Integrating multi-category metrics for automatic dialogue evaluation

Zhihua Jiang[1], **Guanghui Ye**[1], **Dongning Rao**[2*]

[1] Department of Computer Science, Jinan University, Guangzhou 510632, P. R. China
[2]School of Computer, Guangdong University of Technology, Guangzhou 510006, P. R. China
tjiangzhh@jnu.edu.cn, yghljf@stu2020.jnu.edu.cn, raodn@gdut.edu.cn

Abstract

Evaluation metrics shine the light on the best models and thus strongly influence the research directions, such as the recently developed dialogue metrics USR, FED, and GRADE. However, most current metrics evaluate the dialogue data as isolated and static because they only focus on a single quality or several qualities. To mitigate the problem, this paper proposes an interpretable, multi-faceted, and controllable framework IM^2 (Interpretable and Multi-category Integrated Metric) to combine a large number of metrics which are good at accessing different qualities. For doing this, we first divide current popular dialogue qualities into different categories and then apply or propose dialogue metrics to measure the qualities within each category and finally generate a comprehensive IM^2 score. While an early version of IM^2 took the 2^{nd} place in the development set and the 4^{th} place in the test set on the Track5.1@DSTC10 challenge[1], we conducted an extensive study after the competition in this paper. Experimental results show that IM^2 correlates more strongly with human judgments than other 13 compared dialogue metrics[2].

Introduction

Because human evaluation for natural language generation (NLG) systems is both expensive and time-consuming, relevant and meaningful automatic evaluation metrics that strongly correlate with human judgments are crucial. For example, a number of automatic evaluation metrics specifically designed for dialogue have been recently proposed (Lan et al. 2020; Sinha et al. 2020; Huang et al. 2020; Ghazarian et al. 2020; Mehri and Eskénazi 2020a; Li et al. 2021; Mehri and Eskénazi 2020b; Zhang et al. 2020a; Pang et al. 2020; Phy, Zhao, and Aizawa 2020) because the one-to-many nature of dialogue makes standard automatic language evaluation metrics (e.g., BLEU and METEOR) ineffective for evaluating open-domain dialogue systems (Liu et al. 2016).

Although these dialogue metrics correlate with human evaluation better, they focus on a single quality or several qualities, thus evaluating the dialogue data as isolated and static, e.g., GRADE (Huang et al. 2020) evaluates the topic coherence of dialogue and PredictiveEngage estimates the user engagement. Therefore, multi-quality metrics are preferred, e.g., FED (Mehri and Eskénazi 2020a) measures 9 turn-level qualities and 11 dialogue-level qualities for predicting the overall impression score. However, the generalization capability of existing multi-quality metrics is questionable, e.g., FED correlates poorly with human judgments when scoring other dialogues outside the FED data. This year, the purpose of Track5.1@DSTC10 (Zhang et al. 2021), is to develop effective automatic open-ended dialogue evaluation metrics that perform robustly across a range of dialogue evaluation tasks. No single metric will be competitive.

As such, recent work attempted to combine dialogue evaluation metrics: 1) combining USR (Mehri and Eskénazi 2020b), GRADE (Huang et al. 2020), PONE (Lan et al. 2020) and PredictiveEngage (Ghazarian et al. 2020) through simple-averaging has been reported in a comprehensive assessment of dialogue evaluation metrics (Yeh, Eskénazi, and Mehri 2021); 2) USL-H (Phy, Zhao, and Aizawa 2020) divides dialogue qualities into three categories (viz. U, S, L) and hierarchically combines them; 3) the Track5.1@DSTC10 baseline, Deep AM-FM (Zhang et al. 2020a), is a simply-combined metric which measures the Adequacy Metric (AM) and the Fluency Metric (FM) simultaneously. However, since the above combinations are simple, exploring more sophisticated combination mechanisms has been stated as an important direction for future work (Yeh, Eskénazi, and Mehri 2021).

Therefore, we propose a novel mechanism to combine automatic dialogue evaluation metrics, via first collecting popular dialogue qualities and dividing them into three categories according to the cooperation principle in Linguistics and then applying or proposing dialogue metrics (regarded as *sub-metrics*) to measure the qualities within each category and finally generating an overall evaluation score. The three quality categories are: 1) NUF (viz. *natural* and *understandable* and *fluent*), which measures the quality of the response itself; 2) CR (viz. *coherent* and *relevant*), which measures the response quality conditioned on the context; 3) IES (viz. *interesting* and *engaging* and *specific*), which measures the

*Corresponding author.

[1]The actual name of Track5.1@DSTC10 is Automatic Evaluation and Moderation of Open-domain Dialogue Systems (subtask 1) on the AAAI DSTC-10 (Dialog System Technology Challenges 2022) challenge. The leaderboard: https://chateval.org/dstc10.

[2]We will release the code and data after this paper is formally published.

special property of a response. Further, the overall evaluation metric IM^2 leverages the iterative integration, i.e., first merging *sub-metrics* and then merging *categorical metrics*.

The contribution of this paper is two-fold:

1. We propose a novel mechanism for combing automatic dialogue evaluation metrics. The proposed framework IM^2 is: (1) reference-free, which does not need reference responses; (2) interpretable, which integrates fine-grained sub-metrics and meaningful categorical metrics; (3) flexible, which allows every categorical metric to be used independently; (4) controllable, which learns the weight coefficients from training data.
2. We submitted an early version of IM^2 to the Track5.1@DSTC10 challenge and obtained a high average Spearman correlation coefficient 0.394 on the development set (Team ID: T8S1) and 0.282 on the test set (Team ID: T8S5)[3]. After the competition, we further improved the correlation score to 0.4788 (dev) and 0.3756 (test) respectively, via developing more metrics.

Related Work

This section looks at existing work in the space of automatic dialogue evaluation.

Dialogue Evaluation Metrics

This subsection describes individual dialogue metrics. In general, they can be divided into two categories: rule-based and model-based (Yeh, Eskénazi, and Mehri 2021). Rule-based metrics use heuristic rules to evaluate the system response, while model-based metrics are trained on specific dialogue data.

Most rule-based metrics have been proposed for standard language evaluation, e.g., BLEU, METEOR, and ROUGE. BLEU (Papineni et al. 2002) is a popular metric often used to benchmark NLG systems. It computes the n-gram precision of the system responses using human references. METEOR (Banerjee and Lavie 2005) and ROUGE (Lin 2004) have been proposed to address the shortcomings of BLEU. METEOR incorporates stems and synonyms into its calculation, while ROUGE focuses on the n-gram recall instead of precision.

Model-based dialogue metrics have sprung up in recent years, e.g., ADEM, RUBER, BERT-RUBER, PONE, MAUDE, GRADE, PredictiveEngage, FED and FlowScore. ADEM (Lowe et al. 2017) is an early metric designed for dialogue. It uses a recurrent neural network (RNN) to predict the cosine similarity between system and reference responses. RUBER (Tao et al. 2018) uses a hybrid model consisting of both a referenced metric and an unreferenced metric. Later, BERT-RUBER (Ghazarian et al. 2019) is proposed to replace RNN with BERT (Devlin et al. 2019). Based on BERT-RUBER, PONE (Lan et al. 2020) uses a novel algorithm to sample negative examples during training. MAUDE (Sinha et al. 2020) is trained with Noise Contrastive Estimation. GRADE (Huang et al. 2020) models topic transition dynamics in dialogue by constructing a graph representation of the dialogue history. PredictiveEngage (Ghazarian et al. 2020) incorporates an utterance-level engagement classifier. FED (Mehri and Eskénazi 2020a) uses DialoGPT (Zhang et al. 2020b) to measure fine-grained qualities of dialogue. FlowScore (Li et al. 2021) constructs dynamic information flow from the dialogue history.

Metrics Combination

This subsection describes previous work on combining dialogue metrics, including Deep AM-FM, HolisticEval, USR, USL-H. Deep AM-FM (Zhang et al. 2020a) measures two aspects of dialogue quality through adequacy and fluency. HolisticEval (Pang et al. 2020) evaluates more qualities of dialog: context coherence, language fluency, response diversity, and logical self-consistency. However, both deep AM-FM and HolisticEval are simply combined. To the best of our knowledge, the most related work to ours is USR and USL-H. They exploit a comparatively complex combination mechanism. USR (Mehri and Eskénazi 2020b) trains three models to evaluate different dialogue qualities: a language model which measures the fluency; a dialogue retrieval model which determines the relevance; a selection model which checks the knowledge use. USL-H (Phy, Zhao, and Aizawa 2020) splits dialogue qualities into three groups: understandability (U), sensibleness (S), and likability (L). Then it composites these metric groups in the hierarchy (H) through linear regression.

This paper focuses on exploring a sophisticated mechanism for combining dialogue metrics. Better than USR, our proposed framework leverages a number of pre-trained language models (PTMs), including BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019) and DialogGPT (Zhang et al. 2020b), to improve the performance of sub-metrics. Beyond USL-H, we exploit the bi-linear regression for implementing the iterative integration instead of the simple combination. For more details on the above-mentioned dialogue metrics, we refer the readers to (Yeh, Eskénazi, and Mehri 2021).

Problem Statement

The IM^2 framework is reference-free, which scores the system response without human reference(s). Formally, given a dialogue context c and a system response r, the goal is to learn a scoring function $f : (c, r) \rightarrow s$ that evaluates the generated response. Dialogue metrics are assessed by comparing them to human judgments. Concretely, a human annotator or several annotators score the quality of a given response conditioned on the dialogue context: $(c, r) \rightarrow q$. Given the scores produced by a particular metric, $S = \{s_1, ..., s_k\}$, and the corresponding human quality annotations, $Q = \{q_1, ..., q_k\}$, we can measure the performance of the metric by calculating the correlation between S and Q.

The Proposed Framework

Overall Architecture

As shown in Figure 1, the IM^2 framework is primarily composed of three quality categories: NUF, CR, and IES. The

[3]The IM^2 version submitted to Track5.1@DSTC10 only integrated four sub-metrics: VUP, GRADE, AB-BA, and D-MLM. See the details in the subsection 'Sub-metrics'.

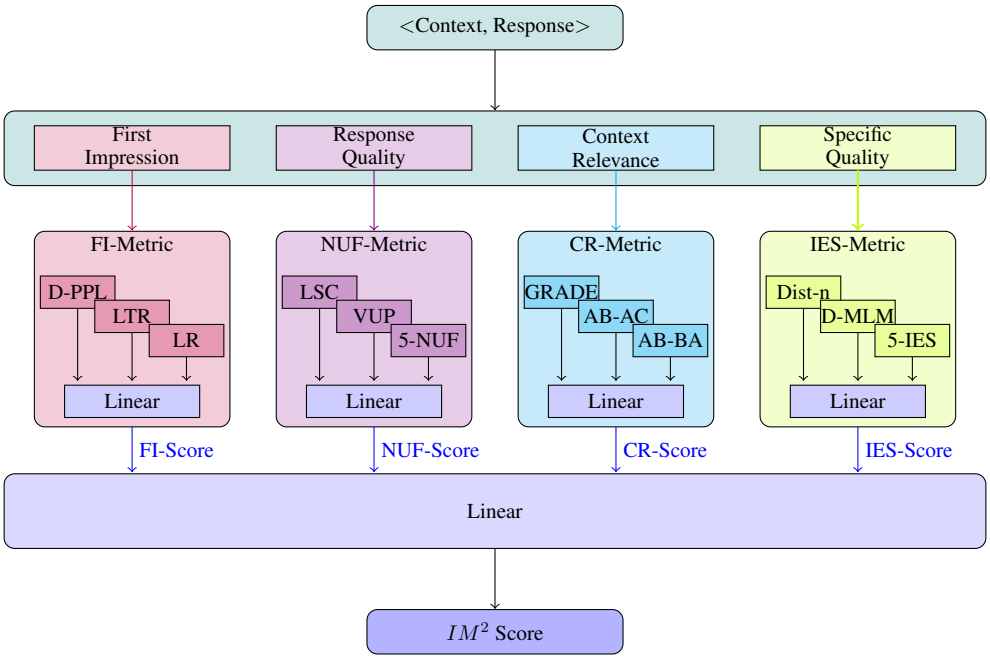

Figure 1: Architecture of IM^2.

NUF category measures the response's naturalness, understandableness, and fluency, while the CR category measures the response's coherency and relevance conditioned on the given context. By contrast, the IES category measures the interestingness, engagement, and specificity of a response. Besides, simulating the habit of human judging, we add the first-impression (FI) category to provide an auxiliary guidance. The FI score can be merged into the final IM^2 score, but, different from the other three categorical scores, it will not be used independently because there are no corresponding qualities on experimental dialugue datasets. Except for FI qualities, all qualities have been used in current dialogue metrics. Finally, the IM^2 framework produces an overall evaluation score (i.e., IM^2-score) for the given context-response pair.

We specify dialogue metrics to measure the qualities within each category. These metrics are regarded as *sub-metrics* in IM^2. Through extensive experiments, we find that only applying or adapting current metrics is far from enough to improve the combined-metric's performance dramatically. Thus, we propose new sub-metrics which can be trained on the evaluated data. Finally, we determine three sub-metrics for each category. Therefore, there are 12 sub-metrics in IM^2, with 2 applied, 3 adapted, and 7 proposed by ourselves. Table 1 presents the summary of sub-metrics. Because a metric can be used to measure a quality or several qualities, the relation between sub-metrics and qualities in IM^2 is shown in Figure 2, where the arrow line indicates which metric measures which quality or qualities.

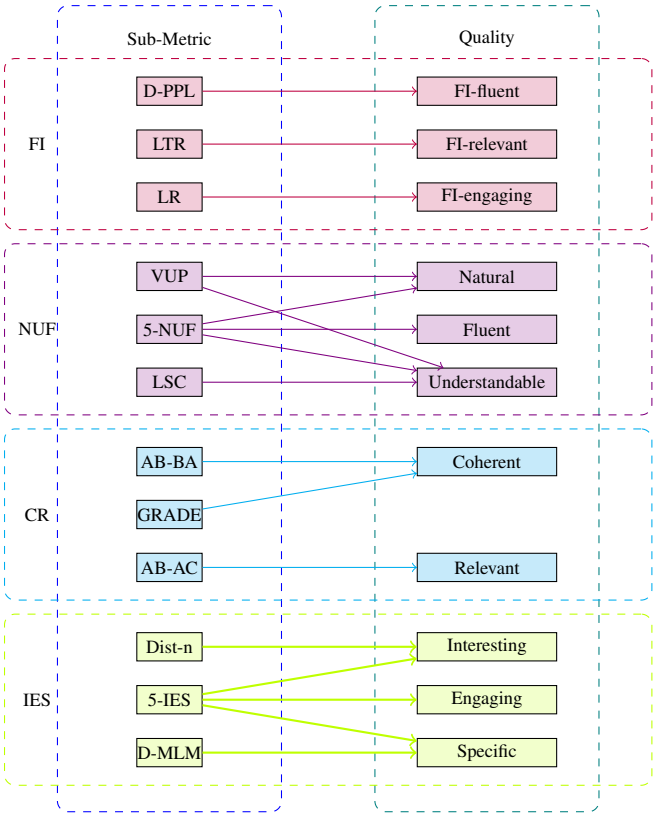

Figure 2: The relation between metrics and qualities in IM^2.

Sub-matric / Categorical	Name	Novelty[1]	Content[2]	PTM[3]	Training Data[4]	Objective[5]
FI-Metric	D-PPL	Adapted	Resp.	DialogGPT-medium	-	-
	LTR	Newly Proposed	Ctx+Resp.	BERT-base	DailyDialog	CE
	LR	Newly Proposed	Ctx+Resp.	-	-	-
NUF-Metric	LSC	Newly Proposed	Resp.	DialogGPT-medium	DailyDialog	-
	VUP	Applied	Resp.	-	-	-
	5-NUF	Newly Proposed	Resp.	RoBERTa-base	NUF-data	MSE
CR-Metric	GRADE	Applied	Ctx+Resp.	-	-	-
	AB-BA	Newly Proposed	Ctx+Resp.	DialogGPT-medium	DailyDialog	CE
	AB-AC	Newly Proposed	Ctx+Resp.	BERT-base	DailyDialog	CE
IES-Metric	Dist-n	Adapted	Resp.	-	-	-
	D-MLM	Adapted	Resp.	RoBERTa-base	PC/TC	MLM
	5-IES	Newly Proposed	Ctx+Resp.	RoBERTa-base	IES-data	MSE

[1] The 'Novelty' column indicates whether the metric is applied, adapted, or proposed by this paper.

[2] The 'Content' column indicates the data content evaluated by the metric. 'Ctx' means the context, 'Resp.' means the response, and '+' means concatenation.

[3] The 'PTM' column indicates the pre-trained language models used for training. '-' means 'None.'

[4] The 'Training Data' column describes the dialogue dataset used when training. 'PC/TC' means 'PersonaChat/TopicalChat'.

[5] The 'Objective' column describes the objective used when training: CE - CrossEntropy, MSE - Mean Square Error, MLM - Masked Language Model.

Table 1: The metrics used in IM^2.

Categorical data

For better training new sub-metrics, we generate three categorical datasets (viz. the NUF/CR/IES data), and an *overall* dataset (viz. the Overall data), from the Track5.1@DSTC10 development data. Specifically, for a NUF/CR/IES category, if the original dataset is human-annotated with at least one member quality, all of its dialogues will be collected into the corresponding categorical data. Comparatively, the NUF/CR/IES data is used to train the sub-metrics, while the Overall data is used to produce the overall IM^2 metric.

Sub-metrics

This subsection describes how to implement sub-metrics used in IM^2. As shown in Table 1, a sub-metric can be applied (viz. without any change), adapted (viz. with a little modification such as using other PTMs or new data), or newly-proposed (viz. proposed by ourselves).

- D-PPL (PPL for dialogue). Similar to *context coherence* in HolisticEval (Pang et al. 2020), we define the initial fluency score as the log-likelihood of generating the next token conditioned the previous tokens. For adaption, we use DialogGPT instead of GPT-2 (Radford et al. 2019).
- LTR (last-turn relevance). This sub-metric is used only if the dialogue is multi-turn. We extract the last-turn context utterance and the generated response into pairs and fine-tune BERT on them.
- LR (length ratio). This sub-metric estimates the user engagement through the simple length analysis. We calculate the ratio between the response length and the last-turn context utterance length.
- GRADE (Huang et al. 2020). We run GRADE via following its original setting.
- AB-BA. This sub-metric identifies the sentence-order coherence instead of the topic coherence like GRADE. We shuffle the context utterance *A* and the response utterance *B*, thus generating a false example which is incoherent with respect to the original sentence order. This metric model is obtained by training DialogGPT to predict the coherence of a context-response pair on DailyDialog[4].
- AB-AC. Similar to AB-BA, we use negative sampling to enhance the relevance prediction between the context *A* and the response *B*. Instead of randomly generating false responses, we use the other dialogue's response *C* which is most cosine-similar to *B*, as a false response. This metric model is obtained by training BERT on DailyDialog.
- LSC (logical self-consistency). This sub-metric heavily depends on AB-BA. We first split the response into clauses, put every two adjacent clauses into pairs, and send them to the AB-BA model to estimate the between-clause coherence.
- VUP (valid utterance prediction). This sub-metric comes from USL-H (Phy, Zhao, and Aizawa 2020). We adopt the same negative sampling technique.
- 5-NUF (5-class NUF metric). Consistent with the 5-point human annotation scheme, we get this sub-metric by training a 5-class classifier on the NUF dataset. Specifically, we train RoBERTa and construct a three-layer fully-connected network.
- Dist-n. Dist-n measures the response's interestingness by detecting fresh words. The more unique words there are, the more interesting the response is. We build a word list for each dataset to calculate the n-gram entropy.
- D-MLM (MLM for dialogue). Guided by the masked language model (MLM) metric, we fine-tune RoBERTa on Topical-Chat and Dailydialog. One word at a time, each word in the response is masked, and its log-likelihood is computed.
- 5-IES (5-class IES metric). Similar to 5-NUF, we train a 5-class classifier on the IES dataset.

Because D-PPL, D-MLM, 5-NUF, and 5-IES are not ranged in [0, 1], we use the Min-Max method (Jain, Nandakumar, and Ross 2005) to ensure the score normalization.

[4]Since DailyDialog is a multi-turn dialogue dataset, we extract conversation utterances in every turn to produce the context-response pairs in pre-processing.

The Integration Mechanism

Based on bi-linear regression, the IM^2 score is derived by:

$$
\begin{aligned}
FI &= w_1 * \text{D-PPL} + w_2 * LTR + w_3 * LR \\
NUF &= w_4 * LSC + w_5 * VUP + w_6 * \text{5-NUF} \\
CR &= w_7 * GRADE + w_8 * \text{AB-AC} + w_9 * \text{AB-BA} \\
IES &= w_{10} * \text{Dist-n} + w_{11} * \text{D-MLM} + w_{12} * \text{5-IES} \\
IM^2 &= \alpha_1 * FI + \alpha_2 * NUF + \alpha_3 * CR + \alpha_4 * IES
\end{aligned} \tag{1}
$$

Where the weight coefficients of sub-metrics, w_1 - w_{12}, and the weight coefficients of categorical metrics, α_1 - α_4, are learnable. Table 3 compares IM^2 against the above-mentioned combined metrics from the number of sub-metrics, qualities, PTMs, and training datasets.

Experiments

Dialogue Data

We use all development data on the Track5.1@DSTC10 challenge. It consists of the following 14 datasets: DSTC6-Eval (D6) (Hori and Hori 2017), DSTC7-Eval (D7) (Galley et al. 2019), Persona-Chatlog (PC) (See et al. 2019), PersonaChat-USR (UP) (Mehri and Eskénazi 2020b), TopicalChat-USR (TP) (Mehri and Eskénazi 2020b), FED-Turn (FT) (Mehri and Eskénazi 2020a), FED-Conversation (FC) (Mehri and Eskénazi 2020a), DailyDialog-Eval (GD) (Gupta et al. 2019), DailyDialog-Eval (ZD) (Zhao, Lala, and Kawahara 2020), PersonaChat-Eval (ZP) (Zhao, Lala, and Kawahara 2020), DailyDialog-Eval (ED) (Huang et al. 2020), Empathetic-Eval (EE) (Huang et al. 2020), ConvAI2-Eval (EC) (Huang et al. 2020), HUMOD (HU) (Merdivan et al. 2020). All experiments are conducted on a workstation equipped with a single NVIDIA Tesla 32GB GPU.

Primary Results

First of all, the weight coefficients of all metrics on the best-performing IM^2 model are shown in Table 4, where the meaning of each coefficient is added as an index for readability. It reveals that each component has a contribution on the overall performance of IM^2. Comparatively, 5-NUF contributes most among sub-metrics and CR-metric contributes most among categorical metrics. The setting of our best model is fixed for all references and testing.

Second, since the IM^2 metric is derived by iterative integration, we design two different strategies to use its categorical metrics and itself for reference:

- OVERALL. For any quality, we use the IM^2-metric as a whole to measure it.
- SELECTIVE. For a specific quality q, we select the most appropriate metric to measure it. The selection rule is:
  - if q is in the NUF category, we use the NUF-metric;
  - if q is in the CR category, we use the CR-metric;
  - if q is in the IES category, we use the IES-metric;
  - otherwise, we use the IM^2-metric.

Particularly, based on the above rule, when q is *overall* or a new quality on unseen datasets, we will use the IM^2-metric to measure it. Further, both these two strategies can be applied to other combined metrics only if their sub-metrics can be used independently. As an example, we experiment with this on USL-H, because it is most related to IM^2. A slight difference is that, for the SELECTIVE strategy, we select a sub-metric for USL-H while select a categorical metric for IM^2. Table 2 presents the correlation to human judgments when comparing IM^2 against other 13 metrics (8 not-combined and 5 combined). Particularly, the correlation score of each dataset is the average of correlation scores with respect to all evaluation qualities and the 'AVG' column indicates the average correlation score on all 14 development datasets.

The primary experimental findings are:

- None of metrics performed well on all 14 datasets (e.g., the Pearson correlation of USR on UP is 0.4452, but 0.0974 on ED). However, IM^2 is stable and has a strong correlation with human judgments on each dataset.
- According to the last 'AVG' column, the top-3 metrics are IM^2-selective, IM^2-overall, and USL-H-selective. It demonstrates that the SELECTIVE strategy is more effective than the OVERALL strategy, not only for IM^2, but also for other combined metrics such as USL-H.
- Even though much inferior to IM^2, both PE+GRADE+USR and GRADE performs better than most other metrics. Due to the fact that CR-metric contains GRADE, it can partly explain why CR-metric contributes most to IM^2 among categorical metrics.

Extended Study

This part presents the experimental findings of the correlation to different qualities and the generalization ability of tested metrics on unseen datasets.

Correlation to Quality. We select the FED data because it contains the largest number of qualities. As shown in Figure 3, in the upper graph, IM^2 better evaluates most qualities than other metrics; in the lower graph, we find each metric has a limited ability for evaluation. Categorical metrics perform best on their specific qualities. For example, the IES-metric scores best at *interesting*, *engaging*, and *specific*.

Performance on Unseen Datasets. We tested three best-performing metrics in Table 2 with respect to the average correlation score, on all five unseen test datasets of Track5.1@DSTC10. As shown in Table 5, IM^2-selective performed best on these unseen datasets and far exceeded the highest Spearman correlation score 0.296 in the leaderboard of the test set on Track5.1@DSTC10.

Conclusion

This paper explores the sophisticated mechanism for combining dialogue metrics. A multi-category integrated metric framework IM^2 is proposed for doing this. In IM^2, we integrate a large amount of different sub-metrics and train the meaningful categorical metrics. Besides, we propose two different strategies to apply IM^2 and its components. We conduct excessive experiments on the Track5.1@DSTC10 data. The results show that IM^2 strongly correlates with human judgments and is superior to all compared metrics.

Dataset[1] / Metric	DSTC6-Eval (D6)		DSTC7-Eval (D7)		Persona-Chatlog (PC)		PersonaChat-USR (UP)		TopicalChat-USR (TP)	
	P	S	P	S	P	S	P	S	P	S
Not-combined:										
BERT-RUBER	0.3390	0.2878	**0.3064**	0.2448	–[7]	–	0.2578	0.2429	0.4023	0.4065
PONE	0.3382	0.2878	**0.3064**	0.2458	–	–	0.2565	0.2394	0.3972	0.4049
MAUDE	0.1953	0.1279	-0.0819	-0.0859	-0.0073*	-0.0065*	0.2541	0.1784*	-0.0083*	-0.0106*
GRADE	0.1105	0.1204	**0.3096**	**0.3207**	0.0245	-0.0172	0.2749	0.2335	0.1503	0.1443
ADEM	0.1510	0.1187	-0.0681	-0.0732	–	–	-0.1419	-0.0851*	-0.0604*	-0.0614*
FED	-0.1128*	-0.0954*	-0.1230*	-0.0862*	-0.0163*	-0.0235*	-0.0282*	-0.002*	-0.1132	-0.0893
FlowScore	-0.0980	-0.1036	-0.0122*	-0.0185*	0.0363*	0.0351*	-0.0102*	-0.0154*	-0.0238*	-0.0236*
BERTScore	**0.3586**	**0.3257**	0.0135*	0.0115*	–	–	0.1462*	0.1325*	0.2886	0.3155
Combined:										
Deep AM-FM	0.1051	0.0615	-0.0335	0.0315	0.0826	0.0295*	0.1314	0.1507	0.1315	0.1878
USR	0.1821	0.1658	0.1223	0.0984*	0.0268	0.0258	0.4452	0.4075	0.4145	0.4386
HolisticEval[2]	0.0011	-0.0038	-0.0658	-0.0613	-0.0658*	-0.0613*	0.0871*	0.1128*	-0.1468	-0.1231
PE+GRADE+USR[3]	0.2136	0.1890	0.2498	0.2143	0.0048*	0.0045*	**0.4682**	**0.4325**	**0.4479**	**0.4698**
USL-H-overall	0.1515	0.1624	0.2405	0.2598	0.0120	0.0074*	0.3149	0.3090	0.2307*	0.2292
USL-H-selective	0.1515	0.1624	0.2849	**0.2969**	**0.0884**	**0.0896**	0.3612	0.3584	0.3375	0.3186
IM^2-overall[4]	**0.3629**[6]	**0.3571**	0.2716	0.2903	**0.1201**	**0.1045**	**0.4569**	**0.4320**	**0.4878**	**0.4657**
IM^2-selective[5]	**0.3629**	**0.3571**	**0.4297**	**0.4008**	**0.1737**	**0.1625**	**0.5640**	**0.5799**	**0.5465**	**0.5567**

Dataset / Metric	FED-Turn (FT)		FED-Conversation (FC)		DailyDialog-Eval (GD)		DailyDialog-Eval (ZD)		PersonaChat-Eval (ZP)	
	P	S	P	S	P	S	P	S	P	S
Not-combined:										
BERT-RUBER	–	–	–	–	0.0895	0.1037	0.2787	0.2279*	0.3325	0.3341
PONE	–	–	–	–	0.0849	0.1040	0.2718	0.2264*	0.2640	0.2744
MAUDE	0.0214*	-0.0023*	-0.0228*	-0.2329	0.1824	0.2567	0.1123*	0.0981	0.2496	0.3646
GRADE	0.0540	0.0375	-0.0910	-0.1301	**0.6044**	**0.5947**	0.3819	0.4025	**0.5777**	**0.5841**
ADEM	–	–	–	–	0.2002	0.1012	0.1014	0.0683	0.1541	0.0324
FED	0.1198	0.0865	0.2222	**0.2954**	0.2104	0.1056	0.1036	0.0785	0.2698	0.1537
FlowScore	0.0729*	0.0552*	0.0640*	0.0232*	-0.0530*	-0.0669*	–	–	-0.0847	-0.0898
BERTScore	–	–	–	–	0.1024	0.0846	0.0541	0.0125	0.1563	0.1152
Combined:										
Deep AM-FM	0.0465	0.0324	0.1212	0.0854	-0.0457	0.1362	0.1982	0.2257	0.2365	**0.4459**
USR	0.1140	0.1170	0.0930*	0.0620*	0.5247	0.4986	0.3045	0.2968	**0.4479**	0.4076
HolisticEval	0.1223	0.1251	-0.2762	-0.3141	0.2085	0.1127	0.1013	0.0607	0.1501	0.0661
PE+GRADE+USR	0.0756*	0.0664*	-0.1301	-0.0984*	0.2575	0.2045	0.2466	0.1835	0.3339	0.2809
USL-H-overall	0.1085*	0.0861	0.1640	0.1780*	**0.5348***	**0.5173**	0.3749	0.3433	0.4266	0.4151
USL-H-selective	**0.1921**	**0.1879**	**0.2498**	**0.2516**	**0.5348***	**0.5173**	**0.4487**	**0.4505**	0.4266	0.4151
IM^2-overall	**0.1587**	**0.2045**	**0.2378**	0.2284*	**0.6466**	**0.6610**	**0.4179**	**0.4691**	**0.6279**	**0.6135**
IM^2-selective	**0.2960**	**0.3729**	**0.3388**	**0.3617**	**0.6466**	**0.6610**	**0.5364***	**0.5728**	**0.6279**	**0.6135**

Dataset / Metric	DailyDialog-Eval (ED)		Empathetic-Eval (EE)		ConvAI2-Eval (EC)		HUMOD (HU)		AVG	
	P	S	P	S	P	S	P	S	P	S
Not-combined:										
BERT-RUBER	0.0339*	0.0155*	0.0599*	0.0198*	0.2256	0.2279	0.1126*	0.1143*	0.2217	0.2023
PONE	0.0380*	0.0173*	0.0611*	0.0082*	0.2247	0.2255	0.1123*	0.1143*	0.2141	0.1953
MAUDE	0.0154*	-0.0257*	0.0598*	0.0635*	0.2511	0.2232	0.0193*	0.0524*	0.0886	0.0715
GRADE	**0.2896**	**0.2531**	0.2970	**0.2960**	0.5505	**0.5718**	**0.3347**	**0.3072**	0.2833	0.2662
ADEM	0.0640*	0.0713*	-0.0365*	-0.0280*	-0.0603*	-0.0574*	0.0617	0.0501	0.0332	0.0124
FED	-0.0234*	-0.0451*	-0.0863	-0.0812*	0.0826*	0.0524*	0.0684	0.0452	0.1150	0.1379
FlowScore	0.0253*	0.0259*	0.1239*	0.1609	0.0613*	0.0858*	0.0401*	0.0356*	0.0109	0.0080
BERTScore	0.1288	0.1013*	0.0351*	0.0286*	0.2458	0.2156	0.0354	0.0257	0.1423	0.1244
Combined:										
Deep AM-FM	0.1649	0.1703	-0.0274	0.0497	0.0947	0.0721	0.0117	0.0969	0.0879	0.1480
USR	0.0974*	0.1457*	**0.2984**	0.2560	0.5424	0.5076	0.1920	0.2253	0.2718	0.2609
HolisticEval	-0.0271*	-0.0203*	0.1956	0.2032	-0.0292*	-0.0184*	0.0201*	0.0374*	0.0197	0.0083
PE+GRADE+USR	0.1570	0.1786	**0.3323**	**0.3922**	0.5484	0.5370	0.1659	0.1556	0.2743	0.2943
USL-H-overall	0.1112	0.1283*	0.1879*	0.1963	0.4787	0.4603	0.2262	0.2172	0.2544	0.2507
USL-H-selective	**0.2812**	**0.2767**	0.2710	0.2757	**0.5511**	0.5381	0.2516	0.2678	**0.2890**	**0.3147**
IM^2-overall	0.2309	0.2075*	0.2845	0.2840	**0.5749**	**0.5660**	**0.2880**	**0.3377***	**0.3690**	**0.3729**
IM^2-selective	**0.3950**	**0.3980**	**0.4715**	**0.4822**	**0.6679***	**0.6857**	**0.4960**	**0.4993**	**0.4680**	**0.4788**

[1] All values are statistically significant to $p < 0.05$, unless marked by *. The 'P' column indicates the Pearson correlation coefficients. The 'S' column indicates the Spearman correlation coefficients. The last 'AVG' column indicates the average correlation coefficient on all 14 datasets.

[2] Referred to (Yeh, Eskénazi, and Mehri 2021), we calculate the average of *context coherence*, *language fluency* and *logical self-consistency*, as the overall score for HolisticEval, because *response diversity* is not available on Track5.1@DSTC10 datasets.

[3] We reproduced 'PE+GRADE+USR' according to (Yeh, Eskénazi, and Mehri 2021).

[4] The 'overall' indicates the OVERALL strategy which uses the metric as a whole to measure every quality.

[5] The 'selective' indicates the SELECTIVE strategy which selects the most appropriate metric to measure a specific quality.

[6] The results of 'overall' and 'selective' are same on D6, GD, and ZP because these three datasets are only quality-annotated with the 'overall' quality.

[7] '-' means no score. The reasons are as follows: (1) All of BERT-RUBER, PONE, ADEM and BERTScore cannot score on PC, FT and FC because these three datasets have no reference responses; (2) FlowScore only scores dialogues with more than 3 utterances, so it cannot be used on ZD.

Table 2: The Pearson and Spearman correlation coefficients with human evaluation scores on all 14 development datasets of Track5.1@DSTC10. The correlation scores of the top-3 metrics on each dataset have been highlighted in bold.

Combined Metric[1]	Sub-metrics	Qualities	PTMs	Training Datasets
Deep AM-FM	Adequacy-metric Fluency-metric	Adequate Fluent	BERT	Twitter
HolisticEval	Context coherence Language fluency Response diversity Logical self-consistency	Coherent Fluent Diverse Consistent	GPT-2	DailyDialog
USR	Fluency Relevance Knowledge use	Fluent Relevant Knowledge use	RoBERTa	PersonaChat TopicalChat
USL-H	U-metric S-metric L-metric	Understandable Sensible Specific	BERT	DailyDialog
IM^2 (ours)	See Table 1; 12 in total	See Figure 2; 11 in total	See Table 1; 3 in total	See Table 1; 5 in total

[1] Both HolisticEval and USR treat quality as metric. Thus, the 'metric' column is identical to the 'quality' column for these two metrics.

Table 3: Comparison on combined dialogue metrics.

w_{PPL} (w_1)	w_{LTR} (w_2)	w_{LR} (w_3)	w_{LSC} (w_4)	w_{VUP} (w_5)	w_{5NUF} (w_6)	w_{GRADE} (w_7)	w_{ABAC} (w_8)	w_{ABBA} (w_9)	w_{Dist} (w_{10})	w_{MLM} (w_{11})	w_{5IES} (w_{12})	α_{FI} (α_1)	α_{NUF} (α_2)	α_{CR} (α_3)	α_{IES} (α_4)
0.25	0.5	0.25	0.2	0.2	0.6	0.45	0.35	0.2	0.33	0.33	0.33	0.2	0.15	0.50	0.15

Table 4: The weight coefficients of the best-performing IM^2.

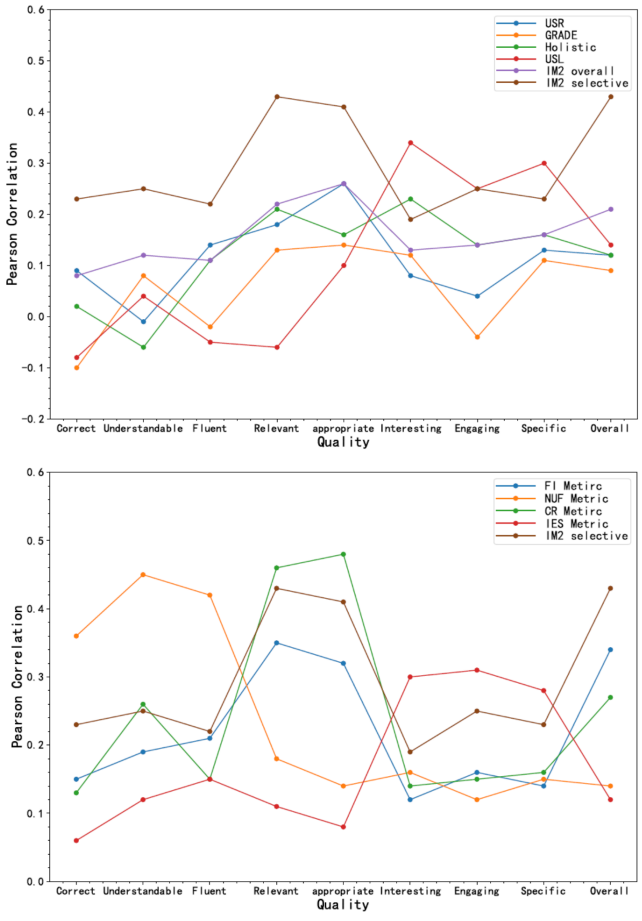

Figure 3: The correlation to different qualities on FED data.

	JSALT		ESL	
Combined Metric	P	S	P	S
USL-H-selective	0.1325	0.1283	0.3107	0.3144
IM^2-overall	0.1956	0.1807	0.4296	0.4325*
IM^2-selective	0.1956	0.1807	0.4296	0.4325*
	NCM		DSTC10-Topical	
Combined Metric	P	S	P	S
USL-H-selective	0.2581	0.2490	0.3364	0.3698
IM^2-overall	0.3679	0.3588	0.2877	0.2690
IM^2-selective	0.3679	0.3588	0.4428	0.4325
	DSTC10-Persona		AVG	
Combined Metric	P	S	P	S
USL-H-selective	0.4210	0.4059	0.2917	0.2935
IM^2-overall	0.3841	0.3903	0.3329	0.3262
IM^2-selective	0.4810	0.4736	0.3834	0.3756

Table 5: The correlation scores on unseen test datasets.

Also, our work reveals that aiming at perfection in learning a metric model for all dialogue datasets is ineffective. Instead, selectively applying the most appropriate metric(s) for different dialogues is promising.

There is still much future work. First, we will pay more attention on challenge dialogue datasets, such as those with lengthy context. Second, we will merge qualities for other competition tasks, such as How to Generate the Safe and Polite Dialogue. Third, we will attempt more powerful dialogue systems, such as PLATO-2 (Bao et al. 2021) which directs towards building an open-domain Chatbot via curriculum learning, to improve the performance of sub-metrics. Last but not least, we will add the IM^2 evaluation scores into the dialogue data for training dialogue systems to generate more human-style responses.

Acknowledgment

This paper is supported by Guangdong Basic and Applied Basic Research Foundation, China (Grant No. 2021A1515012556).

References

Banerjee, S.; and Lavie, A. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, 65–72.

Bao, S.; He, H.; Wang, F.; Wu, H.; Wang, H.; Wu, W.; Guo, Z.; Liu, Z.; and Xu, X. 2021. PLATO-2: Towards Building an Open-Domain Chatbot via Curriculum Learning. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, 2513–2525. Association for Computational Linguistics.

Devlin, J.; Chang, M.; 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186.

Galley, M.; Brockett, C.; Gao, X.; Dolan, B.; and Gao, J. 2019. Grounded response generation task at dstc7. In *In AAAI Dialog System Technology Challenges Workshop*.

Ghazarian, S.; Wei, J. T.; Galstyan, A.; and Peng, N. 2019. Better Automatic Evaluation of Open-Domain Dialogue Systems with Contextualized Embeddings. *CoRR*, abs/1904.10635.

Ghazarian, S.; Weischedel, R. M.; Galstyan, A.; and Peng, N. 2020. Predictive Engagement: An Efficient Metric for Automatic Evaluation of Open-Domain Dialogue Systems. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, 7789–7796.

Gupta, P.; Mehri, S.; Zhao, T.; Pavel, A.; Eskénazi, M.; and Bigham, J. P. 2019. Investigating Evaluation of Open-Domain Dialogue Systems With Human Generated Multiple References. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, SIGdial 2019, Stockholm, Sweden, September 11-13, 2019*, 379–391.

Hori, C.; and Hori, T. 2017. End-to-end Conversation Modeling Track in DSTC6. *CoRR*, abs/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 2020, Online, November 16-20, 2020*, 9230–9240.

Jain, A. K.; Nandakumar, K.; and Ross, A. 2005. Score normalization in multimodal biometric systems. *Pattern Recognit.*, 38(12): 2270–2285.

Lan, T.; Mao, X.; Wei, W.; Gao, X.; and Huang, H. 2020. PONE: A Novel Automatic Evaluation Metric for Open-domain Generative Dialogue Systems. *ACM Trans. Inf. Syst.*, 39(1): 7:1–7:37.

Li, Z.; Zhang, J.; Fei, Z.; Feng, Y.; and Zhou, J. 2021. Conversations Are Not Flat: Modeling the Dynamic Information Flow across Dialogue Utterances. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, 128–138.

Lin, C. 2004. Rouge: a package for automatic evaluation of summaries. In *[online] Barcelona, Spain: Association for Computational Linguistics*, 74–81.

Liu, C.; 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, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, 2122–2132.

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. *CoRR*, abs/1907.11692.

Lowe, R.; Noseworthy, M.; Serban, I. V.; Angelard-Gontier, N.; Bengio, Y.; and Pineau, J. 2017. Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, 1116–1126.

Mehri, S.; and Eskénazi, 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, SIGdial 2020, 1st virtual meeting, July 1-3, 2020*, 225–235.

Mehri, S.; and Eskénazi, 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, ACL 2020, Online, July 5-10, 2020*, 681–707.

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(3).

Pang, B.; Nijkamp, E.; Han, W.; Zhou, L.; Liu, Y.; and Tu, K. 2020. Towards Holistic and Automatic Evaluation of Open-Domain Dialogue Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, 3619–3629.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, 311–318.

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, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, 4164–4178.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners.

See, A.; Roller, S.; Kiela, D.; and Weston, J. 2019. What makes a good conversation? How controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 1702–1723.

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, ACL 2020, Online, July 5-10, 2020*, 2430–2441.

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 Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, 722–729.

Yeh, Y.; Eskénazi, M.; and Mehri, S. 2021. A Comprehensive Assessment of Dialog Evaluation Metrics. *CoRR*, abs/2106.03706.

Zhang, C.; D'Haro, L. F.; Banchs, R. E.; Friedrichs, T.; and Li, H. 2020a. Deep AM-FM: Toolkit for Automatic Dialogue Evaluation. In *Conversational Dialogue Systems for the Next Decade - 11th International Workshop on Spoken Dialogue Systems, IWSDS 2020, Madrid, Spain, 21-23 September, 2020*, 53–69.

Zhang, C.; Sedoc, J.; D'Haro, L. F.; Banchs, R. E.; and Rudnicky, A. 2021. Automatic Evaluation and Moderation of Open-domain Dialogue Systems. *CoRR*, abs/2111.02110.

Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, B. 2020b. 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, ACL 2020, Online, July 5-10, 2020*, 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, ACL 2020, Online, July 5-10, 2020*, 26–33.