PRELIMINARY PREPRINT VERSION: DO NOT CITE
The AAAI Digital Library will contain the published
version some time after the conference.

MDD-Eval: Self-Training on Augmented Data for Multi-Domain Dialogue Evaluation

Chen Zhang[1,2], **Luis Fernando D'Haro**[4], **Thomas Friedrichs**[2], **Haizhou Li**[1,3,5]

[1] National University of Singapore [2] Robert Bosch (SEA), Singapore [3] Kriston AI Lab, China
[4] Universidad Politécnica de Madrid, Spain [5] The Chinese University of Hong Kong (Shenzhen), China
{chen_zhang,haizhou.li}@u.nus.edu, luisfernando.dharo@upm.es, thomas.friedrichs@sg.bosch.com

Abstract

Chatbots are designed to carry out human-like conversations across different domains, such as general chit-chat, knowledge exchange, and persona-grounded conversations. To measure the quality of such conversational agents, a dialogue evaluator is expected to conduct assessment across domains as well. However, most of the state-of-the-art automatic dialogue evaluation metrics (ADMs) are not designed for multi-domain evaluation. We are motivated to design a general and robust framework, MDD-Eval, to address the problem. Specifically, we first train a teacher evaluator with human-annotated data to acquire a rating skill to tell good dialogue responses from bad ones in a particular domain and then, adopt a self-training strategy to train a new evaluator with teacher-annotated multi-domain data, that helps the new evaluator to generalize across multiple domains. MDD-Eval is extensively assessed on six dialogue evaluation benchmarks. Empirical results show that the MDD-Eval framework achieves a strong performance with an absolute improvement of 7% over the state-of-the-art ADMs in terms of mean Spearman correlation scores across all the evaluation benchmarks.

1 Introduction

Recent years have witnessed growing interests in open-domain dialogue systems (Adiwardana et al. 2020; Zhang et al. 2020; Roller et al. 2021). With the increasing availability of high-quality dialogue corpora (Li et al. 2017; Zhang et al. 2018) and advancement of neural architectures (Devlin et al. 2019; Radford et al. 2019), learning-based dialogue systems are becoming possible. The applications call for dialogue technology capable of generating appropriate responses to users' prompts in a diverse range of scenarios, such as general chit-chat (Li et al. 2017), knowledge exchange (Gopalakrishnan et al. 2019), persona-based chat (Zhang et al. 2018), and emotion disclosure (Rashkin et al. 2019).

However, the dialogue research heavily relies on the ability to evaluate system performance with automatic dialogue evaluation metrics (ADMs). Common natural language generation (NLG) metrics used in the dialogue system literature, such as BLEU (Papineni et al. 2002) and ROUGE (Lin 2004), are unsuitable for the multi-domain dialogue evaluation task as they are shown to correlate poorly with human judgements (Liu et al. 2016) due to the one-to-many context-response mapping in dialogues (Zhao, Zhao, and Eskenazi 2017) as well as the multi-faceted nature of dialogue evaluation (Mehri and Eskenazi 2020b).

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Metric	DailyDialog-Eval	Topical-Eval
DEB	0.486	0.116
GRADE	0.533	0.217
USR	0.367	0.423

Table 1: Spearman correlation scores of three state-of-the-art model-based metrics on two dialogue evaluation benchmarks.

An alternative solution is to design model-based ADMs that explicitly learn to discriminate dialogue responses of varying quality. Lately, many model-based ADMs leveraging self-supervised learning are proposed to address the weaknesses of the standard NLG metrics (Sai et al. 2020; Ghazarian et al. 2019; Mehri and Eskenazi 2020b; Huang et al. 2020; Zhang et al. 2021c). While these ADMs have demonstrated strong correlations with human judgements, they lack a generalized skill to evaluate dialogues across multiple domains. For example, in Table 1, DEB (Sai et al. 2020) and GRADE (Huang et al. 2020) are pretrained on the DailyDialog dataset (Li et al. 2017). They perform well on the DailyDialog-Eval (Zhao, Lala, and Kawahara 2020) benchmark that contains responses from dialogue systems trained on chit-chat content. However, their performance significantly drops when assessed on the Topical-Eval (Mehri and Eskenazi 2020b) benchmark, which is close in domain with TopicalChat (Gopalakrishnan et al. 2019) and contains dialogue responses from knowledge-grounded conversations. The reverse is true for USR (Mehri and Eskenazi 2020b), which is pretrained on the TopicalChat dataset.

To design robust ADMs for the multi-domain dialogue evaluation task, we consider two research questions. (1) How to equip the ADM with a rating skill to discriminate responses of varying quality? In other words, the ability to assign a high score to relevant responses and a low score otherwise. (2) How can an ADM learn the general knowledge across dialogue domains so as to generalize the evaluation skill? For the first question, the most direct and ef-

fective way is to learn from humans, i.e., the ADM can be trained with human-annotated dialogue data. As for the second question, the general knowledge can be learned on a large-scale multi-domain dialogue dataset. Ideally, If human annotations are available, an oracle multi-domain dialogue evaluator can be learned. However, performing large-scale human annotations is extremely expensive. Thus, we are motivated to explore semi-supervised learning for our task.

More specifically, we propose a multi-domain dialogue evaluation (MDD-Eval) framework under the self-training paradigm (Scudder 1965; Yarowsky 1995) where a teacher model, trained on human-annotated dialogue evaluation data, creates pseudo labels for unlabeled dialogue data. Then, the synthetically-labeled data are used to train a student model. To obtain the large-scale multi-domain unlabeled dialogue data, we leverage the dialogue data augmentation techniques that have been successfully applied in the self-supervised learning of ADMs, such as random utterance selection (Tao et al. 2018; Zhang et al. 2021c), mask-and-fill (Donahue, Lee, and Liang 2020; Gupta, Tsvetkov, and Bigham 2021) and back-translation (Edunov et al. 2018; Sinha et al. 2020). In this way, we expect that the student model carries the rating skill of the teacher model, and it can generalize across domains after being adapted on a large-scale multi-domain dataset with pseudo labels.

Overall, we make the following contributions:

- A model-based framework, named MDD-Eval, is proposed with a self-training scheme on augmented data. Its rating skill is trained on human-annotated data, and its cross-domain general knowledge is trained on machine-annotated data.
- We release a large-scale multi-domain dialogue dataset with machine annotations that facilitate ADM training. We name the dataset, MDD-Data.
- MDD-Eval attains an absolute improvement of 7% over the state-of-the-art ADMs in terms of mean Spearman correlation over six dialogue evaluation benchmarks.
- MDD-Data, MDD-Eval implementation, and pretrained checkpoints will be released to the public[1]. This allows practitioners and researchers to use and adapt MDD-Eval for automatic evaluation of their dialogue systems.

2 Related Work

2.1 Dialogue Evaluation Metrics

Human evaluation reflects the perceived quality of dialogue systems. However, it is expensive and time-consuming. For system development, we rely on ADMs for model design, hyperparameter tuning and system benchmarking (Yeh, Eskenazi, and Mehri 2021). The current trend of open-domain ADMs is shifting from the reference-based approach towards the model-based approach that is reference-free (Mehri and Eskenazi 2020a; Zhang et al. 2021a). In many ADM solutions, we predict the relatedness between a dialogue context and the generated responses by training a discriminative network to distinguish the original response from negative samples in a self-supervised fashion. Typical examples include RUBER (Tao et al. 2018), BERT-RUBER(Ghazarian et al. 2019), USR (Mehri and Eskenazi 2020b), GRADE (Huang et al. 2020), MaUdE (Sinha et al. 2020) and D-score (Zhang et al. 2021c).

A problem with the metrics learned with self-supervised learning is that the random negative-sampling strategy is likely to produce false-negative or over-simplistic candidates, thus introducing unwanted biases to the ADMs. One idea is to introduce adversarial irrelevant responses to increase the ADMs' discrimination capability (Sai et al. 2020; Gupta, Tsvetkov, and Bigham 2021; Park et al. 2021). In this way, the evaluation model will greatly benefit from a dataset of multiple relevant and adversarial irrelevant responses from diverse dialogue context. The existing methods are focused very much on how to design such a dataset. Along this line of thought, this work presents a novel strategy to learn the rating skill from one dataset first, then generalize the skill across multiple domains.

2.2 Self-Training

Self-training is a simple and effective semi-supervised approach, which incorporates a model's prediction on unlabeled data to obtain additional information. It has been shown effective in many tasks, such as image recognition (Yalniz et al. 2019), text generation (He et al. 2020), automatic speech recognition (Kahn, Lee, and Hannun 2020), and parsing (McClosky, Charniak, and Johnson 2006). There are two key ideas that contribute to the success of self-training: pseudo-labeling and consistency regularization.

Pseudo-labeling refers to the process of converting model predictions to hard labels (Lee et al. 2013). Usually, a confidence-based threshold is imposed to retain unlabeled examples only when the classifier is sufficiently confident (Sohn et al. 2020). In MDD-Eval, we apply pseudo-labeling together with the confidence-based threshold to bootstrap high-quality adversarial and random negative samples from the unlabeled data.

Consistency regularization was first proposed by (Bachman, Alsharif, and Precup 2014). It means that the prediction made by the classification model remains consistent even when the input or the model function is perturbed by a small amount of noise. Recently, the use of consistency regularization to modulate the self-training process has been shown to boost model performance on many image and text classification tasks (Xie et al. 2020a; Berthelot et al. 2020). We are motivated to incorporate consistency regularization into the learning of our dialogue evaluator, which is essentially learned with a text classification task.

Xie et al. (2020b) proposes Noisy Student and Sohn et al. (2020) proposes FixMatch frameworks. Both incorporate pseudo-labeling and consistency regularization into a unified framework. Noisy Student and FixMatch have demonstrated remarkable performance on image classification tasks, that motivates us to unify the pseudo-labeling and consistency regularization ideas in open-domain ADM training for the first time.

[1] https://github.com/e0397123/MDD-Eval

3 Methodology

In this section, we first define the multi-domain dialogue evaluation task (Section 3.1), then formulate MDD-Eval framework in three steps: (a) We pretrain a teacher model (Section 3.2) from a human-annotated dataset, to learn the rating skill to distinguish relevant responses from irrelevant ones. (b) We augment a large-scale multi-domain dataset for MDD-Eval self-training (Section 3.3). (c) We generalize the pretrained teacher model with the augmented data to derive a student model, which carries a generalized rating skill learned from the augmented data. (Section 3.4).

3.1 Problem Formulation

Formally, a dialogue context and the corresponding dialogue response can be denoted as c_i^j and r_i^j respectively. c_i^j and r_i^j are the i^{th} data pair drawn from the j^{th} dialogue evaluation benchmark D^j, where $j \in \{1, ..., J\}$, and $D^j \in D^J$ and $i \in \{1, ..., I\}$. There are J domains, each of which has I data pairs.

Our goal is to learn a metric, $M : (c_i^j, r_i^j) \rightarrow s_i^j$ where s_i^j is the metric score that indicates the quality of (c_i^j, r_i^j) as perceived by M. In addition, each (c_i^j, r_i^j) is annotated by several human judges and each human judge will provide a quality score based on the Likert scale[2] to indicate his or her perception of the quality of (c_i^j, r_i^j). We denote the mean human score given to (c_i^j, r_i^j) as q_i^j. Due to the multi-faceted nature of dialogue evaluation, the quality can refer to language fluency, coherence, topic relevance, logical consistency etc. Since the focus of our work is multi-domain dialogue evaluation instead of multi-dimensional evaluation, we fix the quality as *response appropriateness* here.

To assess the performance of M on D^j, the correlation score between $S = \{s_i^j, \ldots, s_I^j\}$ and $Q = \{q_i^j, \ldots, q_I^j\}$ are calculated. We use ρ_j to represent the correlation score on D^j. Higher ρ_j indicates better performance of the metric on D^j. In the multi-domain dialogue evaluation task, an effective M should achieve good correlation scores across all J domains. In other words, the desired M should obtain a good average correlation $\tilde{\rho} = \frac{1}{J}\sum_{j=1}^{J} \rho_j$.

3.2 Teacher Model

We first pretrain a model on human-annotated data in one particular domain, i.e., the teacher model, $M_{teacher}$, defined by the parameters $\theta_{teacher}$. Given a dialogue context-response pair, $M_{teacher}$ should accurately determine the degree of relevance between the context and the corresponding response. To equip the teacher model with a solid rating skill, we rely on a high-quality human-annotated base dataset $D^b \in D^J$. Note that D^b is from a single-domain, and of much smaller size than the data we would like to augment.

In dataset D^b, there are three categories of responses for a given context: random, adversarial and relevant. The relevant and adversarial responses are generated by human annotators. $M_{teacher}$ is trained on D^b to classify a context-response pair into one of the three categories:

$$\tilde{y}_i^b = f_{\theta_{teacher}}([c_i^b \circ r_i^b]) \quad (1)$$

with the objective function:

$$\min_{\theta_{teacher}} \frac{1}{|D^b|} \sum_{(c_i^b, r_i^b, y_i^b) \in D^b} \mathcal{L}_{CE}(\tilde{y}_i^b, y_i^b) \quad (2)$$

where $\circ$ denotes the concatenation operation. $\tilde{y}_i^b$ is the predicted class, y_i^b is the gold label for (c_i^b, r_i^b) and $\mathcal{L}_{CE}$ is the cross entropy loss.

$M_{teacher}$ plays three key roles: (1) providing pseudo labels to unlabeled context-response pairs, (c_i^*, r_i^*)[3], which are obtained with different dialogue data augmentation techniques. (2) facilitating the data selection process whereby false negatives and adversarial or random samples with low confidence scores as determined by $M_{teacher}$ are removed. (3) serving as a baseline in the evaluation task.

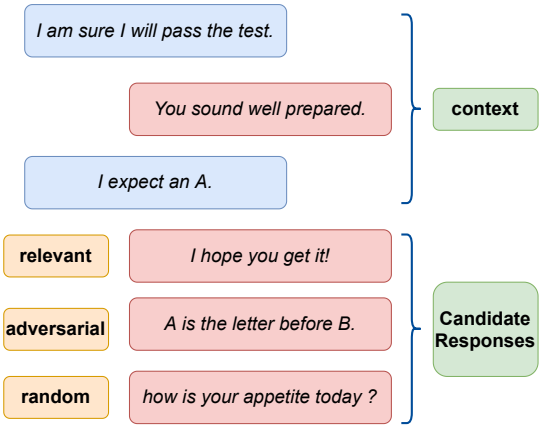

Figure 1: An example of a dialogue context with three candidate responses. $M_{teacher}$ is expected to annotate the context-response pairs as either relevant, adversarial or random.

3.3 Dialogue Data Augmentation

To generalize the teacher model across domains, we collect a multi-domain dataset, denoted as D^*, that contains a large amount of unlabeled context-response pairs. The unlabeled pairs will be automatically annotated in the same way as D^b by $M_{teacher}$. An example of a dialogue context with three candidate responses for annotation is presented in Figure 1. To construct such a dataset, we leverage the following dialogue data augmentation techniques:

Syntactic Perturbation Motivated by (Sinha et al. 2020), we have considered three variants of perturbations at the syntax level: (1) word-drop (a random portion of tokens in the response is dropped). (2) word-shuffle (the ordering of tokens in the response is randomly shuffled). (3) word-repeat (a random portion of tokens in the response is

[2] In the evaluation benchmarks used in our experiments, the Likert scale is from 1 to 5. The higher the better.

[3] $*$ means that the context-response pair can be drawn from dialogue corpora of any domain.

repeated multiple times). The syntactic perturbations are intended to simulate erroneous behaviours of some generative models in generating unnatural dialogue responses.

Back-Translation Back-translation (Edunov et al. 2018) augments a response by generating its syntactic variants. In practice, we adopt the pretrained WMT'19 English-German and German-English ensemble model to perform back-translation.

Generative Model Output State-of-the-art dialogue generators, such as DialoGPT (Zhang et al. 2020) and BlenderBot (Roller et al. 2021), have been pretrained on a large amount of conversation data and are demonstrating strong capability in generating fluent and on-topic responses. They help generate semantic variants of a response conditioned on the respective dialogue contexts.

Random Utterance Selection The random utterance selection is a simple and effective strategy that has been widely adopted in the self-supervised learning of dialogue evaluation metrics (Mehri and Eskenazi 2020b; Huang et al. 2020; Sai et al. 2020) to introduce irrelevant responses w.r.t. a dialogue context. Given a dialogue context, three variants of random utterance selection are adopted: (1) randomly sample a response from a different dialogue. (2) randomly sample a response from the entire pool of responses produced by the generative models. (3) randomly sample a response from the entire pool of responses obtained via back-translation.

Mask-and-fill Above-mentioned techniques tend to produce response candidates for the relevant and random class. The mask-and-fill strategy is adopted to automatically construct candidates for the adversarial class. Specifically, we adopt the Infilling by Language Modeling (ILM) framework (Donahue, Lee, and Liang 2020) to perform the mask-and-fill response augmentation. The process is as follows: given a context-response pair extracted from a natural human-human dialogue, one or a few contiguous tokens in the response are randomly replaced by the $[blank]$ placeholder. The modified response is input into the pretrained ILM model, which then generate tokens in an autoregressive manner. Subsequently, the $[blank]$ placeholder is substituted with the generated tokens to obtain a reconstructed view of the original response. The reconstructed response serves as an adversarial sample w.r.t. the dialogue context.

After obtaining the large number of context-response pairs, we apply the pretrained $M_{teacher}$ to provide soft pseudo labels to all the pairs. The soft pseudo label is a probability distribution over the three classes (random, adversarial and relevant). Then, a filtering process is implemented to improve the quality of pseudo-labeled D^*. A confidence threshold of 70% is applied to exclude pairs classified by $M_{teacher}$ with low confidence. Emprical evidence suggests that the 70% threshold provides a good balance between the quality and quantity of augmented data. Within D^*, the relevant

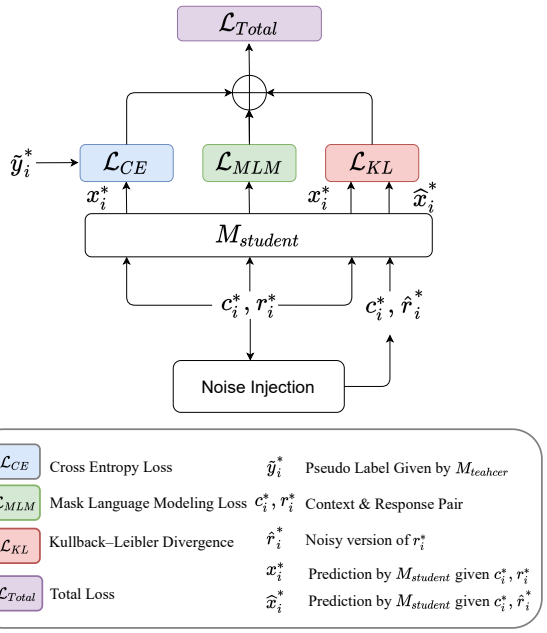

Figure 2: The training process of $M_{student}$. $\mathcal{L}_{Total}$ is the sum of three components: (1) The cross entropy loss $\mathcal{L}_{CE}$, which is computed between $\tilde{y}_i^*$ generated by $M_{teacher}$ and the prediction by $M_{student}$ for an input pair (c_i^*, r_i^*) . (2) The self-supervised MLM loss $\mathcal{L}_{MLM}$, for domain adaptation. (3) The KL Loss $\mathcal{L}_{KL}$, for consistency regularization.

set consists of filtered pairs obtained with back-translation and generative models in addition to the original context-response pairs extracted from dialogues of different dialogue corpora. The adversarial set mainly include filtered pairs that are constructed via syntactic perturbation and mask-and-fill strategy. For the random set, the context-response pairs are mainly obtained with random utterance selection.

3.4 Student Model

Once D^* is ready, we can learn a student model, $M_{student}$ parameterized by $\theta_{student}$, on D^* by performing the following classification task:

$$x_i^* = f_{\theta_{student}}([c_i^* \circ r_i^*]) \tag{3}$$

Figure 2 is a graphical illustration of the training objective of $M_{student}$ and the equation is as follows:

$$\min_{\theta_{student}} \frac{1}{|D^*|} \sum_{(c_i^*, r_i^*, \tilde{y}_i^*) \in D^*} \mathcal{L}_{CE}(x_i^*, \tilde{y}_i^*) + \mathcal{L}_{KL}(x_i^*, \hat{x}_i^*) + \mathcal{L}_{MLM}([c_i^* \circ r_i^*]) \tag{4}$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss, $\mathcal{L}_{KL}$ is the KL divergence and $\mathcal{L}_{MLM}$ is the self-supervised masked language modeling (MLM) loss. x_i^* and $\tilde{y}_i^*$ are the logits output from $M_{student}$ and the pseudo label generated by the pretrained $M_{teacher}$ respectively given the input pair (c_i^*, r_i^*).

$\mathcal{L}_{KL}$ is introduced to enforce consistency regularization, with which $M_{student}$ is less sensitive to noise and hence

smoother w.r.t. perturbations in the input space (Xie et al. 2020a). We denote the noisy version of r_i^* after noise injection as $\hat{r}_i^*$. In the practical implementation, we follow (He et al. 2020) to generate $\hat{r}_i^*$ based on r_i^*. $\hat{x}_i^*$ is the corresponding logits from $M_{student}$ after inputting $(c_i^*, \hat{r}_i^*)$. The KL divergence between the respective post-softmax probability distributions of x_i^* and $\hat{x}_i^*$ is minimized during training.

The last term, $\mathcal{L}_{MLM}$, is intended to help $M_{student}$ extract additional domain-specific knowledge so as to better adapt to the multi-domain synthetic dataset. The MLM implementation follows the standard BERT (Devlin et al. 2019) practice whereby a random portion of tokens in the concatenated sequence, $[c_i^* \circ r_i^*]$, are masked. $M_{student}$ is expected to make predictions on the masked tokens.

3.5 Run-time Scoring Process

The learned student model serves as the backbone of MDD-Eval for performing the multi-domain dialogue evaluation task, that derives the metric score s_i^j for a given context-response pair $(c_i^j, r_i^j) \in D^j$ as mentioned in Section 3.1. We formulate the scoring process by $M_{student}$ as follows:

$$s_i^j = P(\tilde{y}_i^j = \text{relevance}|(c_i^j, r_i^j)) \tag{5}$$

which is the post-softmax probability w.r.t. the relevant class output by $M_{student}$ given the input, (c_i^j, r_i^j).

4 Experiment Setup

We first discuss the dialogue corpora (Section 4.1) used in the experiments. Then, the evaluation benchmarks used for assessing the performance of MDD-Eval are discussed in Section 4.2. Next, Section 4.3 is about the architecture choice for both the teacher and the student model. Finally, the choices of baselines are outlined in Section 4.4.

4.1 Dialogue Corpora

Base for Teacher Training DailyDialog++ (Sai et al. 2020) is a multi-reference dialogue evaluation dataset developed based on DailyDialog (Li et al. 2017); In this work, it is selected as the base dataset. In total, DailyDialog++ contains 11,429 dialogue contexts and the average number of turns per context is 3.31. There are three categories of responses: random, adversarial and relevant. For each context, the authors collected five different responses per category. Both the relevant and adversarial responses are written by human annotators. The adversarial responses share certain degree of lexical or semantic overlap with the corresponding dialogue contexts, but are still deemed as inappropriate responses. They are introduced to avoid model decision-making based on spurious features in the context-response pairs.

Multi-domain Dialogue Corpora for Augmentation We make use of four publicly-available, high-quality and human-written conversation corpora to form a multi-domain synthetic dataset: DailyDialog (Li et al. 2017), ConvAI2 (Dinan et al. 2020), EmpatheticDialogues (Rashkin et al. 2019) and TopicalChat (Gopalakrishnan et al. 2019). The detailed statistics of the four dialogue corpora are presented in Table 2. We only use the training and validation sets of the dialogue corpora since some dialogue contexts in the evaluation benchmarks are sampled from their test sets.

To extract context-response pairs from the human-human dialogues, we take the dialogue history and current response as an original context-response pair. The number of utterances per context is kept between one and four. For each original context-response pair, we sample ten different augmented pairs per augmentation technique. After the filtering process, we end up with a class-balanced and multi-domain synthetic dataset of around 2.6 million context-response pairs. We name this synthetic dataset, MDD-Data. In our experiment, for quick turn-around, we sub-sample 600K context-response pairs from MDD-Data to train the final student model.

4.2 Evaluation Datasets

Guided by (Yeh, Eskenazi, and Mehri 2021), we use publicly-available dialogue evaluation datasets to assess the performance of the ADMs. Additionally, we propose a few criteria for the selection of high-quality dialogue evaluation datasets. First, we select the ones that cover as many domains as possible. Second, the size of the datasets should be sufficiently large to provide statistically significant analysis. Third, most of the state-of-the-art metrics should have achieved relatively good correlation results on the datasets. This is to avoid inclusion of any biased evaluation dataset. Next, the inter-annotator agreement should be relatively good ($\sim$0.6). Lastly, the evaluation datasets should cover responses of a wide quality spectrum. In total, we have adopted six different publicly-available dialogue evaluation datasets with each accounting for a dialogue domain for assessing MDD-Eval[4]: DailyDialog-Eval (Zhao, Lala, and Kawahara 2020), Persona-Eval (Zhao, Lala, and Kawahara 2020), Topical-Eval (Mehri and Eskenazi 2020b), Movie-Eval (Merdivan et al. 2020), Empathetic-Eval (Huang et al. 2020) and Twitter-Eval (Hori and Hori 2017). Detailed statistics of each evaluation dataset is listed in Table 3.

4.3 Model Architecture Choice

We choose RoBERTa-Large (Liu et al. 2019) for both the teacher and the student model in MDD-Eval. There are two reasons. First, RoBERTa has been pretrained on more than 160GB of uncompressed text covering multiple domains including news, stories, books and web text. Therefore, it equips the prediction model with general knowledge of the text with which the prediction model can easily adapt to the downstream dialogue evaluation tasks. Second, it has been proven as a powerful text encoder that are beneficial for the automatic dialogue evaluation task in prior works (Zhao, Lala, and Kawahara 2020; Mehri and Eskenazi 2020b; Zhang et al. 2021c,b).

[4]The names of the datasets are unified in our paper to better distinguish their respective domains.

DailyDialog	**training**	**validation**	**EmpatheticDialog**	**training**	**validation**
#dialogues	11,118	1,000	#dialogues	19,529	2,768
#utterances	87,170	8,069	#utterances	84,158	12,075
#words	1,186,046	108,933	#words	1,127,355	174,786
#avg utterances per dialogue	7.84	8.07	#avg utterances per dialogue	4.31	4.36
#avg words per dialogue	106.68	108.93	#avg words per dialogue	57.73	63.15
ConvAI2	**training**	**validation**	**TopicalChat**	**training**	**validation**
#dialogues	17,878	1000	#dialogues	8,627	538
#utterances	253,698	15,566	#utterances	188,357	11,660
#words	3,024,032	189,374	#words	4,374,304	273,331
#avg utterances per dialogue	14.19	15.57	#avg utterances per dialogue	21.83	21.67
#avg words per dialogue	169.15	189.37	#avg words per dialogue	507.05	508.05

Table 2: Human-Human Dialogue Corpora Statistics

Name	#Instances	Avg.#Utts.	Avg.#Ctx/Hyp Words	Type	#Criteria	#Annotations	Used NLG models
Persona-Eval (2020)	900	5.1	48.8 / 11.5	Turn-level	1	3,600	LSTM Seq2Seq, Random sampling, and GPT-2
DailyDialog-Eval (2020)	900	4.7	47.5 / 11.0	Turn-level	4	14,400	LSTM Seq2Seq, Random sampling, and GPT-2
Topical-Eval (2020b)	360	11.2	236.3 / 22.4	Turn-level	6	6,480	Transformers
Empathetic-Eval (2020)	300	3.0	29.0 / 15.6	Turn-level	1	3,000	Transformer Seq2Seq, Transformer Ranker
Twitter-Eval (2017)	9,990	3.5	35.3 / 11.2	Turn-level	3	29,700	RNN, LSTM Seq2Seq
Movie-Eval (2020)	9,500	3.9	17.0 / 6.1	Turn-level	2	57,000	Random sampling

Table 3: Summary of the evaluation datasets. Some information are obtained from (Yeh, Eskenazi, and Mehri 2021) and (Zhang et al. 2021d). #criteria is the number of response qualities that have been annotated, such as appropriateness, naturalness, etc.

4.4 Baselines

We compare MDD-Eval against state-of-the-art reference-free dialogue metrics, including DEB (Sai et al. 2020), USL-H (Phy, Zhao, and Aizawa 2020), GRADE (Huang et al. 2020), USR (Mehri and Eskenazi 2020b), unreferenced BERT-RUBER (uBERT-R) (Ghazarian et al. 2019), and D-score (Zhang et al. 2021c). The selection of baselines is guided by a recent comprehensive survey on ADMs (Yeh, Eskenazi, and Mehri 2021), which has showcased the strong performance of the above-mentioned metrics. In fact, each selected metric is one of the top-ranking metrics on one or more public dialogue evaluation benchmarks. As in the previous work, we use the publicly-available checkpoints of the selected evaluation metrics for our evaluation tasks. Table 4 summarizes the training details of the model-based evaluation metrics including the teacher (MDD-T) and student models (MDD-S) in MDD-Eval as well.

5 Results & Analysis

Main Correlation Results Table 5 presents the Spearman correlation scores of baseline evaluation metrics, the proposed MDD-Eval metric, and its ablation versions, across six dialogue evaluation benchmarks. For each MDD-Eval variant, we train the model five times with different random seeds and report the average results across the five runs. It can be observed that the full student model, MDD-S, performs generally well across all the evaluation benchmarks with an average Spearman correlation score of 0.476. MDD-S outperforms all the state-of-the-art model-based evaluation metrics. Remarkably, it outperforms the best baseline, DEB, by roughly 7% in absolute terms. This confirms that MDD-Eval is a robust framework for the multi-domain dialogue evaluation task.

Ablation Study To better understand the influence of each component of MDD-S. The results w.r.t three ablation versions, MDD-T, MDD-C, and MDD-CM, are presented in Table 5. MDD-T is the teacher model, which is trained only on the single-domain human-annotated dataset, DailyDialog++. It performs well on the DailyDialog-Eval and Empathetic-Eval benchmarks, which are close in domain w.r.t its training data source, compared to the baselines. This confirms our statement in Section 1 that learning from humans is an effective approach to equip ADMs with a rating skill to discriminate responses of varying quality.

MDD-C brings significant performance improvement over MDD-T (7.1% Spearman correlation score). Note that MDD-C is learned with the vanilla self-training setup without consistency regularization and domain adaptation. The performance improvement showcases that the student model can generalize the rating skill of the teacher through the MDD-Data alone without any additional inductive bias.

MDD-CM brings a further improvement of 2.4% Spearman correlation score. This confirms the usefulness of the self-supervised MLM objective in helping the student model to extract additional domain-specific knowledge.

Finally, the full model MDD-S achieves the highest average Spearman correlation score of 0.476. This showcases the effectiveness of consistency regularization in our self-training setup.

MDD-Eval vs uBERT-R Unreferenced BERT-RUBER (uBERT-R) can be considered the fundamental representative of the recent family of ADMs based on self-supervised

	Training Dataset	Size	Pretrained Model	Objective	External Knowledge	Single
DEB	DailyDialog++	∼139K	BERT	CrossEntropy	Reddit Conversations	Yes
GRADE	DailyDialog	∼178K	BERT	Triplet	ConceptNet	Yes
USR	TopicalChat / PersonaChat	Unknown	RoBERTa	MLM / CrossEntropy	Persona Profiles / Wikipedia	No
D-score	PersonaChat / Twitter / DailyDialog	∼1.31M / ∼1.16M	RoBERTa	MLM / CrossEntropy	None	No
USL-H	DailyDialog	∼138K	BERT	MLM / CrossEntropy	None	No
uBERT-R	DailyDialog / PersonaChat	Unknown	BERT	Triplet	None	Yes
MDD-T	DailyDialog++	∼139K	RoBERTa	CrossEntropy	None	Yes
MDD-S	MDD-Data	∼600K	RoBERTa	CrossEntropy/ MLM / KL	None	Yes

Table 4: Training details of model-based metrics. 'Training Dataset' and 'Size' indicate the training dialogue corpora and training data size in terms of the number of context-response pairs respectively. 'Pretrained Model' and 'Objective' refer to the backbone pretrained language model and the loss function used by the metrics accordingly. 'External Knowledge' means whether the training process leverages additional knowledge sources. 'Single' denotes that whether a metric is a single evaluation model or a combination of multiple evaluation models. 'Unknown' means that information is not publicly available.

	Baselines						**Ablation Metrics**			**Final**
Benchmarks	**DEB**	**USL-H**	**GRADE**	**USR**	**uBERT-R**	**D-score**	**MDD-T**	**MDD-C**	**MDD-CM**	**MDD-S**
DailyDialog-Eval	0.486	0.391	0.533	0.367	0.285	0.426	0.501	0.482	0.546	**0.579**
Persona-Eval	0.579	0.407	0.583	0.571	0.384	0.511	0.528	0.580	0.594	**0.621**
Topical-Eval	0.116	0.340	0.217	0.423	0.348	0.233	0.218	0.373	0.484	**0.520**
Empathetic-Eval	0.395	0.235	0.297	0.255	0.148	<u>0.087</u>	0.345	0.404	**0.404**	0.374
Movie-Eval	**0.649**	0.531	0.612	0.366	0.388	0.340	0.383	0.556	0.524	0.537
Twitter-Eval	0.214	0.179	0.122	0.166	0.217	**0.301**	0.249	0.258	0.241	0.227
Average	0.407	0.347	0.394	0.358	0.295	0.316	0.371	0.442	0.466	**0.476**

Table 5: Spearman correlation scores of state-of-the-art ADMs and MDD-Eval variants on the six dialogue evaluation benchmarks. Scores with p-values larger than 0.05 are underlined (indicating statistical insignificance). The best score for each benchmark is highlighted in bold. The ablation metrics include MDD-T, MDD-C and MDD-CM, which refer to the teacher model, the student model optimized with only $\mathcal{L}_{CE}$, and the student model optimized with both $\mathcal{L}_{MLM}$ and $\mathcal{L}_{CE}$ respectively. MDD-S is the full student model optimized with all three losses.

learning and pretrained language models. It can be observed that uBERT-R performs much worse than the MDD-Eval variants. There are two major reasons. First, uBERT-R acquires the rating skill to discriminate varying-quality responses in a self-supervised manner. The random sampling strategy adopted by uBERT-R is prone to introduction of false-negative and over-simplistic samples that can negatively impact the evaluation performance. The better performance of MDD-T than uBERT-R indicates that the human-designed sampling strategy is much more useful than the automatic random sampling scheme for equipping ADMs with the rating skill. Second, MDD-S generalizes the rating skill to multiple domains with a self-training framework in which additional inductive biases are incorporated, including mask language modeling and consistency regularization. The much better performance of MDD-S than uBERT-R showcases that semi-supervised learning is a promising option in improving dialogue evaluation performance compared to purely unsupervised learning.

MDD-Eval vs DEB DEB and MDD-Eval variants are learned with the same classification task and their backbone model architectures are also similar. The only difference between MDD-T and DEB is that DEB is equipped with the general knowledge about dialogues across multiple domains through pretraining on 727M Reddit conversations with the MLM objective. As a result, DEB outperforms MDD-T by 3.5% in terms of the average Spearman correlation score. This shows that pretraining on large-scale conversations is useful for the multi-domain dialogue evaluation task. However, DEB performs worse than MDD-S, which is trained only on 600K context-response pairs. The key difference between MDD-S and DEB is the generalization strategy. DEB adopts the pretrain-and-finetune paradigm whereas MDD-S adopts self-training. The more superior performance of MDD-S confirms that the self-training strategy is more effective and data-efficient.

MDD-Eval vs Other Metrics Even though GRADE, USL-H, USR, and D-score, have different training configurations, each of them have its unique strengths. Unlike uBERT-R, the four metrics have additional knowledge to generalize their evaluation skill. GRADE leverages Conceptnet Numberbatch (Speer, Chin, and Havasi 2017), which provides additional commonsense knowledge and topic information, to aid the self-supervised learning process. USR, USL-H and D-score consist of multiple model-based sub-metrics, and hence, they leverage more inductive biases for the task. It can be observed that MDD-S significantly outperforms the four state-of-the-art-metrics, confirming the effectiveness of the proposed self-training strategy for evaluation skill generalization. Since none of the current state-of-the-art metrics is explicitly designed to target the multi-domain dialogue evaluation problem, MDD-Eval helps bridge this gap.

Effects of Combining Data of Different Domains There

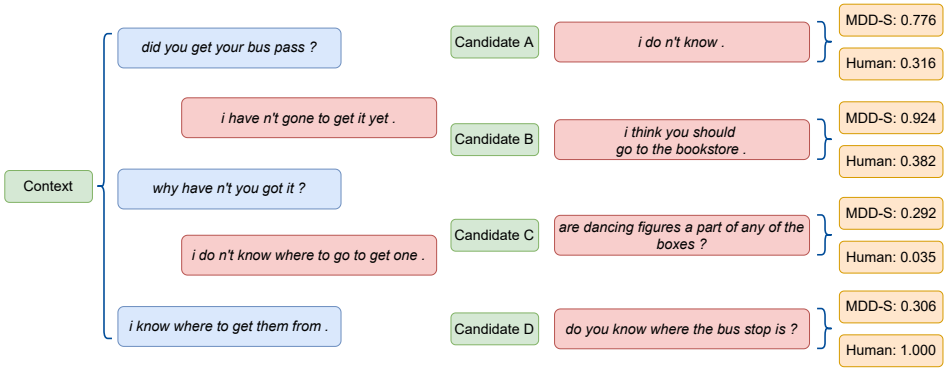

Figure 3: A case study to examine the limitations of MDD-Eval. The ordinal scores of both human and MDD-S are normalized to be within the [0, 1] range, and presented in the yellow box.

Benchmarks	USR	USL-H	GRADE
DailyDialog-Eval	0.491	0.358	0.485
Persona-Eval	0.174	0.431	0.551
Topical-Eval	0.159	0.376	0.271
Empathetic-Eval	0.378	0.377	0.326
Movie-Eval	0.505	0.515	0.559
Twitter-Eval	0.196	0.158	0.080
Average	0.317	0.369	0.379

Table 6: Spearman correlation scores of USR, USL-H, and GRADE trained on the combined dataset.

may be concern that some state-of-the-art metrics are trained on much less data or fewer dialogue domains compared to MDD-S. We presents the results w.r.t USL-H, USR, and GRADE in Table 6. These three metrics are trained on a combined dataset, which contains the training data of all four dialogue corpora used to construct MDD-Data. We didn't include DEB (the best performing baseline) here, because DEB has already been pre-trained on large-scale Reddit conversations (∼767M), and then finetuned on the high-quality DailyDialog++ dataset. We hypothesize that further finetuning DEB an mixed data will suffer from catastrophic forgetting. In addition, it can be observed that our MDD-C approach, which has similar model architecture and objective function as DEB, outperforms DEB on average, but performs worse compared to the final MDD-S metric.

It can be observed that simply combining dialogue data from different domains and training ADMs on the combined data in a self-supervised fashion don't bring robust performance for multi-domain dialogue evaluation. Hence, we need mechanism to filter undesirable data while keeping the ones useful to the evaluation task in order to construct a high-quality multi-domain dataset. MDD-Eval offers a simple, yet effective way to realize that.

Error Analysis We can observe in Table 5 that DEB outperforms MDD-S on Movie-Eval by a large margin. Similarly, D-score also outperforms MDD-S on Twitter-Eval by a large margin. We hypothesize this is because DEB has been pre-trained on 767M Reddit conversations (that could contain information about movies). In addition, we directly use a D-score checkpoint, which is trained on Twitter dialogues, for evaluating D-score's performance on Twitter-Eval. However, for MDD-S, the data distributions of Movie-Eval and Twitter-Eval are very different from its training datasets. This problem can be easily addressed by extend the MDD-data to both the movie, and the twitter domains.

To further analyze the limitations of MDD-Eval, we select a dialogue context and four candidate responses from the DailyDialog-Eval, and then, perform a case study on how MDD-S score the responses. The case study is presented in Figure 3. Firstly, it can be observed that MDD-S provides a high score to a generic response, "I do n't know" while human annotators deem it inappropriate. Future work on MDD-Eval needs to consider modeling specificity.

In addition, the metric focuses more on the neighbouring context, and struggles to capture key information in the longer context. For example, it doesn't recognize that the conversation is about "bus pass", which has little association with "book store" in candidate B. However, candidate B is somehow directive w.r.t its previous utterance (making a suggestion). Hence, MDD-S assigns a high score to candidate B. Future work may consider explicit modeling of speaker dependency, utterance dependency (Zhang et al. 2021a), and the entity transition pattern within the dialogues.

6 Conclusion

We target the multi-domain dialogue evaluation problem and approach the problem with two research questions: (1) How can an ADM learn the rating skill to discriminate responses of varying quality? (2) How can the ADM acquire the general knowledge across different dialogue domains so as to generalize the evaluation skill? We propose MDD-Eval to address the two research questions. Specifically, a teacher evaluator is trained with human-annotated data to acquire the skill to distinguish good context-respons pairs from bad ones in a particular domain. Then, a new evaluator is trained with the teacher-annotated multi-domain data so as to generalizes the evaluation skill across multiple domains. Empirical results demonstrate that MDD-Eval is effective and robust for the multi-domain dialogue evaluation task.

Acknowledgement

We would like to thank all the reviewers for their constructive comments. This work is supported by Science and Engineering Research Council, Agency of Science, Technology and Research (A*STAR), Singapore, through the National Robotics Program under Human-Robot Interaction Phase 1 (Grant No. 192 25 00054); Human Robot Collaborative AI under its AME Programmatic Funding Scheme (Project No. A18A2b0046); Robert Bosch (SEA) Pte Ltd under EDB's Industrial Postgraduate Programme – II (EDB-IPP), project title: Applied Natural Language Processing; The work leading to these results is also part of the project GOMINOLA (PID2020-118112RB-C22) funded by MCIN/AEI/10.13039/501100011033 and project AMIC-PoC (PDC2021-120846-C42) funded by MCIN/AEI/10.13039/501100011033 and by "the European Union "NextGenerationEU/PRTR".

References

Adiwardana, D.; Luong, M.-T.; So, D. R.; Hall, J.; Fiedel, N.; Thoppilan, R.; Yang, Z.; Kulshreshtha, A.; Nemade, G.; Lu, Y.; et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Bachman, P.; Alsharif, O.; and Precup, D. 2014. Learning with pseudo-ensembles. *Advances in Neural Information Processing Systems*, 27: 3365–3373.

Berthelot, D.; Carlini, N.; Cubuk, E. D.; Kurakin, A.; Sohn, K.; Zhang, H.; and Raffel, C. 2020. ReMixMatch: Semi-Supervised Learning with Distribution Matching and Augmentation Anchoring. In *International Conference on Learning Representations*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186.

Dinan, E.; Logacheva, V.; Malykh, V.; Miller, A.; Shuster, K.; Urbanek, J.; Kiela, D.; Szlam, A.; Serban, I.; Lowe, R.; et al. 2020. The second conversational intelligence challenge (convai2). In *The NeurIPS'18 Competition*, 187–208. Springer.

Donahue, C.; Lee, M.; and Liang, P. 2020. Enabling Language Models to Fill in the Blanks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2492–2501.

Edunov, S.; Ott, M.; Auli, M.; and Grangier, D. 2018. Understanding Back-Translation at Scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 489–500.

Ghazarian, S.; Wei, J.; Galstyan, A.; and Peng, N. 2019. Better Automatic Evaluation of Open-Domain Dialogue Systems with Contextualized Embeddings. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*.

Gopalakrishnan, K.; Hedayatnia, B.; Chen, Q.; Gottardi, A.; Kwatra, S.; Venkatesh, A.; Gabriel, R.; Hakkani-Tür, D.; and AI, A. A. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *INTERSPEECH*, 1891–1895.

Gupta, P.; Tsvetkov, Y.; and Bigham, J. 2021. Synthesizing Adversarial Negative Responses for Robust Response Ranking and Evaluation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 3867–3883.

He, J.; Gu, J.; Shen, J.; and Ranzato, M. 2020. Revisiting Self-Training for Neural Sequence Generation. In *International Conference on Learning Representations*.

Hori, C.; and Hori, T. 2017. End-to-end conversation modeling track in DSTC6. *arXiv preprint arXiv:1706.07440*.

Huang, L.; Ye, Z.; Qin, J.; Lin, L.; and Liang, X. 2020. GRADE: Automatic Graph-Enhanced Coherence Metric for Evaluating Open-Domain Dialogue Systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Kahn, J.; Lee, A.; and Hannun, A. 2020. Self-training for end-to-end speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7084–7088. IEEE.

Lee, D.-H.; et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 896.

Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the eight International Joint Conference on Natural Language Processing*, 986–995.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*.

Liu, C.-W.; Lowe, R.; Serban, I.; Noseworthy, M.; Charlin, L.; and Pineau, J. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

McClosky, D.; Charniak, E.; and Johnson, M. 2006. Effective Self-Training for Parsing. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, 152–159.

Mehri, S.; and Eskenazi, M. 2020a. Unsupervised Evaluation of Interactive Dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 225–235.

Mehri, S.; and Eskenazi, M. 2020b. USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Merdivan, E.; Singh, D.; Hanke, S.; Kropf, J.; Holzinger, A.; and Geist, M. 2020. Human Annotated Dialogues Dataset for Natural Conversational Agents. *Applied Sciences*, 10.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Park, C.; Jang, E.; Yang, W.; and Park, J. 2021. Generating Negative Samples by Manipulating Golden Responses for Unsupervised Learning of a Response Evaluation Model. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1525–1534.

Phy, V.; Zhao, Y.; and Aizawa, A. 2020. Deconstruct to Reconstruct a Configurable Evaluation Metric for Open-Domain Dialogue Systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, 4164–4178.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Rashkin, H.; Smith, E. M.; Li, M.; and Boureau, Y.-L. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5370–5381.

Roller, S.; Dinan, E.; Goyal, N.; Ju, D.; Williamson, M.; Liu, Y.; Xu, J.; Ott, M.; Smith, E. M.; Boureau, Y.-L.; and Weston, J. 2021. Recipes for Building an Open-Domain Chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 300–325.

Sai, A. B.; Mohankumar, A. K.; Arora, S.; and Khapra, M. M. 2020. Improving Dialog Evaluation with a Multi-reference Adversarial Dataset and Large Scale Pretraining. *Transactions of the Association for Computational Linguistics*, 8.

Scudder, H. 1965. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3): 363–371.

Sinha, K.; Parthasarathi, P.; Wang, J.; Lowe, R.; Hamilton, W. L.; and Pineau, J. 2020. Learning an Unreferenced Metric for Online Dialogue Evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. A.; Cubuk, E. D.; Kurakin, A.; and Li, C.-L. 2020. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *Advances in Neural Information Processing Systems*, 33.

Speer, R.; Chin, J.; and Havasi, C. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. 4444–4451.

Tao, C.; Mou, L.; Zhao, D.; and Yan, R. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Xie, Q.; Dai, Z.; Hovy, E.; Luong, T.; and Le, Q. 2020a. Unsupervised Data Augmentation for Consistency Training. *Advances in Neural Information Processing Systems*, 33.

Xie, Q.; Luong, M.-T.; Hovy, E.; and Le, Q. V. 2020b. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10687–10698.

Yalniz, I. Z.; Jégou, H.; Chen, K.; Paluri, M.; and Mahajan, D. 2019. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*.

Yarowsky, D. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*.

Yeh, Y.-T.; Eskenazi, M.; and Mehri, S. 2021. A Comprehensive Assessment of Dialog Evaluation Metrics. *arXiv preprint arXiv:2106.03706*.

Zhang, C.; Chen, Y.; D'Haro, L. F.; Zhang, Y.; Friedrichs, T.; Lee, G.; and Li, H. 2021a. DynaEval: Unifying Turn and Dialogue Level Evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 5676–5689.

Zhang, C.; D'Haro, L. F.; Chen, Y.; Friedrichs, T.; and Li, H. 2021b. Investigating the Impact of Pre-trained Language Models on Dialog Evaluation. *arXiv preprint arXiv:2110.01895*.

Zhang, C.; Lee, G.; D'Haro, L. F.; and Li, H. 2021c. D-Score: Holistic Dialogue Evaluation Without Reference. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 2502–2516.

Zhang, C.; Sedoc, J.; D'Haro, L. F.; Banchs, R.; and Rudnicky, A. 2021d. Automatic Evaluation and Moderation of Open-domain Dialogue Systems. *arXiv preprint arXiv:2111.02110*.

Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; and Weston, J. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2204–2213.

Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.-C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, B. 2020. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 270–278.

Zhao, T.; Lala, D.; and Kawahara, T. 2020. Designing Precise and Robust Dialogue Response Evaluators. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 26–33.

Zhao, T.; Zhao, R.; and Eskenazi, M. 2017. Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.