Assessing Dialogue Systems with Distribution Distances

Jiannan Xiang[1][*]   **Yahui Liu**[2][*]   **Deng Cai**[3]   **Huayang Li**[4]   **Defu Lian**[1]   **Lemao Liu**[4]

[1]University of Science and Technology of China   [2]University of Trento, Italy
[3]The Chinese University of Hong Kong   [4]Tencent AI Lab, China

Abstract

An important aspect of developing dialogue systems is how to evaluate and compare the performance of different systems. Existing automatic evaluation metrics are based on turn-level quality evaluation and use average scores for system-level comparison. In this paper, we propose to measure the performance of a dialogue system by computing the distribution-wise distance between its generated conversations and real-world conversations. Specifically, two distribution-wise metrics, FBD and PRD, are developed and evaluated. Experiments on several dialogue corpora show that our proposed metrics correlate better with human judgments than existing metrics.

1 Introduction

Dialogue generation is a special text generation task, which has drawn booming attention in the natural language processing community. It is widely agreed that one single input query is often associated with multiple valid responses in this task, which is termed as a 1-to-n relationship between a query and its responses (Vinyals and Le, 2015; Zhou et al., 2017; Zhao et al., 2017; Liu et al., 2018; Chen et al., 2021; Chan et al., 2021; Gao et al., 2021). It increases the challenges of automatically evaluating the performance of dialogue systems.

In general, the previous evaluation metrics mainly focus on turn-level quality. For example, unsupervised word-overlapping or embedding-based metrics (Papineni et al., 2002; Lin, 2004; Mitchell and Lapata, 2008; Zhang et al., 2020) calculate the similarity or alignment between generated responses and reference responses, which is not well-suited for open-end dialogue tasks. Learned classification or regression systems (Lowe et al., 2017;

Tao et al., 2018; Sellam et al., 2020; Ghazarian et al., 2019) are corpus-dependent because of requiring additional task-specific training or tuning, which run the risk of assigning lower quality to a better model in the overfitting or underfitting cases.

In this paper, we provide a new perspective that distribution distance between generated conversations and real conversations can be applied to measure the performances of dialogue systems. There are three main *contributions*: (1) We firstly propose two unsupervised distribution-wise metrics (i.e., FBD and PRD) to solve the evaluation issue in this field. (2) The experimental results show that the proposed distribution-wise metrics perform well. Particularly, FBD achieves compelling performances on most evaluation corpora, which shows a promising direction for designing evaluation metrics. (3) We collect the typical evaluation corpora and existing evaluation metrics in order to better assess the performance of dialogue systems, which could be useful for researchers in this community [1].

2 Related Work

In this section, we focus on *unsupervised* automatic evaluation metrics for dialogue system evaluation. In general, existing unsupervised metrics mainly measure *turn-level* qualities, which can be categorized into two main classes: word overlapping metrics and embedding-based metrics:

Word-overlapping Metrics Such metrics quantify the amount of word-overlap between generated response and reference responses. For example, BLEU (Papineni et al., 2002) calculates the geometric mean of the precision for n-gram. ROUGE (Lin, 2004) is a recall-oriented metric. METEOR (Banerjee and Lavie, 2005) computes the harmonic mean of precision and recall with stemming and syn-

---

[1]The source codes and data are available at `https://github.com/yhlleo/frechet-bert-distance`.

onyms.

Embedding-based Metrics Embedding-based metrics align the generated response and the reference in latent semantic space. Some adopt the vector similarity of sentence embeddings as a quality measure. For example, Embedding Average (Foltz et al., 1998; Mitchell and Lapata, 2008) calculates sentence-level embeddings by averaging word representations. Vector Extrema (Forgues et al., 2014) computes sentence-level embeddings by taking the most extreme value for each dimension in all word vectors. Others adopt more fine-grained semantic matching. For example, Greedy Matching (Rus and Lintean, 2012) greedily matches each word in a generated response to a word in the reference response, and the final score is defined as the average of word-level similarity scores. Zhang et al. (2020) introduced a better embedding-based metric BERTScore that computes word similarity using contextual embeddings from pre-trained language models.

Our proposed methods are best placed in the literature of embedding-based metrics. However, there are two main differences from previous metrics in this field: (1) We compute the distribution distance between embedding sets as the system-level performance of a dialogue system, which does not require task-specific training/tuning; (2) We propose to extract sentence-level semantic representations directly from pre-trained language models (Devlin et al., 2019; Liu et al., 2019), where there are no operations of converting the wold-level embeddings to sentence-level embeddings.

3 Proposed Methods

Given a collect of sentence pairs $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}^N$, we assume that the corresponding semantic representations $\{\boldsymbol{v}_i\}^N$ can be extracted in this manner:

$$\boldsymbol{v}_i = LM([\boldsymbol{x}_i, \boldsymbol{y}_i]) \tag{1}$$

where $LM(\cdot)$ refers to pre-trained language models(i.e., (Devlin et al., 2019; Liu et al., 2019; Yang et al., 2019; Clark et al., 2020)), $[\cdot, \cdot]$ refers to the concatenation operation. Intuitively, the differences between the distribution R of real samples and the distribution G of generated samples can be applied to measure the performances of dialogue systems (i.e., $d(R, G)$). Therefore, we propose two distribution-based methods to automatically evaluate the performance of dialogue systems, which are presented in this section in detail.

3.1 Fréchet Bert Distance

Semantic representations $\{\boldsymbol{v}_i\}^N$ are extracted by a pre-trained language model, which encodes the contextual information of the sentences. The main intuition is that the distribution of semantic representations of generated sentences should be as close as possible to the distribution of semantic representations of real sentences in a successful system. To measure this, we assume that such semantic representations follow a multi-dimensional Gaussian, which can be represented by variables: *mean* and *covariance*. The difference between two Gaussians (generated and real sentence pairs) is measured by the Fréchet distance (Dowson and Landau, 1982). We call the Fréchet distance between the distribution R with mean $(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$ obtained from real sentence pairs and the distribution G with mean $(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ obtained from generated sentence pairs as "Fréchet Bert Distance" (FBD), which is formulated as:

$$d_{\text{FBD}}(R, G) = \|\boldsymbol{\mu}_r - \boldsymbol{\mu}_g\| + \\ \text{Tr}(\boldsymbol{\Sigma}_r + \boldsymbol{\Sigma}_g - 2(\boldsymbol{\Sigma}_r \boldsymbol{\Sigma}_g)^{1/2}) \tag{2}$$

Once the distribution of generated data closes to the distribution of real data, the model indeed achieves low FBD scores. Similarly, such distance (Heusel et al., 2017) has been widely verified in various Generative Adversarial Networks (GANs) in computer vision tasks (Karras et al., 2017; Zhang et al., 2018a; Park et al., 2019), which is consistent with increasing disturbances and human judgment. Surprisingly, we observed that FBD works well in evaluating open-end dialogue systems.

3.2 Precision-Recall Distance

We notice that FBD is based on the estimated Gaussian parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. There is an optional strategy to get rid of estimating the parameters. Inspired by (Sajjadi et al., 2018), we apply a precision-recall-based method, named as Precision-Recall Distance (PRD), to evaluate the distance between two distributions.

The key intuition is that precision should measure how much of G can be generated by a "part" of R while recall should measure how much of R can be generated by a "part" of G. In general, (a) If R is bimodal and G only captures one of the modes, we should have perfect precision but only limited recall; (b) In the opposite case, we should have perfect recall but only limited precision; (c) If $R = G$, we should have perfect precision and

recall; (d) If the supports of R and G are disjoint, we should have zero precision and recall. The BPD is formulated as:

$$d_{\text{PRD}}(R, G) = \max\left\{\left.\frac{2\alpha(\lambda)\beta(\lambda)}{\alpha(\lambda) + \beta(\lambda)}\right| \lambda \in \Lambda\right\}$$
(3)

where $\Lambda = \{\tan(\frac{i}{m+1}\frac{\pi}{2})|i = 1, \cdots, m\}$, $m \in \mathbb{N}$ refers to a given angular resolution, $\alpha(\lambda) = \sum_{\boldsymbol{v} \in \mathcal{V}} \min(\lambda R(\boldsymbol{v}), G(\boldsymbol{v}))$ and $\beta(\lambda) = \sum_{\boldsymbol{v} \in \mathcal{V}} \min(R(\boldsymbol{v}), \frac{G(\boldsymbol{v})}{\lambda})$ [2]. Therefore, the better dialogue systems will achieve higher PRD scores.

4 Experiments

4.1 Datasets & Systems

To verify the two proposed metrics, we conduct experiments on six public dialogue corpora.

Baseline Metrics. We mainly compare with several widely-used metrics in text generation field: a) three word-overlapping metrics: BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Denkowski and Lavie, 2014); b) four embedding-based metrics: Greedy Matching (Rus and Lintean, 2012), Embedding Average (Wieting et al., 2015), Vector Extrema (Forgues et al., 2014) and BERTScore (Zhang et al., 2020). All these metrics do not require task-specific training.

Datasets. We collect three recently released evaluation corpora which consist of dialogue query and response samples of different systems, and the corresponding human annotations:

- **Persona(M)**: USR (Mehri and Eskénazi, 2020) built an evaluation corpus based on PersonaChat (Zhang et al., 2018b), in which both four system outputs and the corresponding human evaluation scores were collected.

- **Daily(H), Convai2, and Empathetic**: GRADE (Huang et al., 2020) used three dialogue corpora, including DailyDialog (Lowe et al., 2017), Convai2 (Dinan et al., 2019) and EmpatheticDialogues (Rashkin et al., 2018), to do the evaluations and compared two dialogue models: Transformer-Ranker and Transformer-Generator collected from the ParlAI platform (Miller et al., 2017).

- **Daily(Z) and Persona(Z)**: Dialogue Evaluation (Zhao et al., 2020) used Dai-

Source	Address
Transformers	`https://github.com/huggingface/transformers`
USR	`https://github.com/Shikib/usr`
GRADE	`https://github.com/li3cmz/GRADE/tree/main/evaluation`
Daily(Z) Persona(Z)	`https://github.com/ZHAOTING/dialog-processing/tree/master/src/tasks/response_eval`
ParlAI	`https://github.com/facebookresearch/ParlAI`
BLEU METEOR	`https://github.com/nltk/nltk`
ROUGE-L Greedy Average Extrema	`https://github.com/Maluuba/nlg-eval`
BERTScore	`https://github.com/Tiiiger/bert_score`

Table 1: All the public resources in our experiments.

Corpus	Num. of Samples	Systems
Persona(M)	60	Seq2Seq LSTM language model Key-Value Profile Memory Network Generated Human-written
Daily(H) Convai2 Empathetic	150 150 150	Transformer-Ranker Transformer-Generator
Daily(Z) Persona(Z)	100 150	Seq2Seq Attentional Seq2Seq HRED VHRED GPT2-sm GPT2-md

Table 2: The details of each evaluation corpus.

lyDialog (Lowe et al., 2017) and PersonaChat (Zhang et al., 2018b) to build two evaluation corpora and collected outputs of six generative models (with three decoding strategies). The appropriateness of each response was obtained by human annotation.

Implementation Details. We leverage pre-trained BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) for utterance-level contextualized encoding [3] without additional tuning or training the language models. For each query and response pair

---

[2]For a distribution P with a finite state space $\mathcal{V}$, we have $\boldsymbol{v} \in \mathcal{V}$ and $P(\boldsymbol{v}) > 0$.

[3]Based on the public project: `https://github.com/huggingface/transformers`

Metric	Persona(M)		Daily(H)		Convai2		Empathetic		Daily(Z)		Persona(Z)	
	Spr.	Pr.	Spr.	Pr.	Spr.	Pr.	Spr.	Pr.	Spr.	Pr.	Spr.	Pr.
Word-Overlapping Metrics												
BLEU	.400	.672	.445	.444	.800	.801	.136	.331	.595	.421	.400	.390
METEOR	.800	.860	.018	.050	.800	.767	.382	.133	.643	.689	.700	.936
ROUGE-L	.600	.289	.545	.417	.200	.061	.391	.472	.738	.725	.400	.915
Embedding-Based Metrics												
Greedy	.600	.260	.855	.764	.600	.794	.736	.864	.690	.726	.100	.835
Average	.800	**.863**	.209	.209	.600	.879	.664	.725	.548	.769	.300	.861
Extrema	.600	.435	.745	.761	.800	.766	.618	.722	.595	.746	.500	.834
BERTScore^B	.800	.590	.137	.082	.800	.817	.300	.077	.857	.883	.900	.961
BERTScore^R	.800	.517	.855	.857	.800	.939	.600	.697	.714	.860	.700	.918
Distribution-Based Metrics												
FBD^B	**1.00**	.853	.564	.717	.800	.854	.427	.623	.786	.763	.400	.923
FBD^R	**1.00**	.802	**.891**	**.926**	.800	.747	**.864**	**.951**	**.929**	**.963**	**1.00**	.860
PRD^B	.800	.637	.221	.409	**1.00**	**.972**	.227	.399	.690	.914	.900	**.984**
PRD^R	.800	.660	.591	.578	**1.00**	.913	.545	.583	.762	.906	.900	.932

1. "Spr." and "Pr." refer to Spearman and Pearson correlation coefficients, respectively.
2. ^{B} and ^{R} mean using BERT (base) and RoBERTa (base) as language models, respectively.

Table 3: Correlations of all the metrics with overall quality ratings the six dialogue corpora.
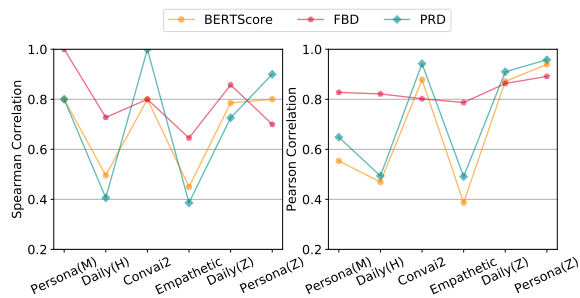

Figure 1: The comparisons between BERTScore, FBD and PRD metrics on various corpora. For each corpus, we average the performances of BERT and RoBERTa language models.
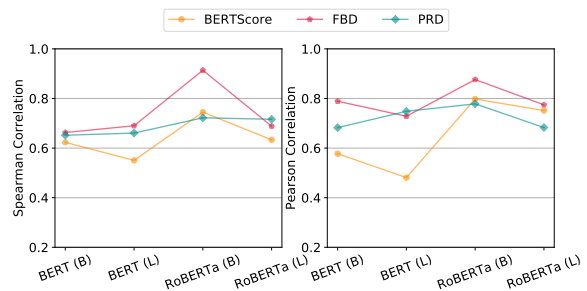

Figure 2: Comparisons between BERTScore, FBD and PRD metrics on various language models. For each model, we average the performances on all the six evaluation corpora.

$(\boldsymbol{x}_i, \boldsymbol{y}_i)$, we use the last hidden output of [CLS] as its semantics representation without tuning or training the language models. To assess the system-level performances of dialogue systems, we calculate the Spearman and Pearson correlations between the rankings of human evaluation and the rankings of evaluation metrics. If a evaluation metric is designed for *turn-level* evaluation, we average the all turn-level scores as the performance of the corresponding dialogue system.

Public Resources. All the compared evaluation corpora and evaluation metrics are available in Table 1. Once the official implementations are not available, we use the repositories with highest "stars" on GitHub. The details of each evaluation corpus, including number of samples and compared

dialogue systems in each corpus, are presented in Table 2.

4.2 Results

We compute the system-level correlation between all automatic metrics and the quality ratings by using Spearman and Pearson correlation coefficients.

Model	P(M)	D(H)	C2	EM	D(Z)	P(Z)
BERT	.57$_{\pm.44}$	.65$_{\pm.39}$	.60$_{\pm.42}$	.65$_{\pm.39}$	.67$_{\pm.38}$	.65$_{\pm.38}$
RoBERTa	.71$_{\pm.38}$	.68$_{\pm.37}$	.73$_{\pm.34}$	.68$_{\pm.37}$	.68$_{\pm.37}$	.71$_{\pm.33}$

1. $\{$P(M), D(H), $\cdots$, P(Z)$\}$ refer to the six evaluation corpora.
2. The reported values are mean$_{\pm\text{standard deviation}}$.

Table 4: Comparisons of the normality on various evaluation corpora. The normality is calculated over each dimension of the extracted semantic representations.

The performances of various evaluation metrics on different public corpora are reported in Table 3. Our proposed two metrics (i.e., FBD and PRD) show comparable performances over the baseline metrics. Especially, FBD[R] achieves compelling performances on five corpora, which indicates a good ability of generalization and robustness on various corpora. In addition, most evaluation metrics are sensitive to the evaluation corpora. For example, BLEU performs well in **Convai2** but fails in **Empathetic**. Similarly, BERTScore[B] performs well in **Convai2** and **Persona(Z)** but fails in **Daily(H)** and **Empathetic**. It indicates that the selection of evaluation corpora has a great influence on assessing the performances of evaluation metrics. Hence, it's better to use multiple corpora to do the comparisons between metrics. Obviously, our proposed FBD[R] outperforms the existing evaluation metrics in the view of robustness.

In Figure 1, we compare the evaluation metrics in the perspective of various evaluation corpora, where the results of BERT and RoBERTa language models are averaged. It confirms the superiority of the FBD metric. Compared to the performance of USR (Mehri and Eskénazi, 2020) (1.000/.820 on Persona(M)), a reference-free metric that relies on task-specific training/tuning with task-specific data, the performances of our proposed methods are comparable without any training/tuning. Therefore, we believe that it's a promising direction to explore distribution-wise metrics for assessing dialogue systems in this field.

As shown in Figure 2, we average the performances of each metric on all evaluation corpora. It shows that our proposed FBD has higher performance expectations that outperform BERTScore with different language models. The large models do not show improvements in average performance compared to the base models. In general, FBD metric achieves better Spearman and Pearson correlations compared to PRD. Surprisingly, RoBERTa-based metrics, including BERTScore, the proposed FBD and PRD, perform better than the corresponding BERT-base ones. Given that our FBD metric lies on the assumption of multivariate Gaussian distribution, we hypothesize that the semantic representations extracted by RoBERTa model fit Gaussian distribution better than BERT model. To verify this point, as shown in Table 4, we use "Shapiro–Wilk test" (Razali et al., 2011) to calculate the normality in statistics, where small

Model	Relevance		Grammar		Content		Overall	
	Spr.	Pr.	Spr.	Pr.	Spr.	Pr.	Spr.	Pr.
Word-Overlapping Metrics								
BLEU	.857	.454	.167	.213	.143	.270	.595	.421
METEOR	.810	.736	.119	.044	.024	.349	.643	.689
ROUGE-L	.857	.758	.238	.075	.190	.375	.738	.725
Embedding-Based Metrics								
Greedy	.881	.769	.214	.133	.119	.350	.690	.726
Average	.762	.808	.071	.197	.238	.399	.548	.769
Extrema	.714	.776	.143	.009	.476	.539	.595	.746
BERTScore[B]	.976	.908	.333	.111	.524	.593	.857	.883
BERTScore[R]	.857	.889	.119	.115	.310	.604	.714	.860
Distribution-Based Metrics								
FBD[B]	.762	.790	.190	.203	**.667**	.404	.786	.763
FBD[R]	**.976**	**.965**	**.429**	**.472**	.524	**.839**	**.929**	**.963**
PRD[B]	.833	.932	.238	.156	.571	.708	.690	.914
PRD[R]	.881	.925	.238	.235	.452	.729	.762	.906

1."Spr." and "Pr." refer to Spearman and Pearson correlation coefficients, respectively.
2. [B] and [R] mean using BERT (base) and RoBERTa (base) as language models, respectively.

Table 5: Fine-grained comparisons on Daily(Z) (Zhao et al., 2020) corpus.

values lead to the rejection of normality whereas a value of one indicates normality of the data.

4.3 Fine-Grained Performances

In the evaluation corpus Daily(Z) (Zhao et al., 2020), it provides four fine-grained human evaluation scores, including *relevance*, *grammar*, *content* and *overall*, which can be used to dive more insights of different evaluation metrics.

As shown in Table 5, our proposed metric FBD[R] achieves the best performances on most evaluations in the fine-grained comparisons. It indicates the distribution-wise metric correlate better with human judgements on various aspects.

5 Conclusions

In this paper, we propose to measure the performance of a dialogue system by computing the distribution-wise difference between its generated conversations and real-world conversations. Specifically, two distribution-wise metrics, FBD and PRD, are developed on pre-trained language models. Experiments on six public dialogue corpora show that our proposed metrics correlate better with human judgments than existing metrics.

References

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.*

Zhangming Chan, Lemao Liu, Juntao Li, Haisong Zhang, Dongyan Zhao, Shuming Shi, and Rui Yan. 2021. Enhancing the open-domain dialogue evaluation in latent space. In *Findings of ACL.*

Wang Chen, Piji Li, and Irwin King. 2021. A training-free and reference-free summarization evaluation metric via centrality-weighted relevance and self-referenced redundancy. In *ACL.*

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *ICLR.*

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the workshop on statistical machine translation.*

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL.*

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.*

DC Dowson and BV Landau. 1982. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455.

Peter W Foltz, Walter Kintsch, and Thomas K Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25(2-3):285–307.

Gabriel Forgues, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay. 2014. Bootstrapping dialog systems with word embeddings. In *NeurIPS, modern machine learning and natural language processing workshop.*

Jun Gao, Wei Bi, Ruifeng Xu, and Shuming Shi. 2021. REAM♯: An enhancement approach to reference-based evaluation metrics for open-domain dialog generation. In *Findings of ACL.*

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

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS.*

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. Grade: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *EMNLP.*

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196.*

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL.*

Yahui Liu, Wei Bi, Jun Gao, Xiaojiang Liu, Jian Yao, and Shuming Shi. 2018. Towards less generic responses in neural conversation models: A statistical re-weighting method. In *EMNLP.*

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

Ryan Lowe, Michael Noseworthy, I. Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic turing test: Learning to evaluate dialogue responses. In *ACL.*

Shikib Mehri and M. Eskénazi. 2020. Usr: An unsupervised and reference free evaluation metric for dialog generation. In *ACL.*

Alexander H Miller, Will Feng, Adam Fisch, Jiasen Lu, Dhruv Batra, Antoine Bordes, Devi Parikh, and Jason Weston. 2017. Parlai: A dialog research software platform. *arXiv preprint arXiv:1705.06476.*

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *ACL.*

Kishore Papineni, S. Roukos, T. Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL.*

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 2019. Semantic image synthesis with spatially-adaptive normalization. In *CVPR.*

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207.*

Nornadiah Mohd Razali, Yap Bee Wah, et al. 2011. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of statistical modeling and analytics*, 2(1):21–33.

Vasile Rus and Mihai Lintean. 2012. An optimal assessment of natural language student input using word-to-word similarity metrics. In *International Conference on Intelligent Tutoring Systems*. Springer.

Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. 2018. Assessing generative models via precision and recall. In *NeurIPS*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *ACL*.

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

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

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Towards universal paraphrastic sentence embeddings. *arXiv preprint arXiv:1511.08198*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2018a. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 41(8):1947–1962.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018b. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *ICLR*.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *ACL*.

Tianyu Zhao, Divesh Lala, and Tatsuya Kawahara. 2020. Designing precise and robust dialogue response evaluators. In *ACL*.

Ganbin Zhou, Ping Luo, Rongyu Cao, Fen Lin, Bo Chen, and Qing He. 2017. Mechanism-aware neural machine for dialogue response generation. In *AAAI*.