IM^2: an Interpretable and Multi-category Integrated Metric Framework for Automatic Dialogue Evaluation

Zhihua Jiang[1], **Guanghui Ye**[1], **Dongning Rao**[2]*, **Di Wang**[1], **Xin Miao**[3]
[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
[3] School of Computer Science, Wuhan University, Wuhan 430072, P.R. China
tjiangzhh@jnu.edu.cn, yghljf@stu2020.jnu.edu.cn, raodn@gdut.edu.cn, windi@stu2020.jnu.edu.cn, miaoxin@whu.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 measuring different qualities. The IM^2 framework first divides current popular dialogue qualities into different categories and then applies or proposes dialogue metrics to measure the qualities within each category and finally generates an overall IM^2 score. An initial version of IM^2 was submitted to the AAAI 2022 Track5.1@DSTC10 challenge[1] and took the 2^{nd} place on both of the development and test leaderboard. After the competition, we develop more metrics and improve the performance of our model. We compare IM^2 with other 13 current dialogue metrics and experimental results show that IM^2 correlates more strongly with human judgments than any of them on each evaluated dataset[2].

1 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. However, as the one-to-many natures of dialogue makes standard automatic language evaluation metrics (e.g., BLEU and METEOR) ineffective for evaluating open-domain dialogue systems (Liu et al., 2016), many 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; Li et al., 2021; Mehri and Eskénazi, 2020b; Zhang et al., 2020a; Pang et al., 2020; Phy et al., 2020).

Although these dialogue metrics correlate with human evaluation, they focus on a single quality or a few 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 (Ghazarian et al., 2020) 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 its own data. Recently, the Track5.1@DSTC10 challenge (Zhang et al., 2021c) just ended, whose purpose is to develop effective automatic open-ended dialogue evaluation metrics that perform robustly across a range of dialogue tasks. No individual metric will be competitive.

Therefore, 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 et al., 2021); 2) USL-H (Phy et al., 2020) divides dialogue qualities into three categories (viz. U, S, L) and linearly combines them; 3) the Track5.1@DSTC10 baseline, Deep AM-FM

*Corresponding author: Dongning Rao.

[1]The full 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]Our code and data are available at: https://github.com/Jnunlplab/IM2.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 11091–11103
December 7-11, 2022 ©2022 Association for Computational Linguistics

(Zhang et al., 2020a), is a simply combined metric which measures the Adequacy Metric (AM) and the Fluency Metric (FM) simultaneously. However, the above combinations are straightforward, and thus exploring more sophisticated combination mechanisms has been claimed as an important direction for future work (Yeh et al., 2021).

On that ground, this paper proposes a novel metric framework named IM^2 (Interpretable and Multi-category Integrated Metric), which first divides current dialogue qualities into three categories, and then applies or proposes dialogue metrics (named *sub-metrics*) to measure the qualities within each category, and finally generates an overall evaluation score. The three quality categories are: 1) NUF (Natural, Understandable, and Fluent), which measures the basic quality of the response; 2) CR (Coherent and Relevant), which measures the response's quality conditioned on the context; 3) IES (Interesting, Engaging, and Specific), which measures the special property of the response. Particularly, IM^2 leverages the multi-level integration, i.e., first producing categorical metrics by integrating on sub-metrics and then producing the overall metric by integrating on categorical metrics.

The contribution of this paper is two-fold:

1. We proposed a novel framework for combing automatic dialogue evaluation metrics. The proposed 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 categorical metrics to be used independently.
2. We submitted an early version of IM^2 to the AAAI 2022 Track5.1@DSTC10 challenge and obtained a high average Spearman correlation coefficient 0.3937 on the development datasets and 0.2819 on the test datasets[3]. After the competition, we further improved the correlation score to 0.4645 and 0.3510 respectively, via developing more metrics.

2 Related Work

2.1 Dialogue Evaluation Metrics

This subsection describes individual dialogue metrics, which can be divided into two categories: rule-based and model-based (Yeh et al., 2021), where rule-based metrics use heuristic rules to evaluate the system response while model-based metrics are trained on specific dialogue data.

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

In contrast, model-based dialogue metrics have sprung up in recent years, e.g., ADEM, RUBER, BERT-RUBER, PONE, MAUDE, GRADE, PredictiveEngage, FED, FlowScore, and DynaEval. ADEM (Lowe et al., 2017) is an early metric designed for dialogue, which 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 which comprised 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. DynaEval (Zhang et al., 2021a) evaluates the dialogue in both turn-level and dialogue-level.

2.2 Metrics Combination

This subsection describes previous studies on combining dialogue metrics, including Deep AM-FM, HolisticEval, USR, and 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 dialogue: context coherence, language fluency, response diversity, and logical self-consistency.

[3]The competition version of IM^2 only integrated four sub-metrics: VUP, GRADE, AB-BA, and D-MLM, and used the SELECTIVE strategy. See Appendix A.1 for the details.

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 model which determines the relevance; a selection model which checks the knowledge use. USL-H (Phy et al., 2020) splits dialogue qualities into three groups: Understandability (U), Sensibleness (S), and Likability (L). Then it composites these groups in a linear hierarchy (H). For more details on the above-mentioned dialogue metrics, we refer the readers to (Yeh et al., 2021).

Although both USL-H and IM^2 divide dialogue metrics into three categories, the differences are specific qualities in categories, the relationship between categories, and the integration mechanism. USL-H decomposes the structure of a response quality in a hierarchy and supposes that understandability is the basis of the whole dialogue quality. If a dialogue is not understandable, then one cannot measure its sensibleness or likability. On the contrary, our categories are designed independently and integrated at multiple levels. See Table 13 in Appendix A.3 for more comparisons.

3 Problem Statement

The proposed 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 response conditioned on the dialogue context: $(c, r) \rightarrow q$. Given the scores produced by a 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.

4 The IM^2 Framework

4.1 The Overall Architecture

As shown in Figure 1, the IM^2 framework produces an overall evaluation score given by a context-response pair. Training and evaluating our model with the standard development data of Track5.1@DSTC10, we divide the quality metrics of the released development datasets into three categories: NUF, CR, and IES. The NUF category measures the response's naturalness, understandableness, and fluency, the CR category measures the response's coherency and relevance conditioned on the context, and the IES category measures the response's interestingness, engagement, and specificity. Table 12 in Appendix A.3 exhibits more detailed descriptions of these qualities.

Through extensive experiments that specify dialogue metrics (i.e., *sub-metrics*) to measure the qualities within each category, we notice that applying or adapting existing metrics is not sufficient to improve the combined-metric's performance greatly. Therefore, we proposed new sub-metrics that can be trained on the evaluation data and determine three sub-metrics for each quality category, as shown in Table 1[4]. The many-to-many relationships between sub-metrics and qualities are also illustrated in Figure 1.

4.2 The Categorical Data

For better training new metrics models, we generate three categorical datasets named the NUF, CR, and IES data, and one Overall data, from the 14 released development datasets of Track5.1@DSTC10. Specifically, for any category, if an original dataset is human-annotated with at least one member quality, all of its dialogue will be collected into the corresponding categorical data. Comparatively, the NUF/CR/IES data is used to train sub-metrics, while the Overall data is used to train the overall metric. See Appendix A.4 for more details of categorical data generation.

4.3 The Sub-metrics

This subsection describes how to train sub-metrics used in IM^2. As shown in Table 1, a sub-metric can be directly applied, adapted with a little modification, or proposed by ourselves. There are three pre-trained language models (PTMs) used in our training: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and DialogGPT (Zhang et al., 2020b). For most sub-metrics, we try each PTM and choose the best-performing one as the final

[4] We tested a lot of sub-metrics and their combinations and found that the combination of sub-metrics listed in Table 1 performed best. I.e., combining most or strongest metrics (e.g., using PredictiveEngage (Ghazarian et al., 2020) for the engagement quality) will not necessarily lead to the best result. Sometimes, the gains of different metrics can be canceled. We will conduct a deeper-in analysis on cooperations and conflicts between metrics in the future.

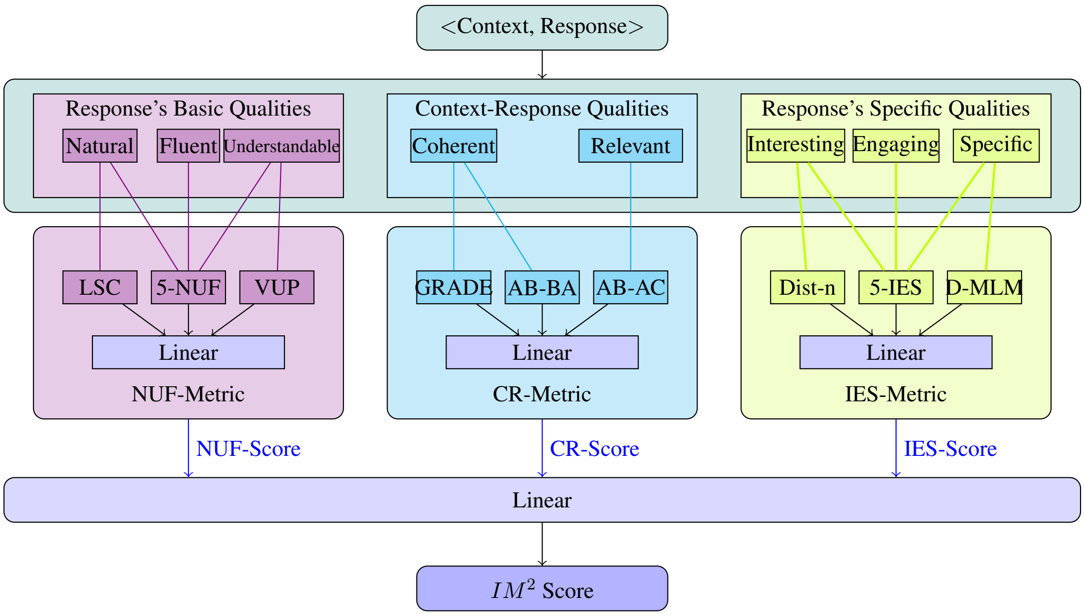

Figure 1: The Architecture of IM^2.

choice. The further discussion on how the PTM choice affects our training result (Zhang et al., 2021b) will be left as a future work.

•GRADE (Huang et al., 2020). We run GRADE via following its original settings.

•AB-BA. We propose this metric to enhance the coherence prediction by using the negative sampling. Given a positive example <A, B> composed of the context A and the response B, we construct a negative example <B, A> by shuffling A and B. The new pair <B, A> is incoherent regarding the original sentence order. Specifically, we train DialogGPT on the pre-processed DailyDialog[5]. Unlike GRADE, AB-BA predicts the sentence-level coherence instead of the topic-level coherence.

•AB-AC. Similar to AB-BA, we propose this metric to enhance the relevance prediction by using negative sampling. Given the context A and its true response B, instead of random generation, we select other dialogue's response C which has the largest cosine similarity[6] regarding B, as a false response. Because the chosen C is coming from different dialogue, it is statistically but not assured to be false. We train BERT on the same pre-processed DailyDialog as that for AB-BA. Figure 3 in Appendix A.5 shows an example for training AB-BA.

•LSC (logical self-consistency). We propose this metric to evaluate the naturalness. It is difficult to give a clear definition of naturalness, e.g., for human annotators with different culture background. In our opinion, a sentence will be natural if it is smooth and does not contain cause-and-effect errors. Thus, we split a response sentence r into sub-sentences $\{r_1, ...r_n\}$ separated by punctuation marks and pack every two adjacent sub-sentences r_i and r_{i+1} into a pair <r_i, r_{i+1}>. We send these pairs to the well-trained AB-BA model, which uses the coherence to check the smoothness, and take the average of all AB-BA scores as the LSC score.

•5-NUF (5-class NUF metric). We propose this metric to evaluate the NUF categorical quality, by simulating the human's 5-point annotation scheme. We train a 5-class classifier on the NUF-data instead of the released development data. Specifically, we train RoBERTa via adding a top three-layer fully-connected network and use Mean Square Error as the training objective.

•VUP (valid utterance prediction). This metric was proposed by USL-H (Phy et al., 2020). The authors trained a model based on BERT to capture the understandability of an utterance by classifying whether it is valid. For doing this, they applied many rules to get a negative sample, e.g., word reorder, word drop, and words repeat. We run VUP via following the original setting.

•Dist-n. Dist-n measures the response's interestingness by detecting unique words, where the more unique words there are, the more interesting the response is. Our adaption for this metric is to build

[5] Since DailyDialog is a multi-turn dataset, we extract every-turn conversation as a positive example <A, B>.

[6] We observed that using a largest similar C performed better than randomly selecting C as a false response. Similar results were reported by PONE (Lan et al., 2020) and USL-H (Phy et al., 2020).

a word list for each dialogue dataset, which records the occurrences of each word in dialogue utterances and thus is used to calculate the n-gram entropy of the response, i.e., the Dist-n score.

- D-MLM (MLM for dialogue). Inspired by the masked language model (MLM) prediction task of BERT, we propose the D-MLM metric to measure the specificity. One word at a time, each word in the response is masked, and its log-likelihood is computed. Then, the normalized scores on all words is the D-MLM score of the response sentence. We fine-tune RoBERTa on PersonaChat and TopicalChat, the joint use of which brings a higher gain than using a single one.
- 5-IES (5-class IES metric). Similar to 5-NUF, we propose this metric to evaluate the IES categorical quality. The training details are identical except that using the IES-data.

4.4 The Integration Mechanism

Using bi-linear regression, the IM^2 score is:

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

Where the weight coefficients w_1 - w_9 and α_1 - α_3 are learnable. The linear function describes the interpretability of the proposed framework.

4.5 The Selection Mechanism

Since IM^2 contains categorical metrics which can be integrated separately, we design two different strategies to use metrics for evaluation:

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

Particularly, when q is *overall* or an unseen quality, we will use the IM^2-metric. Further, the SELECTIVE strategies can be applied to other combined metrics only if their metric members can be used independently. Table 13 in Appendix A.3 compares IM^2 with other combined metrics.

5 Experiments

5.1 Datasets and Setup

There are 14 released development datasets and 5 hidden test datasets on the Track5.1@DSTC10 challenge (Sedoc et al., 2019; Zhang et al., 2021c). We train and evaluate on the development data and verify the model's generality on the test data. See Appendix A.2 for the details of the datasets.

We ran all metrics on a workstation which is equipped with Linux, a single NVIDIA Tesla 32GB GPU, and Python 3.7. About the training time, all sub-metrics were trained within 20-40 minutes, e.g., AB-BA (35 minutes) or 5-IES (20 minutes). The training time is pertinent to the dataset size. About the running time, all sub-metrics ran for about 2 minutes on a single dataset, except for GRADE which ran longer (5 minutes).

5.2 Primary Results

We report our experimental results on the released development datasets in Table 3, along with the official results (the SOTA teams and our team) in Table 4, for comparison. The weight coefficients which lead to the results of IM^2 are shown in Table 2. It reveals that each component has a contribution on the overall performance. There is 13 compared other metrics in Table 3, including 8 single metrics and 5 combined metrics. **All of them have been introduced in Related Work Section**.

We ran all metrics, including our IM^2 and other compared metrics, on each dataset. Some reproduction details are stated as follows:

- The correlation score on each dataset is the average of correlation scores on evaluated qualities.
- Referred to (Yeh et al., 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.
- We reproduced 'PE+GRADE+USR' according to (Yeh et al., 2021).
- We experimented with the SELECTIVE strategy on USL-H. The variant is named USL-H-selective, while the original is named USL-H-overall.
- The results of OVERALL and SELECTIVE are same on D6, GD, and ZP because they only contain the 'overall' quality.
- '-' means no score. The reasons are: (1) PONE and BERT-RUBER cannot score on PC because when using their unreferenced-metrics, the correlation

Categorical \ Sub-metric	Name	Novelty[1]	Content[2]	PTM[3]	Training Data[4]	Objective[5]
NUF-Metric	LSC	Proposed	Resp.	DialogGPT-medium	DailyDialog	CE
	VUP	Applied	Resp.	-	-	-
	5-NUF	Proposed	Resp.	RoBERTa-base	NUF-data	MSE
CR-Metric	GRADE	Applied	Ctx+Resp.	-	-	-
	AB-BA	Proposed	Ctx+Resp.	DialogGPT-medium	DailyDialog$^+$	CE
	AB-AC	Proposed	Ctx+Resp.	BERT-base	DailyDialog$^+$	CE
IES-Metric	Dist-n	Adapted	Resp.	-	-	-
	D-MLM	Adapted	Resp.	RoBERTa-base	PC/TC	MLM
	5-IES	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 context, 'Resp.' means response, and '+' means concatenation.
[3] The 'PTM' column indicates the pre-trained language models used for training. '-' means 'None.'
[4] The 'Training Data' column: 'PC/TC' means 'PersonaChat/TopicalChat'. 'DailyDialog$^+$' means 'the pre-processed DailyDialog'.
[5] The 'objective' column: 'CE' means CrossEntropy, 'MSE' means Mean Square Error, and 'MLM' means Masked Language Model.

Table 1: The summary of metrics used in IM^2.

	w_{LSC} (w_1)	w_{VUP} (w_2)	w_{5NUF} (w_3)	w_{GRADE} (w_4)	w_{ABAC} (w_5)	w_{ABBA} (w_6)	w_{Dist} (w_7)	w_{MLM} (w_8)	w_{5IES} (w_9)	α_{NUF} (α_1)	α_{CR} (α_2)	α_{IES} (α_3)
weight	0.2	0.2	0.6	0.45	0.35	0.2	0.33	0.33	0.33	0.22	0.65	0.13

Table 2: The weight coefficients which lead to the results of IM^2 in Table 3.

coefficient can be calculated only if the dialogue has a human-annotated "relevance" or "coherence" score; (2) FlowScore only scores dialogues with more than 3 utterances, so it cannot be used on ZD.

Experimental findings for Table 3 and 4 are:

- Even though not outperforming the SOTA-dev team, which performed very poorly on test data in Table 6, IM^2 performed much better than all other compared metrics on each dataset.
- The 'AVG' column reveals that the top-3 metrics are IM^2-selective, IM^2-overall, and USL-H-selective, showing that SELECTIVE is more effective than OVERALL, even for USL-H.
- Apart from IM^2, PE+GRADE+USR and GRADE performed better than the others. However, they are not stable, e.g., the Pearson correlation of USR is 0.4452 on UP, but 0.0974 on ED.

5.3 Ablation Studies

Performance on Hidden Test Datasets. We report equivalent results on the hidden test datasets in Table 5, along with the official results (the SOTA teams and our team) in Table 6, for comparison. The weight coefficients which lead to the results of IM^2 on test data are same as those on development data. We excluded 4 out of 14 previous metrics because they performed badly on development data (e.g., their average Spearman correlation score was smaller than 0.1). There are two interesting findings: 1) both IM^2-overall and IM^2-selective outperformed the SOTA-test team; 2) the gain of SELECTIVE over OVERALL on test data is not as significant as on development data. It is because the test data were unseen during the training.

Further, to validate the transferability of IM^2 across domains, we evaluate IM^2 and other 6 competitive metrics on 2 truly unseen test sets: Holistic (Pang et al., 2020) and dstc9 (Gunasekara et al., 2020). The former was proposed by the HolisticEval metric and the latter was used for Track3@DSTC9. As in Table 7, new results show that IM^2 outperforms all the others significantly, justifying its generalization performance.

Categorical Metrics. To verify the effectiveness of categorical metrics, we conducted the ablation study on categorical datasets. As shown in Table 8, each categorical metric performed better than its sub-metrics on categorical data.

Correlation to Qualities. We tested the correlations of metrics to different annotation qualities on one test dataset (DSTC10-Persona) and one development dataset (FED), respectively. Take DSTC10-Persona as an example. Specifically, we select the NUF metric for the *grammar* quality, the CR metric for the *relevance* quality, the IES metric for the *content* quality, and the IM^2 metric for the *appropriateness* quality. The results on DSTC10-Persona are shown in Figure 2. For the space limit, the results on FED are shown in Figure 4 in Appendix A.6. Results show that categorical metrics were good at evaluating their specific qualities and IM^2 strongly correlated to most qualities.

Most-appropriate Metrics. We conducted the most-appropriate-metric test in this part. The result is shown in Table 9. Each most-appropriate metric was parenthesized following the combined metric. This test validated the effectiveness of the SELECTIVE strategy.

Dataset[1] / Metric[2]	Twitter-DSTC6 (D6)		Reddit-DSTC7 (D7)		Persona-See (PC)		Persona-USR (UP)		Topical-USR (TP)	
	P[3]	S[4]	P	S	P	S	P	S	P	S
BERT-RUBER	33.90	28.78	**30.64**	24.48	–	–	25.78	24.29	40.23	40.65
PONE	33.82	28.78	**30.64**	24.58	–	–	25.65	23.94	39.72	40.49
MAUDE	19.53	12.79	-8.19	-8.59	*-0.73*	*-0.65*	25.41	*17.84*	*-0.83*	*-1.06*
GRADE	11.05	12.04	**30.96**	**32.07**	2.45	-1.72	27.49	23.35	15.03	14.43
ADEM	15.10	11.87	-6.81	-7.32	–	–	-14.19	*-8.51*	*-6.04*	*-6.14*
FED	*-11.28*	*-9.54*	*-12.30*	*-8.62*	*-1.63*	*-2.35*	*-2.82*	*-0.20*	-11.32	-8.93
FlowScore	-9.80	-10.36	*-1.22*	*-1.85*	*3.63*	*3.51*	*-1.02*	*-1.54*	*-2.38*	*-2.36*
BERTScore	**35.86**	**32.57**	*1.35*	*1.15*	–	–	*14.62*	*13.25*	28.86	31.55
Deep AM-FM	10.51	6.15	-3.35	3.15	8.26	*2.95*	13.14	15.07	13.15	18.78
USR	18.21	16.58	12.23	*9.84*	2.68	2.58	**44.52**	40.75	41.45	43.86
HolisticEval	0.11	-0.38	-6.58	-6.13	*-6.58*	*-6.13*	*8.71*	*11.28*	-14.68	-12.31
PE+GRADE+USR	21.36	18.90	24.98	21.43	*0.48*	*0.45*	**46.82**	**43.25**	**44.79**	**46.98**
USL-H-overall	15.15	16.24	24.05	25.98	1.20	*0.74*	31.49	30.90	*23.07*	22.92
USL-H-selective	15.15	16.24	28.49	**29.69**	**8.84**	**8.96**	36.12	35.84	33.75	31.86
IM^2-overall	**34.58**	**34.15**	26.05	28.76	**11.43**	**10.23**	43.75	**43.10**	**46.22**	**46.11**
IM^2-selective	**34.58**	**34.15**	**40.61**	**38.76**	**16.69**	**15.43**	**55.98**	**56.90**	**54.82**	**53.21**

Dataset / Metric	FED-Turn (FT)		FED-Dial (FC)		Persona-Zhao (ZD)		DailyDialog -Zhao(ZP)		DailyDialog -Gupta(GD)	
	P	S	P	S	P	S	P	S	P	S
BERT-RUBER	11.96	13.61	22.47	18.46	27.87	*22.79*	33.25	33.41	0.0895	10.37
PONE	14.68	16.55	21.01	20.44	27.18	*22.64*	26.40	27.44	0.0849	10.40
MAUDE	*2.14*	*-0.23*	*-2.28*	-23.29	*11.23*	9.81	24.96	36.46	18.24	25.67
GRADE	5.40	3.75	-9.10	-13.01	38.19	40.25	**57.77**	**58.41**	**60.44**	**59.47**
ADEM	–	–	–	–	10.14	6.83	15.41	3.24	20.02	10.12
FED	11.98	8.65	22.22	**29.54**	10.36	7.85	26.98	15.37	21.04	10.56
FlowScore	*7.29*	*5.52*	*6.40*	*2.32*	–	–	-8.47	-8.98	*-5.30*	*-6.69*
BERTScore	–	–	–	–	5.41	1.25	15.63	11.52	10.24	8.46
Deep AM-FM	4.65	3.24	12.12	8.54	19.82	22.57	23.65	**44.59**	-4.57	13.62
USR	11.40	11.70	*9.30*	*6.20*	30.45	29.68	**44.79**	40.76	52.47	49.86
HolisticEval	12.23	12.51	-27.62	-31.41	10.13	6.07	15.01	6.61	20.85	11.27
PE+GRADE+USR	*7.56*	*6.64*	-13.01	*-9.84*	24.66	18.35	33.39	28.09	25.75	20.45
USL-H-overall	*10.85*	8.61	16.40	*17.80*	37.49	34.33	42.66	41.51	***53.48***	**51.73**
USL-H-selective	**19.21**	**18.79**	**24.98**	**25.16**	**44.87**	**45.05**	42.66	41.51	***53.48***	**51.73**
IM^2-overall	**14.99**	**19.32**	**20.73**	*20.48*	**39.89**	**46.89**	**59.76**	**58.67**	**62.55**	**61.33**
IM^2-selective	**28.69**	**36.95**	**31.16**	**35.48**	***52.98***	**53.01**	**59.76**	**58.67**	**62.55**	**61.33**

Dataset / Metric	DailyDialog -Huang(ED)		ConvAI2 -GRADE(EC)		Empathetic -GRADE(EE)		HUMOD (HU)		AVG[5]	
	P	S	P	S	P	S	P	S	P	S
BERT-RUBER	*3.39*	*1.55*	22.56	22.79	*5.99*	*1.98*	*11.26*	*11.43*	22.17	20.23
PONE	*3.80*	*1.73*	22.47	22.55	*6.11*	*0.82*	*11.23*	*11.43*	21.41	19.53
MAUDE	*1.54*	*-2.57*	25.11	22.32	*5.98*	*6.35*	*1.93*	*5.24*	8.86	7.15
GRADE	**28.96**	**25.31**	**55.05**	**57.18**	29.70	**29.60**	**33.47**	**30.72**	28.33	26.62
ADEM	*6.40*	*7.13*	*-6.03*	*-5.74*	*-3.65*	*-2.80*	6.17	5.01	3.32	1.24
FED	*-2.34*	*-4.51*	*8.26*	*5.24*	-8.63	*-8.12*	6.84	4.52	11.50	13.79
FlowScore	*2.53*	*2.59*	*6.13*	*8.58*	*12.39*	16.09	*4.01*	*3.56*	1.09	0.80
BERTScore	12.88	*10.13*	24.58	21.56	*3.51*	*2.86*	3.54	2.57	14.23	12.44
Deep AM-FM	16.49	17.03	9.47	7.21	-2.74	4.97	1.17	9.69	8.79	14.80
USR	*9.74*	*14.57*	54.24	50.76	**29.84**	25.60	19.20	22.53	27.18	26.09
HolisticEval	*-2.71*	*-2.03*	*-2.92*	*-1.84*	19.56	20.32	*2.01*	*3.74*	1.97	0.83
PE+GRADE+USR	15.70	17.86	54.84	53.70	**33.23**	**39.22**	16.59	15.56	27.43	29.43
USL-H-overall	11.12	*12.83*	47.87	46.03	*18.79*	19.63	22.62	21.72	25.44	25.07
USL-H-selective	**28.12**	**27.67**	**55.11**	53.81	27.10	27.57	25.16	26.78	**28.90**	**31.47**
IM^2-overall	20.85	*20.56*	54.70	**55.68**	25.70	28.11	**28.16**	***33.87***	**34.95**	**36.23**
IM^2-selective	**39.50**	**39.80**	***66.79***	**68.57**	**47.15**	**48.22**	**49.60**	**49.93**	**45.78**	**46.45**

[1] All values are statistically significant to $p < 0.05$, unless in italic.
[2] The 'P' column indicates the Pearson correlation coefficients.
[3] The 'S' column indicates the Spearman correlation coefficients.
[4] The 'overall' label indicates the OVERALL strategy, while the 'selective' label indicates the SELECTIVE strategy.
[5] The last 'AVG' column indicates the average correlation coefficient on all 14 development datasets.

Table 3: The comparison of 14 metrics on the Pearson and Spearman correlation coefficients (%) with human evaluation scores on all 14 development datasets. The top-3 scores on each dataset have been highlighted in bold.

Team[2] \ Dataset[1]	D6	D7	PC	UP	TP	FT	FC	ZD	ZP	GD	ED	EC	EE	HU	AVG
T7(SOTA-dev)	61.63	31.30	27.52	47.88	45.49	35.15	77.42	76.40	54.50	78.85	64.42	57.00	50.10	22.45	52.15
T5 (SOTA-test)	17.94	32.48	8.78	40.36	39.08	30.38	46.89	61.32	48.03	63.25	33.42	58.43	30.57	33.20	38.87
T8 (our team)	18.31	34.12	12.92	36.17	40.24	32.88	49.31	64.58	52.79	60.84	30.06	60.43	24.65	33.83	39.37

[1] The official results only reported the Spearman coefficients (%).
[2] On the leaderboard (development set), T7 ranked first, our team ranked second, and T5 ranked fourth.

Table 4: The official results on the development data reported by (Zhang et al., 2021c).

Metric \ Dataset[1]	JSALT		ESL		NCM		Topical		Persona		AVG	
	P	S	P	S	P	S	P	S	P	S	P	S
BERT-RUBER	*-1.25*	*-0.70*	-5.84	-7.44	6.65	7.28	6.03	*5.29*	8.42	7.80	2.80	2.45
PONE	*0.62*	*1.27*	7.21	5.74	11.67	10.98	17.04	15.89	16.98	15.33	10.70	9.84
GRADE	**13.61**	**12.93**	33.14	30.04	22.14	21.87	28.08	24.72	35.35	34.09	26.46	24.73
FED	2.46	1.99	-1.03	-2.31	10.85	-0.24	8.77	7.18	10.43	9.78	6.30	3.28
BERTScore	-3.65	-4.25	23.63	22.91	10.21	9.07	19.30	18.66	12.82	12.16	12.46	11.71
Deep AM-FM	6.28	5.13	31.45	32.39	15.83	16.50	17.40	17.56	18.96	19.68	17.98	18.24
USR	**11.37**	**11.20**	30.29	29.08	23.75	23.41	25.06	24.33	32.06	31.49	24.51	23.90
PE+GRADE+USR	7.93	8.04	**38.42**	**35.25**	**23.96**	**24.06**	27.50	26.38	31.45	30.88	25.85	24.92
USL-H-overall	8.78	8.18	**40.93**	**36.71**	**25.81**	**24.90**	25.50	23.96	32.40	31.55	26.68	25.06
USL-H-selective	8.78	8.18	**40.93**	**36.71**	**25.81**	**24.90**	**33.64**	**36.98**	**42.10**	**40.59**	**29.17**	**29.35**
IM^2-overall	**16.69**	**14.03**	**40.77**	**40.36**	**33.28**	**32.90**	**29.01**	**27.47**	**37.77**	**38.42**	**31.50**	**30.63**
IM^2-selective	**16.69**	**14.03**	**40.77**	**40.36**	**33.28**	**32.90**	**43.06**	**42.95**	**45.58**	**45.26**	**35.87**	**35.10**

[1] All values are statistically significant to $p < 0.05$, unless in italic.

Table 5: The comparison of 10 metrics on the Pearson and Spearman coefficients (%) with human scores on all 5 hidden test datasets. The top-3 scores on each dataset have been highlighted in bold.

Team[2] \ Dataset[1]	JSALT	ESL	NCM	Topical	Persona	AVG
T7(SOTA-dev)	4.07	3.28	2.01	1.43	2.54	2.30
T5 (SOTA-test)	11.66	40.01	29.60	23.68	37.50	29.63
T8 (our team)	8.75	36.10	25.57	22.77	37.22	28.19

[1] The official results only reported the Spearman coefficients (%).
[2] On the leaderboard (test set), T5 ranked first, our team ranked second, while T7 ranked last.

Table 6: The official results on the test data reported by (Zhang et al., 2021c).

Metric \ Dataset	Holistic		dstc9	
	P	S	P	S
MAUDE	27.50	36.44	5.91	*4.23*
FED	48.56	50.73	12.84	**12.07**
GRADE	67.89	**69.73**	-7.83	7.01
USR	58.97	64.55	*1.96*	*2.03*
USL-H	48.63	53.72	10.54	**10.50**
HolisticEval	67.02	**76.48**	*1.51*	*0.27*
IM^2 (ours)	75.63	**79.44**	18.47	**20.60**

Table 7: Experimental results on 2 non-DSTC10 test datasets: Holistic and dstc9.

Linear Weighting vs. Simple Averaging. We compared two approaches for setting weight coefficients: simple averaging and linear weighting. The former took the arithmetic mean, while the latter used the weight distribution in Table 2. As shown in Table 10, either for IM^2 or any categorical metric, the linear regression obtained a higher correlation score. It reveals that linear weighting is more effective than simple averaging.

Metric	P	S
Part-1: results on the NUF data:		
Metric	P	S
USL-H	19.30	15.89
USR	16.28	18.92
LSC	*15.42*	16.22
VUP	20.47	*24.45*
5-NUF	34.12	36.70
NUF-Metirc (average)	38.47	37.26
NUF-Metirc (linear)	**41.20**	**43.15**
Part-2: results on the CR data:		
Metric	P	S
USL-H	36.07	39.03
USR	34.23	37.25
GRADE	48.99	45.65
AB-AC	44.15	43.92
AB-BA	36.10	38.77
CR-Metric (average)	52.61	56.08
CR-Metric (linear)	**59.17**	**61.75**
Part-3: results on the IES data		
Metric	P	S
USL-H	11.70	13.52
USR	9.15	12.88
Dist-n	12.45	11.96
D-MLM	6.11	8.19
5-IES	28.74	26.18
IES-Metric (average)	31.25	29.91
IES-Metric (linear)	**34.79**	**35.60**
Part-4: results on the Overall data		
Metric	P	S
USL-H	32.59	33.10
USR	45.38	42.98
GRADE	37.75	38.64
PE+GRADE+USR	39.10	41.52
IM^2-overall	**51.49**	**49.77**

Table 8: Comparison on categorical datasets.

Test-1: dataset = Persona[1], quality = *Grammer*		
Metric	P	S
Deep AM-FM (Fluency)	*8.76*	9.13
USL-H (Understandability)	17.45	18.40
HolisticEval (Response Fluency)	16.78	15.43
IM^2 (NUF)	**27.14**	**26.75**
Test-2: dataset = Persona, quality = *relevant*		
Metric	P	S
BERT-RUBER (Unreferenced)	22.07	20.49
USL-H (Sensibleness)	39.01	42.31
HolisticEval (Context Coherence)	*21.85*	19.73
IM^2 (CR)	**56.32**	**57.45**
Test-3: dataset = DailyDialog, quality = *engaging*		
Metric	P	S
PredictiveEngage	41.86	42.01
USL-H (Specificity)	*36.95*	37.82
HolisticEval (Response Diversity)	34.27	36.08
IM^2 (IES)	**48.65**	**51.03**

[1] DailyDialog is one of development datasets, while Persona is released as a hidden test dataset on Track5.1@DSTC10.

Table 9: Results of the most-appropriate metric test.

	Linear		Average	
Metric	P	S	P	S
NUP	19.33	20.34	13.45	11.37
CR	38.40	34.59	28.79	29.88
IES	17.70	18.37	12.90	12.35
IM^2-selective	**45.78**	**46.45**	**32.86**	**33.09**

Table 10: Comparison of linear weighting and simple averaging on the 14 development datasets.

6 Conclusion

This paper explores the sophisticated mechanism for combining dialogue metrics and proposed a novel framework, IM^2. The experimental results show that IM^2 strongly correlates with human judgments and outperforms all compared metrics. Further, our work reveals that training a perfect metric model for all dialogue datasets is difficult, but selecting the most appropriate metric for different dialogues is promising.

There are many future works. First, we will pay more attention to challenge dialogue datasets, such as those with lengthy context. Second, we will merge qualities for newest competition tasks, such as Track4@DSTC11 (Robust and Multilingual Automatic Evaluation Metrics for Open-Domain Dialogue Systems)[7]. Third, we will attempt more powerful dialogue systems, such as PLATO-2 (Bao et al., 2021) which directs towards building an open-domain Chatbot, and with the help from the IM^2 evaluation scores more human-style responses might be generated.

[7]https://chateval.org/dstc11

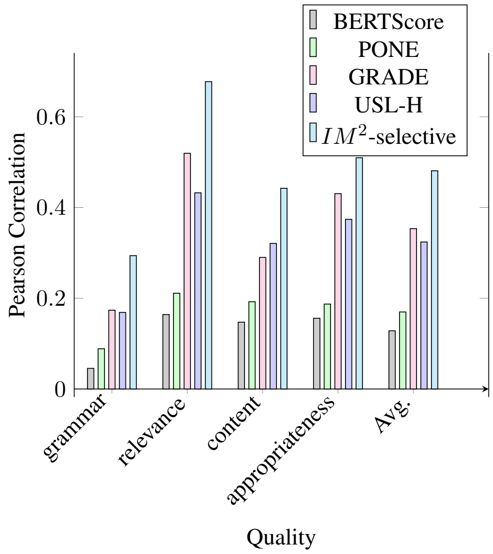

Figure 2: The correlation to different annotation qualities on the DSTC10-Persona data (one of test datasets).

7 Limitations

Conversational AI is one of the most popular NLP applications and developing flexible evaluation frameworks that can emphasize different aspects of quality is important. This paper proposes a novel evaluation framework, which we call IM^2, for delivering exactly that. We conduct a comprehensive set of experiments on this year's DSTC10 challenge data, verifying the effectiveness of our model empirically. However, there are two limitations in our current work: (1) we utilize pretrained models such as DialogGPT, BERT and RoBERTa for training our sub-metrics and choose the best-performing one as the final metric model. While, a deep-in analysis of how the pretrained model choice affects our training result following (Zhang et al., 2021b) is unexplored. (2) we linearly combine various sub-metrics and categorical metrics to generate the final IM^2 score for the interpretability. However, a non-linear combination mechanism such as training a small neural network may bring more promising results, which we leave as one of future works.

Ethics Statement

We use standard datasets which are publicly available. There is no ethics statement for this paper.

Acknowledgements

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

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *CoRR*, abs/2001.09977.

Satanjeev Banerjee and Alon Lavie. 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*, pages 65–72.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2021. PLATO-2: towards building an open-domain chatbot via curriculum learning. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 2513–2525. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186.

Michel Galley, Chris Brockett, Xiang Gao, Bill Dolan, and Jianfeng Gao. 2019. Grounded response generation task at dstc7. In *In AAAI Dialog System Technology Challenges Workshop*.

Sarik Ghazarian, Johnny Tian-Zheng Wei, Aram Galstyan, and Nanyun Peng. 2019. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. *CoRR*, abs/1904.10635.

Sarik Ghazarian, Ralph M. Weischedel, Aram Galstyan, and Nanyun Peng. 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*, pages 7789–7796.

R. Chulaka Gunasekara, Seokhwan Kim, Luis Fernando D'Haro, and et al. 2020. Overview of the ninth dialog system technology challenge: DSTC9. *CoRR*, abs/2011.06486.

Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskénazi, and Jeffrey P. Bigham. 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*, pages 379–391.

Chiori Hori and Takaaki Hori. 2017. End-to-end conversation modeling track in DSTC6. *CoRR*, abs/1706.07440.

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. GRADE: automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9230–9240.

Tian Lan, Xian-Ling Mao, Wei Wei, Xiaoyan Gao, and Heyan Huang. 2020. PONE: A novel automatic evaluation metric for open-domain generative dialogue systems. *ACM Trans. Inf. Syst.*, 39(1):7:1–7:37.

Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021. Conversations are not flat: Modeling the dynamic information flow across dialogue utterances. In *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*, pages 128–138.

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

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2122–2132.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1116–1126.

Shikib Mehri and Maxine Eskénazi. 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*, pages 225–235.

Shikib Mehri and Maxine Eskénazi. 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*, pages 681–707.

Erinc Merdivan, Deepika Singh, Sten Hanke, Johannes Kropf, Andreas Holzinger, and Matthieu Geist. 2020. Human annotated dialogues dataset for natural conversational agents. *Applied Sciences*, 10(3).

Bo Pang, Erik Nijkamp, Wenjuan Han, Linqi Zhou, Yixian Liu, and Kewei Tu. 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*, pages 3619–3629.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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*, pages 311–318.

Vitou Phy, Yang Zhao, and Akiko Aizawa. 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*, pages 4164–4178.

João Sedoc, Daphne Ippolito, Arun Kirubarajan, Jai Thirani, Lyle Ungar, and Chris Callison-Burch. 2019. Chateval: A tool for chatbot evaluation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 60–65. Association for Computational Linguistics.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 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)*, pages 1702–1723.

Koustuv Sinha, Prasanna Parthasarathi, Jasmine Wang, Ryan Lowe, William L. Hamilton, and Joelle Pineau. 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*, pages 2430–2441.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. RUBER: an unsupervised method for automatic evaluation of open-domain dialog systems. In *Proceedings of the 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*, pages 722–729.

Yi-Ting Yeh, Maxine Eskénazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. *CoRR*, abs/2106.03706.

Chen Zhang, Yiming Chen, Luis Fernando D'Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and Haizhou Li. 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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5676–5689. Association for Computational Linguistics.

Chen Zhang, Luis Fernando D'Haro, Rafael E. Banchs, Thomas Friedrichs, and Haizhou Li. 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*, pages 53–69.

Chen Zhang, Luis Fernando D'Haro, Yiming Chen, Thomas Friedrichs, and Haizhou Li. 2021b. Investigating the impact of pre-trained language models on dialog evaluation. *CoRR*, abs/2110.01895.

Chen Zhang, João Sedoc, Luis Fernando D'Haro, Rafael E. Banchs, and Alexander Rudnicky. 2021c. Automatic evaluation and moderation of open-domain dialogue systems. *CoRR*, abs/2111.02110.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 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*, pages 270–278.

Tianyu Zhao, Divesh Lala, and Tatsuya Kawahara. 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*, pages 26–33.

The NUF dataset:	
Original dataset	Quality
dailydialog-zhao	grammar
persona-usr	Natural
topical-usr	Natural
persona-usr	Understandable
topical-usr	Understandable
fed-turn	Correct
fed-turn	Understandable
fed-turn	Fluent
The CR dataset:	
Original dataset	Quality
dailydialog-zhao	relevance
persona-usr	Maintains Context
topical-usr	Maintains Context
fed-turn	Relevant
fed-turn	Semantically appropriate
convai2-grade	relevance
empathetic-grade	relevance
dailydialog-grade	relevance
dstc7	relevance
humod	relevance
The IES dataset:	
Original dataset	Quality
dailydialog-zhao	content
persona-usr	Engaging
topical-usr	Engaging
fed-turn	Interesting
fed-turn	Engaging
fed-turn	Specific
dstc7	informativeness
The Overall dataset:	
Original dataset	Quality
dailydialog-gupta	overall
dailydialog-zhao	overall
persona-usr	overall
topical-usr	overall
persona-zhao	overall
fed-turn	overall
dstc6	overall
dstc7	overall

Table 11: Categorical data.

A Appendix

A.1 The Track5.1@DSTC10 challenge

The challenge goal is to seek effective automatic dialogue evaluation metrics that exhibit the correlation to human judgments and the explainability of the evaluation behaviors. The submitted metric will be ranked according to the average correlation on all 14 open-domain dialogue development datasets. Each team can submit at most five submissions and use at most five metrics in each submission. The metric baseline is Deep AM-FM. The leaderboard (https://chateval.org/dstc10) shows names of submissions and their corresponding Spearman correlation coefficients for each development dataset and each hidden test dataset.

We submitted an early version of IM^2-selective (team ID: T8), which integrates four sub-metrics (VUP, GRADE, AB-BA, and D-MLM).

A.2 Released development datasets

The development datasets of the Track5.1@DSTC10 challenge consist of the following 14 components:

•Twitter-DSTC6 (D6) (Hori and Hori, 2017);
•Reddit-DSTC7 (D7) (Galley et al., 2019);
•Persona-see (PC) (See et al., 2019);
•Persona-USR (UP) (Mehri and Eskénazi, 2020b);
•Topical-USR (TP) (Mehri and Eskénazi, 2020b);
•FED-Turn (FT) (Mehri and Eskénazi, 2020a);
•FED-Dial (FC) (Mehri and Eskénazi, 2020a);
•DailyDialog-Zhao (ZD) (Zhao et al., 2020);
•Persona-Zhao (ZP) (Zhao et al., 2020);
•DailyDialog-Gupta (GD) (Gupta et al., 2019);
•DailyDialog-Huang (ED) (Huang et al., 2020);
•ConvAI2-GRADE (EC) (Huang et al., 2020);
•Empathetic-GRADE (EE) (Huang et al., 2020);
•HUMOD (HU) (Merdivan et al., 2020).

Many of these datasets were collected in different settings. For example, DailyDialog consists of causal conversations about daily life while TopicalChat consists of knowledge-grounded conversations. The FED dataset provides human-system dialogs that were collected in an interactive setting. Specifically, FED data incorporates two state-of-the-art dialogue systems, Meena (Adiwardana et al., 2020) and Mitsuku[8]. For more detailed descriptions on the above-mentioned dialogue datasets, we refer the readers to (Zhang et al., 2021c).

A.3 Comparing IM^2 with other metrics

Table 12 describes all qualities used in our framework. Table 13 compares IM^2 against the above-mentioned combined metrics from the number of sub-metrics, qualities, PTMs, and training datasets.

A.4 Categorical data generation

We collect the dialogues from the Track5.1@DSTC10 datasets to generate the NUF/CR/IES/Overall data. To take the full advantage of the original datasets, we make a slight extension to the NUF/CR/IES category via relaxing the types of qualities, as shown in Table 11. However, the Overall data is only annotated with the *overall* quality. Comparatively, the NUF/CR/IES data is used to train and linear-regress the sub-metrics, while the Overall data is used to linear-regress the categorical metrics.

[8]https://medium.com/pandorabots-blog/mitsuku-wins-loebner-prize-2018-3e8d98c5f2a7.

Quality	Description	Used By Other Metrics
Natural	The response is normal and reasonable.	USR
Understandable	The response is easy to be understood.	FED, USR, USL-H
Fluent	The response is fluently written.	FED, HolisticEval, USR, Deep AM-FM
Coherent	The conversation maintains a good topic flow.	FED, GRADE, FlowScore, HolisticEval
Relevant	The response is relevant to the conversation.	FED, ADEM, USL-H, Deep AM-FM, USR, BUBER, PONE
Interesting	The response is interesting to the average person.	FED
Engaging	The response is engaging.	FED, PredictiveEngage, USR
Specific	The response is specific to the conversation.	FED, USL-H
Overall	The overall impression of the response.	FED, USR

Table 12: The qualities used in IM^2.

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 9 in total	See Table 12 9 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 13: Comparing IM^2 with other combined dialogue metrics.

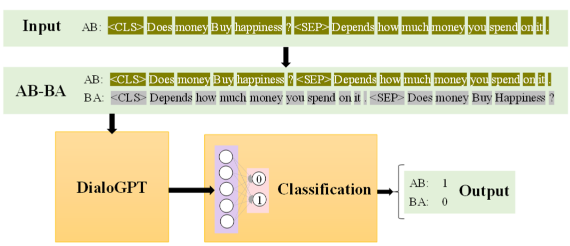

Figure 3: Example for training the AB-BA sub-metric.

A.5 Example for training AB-BA

For AB-BA and AB-AC, we tested three pretrained models (DialogGPT, BERT and RoBERTa) and found that there were only slight differences between the results. We used the best-performing one as the final model for each sub-metric. In particular, we added a fully-connected layer on the top of DialogGPT to determine whether a generated response is coherent. An example for training AB-BA is shown in Figure 3.

A.6 The Correlation-to-qualities Test on FED

We tested the correlations of metrics to different annotation qualities on one test dataset (DSTC10-Persona) and one development dataset (FED). The results are shown in Figure 2 and 4, respectively.

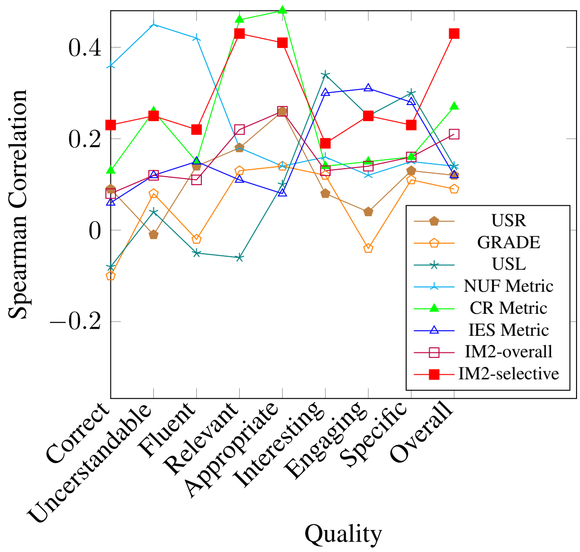

Figure 4: The correlation to different annotation qualities on the FED data (one of development datasets).