a system's response to satisfy a user's request. The dialogue context, rated reference response, and corresponding score are provided in each example. At least three annotators are required for each example. We annotated about 10k dialogues for the three datasets, and the statistics of the collected datasets are listed in Table 2. The ratings achieve reasonable inter-annotator agreements with Fleiss Kappa scores of 0.540, 0.554, and 0.533 on three datasets, respectively. More details about the annotation guideline and details are provided in Appendix A.1.2.

5 Reference-Assisted Automatic Evaluation

We propose RADE, a **R**eference-**A**ssisted Automatic **D**ialogue **E**valuation method under the framework of multi-task learning. Compared with reference-based methods that evaluate based on the distance between the golden and generated response, the proposed RADE explicitly discriminates whether the reference or candidate response fits the dialogue context better. To relieve the one-to-many problem, we augment RADE with a joint response generation task, which aims to perceive the range of feasible candidate responses. To improve the performance of RADE with the limited dataset, we propose a two-stage training strategy, including cross-domain pre-training and task-specific finetune.

5.1 Model architecture

The architecture of RADE is illustrated in Figure 2, which comprises a posterior encoder, a regression layer, and a candidate response generator.

Posterior encoder. The posterior encoder encodes the dialogue context c, reference response r_h, and model-generated response r_a into hidden representation. In particular, we first concatenate c, r_h and r_a together into X with a specific token `[SEP]`:

$$X = \{c \texttt{ [SEP] } r_h \texttt{ [SEP] } r_a\} \tag{1}$$

Then the concatenated sequence is fed into a transformer-based encoder to get the representation $\mathbf{H} \in \mathbb{R}^{|X| \times d}$:

$$\mathbf{H} = \text{Encoder}(X), \tag{2}$$

where d is the hidden size of encoder, $|X|$ is the length of sequence X.

Regression layer. The regression layer aggregates the representation $\mathbf{H}$ and predicts the scores of both reference and candidate response simultaneously. Specifically, a pooling layer aggregates the token-level representation into a sequence-level representation: $\mathbf{h} \in \mathbb{R}^{d \times 1}$:

$$\mathbf{h} = \text{Pooling}(\mathbf{H}) \tag{3}$$

Then, a feedforward network takes $\mathbf{h}$ as input to predict the score of both reference and candidate response:

$$(\hat{s_h}, \hat{s_a}) = \text{FeedForward}(\mathbf{h}), \tag{4}$$

where $\hat{s_h}$ and $\hat{s_a}$ denote the predicted score of r_h and r_a, respectively.

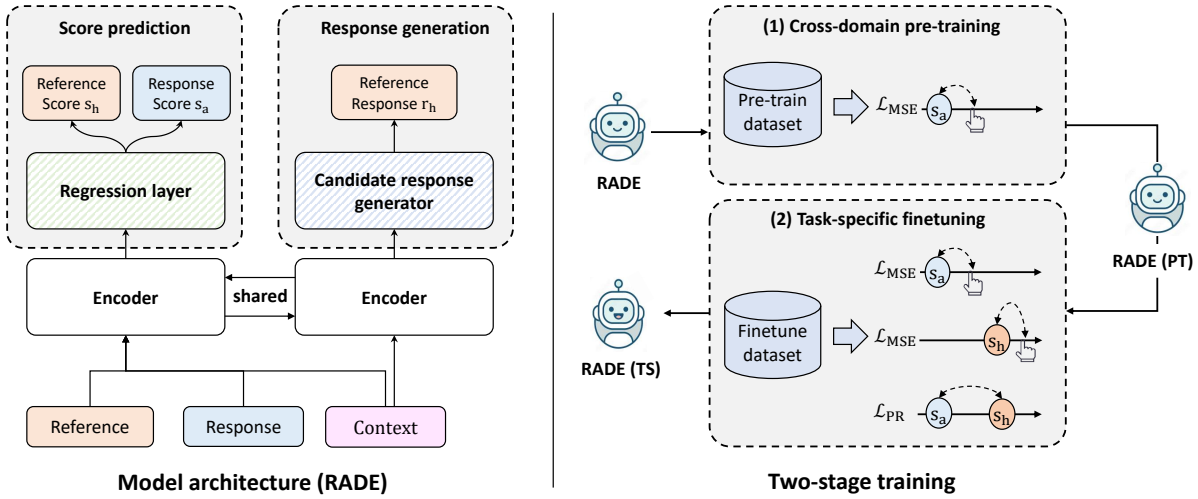

Figure 2: **Left:** An overview of our model which consists of an encoder, a regression layer, and a response generator. **Right:** Our two-stage training process with cross-domain **p**re-**t**raining (PT) and **t**ask-**s**pecific finetuning (TS).

Candidate response generator. To relieve the one-to-many problem, we devise a candidate response generator to perceive the range of feasible candidate responses (Chan et al., 2021). Specifically, a Transformer-based generator learns to generate reference responses autoregressively for a specific context. We first encode the dialogue context c using a encoder:

$$\hat{\mathbf{h}} = \text{Encoder}\,(c), \tag{5}$$

where the Encoder shares the same parameters with the posteriori encoder in Eq. (2). Then, we apply a Transformer-based decoder Decoder to model the generation probability of reference response r_h:

$$P(r_h|c) = \prod_{t=1}^{T} \text{Decoder}(r_h^{(t)}|r_h^{(<t)}, \hat{\mathbf{h}}), \tag{6}$$

where T denotes the length of r_h.

Compared with the previous reference-free methods, which estimate the relation between context and response only with the knowledge acquired from their training data, RADE explicitly takes the pre-created response as a benchmark to reduce the data-induced bias when generalizing to agnostic scenarios. Moreover, different from existing reference-based methods, which use the pre-created response as the golden standard without considering the semantic diversity of the response, we relieve the one-to-many problem via auxiliary response generation tasks. The share encoder enhances the capability of context representation which augment the performance of score-predicting task through multi-task learning.

5.2 Two-stage training

The neural-based model has been proven prone to data-induced bias, but it is costly to annotate a large dataset in every specific task. Therefore, we propose a two-stage strategy that includes: (1) *cross-domain pre-training*, and (2) *task-specific fine-tuning*, keeping a tradeoff of performance between in- and cross-domain. As shown in Figure 2 (right), we pre-train our model based on existing human-annotated datasets from different downstream tasks of open-domain dialogue to improve the generalizability (Ye et al., 2021a). Since the cross-domain datasets suffer from domain gaps and no pair-wised score, we finetune our model in the next stage with newly collected task-specific datasets.

Cross-domain pre-training. The pre-training datasets contain 54,438 dialogue-level examples collected from different downstream tasks, covering a wide range of domains (see more details in Table 7). For learning the coarse-grain judgment of generated response without human-annotated reference scores, our model is first pre-trained by minimizing a new cross-domain pre-training loss $\mathcal{L}_{\text{Cross}}$. Concretely, the $\mathcal{L}_{\text{Cross}}$ is composed of score-prediction loss and generation loss, which can be formulated as:

$$\mathcal{L}_{\text{Cross}} = \mathcal{L}_{\text{MSE}}(\hat{s_a}, s_a) + \mathcal{L}_{\text{GEN}}, \tag{7}$$

where $\hat{s_a}$ and s_a denote the human-annotated score and the predicted score of the candidate response and $\mathcal{L}_{\text{MSE}}(\hat{s_a}, s_a) = (\hat{s_a} - s_a)^2$. $\mathcal{L}_{\text{GEN}}$ is the response generation loss, which is defined as:

$$\mathcal{L}_{\text{GEN}} = -\log P(r_h|c), \tag{8}$$

where $P(r_h|c)$ is the generation probability of r_h defined in Eq. (6).

Task-specific finetuning. We next finetune our model with newly annotated datasets to enhance the performance when evaluating task-specific dialogue agents. The optimize objective $\mathcal{L}_{\text{In}}$ is composed of score-prediction loss, generation loss, and pair-wised ranking loss, which can be formulated as:

$$\begin{aligned}\mathcal{L}_{\text{In}} =& \mathcal{L}_{\text{MSE}}(\hat{s_a}, s_a) + \mathcal{L}_{\text{MSE}}(\hat{s_h}, s_h) + \\ & \mathcal{L}_{\text{GEN}} + \mathcal{L}_{\text{PR}}\end{aligned} \tag{9}$$

where $\mathcal{L}_{\text{MSE}}(\hat{s_a}, s_a)$ and $\mathcal{L}_{\text{MSE}}(\hat{s_h}, s_h)$ are MSE score-prediction loss of reference response and candidte response, respectively. $\mathcal{L}_{\text{GEN}}$ is the generation loss as defined in Eq. (8). $\mathcal{L}_{\text{PR}}$ is the pair-wise ranking loss defined as:

$$\mathcal{L}_{\text{PR}} = -g(s_h, s_a) \log \frac{\mathrm{e}^{\hat{s_a}}}{\mathrm{e}^{\hat{s_h}} + \mathrm{e}^{\hat{s_a}}}, \tag{10}$$

in which $g(s_h, s_a)$ is a labeling function defined as:

$$g(s_h, s_a) = \begin{cases} 0, & s_h \geq s_a \\ 1, & s_h < s_a \end{cases} \tag{11}$$

The $\mathcal{L}_{\text{PR}}$ is introduced to assure that the rank order of the predicted scores satisfies the pre-annotated order. Compared to reference-free models that inherently favor outputs from their underlying models or those trained on similar datasets, RADE is specifically optimized to align with human intentions and effectively alleviate this bias.

6 Experimental Setup

6.1 Dataset and evaluation metrics

We mainly conduct experiments on the three datasets annotated in Section 4. We further evaluate the models on two existing benchmarks, USR-TopicChat and USR-PersonaChat (Mehri and Eskenazi, 2020c), to examine the generalizability of our method. The evaluation metrics include Pearson (r), Spearman (ρ), and Kendall (τ) correlation, which measures the linear relationship, monotonic relationship, and the ordinal association between automatic evaluation and human evaluation, respectively[2]. We abbreviate the Pearson, Spearman, and Kendall correlation as r, ρ, and τ for simplicity.

6.2 Implementation details

We initialize the parameters of the encoder and decoder with BART (Lewis et al., 2019), a Transformer-based pre-trained model. BART is well-suited to our proposed model because it is capable of both text representation tasks and text generation tasks. We optimize the model using Adam optimizer with parameters $\beta_1 = 0.98$, $\beta_2 = 0.97$, and the learning rate of $5e{-}5$. The model is trained up to 10 epochs, and we tune the hyper-parameters and pick the checkpoint on the development set. The training of the model can be done within 5 hours using two 2080Ti GPUs. We denote the RADE model that pre-trained on cross-domain datasets as **RADE (PT)**, and the model that further finetuned on task-specific data as **RADE (TS)**.

6.3 Baselines

We compare our method with two types of baselines: reference-based and reference-free methods.

The reference-free baselines include: *DialoRPT* (Gao et al., 2020a), which trained on large-scale social media feedback data to predict ranking-based scores; *GRADE* (Huang et al., 2020), which enhances the contextualized representations via topic-level commonsense graphs and predicts the score using a regression module; *FED* (Mehri and Eskenazi, 2020a), an unsupervised dialogue evaluation model based on DialogGPT; *UniEval* (Zhong et al., 2022), which evaluates the response from multiple perspectives; *QuesEval* (Scialom et al., 2021), which evaluates the fact-based text using summarizing asks.

The reference-based baselines include: *RUBER* (Tao et al., 2017), an unsupervised evaluation metric considering the similarity of the response with dialog context and reference; *BERTScore* (Zhang et al., 2019a), which employs BERT to greedily match the response and the ground truth at the token level; *BLEURT* (Sellam et al., 2020), which is a BERT-based model pre-trained with millions of synthetic examples; *BARTScore* (De Bruyn et al., 2020), which weights the log-likelihood of the generated response as the score. We also test three reference-based lexical-level metrics: *ROUGE-L*, *BLEU-2*, and *METEOR*.

[2]We use SciPy (`https://scipy.org/`) to calculate the scores.

Methods	ChitChat			Empathetic Dialogue			PersonaChat		
	r	ρ	τ	r	ρ	τ	r	ρ	τ
Reference-free methods									
FED_E (Mehri and Eskenazi, 2020b)	0.241	0.254	0.177	0.202	0.218	0.218	0.138	0.120	0.086
FED_U (Mehri and Eskenazi, 2020b)	0.235	0.248	0.171	0.147	0.156	0.106	0.145	0.162	0.117
QuesEval (Scialom et al., 2021)	0.045	0.021	0.013	0.069	0.084	0.057	-0.003	0.034	0.0237
UniEval (Zhong et al., 2022)	0.456	0.470	0.312	0.403	0.435	0.286	0.306	0.338	0.244
DialoRPT (Gao et al., 2020b)	-0.066*	-0.044*	-0.031*	0.267	0.244	0.166	-0.077*	-0.069*	-0.049*
GRADE (Huang et al., 2020)	0.491	0.434	0.300	0.549	0.568	0.398	-0.031*	-0.005	-0.030*
QuantiDCE (Ye et al., 2021b)	0.348	0.300	0.202	0.498	0.507	0.351	0.162	0.182	0.130
Reference-based lexicon-level methods									
ROUGE-L (Lin, 2004)	0.215	0.178	0.129	0.213	0.214	0.148	0.118	0.114	0.079
BLEU-2 (Papineni et al., 2002)	0.201	0.200	0.158	0.057	0.041*	0.032	0.060	0.039	0.031
METEOR (Banerjee and Lavie, 2005)	0.202	0.188	0.129	0.182	0.194	0.132	0.099	0.051	0.035
Reference-based semantic-level methods									
BERTScore (Zhang et al., 2019a)	0.296	0.243	0.213	0.167	0.243	0.173	0.278	0.292	0.196
BARTScore (Lewis et al., 2019)	0.133	0.057	0.039	0.256	0.253	0.173	0.143	0.168	0.115
RUBER (Tao et al., 2017)	0.332	0.351	0.369	0.252	0.256	0.183	0.122	0.123	0.089
BLEURT (Sellam et al., 2020)	0.353	0.363	0.249	0.343	0.337	0.232	0.105	0.140	0.102
BERT_MLP† (Devlin et al., 2018)	0.304	0.301	0.192	0.501	0.537	0.373	0.331	0.360	0.251
BART_MLP † (Lewis et al., 2019)	0.431	0.440	0.312	0.412	0.447	0.356	0.310	0.335	0.242
Reference-assisted methods									
RADE (Pre-trained model, PT)	0.472	0.491	0.334	0.650	0.601	0.427	0.386	0.390	0.285
RADE (Task-specific model, TS)	**0.601**	**0.569**	**0.409**	**0.863**	**0.849**	**0.685**	**0.470**	**0.465**	**0.347**
Ablation Study									
- w/o $\mathcal{L}_\text{PR}$	0.503	0.514	0.353	0.773	0.756	0.613	0.406	0.403	0.313
- w/o $\mathcal{L}_\text{GEN}$	0.451	0.482	0.332	0.751	0.740	0.602	0.387	0.372	0.272

Table 3: **Results on three benchmarks.** The metrics r, ρ, and τ indicate the Pearson's ρ, Spearman's r, and Kendall'τ. All values are statistically significant to p-value < 0.05 unless marked by *. Methods with † are implemented by ourselves. We underline the best results of each group of baselines methods and **bold** the best results of all methods. The bottom of the table show the ablation study, where the proposed RADE is compared with several variants (-w/o: without). See section 7.2 for details.

Moreover, we implement two reference-based baselines, BERT_MLP and BART_MLP, which are trained with the same human-annotated datasets as RADE, and provide a reasonable comparison with our proposed model. Specifically, we obtain the text representations of the dialogue using BERT or BART and then feed the representations into a multi-layer perception to calculate the scores. For a more comprehensive analysis, we also fine-tune the two strongest baselines, QuantiDCE and GRADE, on our cross-domain datasets as well as our self-collected datasets, respectively.

7 Results and Analysis

7.1 Experimental results

Overall performance. Table 3 shows the experimental performance for all methods. Overall, RADE achieves the best performance in three benchmarks in terms of all metrics. Concretely, the pre-trained model RADE (PT) gets better or comparable correlation with human judgment than the best baseline method on three dialogue tasks. The task-specific model RADE (TS), fine-tuned with the newly collected reference-assisted data, establishes a new state-of-the-art by improving the performance by about 30% on average compared to RADE (PT). For example, RADE (TS) gets $r = 0.601$, $\rho = 0.569$ in the ChitChat domain, and pushes r to 0.863 (0.314 absolute improvements), τ to 0.685 (0.287 absolute improvements) in EmpaDial domain. This result suggests that training with in-domain datasets is critical to enhancing the task-specific evaluation capability of RADE. For a more comprehensive comparison, we also train the two strongest baselines (QuantiDCE and GRADE) with our cross-domain and self-collected datasets, respectively. And the result and analysis are provided in Appendix A.2.3.

Generalizability. We find that the performance of the reference-free method varies dramatically across domains. For example, GRADE and QuantiDCE, trained in the chitchat domain, achieve high correlations with human judgment in ChitChat and EmpaDial but perform poorly in PersonaChat. The

result indicates that the contextual representation capabilities of unsupervised methods are limited by their training data and, therefore, are prone to data-induced bias, decreasing their performance when employing agnostic scenarios. In contrast, the gap between the proposed RADE (PT) methods across different domains is relatively small. These results indicate that RADE has better generalizability than reference-free methods due to the assistance of reference and the proposed cross-domain training strategy.

Results on USR benchmarks. We further examine our methods on two USR datasets (Mehri and Eskenazi, 2020c) to verify the efficiency and robustness of RADE when generalizing to existing dialogue evaluation benchmarks. The results are listed in Table 4. Experiments show that RADE, which has not explicitly trained on these datasets, achieves better or comparable results to previous supervised methods. See Appendix A.2.4 for more results and details.

Methods	USR-Topical		USR-Pearsona	
	r	ρ	r	ρ
GRADE	0.200	0.217	0.358	0.352
USR	0.412	0.423	0.440	0.418
USL-H	0.322	0.340	**0.495**	**0.523**
METEOR	0.336	0.391	0.253	0.271
BERTScore	0.298	0.325	0.152	0.122
BLEURT	0.216	0.261	0.065	0.054
Ours	**0.480**	**0.466**	0.451	0.465

Table 4: Results on USR-TopicalChat and USR-PearsonaChat (Mehri and Eskenazi, 2020c).

7.2 Ablation study

We perform an ablation study to investigate the influence of different components in our methods. We examine two ablative variants: (1) w/o $\mathcal{L}_{\text{PR}}$: we remove the ranking-based loss $\mathcal{L}_{\text{PR}}$ to verify its effectiveness (w/o $\mathcal{L}_{\text{PR}}$); (2) w/o $\mathcal{L}_{\text{GEN}}$: we remove the $\mathcal{L}_{\text{GEN}}$ to verify training with response generation task jointly can improve the predicting correlation with human judgment.

Table 3 presents the results. Overall, the variants of our methods show a decreased performance compared to the base model. For example, Pearson drops 0.10, 0.09, and 0.07 in three benchmarks, respectively, after the $\mathcal{L}_{\text{PR}}$ is removed. This result indicates that ranking-based loss can enhance performance by explicitly building the relation between response and reference. After removing the $\mathcal{L}_{\text{GEN}}$, the correlation in all benchmarks has a prominent decrease, e.g., Spearman correlation drops by 0.15, 0.10, and 0.09, respectively. The results suggest that the auxiliary response generation task improves the representation capability of our method and relieves the one-to-many problem.

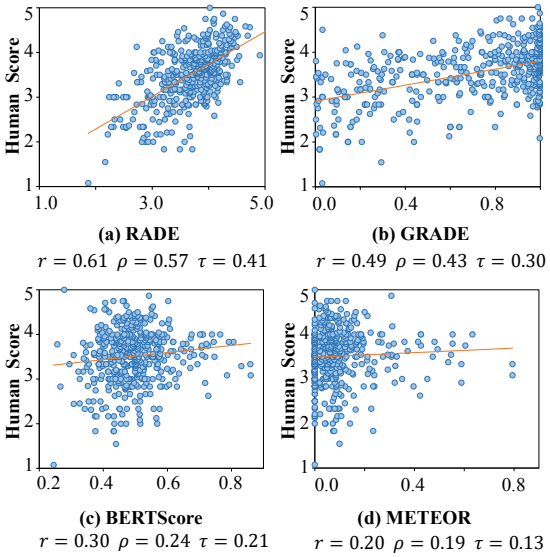

Figure 3: Score correlation of automatic evaluation and human evaluation on the EmpaDial domain. The horizontal axis indicates the different automatic evaluation methods, and the vertical axis indicates human rating.

7.3 Case study

Our case studies demonstrate that RADE is more consistent with human judgment than baselines. Details about our case studies are available in Appendix A.2.5.

7.4 Qualitative analysis

To explain more intuitively, we show the scatter plots against human judgments for different automatic evaluation methods (i.e., RADE, GRADE, BERTScore, METEOR) on the EmpaDial dataset in Figure 3. As shown in Figure 3 (a), our method RADE achieves a stronger correlation with human judgment than the other methods. Figure 3 (d) illustrates that METEOR scores are zero or extremely low for the most response. It results from the one-to-many nature of open-domain dialogue, and word overlapping occasionally occurs. Figure 3 (c) suggests that the BERTScore scores are mainly concentrated in the range of 0.3-0.6, indicating no significant differentiation between the different responses. Figure 3 (b) shows that GRADE achieves a better

correlation with human judgments. However, the distribution of GRADE predicted scores is concentrated in the high-scoring band, resulting in a low distinction of responses; RADE uses reference as a benchmark and thus has a more balanced distribution of predicted scores.

8 Discussions

The impact of the training data scale. To explore the minimum data scale required for our method, we train RADE using different amounts of randomly sampled annotated data. We observe a minor degradation in RADE's performance as the amount of data decreases. For example, when training on 2,400 examples from the EmpatheticDialogue dataset, RADE(TS) achieves Pearman'r=0.837 and Spearman'rho=0.829; whereas with 1,200 examples, it obtains Pearman'r=0.807 and Spearman'rho=0.806. All results are averaged over three runs. Moreover, we find that RADE outperforms all baselines with only 800 training examples in three datasets, respectively.

The difference between golden and candidate Responses. *Golden response* refers to a scenario where there is only one correct response, and any different response is given a low score. For example, BERTScore calculates the cosine similarity between the golden and model-generated response. However, *Candidate responses* implies that there can be multiple correct answers, which is more flexible and human-intuitive. And RADE is optimized to align with this human intention using generative and pairwise-ranking loss. If more references are available, the RADE can consider multiple valid responses to make more reliable evaluations. To achieve this, we can concatenate model-generated responses with different references. However, due to the limitation of our datasets, we concatenate one reference and model-generated response, which are then fed to the encoder.

Employing RADE when the reference response is not available. Considering the reference is not always available in real-world scenarios, we design two alternatives to enable RADE, i.e., constructing a pseudo-reference via retrieval or generative method. We verify the two solutions on the FED dataset and the details can be found in Appendix A.3.

9 Conclusion

We have presented a new reference-assist dialogue evaluation (RADE) method to address the one-to-many problem when evaluating open-domain dialogue systems. RADE evaluates the response generated by open-domain dialogue agents with the assistance of reference response. In addition, we have curated the reference-assisted dialogue evaluation datasets by expanding three existing datasets via a pairwise human annotation. The extended datasets contain over 10K dialogues. Extensive experiments on three extended datasets and two existing benchmarks have verified the effectiveness and robustness of the proposed methods and their generalizability.

Limitations

The main limitation of this paper is the need for human-labeled reference responses. We will explore automated or human-machine collaboration methods to reduce the cost of annotation in the next stage. Another limitation is that we need to explore whether other auxiliary tasks can also enhance the performance of score prediction. In the future, we also plan to reproduce the proposed method for other, less resource-rich languages.

Ethics Statement

The paper proposes a dialogue evaluation method, which is intended to evaluate open-ended dialogue on topics such as books and movies. A new dataset is developed using some existing dialogue systems, such as DialoGPT, which are trained on large-scale web data that is known to contain biased or discriminatory content. The datasets that we trained on may also include subjective knowledge (comments on movies) that may express the bias of the writers.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL*.

Zhangming Chan, Lemao Liu, Juntao Li, Haisong Zhang, Dongyan Zhao, Shuming Shi, and Rui Yan. 2021. Enhancing the open-domain dialogue evaluation in latent space. In *ACL*.

Maxime De Bruyn, Ehsan Lotfi, Jeska Buhmann, and Walter Daelemans. 2020. Bart for knowledge grounded conversations. In *KDD*.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. On the limitations of reference-free evaluations of generated text. *ArXiv*.

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*.

Mihail Eric, Nicole Chartier, Behnam Hedayatnia, Karthik Gopalakrishnan, Pankaj Rajan, Yang Liu, and Dilek Hakkani-Tur. 2021. Multi-sentence knowledge selection in open-domain dialogue. In *ACL*.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020a. Dialogue response ranking training with large-scale human feedback data. In *EMNLP*.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020b. Dialogue response ranking-training with large-scale human feedback data. In *EMNLP*.

Sarik Ghazarian, Johnny Wei, Aram Galstyan, and Nanyun Peng. 2019. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. In *NAACL*.

Sarik Ghazarian, Ralph Weischedel, Aram Galstyan, and Nanyun Peng. 2020. Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems. In *AAAI*.

Chiori Hori and Takaaki Hori. 2017. End-to-end conversation modeling track in dstc6. *arXiv preprint arXiv:1706.07440*.

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. Grade: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *EMNLP*.

Tianbo Ji, Yvette Graham, Gareth Jones, Chenyang Lyu, and Qun Liu. 2022. Achieving reliable human assessment of open-domain dialogue systems. In *ACL*.

Baber Khalid and Sungjin Lee. 2022. Explaining dialogue evaluation metrics using adversarial behavioral analysis. In *NAACL*.

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. *TOIS*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*.

Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2020a. EmpDG: Multi-resolution interactive empathetic dialogue generation. In *COLING*.

Qintong Li, Pijian Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2020b. Knowledge bridging for empathetic dialogue generation. In *AAAI*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *IJCNLP*.

Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021. Conversations are not flat: Modeling the dynamic information flow across dialogue utterances. In *ACL*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *ACL*.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. MoEL: Mixture of empathetic listeners. In *EMNLP*.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike 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. In *EMNLP*.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic Turing test: Learning to evaluate dialogue responses. In *ACL*.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Mime: Mimicking emotions for empathetic response generation. In *EMNLP*.

Shikib Mehri and Maxine Eskenazi. 2020a. Unsupervised evaluation of interactive dialog with dialogpt. In *SIGDIAL*.

Shikib Mehri and Maxine Eskenazi. 2020b. Unsupervised evaluation of interactive dialog with dialogpt. In *ACL*.

Shikib Mehri and Maxine Eskenazi. 2020c. USR: An unsupervised and reference free evaluation metric for dialog generation. In *ACL*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Vitou Phy, Yang Zhao, and Akiko Aizawa. 2020. Deconstruct to reconstruct a configurable evaluation metric for open-domain dialogue systems. In *COLING*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: a new benchmark and dataset. In *ACL*.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *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. *TACL*.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *EMNLP*.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *ACL*.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *CoRR*.

Koustuv Sinha, Prasanna Parthasarathi, Jasmine Wang, Ryan Lowe, William L. Hamilton, and Joelle Pineau. 2020. Learning an unreferenced metric for online dialogue evaluation. In *ACL*.

Fengyi Tang, Lifan Zeng, Fei Wang, and Jiayu Zhou. 2021. Persona authentication through generative dialogue. *ArXiv*.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2017. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *AAAI*.

Zheng Ye, Liucun Lu, Lishan Huang, Liang Lin, and Xiaodan Liang. 2021a. Towards quantifiable dialogue coherence evaluation. In *ACL*.

Zheng Ye, Liucun Lu, Lishan Huang, Liang Lin, and Xiaodan Liang. 2021b. Towards quantifiable dialogue coherence evaluation. In *ACL*.

Chen Zhang, Yiming Chen, Luis Fernando D'Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and Haizhou Li. 2021. DynaEval: Unifying turn and dialogue level evaluation. In *ACL*.

Chen Zhang, L. F. D'Haro, Rafael E. Banchs, Thomas Friedrichs, and Haizhou Li. 2020. Deep am-fm: Toolkit for automatic dialogue evaluation. In *IWSDS*.

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? In *ACL*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019a. Bertscore: Evaluating text generation with bert. *ICLR*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B. Dolan. 2019b. Dialogpt : Large-scale generative pre-training for conversational response generation. In *ACL*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Peng Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *EMNLP*.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *CCL*.

A Appendix

A.1 Human Evaluation Details

A.1.1 Details for Data Preparation

We first employ the generation models to generate one more response for our human annotation proposed in Section 3. The annotators are instructed to rate the newly generated responses. Specifically, we employ the following generation model:

- **Blenderbot** (Roller et al., 2021): Blender is a conversational agent based on the large-scale model that mainly focuses on generating personal, engaging, knowledgeable, and empathetic responses.
- **DialogGPT** (Zhang et al., 2019b): DialogGPT is a large, tunable neural conversational response generation model.
- **KEMP** (Li et al., 2020b): KEMP is an emotional dialogue agent enhanced with a knowledge-enriched context graph.
- **MoEL** (Lin et al., 2019): MoEL is an emotional dialogue agent based on encoder-decoder architecture. MoEL softly combines the response representation from different decoders, each focusing on one type of emotion.
- **MIME** (Majumder et al., 2020): MIME is an empathetic dialogue model considering polarity-based emotion clusters and emotional mimicry.
- **EmpDG** (Li et al., 2020a): EmpDG is a multi-resolution empathetic chatbot enhanced by exploiting user feedback.
- **PersonaGPT** (Tang et al., 2021): PersonaGPT is a GPT2-based open-domain dialogue agent designed to generate personalized responses.

As shown in Table 5, we extend the DSTC dataset with *Blenderbot* and *DialoGPT*, the Empathetic Dialogue dataset with *KEMP*, *MoEL*, *MIME* and *EmpDG*; the Persona-Chat dataset with *Blenderbot* and *PersonaGPT*.

Since Roller et al. points out the length of the utterances is crucial to human judgments, i.e., too short responses are seen as dull, we only sample the example with at least two turn interactions with an average length of utterance no more than 25 vocab. And we randomly split the train-dev-test of collected datasets as Chitchat (1490/300/300, 5/1/1), Empathetic Dialogue (3022/500/500, 6/1/1), Persona Chat (3000/500/500, 6/1/1).

Model	DSTC	EmpaDial	PersonaChat
Blenderbot	812		500
DialoGPT	1278		500
KEMP		3014	
MoEL		231	
MIME		242	
EmpDG		535	
PersonaGPT			3000

Table 5: The data distribution of seven well-performing dialogue models, which are used for extend corresponding dataset.

A.1.2 Annotation Guideline

Table 6 provides detailed instructions for the annotators to s help them understand the setting of our annotation task.

Annotation Guideline
Instruction
You need to read the context for each conversation to understand the specific context. Afterward, compare the two responses and determine which is better on the given metric. Since we have given a score to the reference response, you should take it as the benchmark and rate the generated response.
Dataset
(1) context: The historical interaction between two partners. (2) (reference,s_h): The reference response and corresponding score. (3) response: The response generated via agent which you need to rate.
Rating Details
(1) If the generated responds is better, the scores you give should be more than s_h. (2) If the generated responds is worse, the scores you give should be less than s_h. (3) If there is no significant difference between the two response, you can give the same score as s_h.

Table 6: The guideline used for our human annotation.

A.1.3 User Study

The dialogue can be evaluated from multiple perspectives. Some perspectives are universal to assess all dialogue agents, e.g., fluency, and relevance, while the other metrics are only used for task-specific dialogue agents. For example, the emotion-aware is a critical property for empathetic dialogue but is less important for persona dialogue. Therefore, we first simplify by sorting the possible aspects into two categories, i.e., the general view and the task-specific view. The former contains rel-

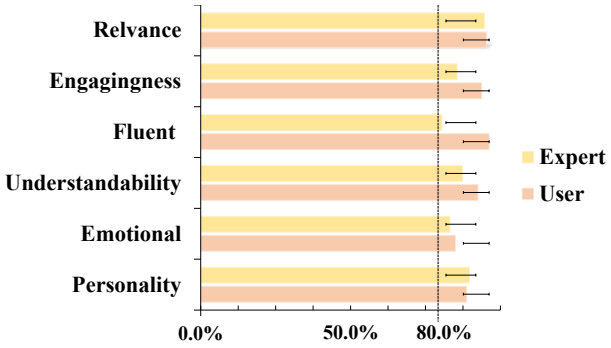

Figure 4: Result of two-role user study.

evance, engagingness, and fluency, while the latter consists of understandability, emotion-aware, and personality-aware, which correspond to chitchat dialogue, emotional dialogue, and persona dialogue. To understand the relation between sub-metrics and overall quality, we conduct a user study to learn their preference for different sub-metrics. Specifically, we invite 20 experts and 80 users, each of whom is asked to select the four most important ones from the sub-metrics. The results are listed in Figure 4. The approval rates reflect the user preference for different sub-metrics, which can be used as a weight to calculate the overall score. Moreover, we apply the softmax function on these weights to make them more interpretable.

A.2 Experiment Details

A.2.1 Datasets for Pre-train Stage

Our training process includes two stages, e.g., cross-domain pre-train and task-specific finetune. We first pre-train the model on diverse open-domain dialogue datasets as listed in Table 7 with the objective $\mathcal{L}_{\text{cross}}$. The next stage relies on task-specific dataset with the objective $\mathcal{L}_{\text{in}}$ (see in section 5).

These datasets are collected from https://github.com/e0397123/dstc10_metric_track, which contain a variety of open-domain dialogue, such as emotional dialogue, personalized dialogue, knowledge-grounded dialogue, and chitchat. Every example in the datasets contains the dialogue *context*, *response* generated by dialogue agent, pre-created *reference* response, and the *score* of generated response which has been annotated for at least three people from several perspectives. We use cross-domain datasets for pre-training to improve the robustness and generalisability of the models across different evaluation scenarios.

Table 7: Statistics of our datasets used for pre-train stage. AVG. Utts: the average of utterances per dialogue; AVG. Words : the average of words per dialogue.

Dataset	Dialogue	AVG. Utts	AsVG. Words
DSTC6-Eval	33,795	2.63	11.36
DSTC7-Eval	9,711	3.83	13.40
DSTC10-Eval	9,291	4.00	14.15
JSALT-Eval	741	3.47	17.12
PersonaChat-Zhao	900	5.13	11.77

A.2.2 Experimental Details on Our Benchmarks

We show the details of our automatic evaluation experiments in Table 9. The BERTScore and BLEURT are computed based on the large version of Roberta. As in Section 6, we implement two reference-based baselines, BERT_{MLP} and BART_{MLP}, using the same human-annotated datasets as RADE for training, and provide a reasonable comparison with our proposed model. Specifically, the BERT_{MLP} is built on the base version of BERT (Devlin et al., 2018), while the BART_{MLP} is built on the base version of BART (Lewis et al., 2019).

A.2.3 More Fair Comparison after Training

For a fair analysis, we pre-train the two strongest baselines (QuantiDCE and GRADE) with our cross-domain dataset. GRADE achieves Pearman'r=0.383, 0.378, -0.122, and QuantiDCE achieves Pearman'r=0.408, 0.522, 0.238 in the ChitChat, EmpatheticDialogue, and Personachat datasets. However, our proposed RADE(PT) remains the best results (Pearman'r=0.472, 0.650, 0.386). We further fine-tune GRADE and QuantiDCE with our self-collected datasets for a more comprehensive analysis. GRADE achieves Pearman'r=0.413, 0.430, -0.013, and QuantiDCE achieves Pearman'r=0.458, 0.589, 0.278 in three datasets, underperforming the proposed RADE(TS) (Pearman'r=0.601, 0.863, 0.470).

We skip pre-training/fine-tuning four baselines for the following reasons: (1) UniEval and QuestionEval have been pre-trained on multiple datasets across various domains. (2) The FED metric is unsupervised (cf. Shikib Mehri et al.) (3) The DialoRPT has been trained on a sizeable human-feedback dataset (133M) covering various domains. These analyses validate the superiority of our method.

Methods	USR-TopicalChat		USR-Pearsonachat		DailyDialogue	
	Pearson'r	Spearman'ρ	Pearson'r	Spearman'ρ	Pearson'r	Spearman'ρ
Reference-free methods						
MAUDE (Sinha et al., 2020)	0.044*	0.083*	0.345	0.298	-0.036*	-0.073*
FED (Mehri and Eskenazi, 2020b)	-0.124	-0.135	-0.028*	-0.000*	-0.080*	0.064*
HolisticEval (Liang et al., 2022)	-0.147	-0.123	0.087*	0.113*	0.025*	0.020*
FlowScore (Li et al., 2021)	0.095*	0.082*	0.118*	0.079*	-	-
QuestEval (Scialom et al., 2021)	0.300	0.338	0.176	0.236	0.020*	0.006*
USR (Mehri and Eskenazi, 2020c)	0.412	0.423	0.440	0.418	0.057*	0.057*
GRADE (Huang et al., 2020)	0.200	0.217	0.358	0.352	0.278	0.253
PredictiveEngage (Ghazarian et al., 2020)	0.222	0.310	-0.003*	0.033*	-0.133*	-0.135
DialogRPT (Gao et al., 2020b)	0.120	0.105*	-0.064*	-0.083*	-0.000*	0.037*
DynaEval (Zhang et al., 2021)	-0.032*	-0.022*	0.149	0.171	0.108*	0.120*
DEB (Sai et al., 2020)	0.180	0.116	0.291	0.373	0.337	0.363
USL-H (Mehri and Eskenazi, 2020c)	0.322	0.340	**0.495**	**0.523**	0.108*	0.093*
Reference-based lexicon-level methods						
BLEU-4 (Papineni et al., 2002)	0.216	0.296	0.135	0.090*	0.075*	0.184
METEOR (Banerjee and Lavie, 2005)	0.336	0.391	0.253	0.271	0.093*	0.010*
ROUGE-L (Lin, 2004)	0.275	0.287	0.066*	0.038*	0.154	0.147
Reference-based semantic-level methods						
RUBER (Tao et al., 2017)	0.247	0.259	0.131	0.190	-0.084*	-0.094*
BERT-RUBER (Tao et al., 2017)	0.342	0.348	0.266	0.248	0.134	0.128
BERTScore (Zhang et al., 2019a)	0.298	0.325	0.152	0.122*	0.129	0.100*
Deep AM-FM (Zhang et al., 2020)	0.285	0.268	0.228	0.219	0.026*	0.022*
ADEM (Lowe et al., 2017)	-0.060*	-0.061*	-0.141	-0.085*	0.064*	0.071*
BLEURT (Sellam et al., 2020)	0.216	0.261	0.065*	0.054*	0.176	0.133
PONE (Lan et al., 2020)	0.271	0.274	0.373	0.375	0.163	0.163
Reference-assist						
Ours (Pretrain-train model, PT)	**0.480**	**0.466**	0.451	0.465	**0.356**	**0.370**

Table 8: **Results on USR-TopicalChat, USR-PearsonaChat and Grade-DailyDialogue.** We divide the methods in Reference-free, Reference-based and REDE, while the reference-based methods including semantic-level and lexicon-level. The metrics r, ρ, and τ indicate the Pearson's ρ, Spearman's r, and Kendall'sτ. All values are statistically significant to p-value < 0.05 unless marked by *. We underline the best results of each group of baselines methods and **bold** the best results of all methods.

A.2.4 Results on Existing Benchmarks

We further examine three existing benchmarks, i.e., USR-TopicalChat, USR-PersonaChat and Grade-DailyDialogue to verify the efficiency and robustness of RADE when generalizing to agnostic scenarios. USR-TopicalChat and USR-PersonaChat datasets are collected to assess dialog evaluation metrics, with examples containing the dialogue *context*, *reference*, *response* and corresponding *scores*, which three people have annotated. The Grade-DailyDialogue contains high-quality open-domain conversations about daily life including diverse topics. And the results are summarized in Table 8.

The experimental results show that RADE outperforms the state-of-the-art reference-free and reference-based methods on the USR-TopicalChat dataset. For example, we push the Pearson correlation to 48.0% (7% definite improvement) and Spearman correlation to 46.6% (4% absolute improvement). Moreover, RADE shows a stronger correlation with human judgment than existing reference-based methods on the second dataset. It achieves comparable, even better results with the reference-free methods except for USL-H. The results demonstrate that our pre-trained model is more robust even under agnostic scenarios.

We also compare the two existing methods, and the results suggest a similar phenomenon as 3. Firstly, the reference-free methods achieve better consistency than reference-based methods, i.e., the former has the highest result of $r = 41.2\%$, $\rho = 42.3\%$ while the latter gets $r = 34.2\%$, $\rho = 34.8\%$ on the USR-TopicalChat dataset. However, the reference-free methods suffer from more significant variance. For example, the MAUDE gets $r = 0.345\%$ and $\rho = 0.298\%$ on the USR-PearsonChat dataset but gets $r = 0.044\%$ and $\rho = 0.083\%$ on the USR-TopicChat dataset. It indicates that reference-free methods are more vulnerable and prone to data-induced bias.

A.2.5 Case Study

To explain more intuitively, we show examples of automatic evaluation and them with human judgment in Table 10, 11, 12, suggesting that the scores of our methods are closer to human ratings.

A.3 Presudo reference

Since the original FED does not provide the reference response, we construct a pseudo-reference via retrieval or generative method. The former retrieves reference from a curated response corpus based on our cross-domain datasets via BM25 with the dialogue context as the query. The latter generates via a large language model GPT-3 based on the dialogue context. The results show that RADE(PT) obtains Pearman'r=0.381 and Spearman'rho=0.368 with the retrieved reference while achieving Pearman'r=0.343, Spearman'rho=0.347 with generative reference, outperforming the state-of-the-art baseline (QuantiDCE, Pearman'r=0.319, Spearman'rho=0.323).

To further validate the generalizability of our method, we evaluate our proposed RADE(PT) on another challenging benchmark, GRADE-Dailydialogue. Our RADE(PT) achieves Pearman'r=0.356 and Spearman'rho=0.370 with 5% and 2% relative improvements compared to state-of-the-art baseline, indicating that our method can generalize to more challenging benchmarks.

Methods	ChitChat			Empathetic Dialogue			PersonaChat		
	r	ρ	τ	r	ρ	τ	r	ρ	τ
Reference-free methods									
FED_E (Mehri and Eskenazi, 2020b)	0.241	0.254	0.177	0.202	0.218	0.218	0.138	0.120	0.086
FED_U (Mehri and Eskenazi, 2020b)	0.235	0.248	0.171	0.147	0.156	0.106	0.145	0.162	0.117
QuesEval (Scialom et al., 2021)	0.045	0.021	0.013	0.069	0.084	0.057	-0.003	0.034	0.0237
UniEval (Zhong et al., 2022)	0.456	0.470	0.312	0.403	0.435	0.286	0.306	0.338	0.244
DialoRPT (Gao et al., 2020b)	-0.066*	-0.044*	-0.031*	0.267	0.244	0.166	-0.077*	-0.069*	-0.049*
GRADE (Huang et al., 2020)	0.491	0.434	0.300	0.549	0.568	0.398	-0.031*	-0.005	-0.030*
QuantiDCE(R) (Ye et al., 2021b)	0.348	0.300	0.202	0.498	0.507	0.351	0.162	0.182	0.130
QuantiDCE(P) (Ye et al., 2021b)	0.408	0.387	0.234	0.522	0.521	0.372	0.238	0.257	0.189
QuantiDCE(F) (Ye et al., 2021b)	0.458	0.427	0.265	0.589	0.577	0.436	0.278	0.326	0.237
Reference-based lexicon-level methods									
ROUGE-1 (Lin, 2004)	0.217	0.192	0.133	0.221	0.217	0.151	0.116	0.101	0.069
ROUGE-2 (Lin, 2004)	0.210	0.145	0.148	0.009*	0.046	0.058	0.065	0.040	0.032
ROUGE-L (Lin, 2004)	0.215	0.178	0.129	0.213	0.214	0.148	0.118	0.114	0.079
BLEU-1 (Papineni et al., 2002)	0.201	0.190	0.131	0.115	0.118	0.076	0.010	0.081	0.055
BLEU-2 (Papineni et al., 2002)	0.201	0.200	0.158	0.057	0.041*	0.032	0.060	0.039	0.031
BLEU-3 (Papineni et al., 2002)	0.201	0.189	0.153	0.049	0.036	0.030*	0.017	-0.001*	-0.001*
BLEU-4 (Papineni et al., 2002)	0.203	0.207	0.169	0.059	0.056	0.046	0.017	-0.005*	-0.004*
METEOR (Banerjee and Lavie, 2005)	0.202	0.188	0.129	0.182	0.194	0.132	0.099	0.051	0.035
Reference-based semantic-level methods									
$\text{Bertscore}_\text{p}$ (Zhang et al., 2019a)	0.347	0.334	0.334	0.229	0.146	0.104	-0.446	-0.089	-0.061*
$\text{Bertscore}_\text{r}$ (Zhang et al., 2019a)	0.296	0.243	0.213	0.167	0.243	0.173	0.278	0.292	0.196
$\text{Bertscore}_\text{f1}$ (Zhang et al., 2019a)	0.229	0.308	0.213	0.211	0.204	0.145	0.133	0.115	0.079
BARTScore (Lewis et al., 2019)	0.133	0.057	0.039	0.256	0.253	0.173	0.143	0.168	0.115
RUBER (Tao et al., 2017)	0.332	0.351	0.369	0.252	0.256	0.183	0.122	0.123	0.089
BLEURT (Sellam et al., 2020)	0.353	0.363	0.249	0.343	0.337	0.232	0.105	0.140	0.102
$\text{BERT}_\text{MLP}^\dagger$ (Devlin et al., 2018)	0.241	0.255	0.173	0.186	0.225	0.153	0.274	0.330	0.202
$\text{BERT}_\text{MLP}^\dagger$ (Devlin et al., 2018)	0.304	0.301	0.192	0.501	0.537	0.373	0.331	0.360	0.251
$\text{Roberta}_\text{MLP}^\dagger$ (Zhuang et al., 2021)	0.275	0.306	0.300	0.285	0.307	0.307	0.317	0.334	0.223
$\text{BART}_\text{MLP}^\dagger$ (Lewis et al., 2019)	0.431	0.440	0.312	0.412	0.447	0.356	0.310	0.335	0.242
Reference-assisted methods									
RADE (Pre-trained model, PT)	0.472	0.491	0.334	0.650	0.601	0.427	0.386	0.390	0.285
RADE (Task-specific model, TS)	**0.601**	**0.569**	**0.409**	**0.863**	**0.849**	**0.685**	**0.470**	**0.465**	**0.347**

Table 9: **Details** of our automatic evaluation experiment on three benchmarks. We divide the methods in Reference-free, Reference-based and RADE, while the reference-based including the semantic-level and lexicon-level methods. Note that r, ρ and τ indicate the Pearson's ρ, Spearman's r and Kendall'sτ. All values are statistically significant to p-value < 0.05, unless marked by*. The FED_E and FED_U indicate two evaluation perspective of FED, i.e., engagement and understandability. Methods with † are implemented by ourselves. We underline the best results of each group of baselines methods and **bold** the best results of all methods.

Topic	Library of Alexandria					
Context	User1: The library of Alexandria had a unique way of gathering books by making all ships visiting give their books for copy. User2: They must have had a impressive collection! User1: How unfair, they would copy their books give them back the copy and keep the originals.					
Reference	I guess that is true. Do you think we will ever have a centralized digital library of all our works and knowledge?					
Response	That is exactly why they didn't stick around and stay put. I wish I could work somewhere where I could share the workload.					
Evaluation	Human (1-5)	QuantiDCE (1-3)	GRADE (0-1)	METEOR (0-1)	BERTScore (0-1)	RADE (1-5)
	3.67	3.00	0.26	0.17	0.48	3.54

Table 10: An example of the DSTC dataset evaluated by five automatic and human evaluation methods. This example shows that our methods correlate more strongly with human judgment than the other methods.

Emotion	Confident					
Context	User1:I love when I need to show someone how to do something and everything goes off without a hitch. User2:Oh yeah, that is always nice. Sometimes they just do not get it, though. User1:Fortunately, it is usually just not having the best answer for what they ask. I have learned and prepared a lot more this time.					
Reference	Wow, you sound pretty confident! You must really know what you are talking about.					
Response	I agree, I am sure you will do great things.					
Evaluation	Human (1-5)	QuantiDCE (1-3)	GRADE (0-1)	METEOR (0-1)	BERTScore (0-1)	RADE (1-5)
	3	3.00	0.45	0.04	0.57	3.64

Table 11: An example of the EmpatheticDialogue dataset evaluated by five automatic and human evaluation methods.

Emotion	Excited					
Context	User1: I am looking forward to my best friend s surprise birthday party next week! User2: That sounds like so much fun! I love parties! User1: I am really happy about it. She is a great friend and she is turning 40, so it is a big one!					
Reference	Hey, I just had that one! What do you have planned?					
Response	That is great. Do you have any other day?					
Evaluation	Human (1-5)	QuantiDCE (1-3)	GRADE (0-1)	METEOR (0-1)	BERTScore (0-1)	RADE (1-5)
	2.98	2.7	0.29	0.34	0.61	2.64

Table 12: An example of the EmpatheticDialogue dataset evaluated by five automatic and human evaluation methods.

Personality	(1): I like to donate time and money to charity. (2): I work as a computer programmer. (3): i make over 200 , 000 per year.					
Context	User1: Hi, how are you ? tell me something about yourself! User2: Well, I love going fishing with my four attractive daughters every day. User1: Sounds fun! I enjoying volunteering and donating to charities. User2: Cool maybe you'd like to run a charity at my new race track. I race cars! User1: Sounds exciting! I am a computer programmer, which pays over 200k a year.					
Reference	Would you like to marry one of my four attractive daughters? I will sell one.					
Response	wow! That's a lot of money. Do you have any hobbies?					
Evaluation	Human (1-5)	QuantiDCE (1-3)	GRADE (0-1)	METEOR (0-1)	BERTScore (0-1)	RADE (1-5)
	3.25	2.87	0.02	0.06	0.53	2.91

Table 13: An example of the PersonaChat dataset evaluated by five automatic and human evaluation methods, showing that our methods have a stronger correlation with human judgment than the other methods.

Personality	(1): I like cupcakes and reading. (2): I work at a grocery store. (3): My favorite color is blue.					
Context	User1: Hi! what are you up to ? I am doing what I like best, eating cupcakes! User2: Hi I am winding down from a day at work. User1: So am I. The local grocery store is where I work. What about you? User2: I also work in the retail industry at the local department store! User1: Other than eating cupcakes, reading is also what I like to do to wind down.					
Reference	I like to read also and play with my dog. Do you have a pet?					
Response	What do you do for fun? My girlfriend and I go to the lake a lot.					
Evaluation	Human (1-5)	QuantiDCE (1-3)	GRADE (0-1)	METEOR (0-1)	BERTScore (0-1)	RADE (1-5)
	2.75	3.00	0.01	0.22	0.58	2.79

Table 14: An example of the PersonaChat dataset evaluated by five automatic and human evaluation methods. This example shows that our methods have a stronger correlation with human judgment than the other methods.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section Limitations

☑ A2. Did you discuss any potential risks of your work?
Section Ethics Statement

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
No response.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

Section 6 and 7

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 6

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 6

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 7

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 6

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 4 and Appendix A

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix A

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 4 and Appendix A

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Section 4

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Section 4

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Left blank.