Unsupervised Evaluation of Interactive Dialog with DialoGPT

Shikib Mehri and Maxine Eskenazi
Dialog Research Center, Language Technologies Institute
Carnegie Mellon University, USA
{amehri,max}@cs.cmu.edu

Abstract

It is important to define meaningful and interpretable automatic evaluation metrics for open-domain dialog research. Standard language generation metrics have been shown to be ineffective for dialog. This paper introduces the **FED** metric (fine-grained evaluation of dialog), an automatic evaluation metric which uses DialoGPT, without any fine-tuning or supervision. It also introduces the FED dataset which is constructed by annotating a set of human-system and human-human conversations with eighteen fine-grained dialog qualities. The FED metric (1) does not rely on a ground-truth response, (2) does not require training data and (3) measures fine-grained dialog qualities at both the turn and whole dialog levels. FED attains moderate to strong correlation with human judgement at both levels.

1 Introduction

Evaluation metrics often define the research direction of a field. As dialog systems begin to demonstrate human-level performance, the development and adoption of meaningful and interpretable automatic evaluation measures is essential (Zhang et al., 2019; Adiwardana et al., 2020). Since standard metrics (e.g., BLEU, METEOR) have been shown to be ineffective for dialog (Deriu et al., 2019; Liu et al., 2016), human evaluation is often used. However, it is typically only used as a final evaluation since it is costly. During development, systems are generally optimized for poorly correlated automatic metrics which can result in sub-par performance (Dinan et al., 2019). Automatic metrics must be meaningful and interpretable so that they can be used to compare dialog systems, understanding their respective strengths and weaknesses, and effectively guide dialog research.

Dialog evaluation is difficult for several reasons: (1) The one-to-many nature of dialog (Zhao

et al., 2017) makes word-overlap metrics ineffective for scoring valid responses that deviate from the ground-truth (Liu et al., 2016; Gupta et al., 2019). (2) Dialog quality is inherently multi-faceted (Walker et al., 1997; See et al., 2019) and an interpretable metric should measure several qualities (e.g., *interesting*, *relevant*, *fluent*). (3) Dialog systems have begun to be evaluated in an interactive setting (Ram et al., 2018; Adiwardana et al., 2020) where a real user has a back-and-forth conversation with a system. Interactive evaluation is not constrained to a static corpus and better captures the performance of a system in a realistic setting. However, the existing automatic metrics compare to a ground-truth response, making them unsuitable for assessing interactive conversations. To address these three problems, this paper presents the **FED** metric (fine-grained evaluation of dialog) which assesses eighteen qualities of dialog without relying on a reference response.

First, a dataset of human quality annotations is collected for the human-system (Meena and Mitsuku) and human-human conversations released by Adiwardana et al. (2020). Dialogs are annotated at both the turn level and the dialog level for eighteen fine-grained dialog qualities. This FED dataset can be used to benchmark the performance of automatic metrics relative to human judgement. Analysis of this data provides insight into the qualities of dialog that are most important to human annotators. It therefore highlights the qualities that should be the focus of attention in dialog research.

The FED dataset is intended only for evaluating automatic metrics relative to human judgement. It does not consist of any training data. As such, this paper addresses the task of developing an automatic evaluation metric which (1) does not compare to a reference response, (2) assesses eighteen different qualities of dialog and (3) relies on no training data or supervision. This paper is the first, to the best

of our knowledge, to address this important and challenging problem.

The FED metric described here leverages a massively pre-trained model, DialoGPT (Zhang et al., 2019), which can generate practically human-level responses. Kocijan et al. (2019) assert that pre-trained models implicitly capture world knowledge and can therefore perform common-sense reasoning. Similarly, we posit that DialoGPT has implicitly captured some notion of dialog quality and can therefore be used for dialog evaluation. Eskenazi et al. (2019) assessed the quality of a system utterance in an interactive setting by looking at the *following user response*. The proposed evaluation metric is based on the same intuition. Given a system response, its quality is measured by computing the likelihood that DialoGPT will respond to it with a particular follow-up utterance (e.g., *"That is really interesting!"*). DialoGPT is more likely to respond in this way to what it believes is an *interesting* system response. A set of follow-up utterances is constructed for each of the eighteen qualities and the likelihoods of these follow-up utterances are used to measure dialog quality.

The FED metric obtains moderate to strong correlation with human judgement for turn-level and dialog-level evaluation without any training data or ground-truth response. Analysis in this paper demonstrates that through large-scale pre-training, DialoGPT has implicitly captured some notion of dialog quality. These results suggest that pre-trained models can be leveraged to further improve dialog evaluation.

The contributions of this paper are as follows: (1) The FED dataset was collected for fine-grained evaluation of interactive dialog, with annotations for eighteen dialog qualities at both the turn- and the dialog-level. (2) Analysis of the FED dataset identifies the dialog qualities most important to human annotators. (3) DialoGPT is shown to implicitly capture an understanding of dialog quality. (4) The FED metric has moderate to strong correlation with human judgement by leveraging DialoGPT, without training data or reference responses.

2 Related Work

2.1 Automatic Dialog Evaluation

Standard automatic metrics for language generation have been shown to correlate poorly with human judgement of dialog (Liu et al., 2016; Lowe et al., 2017; Gupta et al., 2019). This poor per-

formance can largely be explained by the one-to-many nature of dialog (Zhao et al., 2017). To avoid comparing to a single reference response, several authors have proposed using multiple reference responses. Multiple reference responses can be obtained with retrieval models (Galley et al., 2015; Sordoni et al., 2015) or through data collection (Gupta et al., 2019). These multi-reference metrics show performance improvement, but it is infeasible to thoroughly cover the space of all potential responses. The FED metric does not rely on a ground-truth response.

Lowe et al. (2017) train ADEM to produce a quality score conditioned on the dialog context, the reference response and the generated response. Venkatesh et al. (2018) present a framework for evaluating Alexa prize conversations which attains moderate correlation with user ratings. Both methods are trained on explicit quality annotations. In contrast, the FED metric proposed here requires no supervision.

Mehri and Eskenazi (2020) introduce USR, an unsupervised and reference-free evaluation metric for dialog generation. Similar to FED, USR uses pre-trained models to assess several dialog qualities. However, they are limited to five qualities with hand-designed models and unsupervised tasks for each quality. In comparison, FED is more general and encapsulates eighteen dialog qualities.

2.2 Dialog Qualities

Human evaluation in dialog is often limited to only measuring overall quality or response appropriateness. However, dialog quality is multi-faceted and should not be reduced to a single measurement.

PARADISE (Walker et al., 1997), one of the first frameworks for dialog evaluation, measured several different properties of dialog and combined them to estimate user satisfaction. See et al. (2019) used a variety of human judgements for dialog including interestingness, making sense, avoiding repetition, fluency, listening and inquisitiveness. See et al. (2019) emphasize the importance of measuring multiple qualities when evaluating dialog systems. There are several examples of human evaluation of multiple dialog qualities. Gopalakrishnan et al. (2019) annotate system responses using: interesting, comprehensible, on-topic and use of knowledge. Shin et al. (2019) measure empathy, fluency and relevance. Zhang et al. (2019) evaluate responses using relevance, informativeness and

human-likeness. Adiwardana et al. (2020) evaluate in both static and interactive environments using specificity and sensibleness.

2.3 Pre-trained Dialog Models

The success of pre-trained language models (Radford et al., 2018; Devlin et al., 2018) has recently been extended to the domain of dialog. Zhang et al. (2019) pre-train DialoGPT on Reddit and attain human-level performance on the task of response generation. The open-source DialoGPT model was used to construct the FED metric presented in this paper. (Adiwardana et al., 2020) similarly pre-trained their Meena dialog system on an unspecified large conversational dataset.

3 Data Collection

A dataset of human quality annotations was collected to assess automatic metrics by measuring correlation with human judgements. Adiwardana et al. (2020) collected a set of conversations[1] between a human and two open-domain dialog systems, Meena (Adiwardana et al., 2020) and Mitsuku[2]. In addition, they also released human-human dialogs collected in the same environment where one of the humans was selected to play the role of the system. We annotated a subset of these conversations with human quality judgements to create the FED dataset.

Workers on Amazon Mechanical Turk (AMT) annotated 40 Human-Meena conversations, 44 Human-Mitsuku conversations and 40 Human-Human conversations. For each conversation, three system responses were hand-selected to be annotated at the turn level, presented to the worker sequentially. Then the worker was shown the entire conversation and annotated on the dialog level. Five workers annotated each conversation. They did not know which system was involved in a conversation, since all mentions of the system name were replaced with the word "System."

Since dialog quality is inherently multi-faceted it is important to measure several different qualities of dialog. Eighteen fine-grained dialog qualities are measured in the FED dataset: eight at the turn level and ten at the dialog level.

---

[1]https://github.com/google-research/google-research/tree/master/meena

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

3.1 Turn-Level Annotation

Given a dialog context and a system response, the worker assessed the response according to eight fine-grained measures as well as for overall quality. The list of turn-level measures is shown in Table 1. The options for each of the fine-grained qualities were: *No*, *Somewhat*, *Yes*, *N/A*. For *understandable*, the *Somewhat* option was not provided, similar to prior past work (Gopalakrishnan et al., 2019). Responding *N/A* required written justification. The overall impression question was measured on a five-point Likert scale.

The workers were given detailed instructions and examples for each question presented in Table 1. These instructions are provided in the supplementary materials.

3.2 Dialog-Level Annotation

For dialog-level annotation, workers were asked to label the quality of a system over the duration of an entire conversation. The dialog-level questions listed in Table 2 cover ten fine-grained dialog qualities and an additional question on overall impression. The available options for each of the fine-grained qualities were *No*, *Somewhat*, *Yes*, *N/A*. For *consistency*, the *Somewhat* option was not provided because the existence of an inconsistency is binary. Overall impression was measured on a five-point Likert scale.

3.3 Dataset Statistics

A total of 124 conversations were annotated (40 Meena, 44 Mitsuku, 40 Human). Five different workers saw each conversation (HIT). Each conversation had one dialog-level annotation and three turn-level annotations for chosen system responses that were randomly sampled from the conversation. There were 9 questions for turn-level annotation and 11 for dialog-level annotation. In total, the FED dataset includes 3348 turn-level and 1364 dialog-level data points, for a total of 4712. This dataset intended to be used solely for the evaluation of metrics, as the number of annotated conversations is not large enough to accommodate both training and testing.

3.4 Data Processing

Given that each of the 4712 data points was labeled by five annotators, post-processing was used to improve the quality of the data through the removal of outliers. Given five annotations for a given ques-

Question	Used By
To the average person, is the response **interesting**?	See et al. (2019); Gopalakrishnan et al. (2019); Mehri and Eskenazi (2020)
Is the response **engaging**?	Yi et al. (2019)
Is the response **generic** or **specific** to the conversation?	Adiwardana et al. (2020)
Is the response **relevant** to the conversation?	See et al. (2019); Gopalakrishnan et al. (2019); Shin et al. (2019); Zhang et al. (2019); Mehri and Eskenazi (2020)
Is the response **correct** or was there a misunderstanding of the conversation?	None specifically
Is the response **semantically appropriate**?	See et al. (2019)
Is the response **understandable**?	Gopalakrishnan et al. (2019); Mehri and Eskenazi (2020)
Is the response **fluently written**?	See et al. (2019); Shin et al. (2019); Zhang et al. (2019); Ghandeharioun et al. (2019); Mehri and Eskenazi (2020)
Overall impression of the response?	Many

Table 1: The questions asked for turn-level annotation. Examples of prior work that has used each dialog quality are listed. No one has specifically used *Correct*, however its meaning is often encapsulated in *Relevant*.

Question	Used By
Throughout the dialog, is the system **coherent** and maintain a good conversation flow?	See et al. (2019)
Is the system able to **recover from errors** that it makes?	None
Is the system **consistent** in the information it provides throughout the conversation?	Qin et al. (2019)
Is there **diversity** in the system responses?	See et al. (2019); Ghandeharioun et al. (2019)
Does the system discuss topics in **depth**?	Guo et al. (2018)
Does the system display a **likeable** personality?	Shin et al. (2019); Ghandeharioun et al. (2019)
Does the system seem to **understand** the user?	See et al. (2019)
Is the system **flexible and adaptable** to the user and their interests?	Guo et al. (2018)
Is the system **informative** throughout the conversation?	Zhang et al. (2019)
Is the system **inquisitive** throughout the conversation?	See et al. (2019)
Overall impression of the dialog?	Many

Table 2: The qualities annotated at the dialog-level. Examples of prior work that has used each dialog quality are listed. To our knowledge, error recovery has not been used for human evaluation.

Quality	Spearman
Turn-Level	
Interesting	0.819
Engaging	0.798
Specific	0.790
Relevant	0.753
Correct	0.780
Semantically Appropriate	0.682
Understandable	0.522
Fluent	0.714
Overall Impression	0.820
Dialog-Level	
Coherent	0.809
Error Recovery	0.840
Consistent	0.562
Diverse	0.789
Topic Depth	0.833
Likeable	0.838
Understanding	0.809
Flexible	0.816
Informative	0.806
Inquisitive	0.769
Overall Impression	0.830

Table 3: Spearman correlation for each of the dialog qualities. The correlation was measured by correlating each annotation with the mean of the other annotations for the same question.

Quality	Mitsuku	Meena	Human
Turn-Level			
Interesting	2.30	**2.58**	2.35
Engaging	2.53	**2.75**	2.49
Specific	2.48	**2.74**	2.56
Relevant	2.80	**2.88**	2.74
Correct	2.74	**2.84**	2.66
Semantically-Appropriate	2.84	**2.92**	2.85
Understandable	**0.97**	**0.97**	0.94
Fluent	2.83	**2.90**	2.80
Overall	3.81	**4.19**	3.85
Dialog-Level			
Coherent	2.20	2.88	**2.94**
Error Recovery	2.22	2.69	**2.86**
Consistent	0.82	0.95	**0.98**
Diverse	2.23	2.46	**2.88**
Topic Depth	1.80	2.28	**2.78**
Likeable	2.10	2.61	**2.97**
Understanding	2.23	2.86	**2.98**
Flexible	2.22	2.72	**2.97**
Informative	2.10	2.60	**2.85**
Inquisitive	2.35	2.76	**2.88**
Overall	3.10	4.11	**4.60**

Table 4: Performance of each system on the fine-grained qualities. All scores are 1-3, except Understandable and Consistent are 0-1 and Overall is 1-5.

tion, the furthest label from the mean is removed if its distance from the mean is greater than half the standard deviation of the five annotations.

4 Data Analysis

The fine-grained nature of the FED dataset is grounds for a rich analysis. First, inter-annotator agreement is evaluated for all of the dialog qualities. Next, the dataset is used to better understand the comparative strengths and weaknesses of the three systems (Mitsuku, Meena, Human). Finally, detailed analysis of the data provides insight into the fine-grained qualities that most strongly contribute to the annotators' overall impression.

4.1 Inter-Annotator Agreement

To compute inter-annotator agreement, the correlation between each annotation and the mean of the five (or four, after outlier removal) annotations for the same question is measured. The Spearman correlation for each turn-level and dialog-level question is shown in Table 3

Inter-annotator agreement is high for all of the dialog qualities, suggesting that all of the qualities were well-understood by the annotators and relevant and that the instructions removed much of the ambiguity from the task. Two qualities, *understandable* and *consistent*, have slightly lower correlations, in the 0.5 - 0.6 range. These qualities did not include *Somewhat* as an answer. This probably contributed to the lower inter-annotator agreement.

4.2 System Performance

While Adiwardana et al. (2020) presented a performance comparison between Mitsuku, Meena and Humans in an interactive setting, their evaluation only used two qualities: *specificity* and *sensibility*. In contrast, the FED dataset has eighteen fine-grained qualities thus providing more information about the strengths and weaknesses of each system.

The fine-grained performance of each system shown in Table 4. For all of the turn-level qualities,

Meena outperforms both Mitsuku and Human. The strength of Meena is most noticeable for *interesting*, *engaging* and *specific*.

However, turn-level qualities are insufficient to evaluate a dialog system. Dialog is by definition a multi-turn interaction. Thus, in some cases, a sub-optimal system response might result in a better long-term dialog. Humans significantly outperform the two systems for dialog-level qualities. The difference between Meena and Mitsuku is very pronounced at the dialog level, with a 1 point difference in overall score. The higher variance in scores and the stronger performance of human dialogs, shows that dialog-level evaluation is reliable than turn-level. Meena's scores suggest that it is fairly *coherent*, *understanding* and *flexible*. However, it struggles with *diversity*, *topic depth* and *likeable*.

4.3 Fine-Grained Quality Analysis

The FED dataset can be used to examine the relative importance of each fine-grained dialog quality by measuring its contribution to the overall impression. For both turn-level and dialog-level, a regression is trained to predict the overall score given the fine-grained qualities as input. The regression weights provide insight into the fine-grained qualities that most contribute to the overall impression as labeled by human annotators. A softmax is computed over the regression weights to determine the relative contribution of each fine-grained dialog quality. A dialog quality with a higher weight contributes more to the human's overall impression. The results are shown in Table 5.

The most important turn-level qualities are *interesting*, *relevant* and *fluent*. This suggests that developing a system that is consistently interesting, relevant and fluent will result in the highest improvement in the user's overall impression. There is less variance in the importance of dialog-level qualities than in the turn-level qualities possibly because there is less overlap in meaning amongst the qualities and all of the dialog-level qualities seem somewhat important. The most important dialog-level qualities are *coherent*, *likeable* and *understanding*. Improving a system's coherence, understanding of the user and its likeableness would thus be the most likely way to improve the overall impression of a dialog system.

Quality	Importance (%)
Turn-Level	
Interesting	**16.15**
Engaging	7.46
Specific	9.64
Relevant	**18.10**
Correct	13.77
Semantically Appropriate	9.90
Understandable	10.70
Fluent	**14.27**
Dialog-Level	
Coherent	**10.95**
Error Recovery	9.15
Consistent	7.92
Diverse	10.09
Topic Depth	10.51
Likeable	**12.03**
Understanding	**11.01**
Flexible	10.34
Informative	8.00
Inquisitive	9.50

Table 5: Relative importance of each dialog quality for predicting the overall impression. The most important qualities for turn-level and dialog-level are in bold.

5 Methods

The FED (fine-grained evaluation of dialog) metric is an automatic evaluation metric for dialog which (1) does not need to compare to a reference response, (2) measures eighteen fine-grained qualities of dialog, and (3) does not use training data. Capturing a diverse set of fine-grained qualities without supervision is an especially challenging problem.

The development of the FED metric is motivated by two areas of prior work: (1) pre-trained language models and their capabilities and (2) the use of follow-up utterances as a means of evaluation.

5.1 DialoGPT

Zhang et al. (2019) extend GPT-2 (Radford et al., 2018) to train DialoGPT on 147M conversation-like interactions from Reddit. As per their evaluation, DialoGPT outperforms humans at producing relevant, interesting and human-like responses.

Kocijan et al. (2019) show that pre-trained language models, specifically BERT (Devlin et al., 2018), implicitly capture world knowledge and can therefore perform common sense reasoning. By calculating which answer results in a more proba-

ble sentence according to BERT, they strongly out-perform other methods on the Winograd Schema Challenge (Levesque et al., 2012).

Just as BERT has been shown to capture world knowledge, we posit that DialoGPT has implicitly captured some notion of dialog quality. The qualities of a particular dialog context (e.g., *interesting*, *relevant*, *informative*) likely inform DialoGPT's response and, as such, must be captured by the model. If there was training data for the eighteen dialog qualities, this hypothesis could be verified by fine-tuning DialoGPT for the task of dialog evaluation. Without training data, however, the challenge is to devise an unsupervised mechanism for extracting the quality information captured by DialoGPT.

5.2 Follow-Up Utterance for Evaluation

Eskenazi et al. (2019) assess the quality of a system utterance in an interactive setting, by looking at the *following user response*. When users speak to a system, their response to a given system utterance may implicitly or explicitly provide feedback for the system. For example, if a user follows up a system utterance with *"That's not very interesting"*, they are providing information about the quality of the system utterance.

The conversations in the FED dataset were collected in an interactive setting. Thus the use of the follow-up utterance is a valid option. Even if users consistently provided feedback, it would be difficult to interpret without training data.

5.3 Evaluating with DialoGPT

The proposed FED metric is motivated by (1) the intuition that DialoGPT has implicitly learned to reveal dialog quality and (2) that the follow-up utterance can provide valuable information about a system response. To measure the quality of a system response s, we compute the likelihood of the model generating various follow-up utterances (e.g., *"Wow! Very interesting."*) in response to s. DialoGPT will be more likely to respond with a positive follow-up utterance if given a better (e.g., more *interesting/relevant/fluent*) preceding system utterance.

For each of the eighteen fine-grained dialog qualities, a set of positive follow-up utterances, p, and a set of negative follow-up utterances, n, is constructed. Specifically, given a dialog context c, a system response r and a function $\mathcal{D}$ that computes the log-likelihood of DialoGPT generating a particular response, the predicted score for a dialog quality is calculated as:

$$\sum_{i=1}^{|p|} \mathcal{D}(c + r, p_i) - \sum_{i=1}^{|n|} \mathcal{D}(c + r, n_i) \qquad (1)$$

This equation can be modified to predict scores for dialog-level qualities, by simply removing the system response r from the equation.

A response is said to be *interesting* if it is more likely that DialoGPT (acting as the user) responds with a positive follow-up utterance (e.g., *"Wow! Very interesting"*) than with a negative one (e.g., *"That's really boring"*). For each of the eighteen qualities, several positive and negative utterances were hand-written and minimally tuned on a small subset of the dataset (10 conversations). Follow-up utterances for each quality are provided in the supplementary materials.

Generally, negative follow-up utterances are more meaningful than positive ones. For example, if a system response is *irrelevant*, a follow-up utterance of *"That's not relevant"* is reasonable. However, acknowledging the relevance of a system response is less likely. Therefore the log-likelihood produced by DialoGPT will be noisier and less informative. The number of positive utterances for each dialog quality ranges between 0 and 4, and the number of negative utterances ranges between 1 and 4. While the fine-grained qualities are computed in this manner, the overall impression scores are calculated as an average of the scores for either the turn-level or dialog-level qualities.

6 Results

6.1 Experimental Setup

The FED metric was evaluated using four variations of the pre-trained DialoGPT model. The pre-trained DialoGPT models can be either medium size: 345M or large: 762M. They are either fine-tuned from GPT-2 (Radford et al., 2018) or trained from scratch. The follow-up utterances were hand-written and minimally tuned on 10 conversations using the 762M fine-tuned model. The small (117M) DialoGPT model was not used since Zhang et al. (2019) demonstrated its poor performance.

Most of the turn-level qualities were scored using only the last system response as context. For *relevant*, *correct* and dialog-level metrics, the entire conversation was used as context.

6.2 Correlation with Human Judgement

The Spearman correlation was measured between the predicted quality scores and the mean of the annotated scores. Correlations for all the dialog qualities, and all four variations of the underlying DialoGPT model are shown in Table 6.

The best overall turn-level correlation is **0.209** and the best overall dialog-level correlation is **0.443**. To our knowledge, there are presently no other metrics that operate without a ground-truth response, thus these results cannot be directly compared to any existing metrics. However, prior work on dialog evaluation reveals roughly similar correlation. Multi-reference evaluation for dialog achieves correlations in the 0.10 - 0.27 range (Gupta et al., 2019) and ADEM has correlations in the 0.28 - 0.42 range (Lowe et al., 2017). Given neither training data nor ground-truth response, the FED metric performs competitively relative to this prior work.

6.3 Discussion

The FED metric works better for some dialog qualities than others. This is because DialoGPT was trained on Reddit. It is more likely that it has captured certain dialog qualities that Reddit exhibits. For example, it is more likely that DialoGPT learns to measure qualities like *interesting* and *engaging*, than *understandable* and *consistent*. In the Reddit training data, the former two qualities show more variation than the latter. For example, there are interesting and un-interesting utterances, however most utterances on Reddit are generally understandable. The former two qualities are also more likely to influence the system response. Conversely, the latter two qualities are unlikely to be acknowledged in the response. For example, since Reddit is a multi-participant forum and not a one-on-one conversation, inconsistencies in conversation history are unlikely to be reflected in the response. As such, it is unsurprising that this approach struggles to measure the consistency of a dialog.

An optimal generation model (e.g., a human) should exhibit compositionality and be capable of producing utterances that have never been observed. For example, even if *'That is not consistent'* has never appeared in the training data, a compositional model would be capable of generating it. This difference in performance across the different dialog qualities suggests that DialoGPT exhibits some degree of compositionality, as evidenced by its ability to compose some follow-up utterances which are

not frequently observed in the Reddit data (e.g., *'You really don't know much?'*), however it still struggles with follow-up utterances consisting of less frequently observed concepts (e.g., *consistent*, *understandable*).

DialoGPT could be used to better measure these qualities by fine-tuning on additional conversational data from a source other than Reddit or on a training set annotated with human quality judgements. However, even without additional fine-tuning, FED effectively measures many qualities.

This paper has carried out an assessment of the FED metric for three open-domain conversation agents: Meena, Mitsuku and Human. Since these three systems are different in nature and FED exhibits strong correlation with human judgements across all the systems, we believe that the performance of FED will hold for other open-domain dialog systems and will not be restricted to a particular type of model or a specific dataset. However, the FED dataset consists of only open-domain chit-chat conversations. As such, future work is needed to determine whether the FED metric will generalize to goal-oriented dialog. Since DialoGPT has not observed goal-oriented training data, it may be necessary to use self-supervised fine-tuning on the new domain (Mehri and Eskenazi, 2020).

As with all automated metrics, there is the potential to game the FED metric and obtain artificially high scores, especially by having a model produce responses that are likely to result in specific follow-up utterances. To this end, the FED metric is not a replacement for human evaluation. It is instead a means of measuring dialog quality for the purposes of validation and model tuning.

The FED metric is (1) unsupervised, (2) does not rely on a reference response and (3) can be used to assess many dialog qualities. By having DialoGPT play the role of the user and assign probabilities to follow-up utterances, we have devised a mechanism of extracting information about dialog quality without any supervision. This mechanism is versatile and could potentially be extended to other dialog qualities.

7 Conclusion

This paper introduces the FED dataset and the FED metric. The FED dataset is constructed by annotating a set of interactive conversations with eighteen fine-grained dialog qualities. The FED metric can be used to measure fine-grained qualities of dia-

Quality	345M fs	345M ft	762M fs	762M ft
Turn-Level				
Interesting	0.388	**0.431**	0.406	0.408
Engaging	0.268	0.285	0.278	**0.318**
Specific	0.260	**0.326**	0.270	0.267
Relevant	*0.028*	*-0.027*	*0.001*	**0.152**
Correct	*0.000*	*0.037*	*0.020*	**0.133**
Semantically Appropriate	*0.040*	**0.177**	0.141	0.155
Understandable	*0.047*	*0.048*	*0.075*	**0.111**
Fluent	0.157	0.184	0.133	**0.224**
Overall	0.122	*0.092*	*0.094*	**0.209**
Dialog-Level				
Coherent	0.195	*0.151*	*0.149*	**0.251**
Error Recovery	*0.165*	*0.128*	*0.126*	*0.165*
Consistent	*0.041*	*0.011*	*0.006*	*0.116*
Diverse	**0.449**	0.431	0.414	0.420
Topic Depth	**0.522**	0.479	0.470	0.476
Likeable	*0.047*	*0.172*	0.224	**0.262**
Understanding	0.237	0.174	0.192	**0.306**
Flexible	0.260	**0.408**	0.298	0.293
Informative	0.264	0.328	**0.337**	0.288
Inquisitive	*0.137*	*0.143*	**0.298**	0.163
Overall	0.401	0.359	0.355	**0.443**

Table 6: Spearman correlations with human judgement. All values that are not statistically significant ($p > 0.05$) are italicized. The highest correlation for each quality is shown in bold.

log without comparing to a ground-truth response. By having DialoGPT take the role of the user and calculate the likelihood of follow-up utterances, the FED metric attains moderate to strong correlation with human judgement, without the use of any training data. The FED metric is inherently versatile and generalizable, making it applicable to other dialog qualities, domains or tasks. Both the FED dataset and the code for the FED metric will be released upon acceptance of this paper.

This paper sets the groundwork for several directions of future work. (1) The FED dataset can be used to benchmark automatic evaluation metrics on eighteen fine-grained dialog qualities. (2) Building on this paper, future work could identify mechanisms that further leverage pre-trained models for dialog evaluation. (3) Future work can explore strategies for extending the FED metric beyond open-domain chit-chat conversations to goal oriented dialog. (4) The FED metric can be used to evaluate, analyze and improve dialog systems.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2019. Survey on evaluation methods for dialogue systems. *arXiv preprint arXiv:1905.04071*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2019. The second conversational intelligence challenge (convai2). *arXiv preprint arXiv:1902.00098*.

Maxine Eskenazi, Shikib Mehri, Evgeniia Razumovskaia, and Tiancheng Zhao. 2019. Beyond turing: Intelligent agents centered on the user. *arXiv preprint arXiv:1901.06613*.

Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. 2015. deltaBLEU: A discriminative metric for generation tasks with intrinsically diverse targets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 445–450, Beijing, China. Association for Computational Linguistics.

Asma Ghandeharioun, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah Jones, Agata Lapedriza, and Rosalind Picard. 2019. Approximating interactive human evaluation with self-play for open-domain dialog systems. In *Advances in Neural Information Processing Systems*, pages 13658–13669.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, Dilek Hakkani-Tür, and Amazon Alexa AI. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. *Proc. Interspeech 2019*, pages 1891–1895.

Fenfei Guo, Angeliki Metallinou, Chandra Khatri, Anirudh Raju, Anu Venkatesh, and Ashwin Ram. 2018. Topic-based evaluation for conversational bots. *arXiv preprint arXiv:1801.03622*.

Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskenazi, and Jeffrey P Bigham. 2019. Investigating evaluation of open-domain dialogue systems with human generated multiple references. *arXiv preprint arXiv:1907.10568*.

Vid Kocijan, Ana-Maria Cretu, Oana-Maria Camburu, Yordan Yordanov, and Thomas Lukasiewicz. 2019. A surprisingly robust trick for winograd schema challenge. *arXiv preprint arXiv:1905.06290*.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

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 (Volume 1: Long Papers)*, pages 1116–1126, Vancouver, Canada. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020. Usr: An unsupervised and reference free evaluation metric for dialog generation. *arXiv preprint arXiv:2005.00456*.

Libo Qin, Yijia Liu, Wanxiang Che, Haoyang Wen, Yangming Li, and Ting Liu. 2019. Entity-consistent end-to-end task-oriented dialogue system with kb retriever. *arXiv preprint arXiv:1909.06762*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf*.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, et al. 2018. Conversational ai: The science behind the alexa prize. *arXiv preprint arXiv:1801.03604*.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. *arXiv preprint arXiv:1902.08654*.

Jamin Shin, Peng Xu, Andrea Madotto, and Pascale Fung. 2019. Happybot: Generating empathetic dialogue responses by improving user experience lookahead. *arXiv preprint arXiv:1906.08487*.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and William B. Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *HLT-NAACL*.

Anu Venkatesh, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Behnam Hedayatnia, Angeliki Metallinou, et al. 2018. On evaluating and comparing open domain dialog systems. *arXiv preprint arXiv:1801.03625*.

Marilyn A Walker, Diane J Litman, Candace A Kamm, and Alicia Abella. 1997. Paradise: A framework for evaluating spoken dialogue agents. *arXiv preprint cmp-lg/9704004*.

Sanghyun Yi, Rahul Goel, Chandra Khatri, Tagyoung Chung, Behnam Hedayatnia, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tur. 2019. Towards coherent and engaging spoken dialog response generation using automatic conversation evaluators. *arXiv preprint arXiv:1904.13015*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. *arXiv preprint arXiv:1703.10960*.