TOWARDS HOLISTIC AND AUTOMATIC EVALUATION OF OPEN-DOMAIN DIALOGUE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-domain dialogue generation has gained increasing attention in Natural Language Processing. Comparing these methods requires a holistic means of dialogue evaluation. Human ratings are deemed as the gold standard. As human evaluation is inefficient and costly, an automated substitute is desirable. In this paper, we propose holistic evaluation metrics which capture both the quality and diversity of dialogues. Our metrics consists of (1) GPT-2 based context coherence between sentences in a dialogue, (2) GPT-2 based fluency in phrasing, and, (3) n-gram based diversity in responses to augmented queries. The empirical validity of our metrics is demonstrated by strong correlation with human judgments. We provide the associated code, datasets and human ratings.

1 INTRODUCTION

1.1 AUTOMATED DIALOGUE EVALUATION

Learning to communicate is a key capacity of intelligent agents. Research on enabling a machine to have meaningful and natural conversation 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 & Le, 2015; Serban et al., 2016; Ghazvininejad et al., 2018; Serban et al., 2017). Moreover, dialogue generation has a wide range of industrial applications such as Microsoft's Xiaoice and Baidu's Dumi.

Evaluating models of dialogue generation in an efficient manner poses a significant challenge in developing dialogue systems. The prevalent method of dialogue evaluation is human-based rating under a given rubric. This method of evaluation is deemed impracticable, when various variations in the model and sets of hyperparameters are needed. These drawbacks may hinder the research progress and render the human evaluation approach not scalable.

Previous automatic evaluation metrics generally focus on the quality of the dialogue generation (Tao et al., 2018; Ghazarian et al., 2019). In this work, we propose holistic metrics which considers both the quality and diversity of generated dialogues. Specifically, we consider (1) *context coherence* of a dialogue (i.e., the meaningfulness of a response within the prior context of the dialogue), (2) *language fluency* of generated responses (i.e., the quality of phrasing relative to a human native speaker), and, (3) *response diversity* of a set of generated responses (i.e., the variety in meaning and word choice of responses). A strong language model such as GPT-2 (Radford et al., 2019) naturally captures (1) and (2). Therefore, we propose to recruit and fine-tune GPT-2 as a measure of quality. Moreover, we utilize n-gram based entropy to capture (3). Specifically, we propose to measure response diversity under augmented queries with controlled diversity. Two such augmentation strategies are considered. Finally, extensive human evaluations are conducted to substantiate the validity of our proposed metrics.

1.2 PRIOR ART

Evaluation metrics based on heuristics have been shown to align well with human judgments and widely applied in various language generation tasks. For machine translation, BLUE (Papineni et al., 2002) computes n-gram precision, whereas METEOR (Banerjee & 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 task like dialogue generation where a diverse range of generations are acceptable conditional on a query. Indeed, Liu et al. (2016) conduct 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.

Context of Conversation
Speaker A: Hey, what do you want to do tonight?
Speaker B: Why dont we go see a movie?
Model Response
Nah, lets do something active.
Reference Response
Yeah, the film about Turing looks great!

Table 1: An example of low BLEU score and low semantic similarity between model response and reference response while the generated response appears reasonable within the dialogue.

The word-overlap metrics (e.g., BLUE) 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 using 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 the reference responses randomly chosen from the dataset, hence avoiding the need of human-annotated data. After training, the softmax score is utlized 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 are reasonable given a query. Hence a response can be acceptable even if it does not align well with the reference either in terms of word-overlap or semantic embedding. See Table 1 for an example.

Prior art on automatic metrics focuses on the quality, mostly the relevance to the query, of the generated responses. 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 dialogue or story generation (Hashimoto et al., 2019). The current work proposes metrics to holistically evaluate the quality and diversity of open-domain dialogue generation. One key component of dialogue response generation 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. In particular, the query coherence metric is computed as the conditional probability of the response given the query, which reflects whether the response appropriately follows the query under a language model. We adopt a transfer learning approach to obtain a powerful

language model. 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.

While the aforementioned metrics attempt to measure the quality of text generation, some n-gram based metric has also 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 an improper metric 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. Instead we approach diversity evaluation of a dialogue model with controlled queries, whereby we control the diversity of the queries while evaluating the diversity of the responses. Controlling query diversity involves minimizing diversity in both meaning and word use and avoiding feeding the dialogue models identical inputs. A dialogue model with poor diversity always generates responses with the same phrases and words, whereas an ideal model produces varying words and sentence structures. The controlled queries are generated by augmenting the original query with sentences close in meaning and slightly different in word use. For the purpose of generality, we propose WordNet substitution and Conditional Text Generator to generate controlled queries. The n-gram entropy across the responses given the controlled queries is deemed as a diversity measure.

1.3 CONTRIBUTIONS

In this work, we propose a metric to holistically evaluate open-dialogue models by taking into consideration both quality and diversity of generated dialogues. Our contributions are summarized below.

- Both context coherence and response fluency (quality metrics) are naturally captured by metrics based on a strong language model. Empirically, we demonstrate that the language model based metrics clearly outperform previous relevant metrics.
- In view of the complexity of diversity evaluation, we propose two effective approaches to generate augmented utterances with controlled diversity: word substitution and text generator with k-best decoder. Our experiments show that the diversity metric strongly correlates with human judgments on the response diversity. Moreover, our proposed datasets significantly improve the agreement between human evaluation, leading to a more accurate and straightforward human annotation.
- We release the datasets, human ratings and implementation of the metric as open-source contribution to pave the way towards further research.

2 METRICS

2.1 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 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:

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

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 intepreted relative to other values. Also, the unboundedness renders it prone to extreme values. Hence, a normalized score is proposed 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

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

which ranges from 0 to 1.

2.2 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,

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

Due to the negativeness and unboundedness of the raw score, a normalized version, $f(r)$, similar to the normalized context coherence score is proposed,

$$
f(r) = \frac{max(f_{5th}, f_{raw}(r)) - f_{5th}}{-f_{5th}}. \tag{4}
$$

2.3 Response Diversity

We measure response diversity utilizing augmented queries with controlled diversity. Controlling query diversity involves minimizing diversity in both meaning and word use and avoiding feeding the dialogue models identical inputs. We thus aim to augment the original query with sentences close in meaning and slightly different in word use. To achieve so, two augmentation approaches are proposed: (1) WordNet Substitution (WS) and (2) Conditional Text Generator (CTG).

WordNet Substitution (WS) is word-level manipulation method suitable for both single-turn and multi-turn datasets. It is achieved by first using Part-Of-Speech (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.

Different from WS, Conditional Text Generator (CTG) is an approach to testing language diversity using multi-turn datasets. It requires a sequence-to-sequence or a transformer model 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-5 beams, $u_t^{(1)}, ..., u_t^{(5)}$, from the CTG model with the concatenated utterance history $[u_1; ...; u_{t-1}]$ is input into a model to be evaluated.

Given a set of augmented queries for the ith query with controlled diversity, the responses, $\mathcal{R}_i$, are generated by the model under test. Then n-gram entropy for the ith sample is computed as,

$$
d_i = -\frac{1}{|\mathcal{R}_i|} \sum_{w \in \mathcal{R}_i} \log_2 p(w) \tag{5}
$$

where p is the n-gram probability in $\mathcal{R}_i$. The diversity metric is then defined as the averaged entropy over the dataset,

$$d = \frac{1}{N} \sum_{i=1}^{N} d_i. \tag{6}$$

3 EXPERIMENTS

3.1 DATASET

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

3.2 RESPONSE GENERATION

A sequence-to-sequence (seq2seq) 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 $k = \{1, 50, 500\}$.

3.3 LANGUAGE MODEL FINE-TUNING

The base GPT-2 model with 12 layers was used to compute our metrics. 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. The GPT-2 model was fine-tuned on the training and validation data. In fine-tuning, the query and response 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.

3.4 CONTROLLED QUERY GENERATION

WordNet substitution and conditional text generator were used to augment diversity-controlled queries. The Stanford POS tagger (Toutanova & Manning, 2000) and the WordNet by Miller (1998) were utilized to do WordNet substitution. 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-4 beams.

3.5 METRIC EVALUATION

To assess 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 generated query-response pairs (or responses only) 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 in generation quality corresponding to that metric. On both Context Coherence and Fluency metrics, we select 200 datapoints with diverse range of generation quality. There are 200 query-response pairs to be rated for Context Coherence and 200 responses to be rated for Fluency. For Diversity metric, 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 a CTG given the same context. After Amazon Turk results are collected, we then compute Pearson Correlation between our evaluation metrics and human ratings to assess the validity of our metric and selected datasets. We normalize the human rating scores from 0 to 1.

Query	Generated Reply	Human Score	RUBER	Ours
Of course. A two-week paid vacation a year, a five-day workweek.	So, if I get a margin card, I could take a margin card for you to travel to a company as soon as possible.	0.20	0.97	**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.

4 RESULTS

4.1 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 show 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 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 across and taking the maximum all held-out correlation scores. The inter-rater reliability results also support the strong performance our proposed context coherence metric since the correlation between the automatic metric and human evaluation was close to the inter-rater correlations.

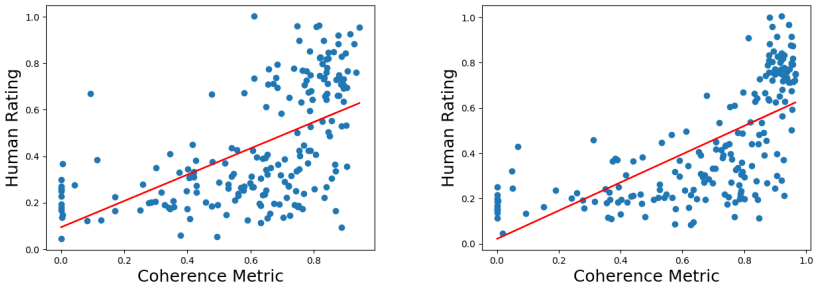
Table 2 displays a case study. Both our coherence metric and the human evaluation agreed that the generated response is not coherent with the given query, while RUBER indicated this reply is coherent. This might be because RUBER simply compares the embeddings of the query and response and business travel 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
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).

4.2 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 the human-ratings in terms of Pearson and Spearman correlation. The importance of fine-tuning GPT-2 (as outlined in Section 3.3) is evident. We observe an increase from 0.43 to 0.82 in Pearson correlation. In addition, Figure 2 details the effect of fine-tuning. Notably, a correction of outliers occurs. Moreover, the consistency of human ratings is demonstrated by high mean pair-wise correlations between pairs of ratings.

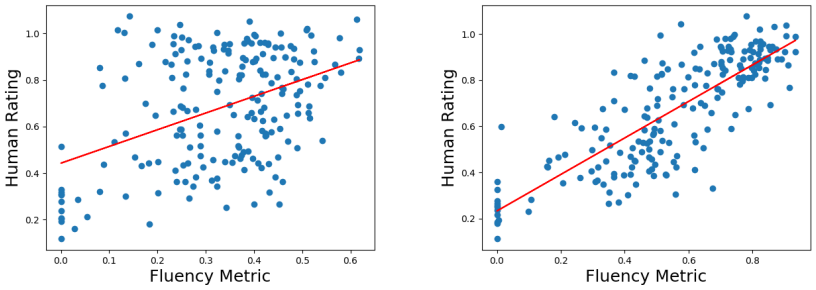


(a) GPT-2 w/o Fine-tune (b) GPT-2 w/ Fine-tune

Figure 1: Correlation between context coherence metric $c(r|q)$ and human ratings without and with fine-tuning of GPT-2.

		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 fluency metric $f(r)$ and human ratings without and with fine-tuning of GPT-2. Pairwise mean and max correlations of human ratings.



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

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

4.3 Response Diversity

Table 5 shows the evaluation of our generated datasets using WS and CTG. Unigram, bigram, and trigram entropy are used to calculate responses' diversity and are compared to human ratings in Pearson and Spearman Correlation. Note that automatic evaluations on our datasets consistently achieve higher correlation compared to the baseline dataset. We also show our datasets evaluated using three different diversity metrics in Figure 3. The figures show correlations between normalized human ratings and corresponding n-gram entropy. A line of best-fit is drawn to indicate their correlations, and for plotting purpose, each datapoint after normalization is added a random noise sampled from $\mathcal{N}(0,\, 0.05^2)$. Clearly, WS and CTG Dataset show more clustered datapoints and slopes closer to 1 than our baseline dataset, a result consistent with the reported correlations.

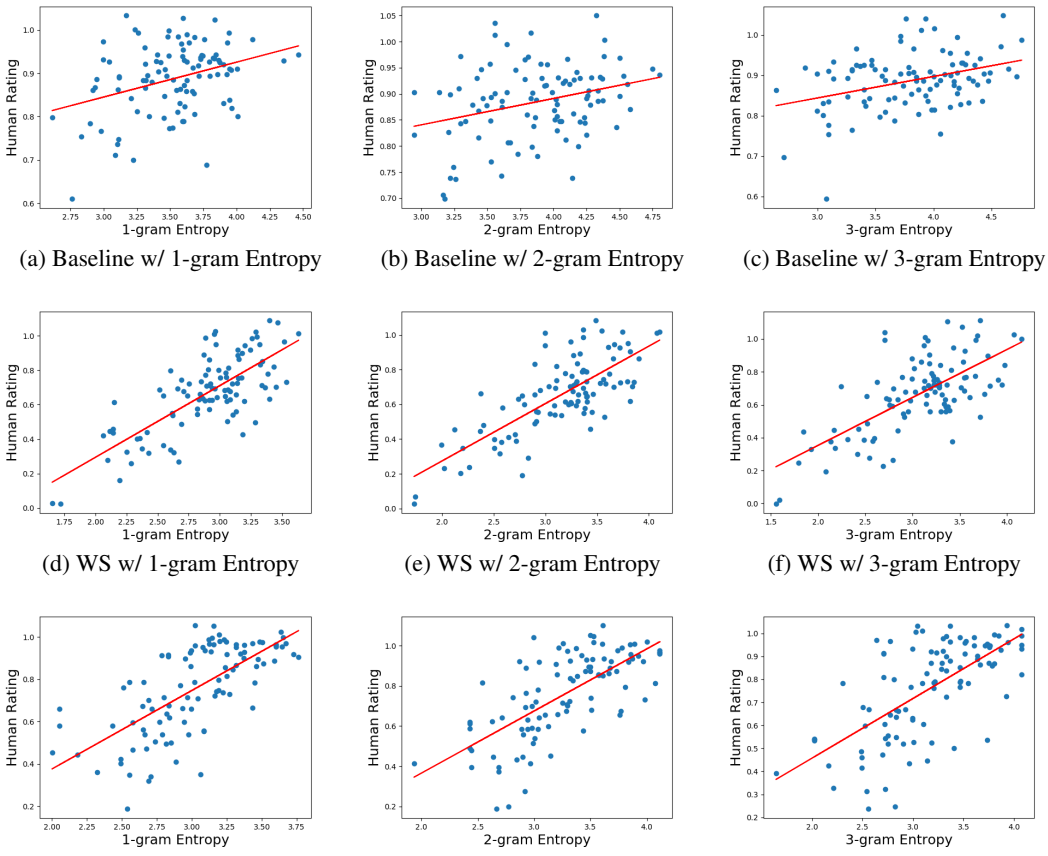
Table 6 shows inter-rater Pearson Correlation, Spearman correlations, and variance in human ratings. Interestingly, both WS Dataset and CTG Dataset display similarly high correlations, indicating that raters generally agree with each other. WS Dataset is also lowest in Human Variance, suggesting human raters are more certain about their ratings. Baseline Dataset, on the other hand, has poor inter-rater correlations. This is most 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 diversity performance of the model. Furthermore, both of our datasets achieve scores close to that of their

	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 between the Baseline Dataset and our generated datasets (WS Dataset and CTG Dataset) using Spearman Correlations, Pearson Correlations and variance with collected human ratings.

	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 between the Baseline Dataset and our generated datasets (WS Dataset and CTG Dataset) using Inter-Rater Spearman and Inter-Rater Pearson correlations.



(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.

corresponding mean inter-rater correlations, indicating that the evaluation metric on our datasets can reveal diversity of a dialog system consistent with humans.

5 CONCLUSION

This paper provides a holistic and automatic evaluation method of open-domain dialogue models. In contrast to prior art, our means of evaluation captures not only the quality of generation, but also the diversity of responses. We recruit GPT-2 as a strong language model to evaluate the fluency and context-coherency of a dialogue. For diversity evaluation, the diversity of queries is controlled while the diversity of responses is evaluated by n-gram entropy. Two methods for controlled diversity are proposed, WordNet Substitution and Conditional Text Generator. The proposed metrics show strong correlation with human judgments.

We are providing the implementations of our proposed metrics, associated fine-tuned models and datasets to accelerate the research on open-domain dialogue systems. It is our hope the proposed holistic metrics may pave the way towards comparability of open-domain dialogue methods.

REFERENCES

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

Satanjeev Banerjee and Alon Lavie. 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*, pp. 65–72, 2005.

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

Fumihiro Bessho, Tatsuya Harada, and Yasuo Kuniyoshi. 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*, pp. 227–231, 2012.

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

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

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

Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. Learning to write with cooperative discriminators. *arXiv preprint arXiv:1805.06087*, 2018.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pp. 67–72, Vancouver, Canada, July 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P17-4012.

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

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W04-1013.

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

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

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

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318. Association for Computational Linguistics, 2002.

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

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

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

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

Lifeng Shang, Zhengdong Lu, and Hang Li. 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)*, pp. 1577–1586, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1152. URL https://www.aclweb.org/anthology/P15-1152.

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

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

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

Kristina Toutanova and Christopher D Manning. 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*, pp. 63–70. Association for Computational Linguistics, 2000.

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

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

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

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