Towards Holistic and Automatic Evaluation of Open-Domain Dialogue Generation

Bo Pang[1]*, **Erik Nijkamp**[1]*, **Wenjuan Han**[2]*§, **Linqi Zhou**[1]*, **Yixian Liu**[3], **Kewei Tu**[3]

[1]Department of Statistics, University of California, Los Angeles
[2] School of Computing, National University of Singapore, Singapore
[3]School of Information Science and Technology, ShanghaiTech University

{bopang, enijkamp, linqi.zhou}@ucla.edu
dcshanw@nus.edu.sg
{liuyx, tukw}@shanghaitech.edu.cn

Abstract

Open-domain dialogue generation has gained increasing attention in Natural Language Processing. Its evaluation requires a holistic means. Human ratings are deemed as the gold standard. As human evaluation is inefficient and costly, an automated substitute is highly desirable. In this paper, we propose holistic evaluation metrics that capture different aspects of open-domain dialogues. Our metrics consist of (1) GPT-2 based context coherence between sentences in a dialogue, (2) GPT-2 based fluency in phrasing, (3) n-gram based diversity in responses to augmented queries, and (4) textual-entailment-inference based logical self-consistency. The empirical validity of our metrics is demonstrated by strong correlations with human judgments. We open source the code and relevant materials. [1]

1 Introduction

Learning to communicate is a key capacity of intelligent agents. Research on enabling a machine to have meaningful and natural conversations with humans plays a fundamental role in developing artificial general intelligence, as can be seen in the formulation of Turing test (Turing, 1950). Recently open-domain or non-task-oriented dialogue systems have attracted a surge of research interest (Bessho et al., 2012; Sordoni et al., 2015; Shang et al., 2015; Vinyals and Le, 2015; Serban et al., 2016, 2017; Ghazvininejad et al., 2018).

Evaluating models of open-domain dialogue generation in an efficient manner poses a significant challenge in developing dialogue systems. The prevalent method of open-domain dialogue evaluation is human-based rating with a given rubric. When various variations in the model and sets of hyper-parameters are needed, the labor-intensive human evaluation is deemed impracticable. This key drawback may hinder the research progress and render the human evaluation approach not scalable.

Context of Conversation	
Speaker A:	Hey, what do you want to do tonight?
Speaker B:	Why don't we go see a movie?
Model Response	
Speaker A:	Nah, let's do something active.
Reference Response	
Speaker A:	Yeah, the film about Turing looks great!
Context of Conversation	
Speaker A:	Do you have maps of downtown area?
Speaker B:	*Yes,* here you are.
Speaker A:	How much is it?
Model Response	
Speaker B:	I've no idea, *I don't have any maps of the area.*

Table 1: Two responses from an dialogue system (Wolf et al., 2019) on Daily Dialogue Dataset. The first generated response appears reasonable within the open-domain dialogue, while its BLEU score and semantic similarity between model response and reference response is low. The second generated response conflicts with its prior utterances. The *italic* text highlights the logical contradiction.

Previous automatic evaluation metrics generally focus on the quality of the dialogue generation: context coherence and fluency. Word-overlap metrics (Papineni et al., 2002; Banerjee and Lavie, 2005; Lin, 2004) or ad-hoc classifiers (Tao et al., 2018; Ghazarian et al., 2019) are designed for measuring the quality. In open-domain dialogue, the relation between two utterances is more critical as shown in the first example of Table 1. Compared with the previous two approaches, a language model, trained on an enormous amount of text, can naturally capture coherence among both words and utterances. On the other hand, a good evaluation metric should not only measure the quality of generation, but also the diversity of generation, which is especially important for open-ended tasks like di-

*Equal contributions.

§Wenjuan Han is the corresponding author. Wenjuan Han contributed to this work when at ShanghaiTech University.

[1] https://github.com/alexzhou907/dialogue_evaluation.

alogue or story generation (Hashimoto et al., 2019). Some n-gram based metrics have been utilized to measure diversity (Mou et al., 2016; Serban et al., 2017). However, this metric might be improper for diversity evaluation since the generated utterances given various queries provided by the benchmark are generally diverse. In our experiments, we observe constantly high diversity in terms of human ratings and n-gram based entropy when evaluating the generated responses directly. In addition to the three aforementioned metrics, logical self-consistency is also a key aspect of dialogue models (Zhang et al., 2018). An dialogue example with logical contradiction is displayed in the second example of Table 1. Welleck et al. (2019) measured logical self-consistency by transferring each sentence into a rule-based triple, (category, relation, category), with the help of human annotators. We are nevertheless unaware of any reliable automatic measure of logical consistency in open-domain dialogue.

In this work, we propose holistic metrics that evaluate distinctive aspects of generated dialogues. Specifically, we consider (1) *context coherence* of a dialogue: the meaningfulness of a response within the context of prior query, (2) *language fluency* of generated responses: the quality of phrasing relative to a human native speaker, (3) *response diversity* of a set of generated sentences: the variety in meaning and word choice of responses, and (4) *logical self-consistency*: the logical consistency of utterances from a dialogue agent. Both *context coherence* and *response fluency* (quality metrics) can naturally be captured by metrics based on strong language models like GPT-2 (Radford et al., 2019). Therefore, we propose to recruit and fine-tune GPT-2 as a basis of our quality metrics. With regard to *response diversity* and *logical self-consistency*, we propose to measure them under augmented utterances with controlled paraphrasing. We leverage two effective approaches to generate augmented utterances: word substitution and text generator with a k-best decoder. Moreover, we utilize n-gram based entropy to capture *response diversity* and entailment based approach to capture *logical self-consistency*. Our experiments show that the proposed metrics strongly correlate with human judgments. Moreover, our augmented datasets allow for a more accurate and straightforward human annotation, significantly improving the agreement between human evaluation. We release the code and relevant materials as open-source contribution to pave the way towards further research.

2 Prior Art

Heuristic-based metrics have been shown to align well with human judgments and widely applied in various language generation tasks. For machine translation, BLEU (Papineni et al., 2002) computes n-gram precision, whereas METEOR (Banerjee and Lavie, 2005) takes into account both precision and recall. For summarization, ROUGE (Lin, 2004) also considers both precision and recall by calculating F-measure. These n-gram based metrics are well-suited for the generation tasks that are more source-determined or low conditional entropy such as translation, image captioning, and summarization. Some dialogue studies adopted these metrics to evaluate the quality of generated conversation responses (Ritter et al., 2011; Su et al., 2018; Sordoni et al., 2015). They nevertheless are not suitable for open-ended generations or high conditional entropy tasks like dialogue generation where a diverse range of generations is acceptable conditional on a query. Indeed, Liu et al. (2016) conducts extensive empirical studies on these metrics (e.g., BLEU, METEOR, and ROUGE) to test their effectiveness on evaluating dialogue generation and find limited relation between these automatic metrics and human judgments.

The word-overlap metrics (e.g., BLEU) fail to capture the semantic similarity between model and reference responses. The following works leverage the distributed representation learned in neural network models to capture semantic similarity among context, model response, and reference response. Lowe et al. (2017) collect a dataset of human scores and train a hierarchical recurrent neural network (RNN) to predict human-like scores to input responses given the context, resulting in an automatic metric that has a medium level correlation with human judgments. Obtaining this metric however requires a large dataset of human-annotated scores, thus rendering this approach less flexible and extensible. Tao et al. (2018) proposes a referenced metric and unreferenced metric blended evaluation routine (RUBER) for open-domain dialogue systems. This blended metric is a combination of two metrics. A referenced metric measures the similarity between model-generated and reference responses on the basis of word-embeddings. An unreferenced metric captures the relevance between the query and response. It is obtained by training a neural network classifier to determine whether a response is appropriate. The positive examples are the references, while the negative examples are reference responses randomly chosen from the dataset, hence avoiding the need of human-annotated data. After training, the Softmax score is utilized to measure whether the generated response is coherent with the query. Attempting to improve RUBER, Ghazarian et al. (2019) explores to use contextualized embeddings from BERT. The BERT-based unreferenced metric improves over the word-embedding-based RUBER unreferenced metric. Interestingly, they show that the combined metric has a reduced correlation with human judgments than the unreferenced metric alone. Although this finding is counterintuitive, it is consistent with the characteristics of open-domain dialogue that a range of diverse responses is reasonable given a query. Hence a response can be acceptable to human annotators even if it does not align well with the reference either in terms of word-overlap or semantic embedding.

Context Coherence. One key component of dialogue response is its coherence to the query as explored in Tao et al. (2018) and Ghazvininejad et al. (2018). Prior work measures the coherence based on the Softmax score of a trained binary classifier. Here we explore an alternative approach based on language modeling (Bengio et al., 2003). A language model can naturally capture the coherence of the response to the query without resorting to an ad-hoc classifier.

Language Fluency. Besides coherence, a good response should be fluent. Fluency is often measured by a language model (Holtzman et al., 2018; Xu et al., 2018). We define the response fluency score as negative perplexity of generated responses.

Response Diversity. In addition to quality metrics, response diversity is also critical, especially for high conditional entropy tasks like dialogue or story generation (Hashimoto et al., 2019). Some n-gram based metric has been utilized to measure diversity. Mou et al. (2016) and Serban et al. (2017) compute unigram entropy across all generated utterances to measure the diversity. This metric might be improper for diversity since the generated utterances given various queries are generally diverse. In our experiments, we observe constantly high diversity in terms of human ratings and n-gram based entropy. In another perspective, the entropy computed across all generated responses is essentially measuring the marginal entropy of the responses, while our actual interest is in the conditional entropy of the responses conditional on the queries.

Logical Self-Consistency. Similar to diversity evaluation, current benchmarks are not suitable for evaluating *logical self-consistency*. The current dataset is well-formed making the system to generate a simple and nonredundant response, but unfortunately, there still exist logical contradictions as shown in Table 1. The natural language inference (NLI) task (Williams et al., 2018) aiming to check whether the sentence is entailed or contradicted by a previous sentence is highly related to logic evaluation on open-domain dialogues.

3 Metrics

3.1 Context Coherence

Language models, which predict the next token given previous tokens, naturally capture the coherence between sentences and particularly the dialogue query and response in our case. GPT-2 (Radford et al., 2019) is a large-scale pre-trained language model based on the transformer architecture (Vaswani et al., 2017). It is trained on a vast amount of diverse data and demonstrates impressive text generation capabilities. In order to better capture the dependence between the queries and responses, GPT-2 can be fine-tuned using the next sentence prediction task on the dialogue dataset of interest.

Suppose a query q contains tokens $\{q_t : t = 1, ..., T_q\}$ and a response r has tokens $\{r_t : t = 1, ..., T_r\}$. Let P denote the fine-tuned GPT-2, then the context coherence is defined as the log-likelihood of the response conditional on the the query normalized by the length of the response length:

$$\begin{aligned} c_{raw}(r|q) =& \frac{1}{T_r} \log \frac{P(q,r)}{P(q)} \\ =& \frac{1}{T_r} \sum_{t}^{T_r} \log P(r_t | r_{<t}, q). \end{aligned} \tag{1}$$

Note that $c_{raw}(r|q)$ is some negative number and unbounded from below. A single value is then hard to explain absolutely and can only be interpreted relative to other values. Also, the unboundedness renders it prone to extreme values. Hence, a normalized score is utilized instead. Since the score distribution varies as a function of the dataset, the lower bound is defined as 5th percentile, denoted as c_{5th}, instead of some arbitrary value. Then the normalized score, $c(r|q)$, is

$$c(r|q) = -\frac{max(c_{5th}, c_{raw}(r|q)) - c_{5th}}{c_{5th}} \quad (2)$$

which ranges from 0 to 1.

3.2 Response Fluency

To capture the fluency of responses, we also adopt the pretrained language model, GPT-2. In particular, the raw response fluency score, $f_{raw}(r)$, is defined as,

$$f_{raw}(r) = \frac{1}{T_r} \sum_{t}^{T_r} \log P(r_t | r_{<t}). \quad (3)$$

Similar to *context coherence*, a normalized version, $f(r)$, of $f_{raw}(r)$ is employed.

3.3 Response Diversity

Prior work (Mou et al., 2016; Serban et al., 2017) measured diversity by computing the n-gram entropy across all generated responses, which essentially reflects the marginal entropy of the responses. Diversity of the responses conditional on the query (e.g., conditional entropy) are however more of interest for dialogue models. On the other hand, if we measure diversity based on responses randomly sampled from a model conditional on a single query, the response quality is generally low (Caccia et al., 2018). The current work instead proposes to measure response diversity utilizing augmented datasets with controlled paraphrasing, which allows for measuring diversity among top-ranked responses conditional on paraphrased queries and hence avoiding the tradeoff or dependency between diversity and quality. In other words, for a given query, we slightly tilt the corresponding element in the query-response joint space along the query dimension (achieved by paraphrasing-augmentation) and then measure the entropy of high-quality responses in the neighbourhood of the targeted query.

While augmenting the queries to measure the conditional entropy of responses, we need to control the diversity of the augmented queries such that the augmented ones stay in the vicinity of the targeted query. Hence the goal of controlled augmentation is to minimize diversity in both meaning and word use and avoid feeding the dialogue model identical inputs. To achieve so, two augmentation approaches are considered: (1) WordNet (Miller, 1998) Substitution (WS) and (2) Conditional Text Generator (CTG).

WordNet Substitution (WS) is a word-level manipulation method that replaces some words with synonyms defined in WordNet. Different from WS, Conditional Text Generator (CTG) is used to augment queries in multi-turn dialogue. It requires a generator to produce augments conditioned on the context, which is defined as the prior utterance history to the selected query. For instance, suppose $[u_1; ...; u_{t-1}]$ denotes the utterance history and u_t indicates the query to be augmented, then the top-K beams, $\{u_t^{(1)}, ..., u_t^{(K)}\}$, from the CTG model conditional on the utterance history are produced.

Given the target query and a set of augmented queries for it with controlled paraphrasing, $\{u_t^{(k)} : k \in 0, ..., K\}$ where $u_t^{(0)} := u_t$, the corresponding responses are generated by the model under test. Then we can calculate the n-gram entropy for samples in the set $\{u_{t+1}^{(k)} : k \in 0, ..., K\}$.

3.4 Logical Self-Consistency

Logical self-consistency measures if a generated response is logically contradictory to what the agent uttered in the multi-turn history. The basic idea is to apply a pretrained Multi-Genre Natural Language Inference (MNLI; Williams et al. 2018) model to label if the relation of the response and the utterance history of the same agent is logically consistent. More specifically, we train a ternary classifier that takes two utterances as input and predicts the relation as either contradiction, entailment or neutral on the MNLI dataset. Then we average the contradiction class probabilities of the current utterance and each prior utterance from this agent as the contradiction score. In order to match the human ratings, we use 1 minus the contradiction score as the final score of *logical self-consistency* evaluation.

Moreover, we measure *logical self-consistency* under augmented datasets with controlled paraphrasing, using WS and CTG introduced in Section 3.3. The main idea is to generate augmented multi-turn utterance history that more likely induces the dialogue system to produce contradictory responses. We assume that it is more likely for the agent producing self-contradictory responses when responding to similar queries. We use WS and CTG to paraphrase the query and then calculate the contradiction score of the current utterance and each prior utterance from this agent.

4 Experiments

4.1 Dataset

To facilitate comparison with prior work (Ghazarian et al., 2019), the DailyDialog dataset (Li et al., 2017) is adopted for the empirical analysis of our proposed metrics. This dataset contains 13,118 high-quality multi-turn dialogue dataset. The dialogue is split into a 42,000 / 3,700 / 3,900 train-test-validation partitions.

4.2 Response Generation

A sequence-to-sequence (seq2seq) model with attention (Bahdanau et al., 2014) was trained with the train and validation partitions to generate dialogue responses. The implementation in OpenNMT (Klein et al., 2017) was used to train the model. The seq2seq consists of a 2-layer LSTM with 500 hidden units on both the encoder and decoder. The model was trained with SGD and learning rate of 1. To obtain responses on a wide spectrum of quality and diversity, we sample the data with top-k sampling where $k = \{1, 10, 100\}$.

4.3 Language Model Fine-tuning

The base GPT-2 model with 12 layers was used to compute our metrics [2]. The GPT-2 model was fine-tuned on the training and validation data. In fine-tuning, the queries and responses were concatenated together as a single sentence to feed into GPT-2. The perplexity of the fine-tuned language model on the test dataset was 16.5.

4.4 Controlled Query Generation

WordNet substitution and conditional text generators were used to augment diversity-controlled queries. The Stanford part-of-speech (POS) tagger (Toutanova and Manning, 2000) and the WordNet by Miller (1998) were utilized to do WordNet substitution. It is achieved by first using Stanford POS tagger to tag tokens in a query. Then four augmented inputs are generated by substituting verbs, nouns, adjectives & adverbs, or all of the above with synonyms in WordNet. As for conditional text generator, we trained an OpenNMT Transformer on the training and validation splits for query augmentation, which was applied to the testing dataset to augment the query with the top-K beams. For *response diversity*, five variants are obtained, the original query and four paraphrased ones; for *logical self-consistency*, two variants are obtained, the original query and one paraphrase.

Context of Conversation	
Speaker A:	Of course. A two-week paid vacation a year, a five-day workweek.
Speaker B:	So, if I get a margin card, I could take a margin card for you to travel to a company as soon as possible.
Human Score:	0.20
RUBER Score:	0.97
Our Score:	0.19

Table 2: Case study. Both our coherence metric and the human evaluation agreed that the generated response is not coherent with the given query, while RUBER indicated this reply is coherent.

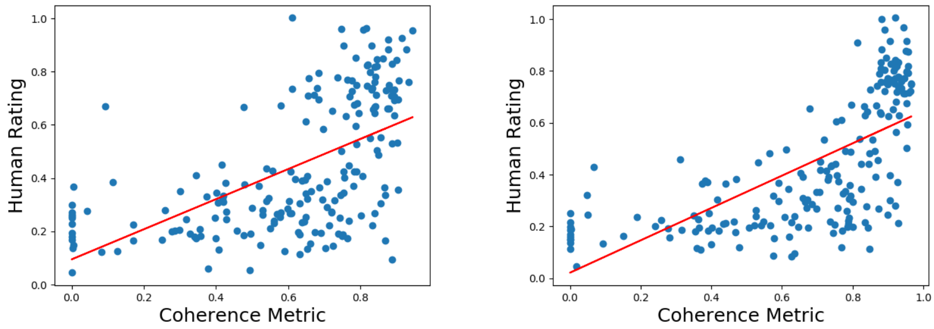
		Pearson	Spearman
RUBER+BERT		0.47	0.51
GPT-2 w/o Fine-tune		0.59	0.65
GPT-2 w/ Fine-tune		**0.67**	**0.76**
Inter-Rater	*Mean*	0.61	0.57
	Max	0.91	0.87

Table 3: Correlation between RUBER+BERT and context coherence metric $c(r|q)$ with human ratings (without and with fine-tuning of GPT-2).

4.5 Metric Evaluation

To assess the validity of our proposed metrics, we utilize Amazon Turk to collect high quality human ratings from 10 subjects. For each metric, we select a set of samples to be presented to humans and each datapoint is to be rated from 1 to 5, with 1 being the worst and 5 being the best on each metric. On both *context coherence* and *response fluency*, we select 200 datapoints with a diverse range of generation quality. There are 200 query-response pairs to be rated for *context coherence* and 200 responses to be rated for *response fluency*. For *response diversity*, we select 100 datapoints, totaling 500 responses, to be rated in groups of 5, all of which are conditioned on the controlled inputs generated by CTG or WS given the same context. For *logical self-consistency*, 100 datapoints are selected independent from *response diversity*. After Amazon Turk results are collected, we compute the Pearson and Spearman correlation between our automatic metrics and human ratings to assess the validity of our metrics. We normalize the human rating scores to be in the range of 0 to 1.

[2] We also experimented with the medium GPT-2 with 24 layers and found that the results were generally the same. And larger models (the 36- and 48-layers GPT-2) might pose computational difficulty for some researchers and thus were not considered.



(a) GPT-2 w/o Fine-tune

(b) GPT-2 w/ Fine-tune

Figure 1: Correlation between *context coherence* metric $c(r|q)$ and human ratings without and with fine-tuning of GPT-2. Note that random jitters sampled from $\mathcal{N}(0,\ 0.05^2)$ are added to human ratings in visualizing scatter plots showed in this paper to overlapping points.

5 Results

5.1 Context Coherence

Table 3 demonstrates the Pearson and Spearman correlations between the proposed *context coherence* metric and human judgments. Also, the results were compared to the previous best-performing automatic metric, RUBER with BERT embeddings (Ghazvininejad et al., 2018). Clearly both our language model based coherence metric shows higher correlation with human judgments than the classifier-based metric, RUBER.

In addition, we compared the proposed metric with a similar metric based on a GPT-2 language model without fine-tuning on the target dataset. The fine-tuned version improved the results, indicating that fine-tuning on the dialogue dataset enables the language model to better capture the dependency between the queries and replies. Interestingly, even the metric based on the language model without fine-tuning correlated with human ratings stronger than RUBER.

We also examined the inter-rater reliability. It is computed by holding out the ratings of one rater at a time, calculating its correlation with the average of other rater's judgments, and finally averaging over or taking the maximum of all held-out correlation scores. The inter-rater reliability results also support the strong performance of our proposed *context coherence* metric in that the correlation between the automatic metric and human evaluation was close to the inter-rater correlations.

In addition, Figure 1 details the effect of fine-tuning on GPT-2. It helps to improve the consistency between human rating and automatic metric.

Table 2 displays a case study. Our coherence metric and the human evaluation agreed that the generated response is not coherent with the given query, while RUBER indicated that this reply is coherent. This might be because RUBER simply compares the embeddings of the query and response and business travel related words in the query such as *vacation*, *workweek* and in the reply such as *travel*, *company* make RUBER judge that they are similar.

		Pearson	Spearman
GPT-2 w/o Fine-tune		0.43	0.32
GPT-2 w/ Fine-tune		**0.82**	**0.81**
Inter-Rater	*Mean*	0.70	0.70
	Max	0.88	0.85

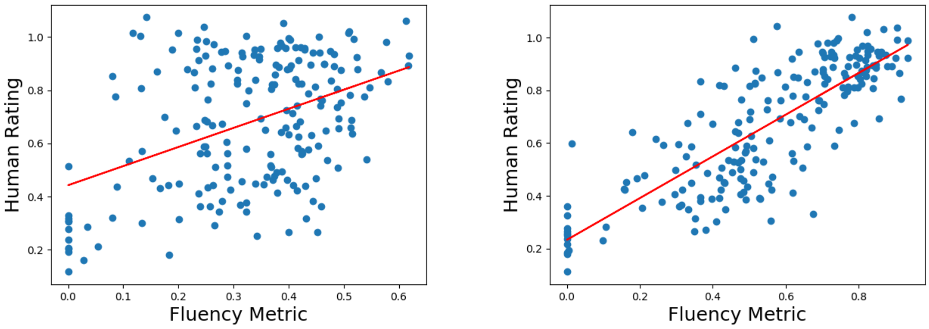
Table 4: Correlation between *response fluency* metric $f(r)$ and human ratings without and with fine-tuning of GPT-2. Pairwise mean and max correlations of human ratings.

5.2 Response Fluency

Our findings show that the proposed fluency metric $f(r)$ is highly correlated with human judgments. Table 4 summarizes the relation between our proposed fluency metric and human ratings in terms of Pearson and Spearman correlation. The importance of fine-tuning GPT-2 (as outlined in Section 4.3) is evident. We observe an increase from 0.43 to 0.82 in Pearson correlation and an enhancement from 0.32 to 0.81 in Spearman correlation. In addition, Figure 2 details the effect of fine-tuning. Notably, a correction of outliers occurs.

5.3 Response Diversity

Table 5 shows the evaluation of the proposed diversity metric on the basis of the augmented datasets with WS and CTG. We also include a baseline dataset which consists of responses from randomly



(a) GPT-2 w/o Fine-tune

(b) GPT-2 w/ Fine-tune

Figure 2: Correlation between *response fluency* metric $f(r)$ and human ratings for GPT-2 without and with fine-tuning.

	1-Gram Entropy		2-Gram Entropy		3-Gram Entropy	
	Pearson	*Spearman*	*Pearson*	*Spearman*	*Pearson*	*Spearman*
Baseline Dataset	0.46	0.32	0.45	0.33	0.43	0.33
WS Dataset	**0.77**	0.69	**0.76**	0.67	**0.71**	0.61
CTG Dataset	0.72	**0.72**	0.72	**0.72**	0.66	**0.66**

Table 5: Comparison of *response diversity* metric between the baseline dataset and our paraphrasing-augmented datasets (WS and CTG datasets) using Spearman and Pearson correlations.

	Inter-Rater Pearson		Inter-Rater Spearman		Human Variance
	mean	*max*	*mean*	*max*	
Baseline Dataset	0.21	0.51	0.23	0.65	0.93
WS Dataset	**0.78**	**0.89**	0.78	**0.92**	**0.68**
CTG Dataset	**0.78**	0.86	**0.79**	0.81	0.69

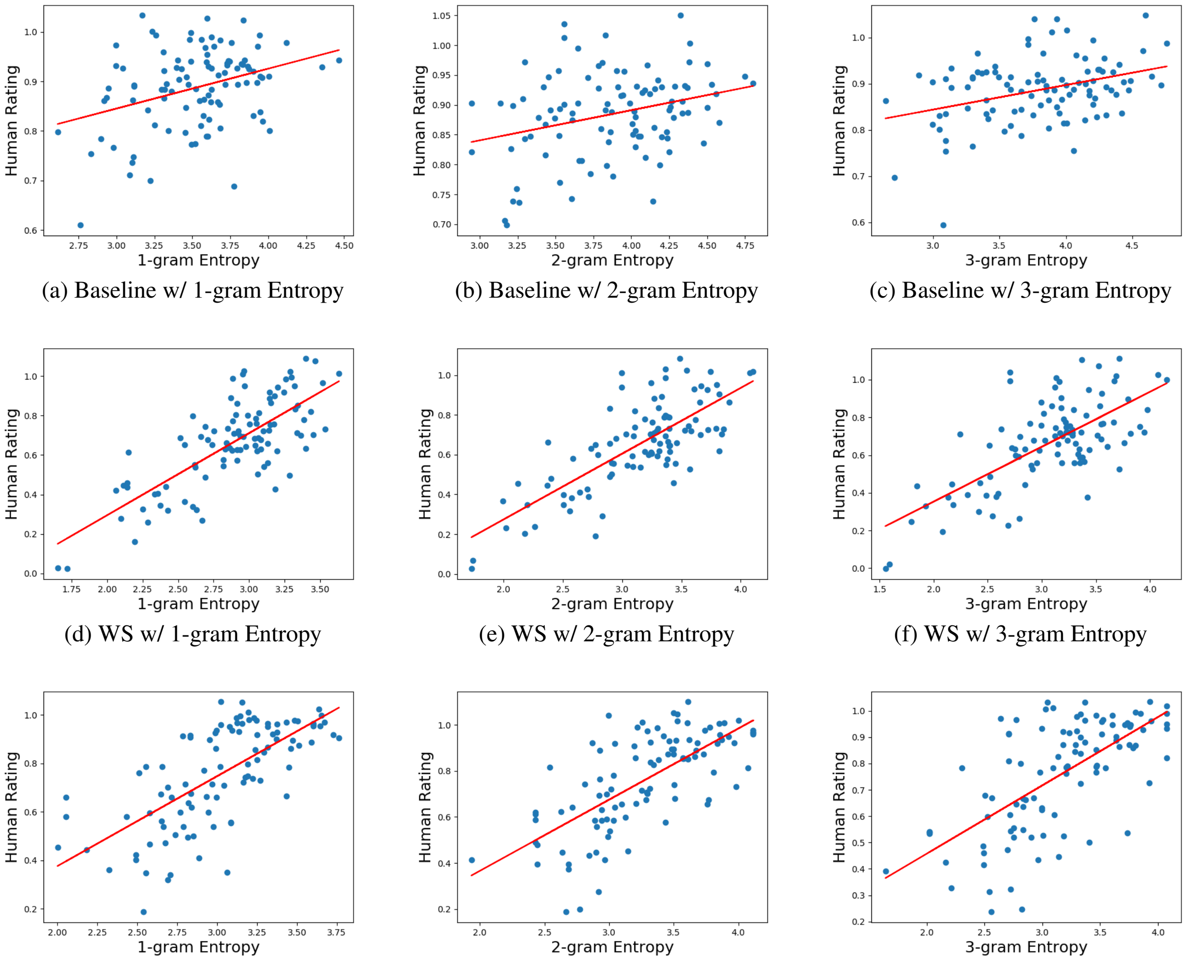
Table 6: Comparison of *response diversity* between the baseline dataset and and our paraphrasing-augmented datasets (WS and CTG datasets) using Inter-Rater Spearman and Pearson correlations.

chosen queries from the testing data. Unigram, bigram, and trigram entropy are utilized to calculate responses' diversity and are compared to human ratings with Pearson and Spearman correlation. It is clear that automatic evaluations with the controlled paraphrasing datasets consistently achieve higher correlation compared to those with the baseline dataset. Figure 3 display correlations between normalized human ratings and corresponding n-gram entropy based on the augmented dataset. Entropy values based on WS and CTG datasets demonstrate stronger relations with human ratings, compared to those based on the baseline dataset, consistent with the reported correlations.

Table 6 displays inter-rater Pearson and Spearman correlations and variance in human ratings. Human ratings based on the paraphrasing augmented datasets show high inter-rater correlations and lower variance, indicating that raters generally agree with each other. The poor baseline performance is likely due to the uncontrolled nature of input sentences such that outputs of evaluated models are generally diverse, making it difficult for humans to judge the diversity performance of the model. Furthermore, our diversity metrics have correlations with human ratings close to the corresponding mean inter-rater correlations, suggesting that the diversity evaluation based on the paraphrasing-augmented data can reveal the diversity of a dialogue system consistent with humans.

Context of Conversation	
Speaker A:	Are you more of a leader or a follower?
Speaker B:	*I don 't try to lead people.* I' d rather cooperate with everybody, and get the job done by working together.
Generated Utterance	
Speaker A:	Are you more of a follower or a leader?
Model Response	
Speaker B:	I like to keep to myself. *I'm a person who does not want to be a follower.*
Our Score:	0.09
Human Score:	0.20

Table 7: Case study of *logical self-consistency*. **Generated Utterance** is generated by CTG. *Blue italic* words highlights the logic contradiction. Our automatic score is low indicating that the logic contradiction can be detected.



(a) Baseline w/ 1-gram Entropy (b) Baseline w/ 2-gram Entropy (c) Baseline w/ 3-gram Entropy

(d) WS w/ 1-gram Entropy (e) WS w/ 2-gram Entropy (f) WS w/ 3-gram Entropy

(g) CTG w/ 1-gram Entropy (h) CTG w/ 2-gram Entropy (i) CTG w/ 3-gram Entropy

Figure 3: Correlation between n-gram entropy and human ratings on the baseline dataset, WS dataset and CTG dataset.

	Pearson	*Spearman*
Baseline Dataset	0.26	0.27
WS Dataset	0.59	0.64
CTG Dataset	**0.65**	**0.66**

Table 8: Comparison of *logical self-consistency* metric between the paraphrasing-augmented data (WS and CTG data) and the baseline data without augmentation using Spearman and Pearson correlations with human ratings.

	Inter-Rater Pearson		Inter-Rater Spearman	
	mean	*max*	*mean*	*max*
Baseline	0.61	0.75	0.62	0.74
WS	0.64	**0.80**	0.64	**0.79**
CTG	**0.65**	0.75	**0.66**	0.76

Table 9: Comparison of *logical self-consistency* metric between the paraphrasing-augmented data (WS and CTG data) and the baseline data without augmentation using Inter-Rater Spearman and Pearson correlations.

5.4 Logical Self-Consistency

Table 8 displays the correlations between the proposed automatic ratings and human ratings on the the paraphrasing augmented data using WS and CTG and a baseline without augmentation. The automatic metric based on augmented data has a stronger relation with that based on the baseline. In particular, the metric based on CTG augmentation aligns with human judgments the closet.

Inter-rater Pearson and Spearman correlations are reported in Table 9. Human ratings on the augmented data are more consistent than those on the baseline, indicating the necessity and efficiency of using a refined dataset instead of the original one. We show a case study in Table 7.

5.5 Relation between the Four Metrics

Although the four proposed metrics are intuitively and theoretically important in evaluating a dialogue system, it is not entirely clear whether they are independent from each other such that it is necessary to measure all of them. We empirically investigate their association. We randomly choose 50 dialogues from the testing dataset and construct the evaluation data for the four metrics. Five human evaluators rate on the four aspects of each dialogue. We then examine the pairwise correlation of human ratings on the four metrics. Response fluency correlates with context coherence ($r = 0.42, p = 0.003$). This is mainly due to the fact that inarticulate responses are often considered incoherent with the context. All other pair-wise correlations are non-significant ($r's < 0.1, p's > 0.25$)[3]. Thus, the four metrics are relatively independent from each other and it is critical to take into account all of them to obtain a holistic evaluation of a dialogue model.

6 Conclusion

This paper provides a holistic and automatic evaluation method for open-domain dialogue models. In contrast to prior art, our means of evaluation captures not only the quality of generation, but also the diversity and logical consistency of responses. We recruit GPT-2 as a strong language model to evaluate the *context coherency* and *response fluency*. For *response diversity* and *logical self-consistency*, we propose to measure these two aspects under augmented utterances with controlled paraphrasing. We leverage two effective approaches to generate augmented utterances: word substitution and text generator with k-best decoder. Moreover, we utilize n-gram based entropy to capture *response diversity* and entailment based approach to measure *logical self-consistency*. The proposed metrics show a strong correlation with human judgments. It is our hope the proposed holistic metrics may pave the way towards the comparability of open-domain dialogue models.

Acknowledgments

Wenjuan Han, Yixian Liu and Kewei Tu were supported by the National Natural Science Foundation of China (61976139).

[3] We do not observe any obvious non-linear dependency either.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Fumihiro Bessho, Tatsuya Harada, and Yasuo Kuniyoshi. 2012. Dialog system using real-time crowdsourcing and twitter large-scale corpus. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 227–231.

Massimo Caccia, Lucas Caccia, William Fedus, Hugo Larochelle, Joelle Pineau, and Laurent Charlin. 2018. Language gans falling short. *arXiv preprint arXiv:1811.02549*.

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

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Tatsunori B Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. *arXiv preprint arXiv:1904.02792*.

Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. 2018. Learning to write with cooperative discriminators. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1638–1649.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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 V Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic turing test: Learning to evaluate dialogue responses. *arXiv preprint arXiv:1708.07149*.

George A Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. 2016. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. *arXiv preprint arXiv:1607.00970*.

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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the conference on empirical methods in natural language processing*, pages 583–593. Association for Computational Linguistics.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1577–1586, Beijing, China. Association for Computational Linguistics.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714*.

Hui Su, Xiaoyu Shen, Pengwei Hu, Wenjie Li, and Yun Chen. 2018. Dialogue generation with gan.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Kristina Toutanova and Christopher D Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*, pages 63–70. Association for Computational Linguistics.

Alan Turing. 1950. Computing machinery and intelligence-am turing. *Mind*, 59(236):433.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *CoRR*, abs/1901.08149.

Jingjing Xu, Xuancheng Ren, Junyang Lin, and Xu Sun. 2018. Dp-gan: diversity-promoting generative adversarial network for generating informative and diversified text. *arXiv preprint arXiv:1802.01345*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.