Improving Dialog Evaluation with a Multi-reference Adversarial Dataset and Large Scale Pretraining

Ananya B. Sai* **and Akash Kumar Mohankumar*** **and**
Siddhartha Arora and Mitesh M. Khapra
{ananya, miteshk}@cse.iitm.ac.in, {makashkumar99, sidarora1990}@gmail.com
Robert-Bosch Centre for Data Science and Artificial Intelligence
Indian Institute of Technology, Madras

Abstract

There is an increasing focus on model-based dialog evaluation metrics such as ADEM, RUBER, and the more recent BERT-based metrics. These models aim to assign a high score to *all* relevant responses and a low score to all irrelevant responses. Ideally, such models should be trained using multiple relevant and irrelevant responses for any given context. However, no such data is publicly available, and hence existing models are usually trained using a *single* relevant response and multiple *randomly selected* responses from other contexts (random negatives). To allow for better training and robust evaluation of model-based metrics, we introduce the *DailyDialog++* dataset, consisting of (i) five relevant responses for each context and (ii) five *adversarially crafted* irrelevant responses for each context. Using this dataset, we first show that even in the presence of multiple correct references, n-gram based metrics and embedding based metrics do not perform well at separating relevant responses from even *random* negatives. While model-based metrics perform better than n-gram and embedding based metrics on random negatives, their performance drops substantially when evaluated on adversarial examples. To check if large scale pretraining could help, we propose a new BERT-based evaluation metric called DEB, which is pretrained on 727M Reddit conversations and then finetuned on our dataset. DEB significantly outperforms existing models, showing better correlation with human judgments and better performance on random negatives (88.27% accuracy). However, its performance again drops substantially when evaluated on adversarial responses, thereby highlighting that even large-scale pretrained evaluation models are not robust to the adversarial examples in our dataset. The dataset[1] and code[2] are publicly available.

1 Introduction

Open-domain conversational systems are increasingly in demand for several applications ranging from personal digital assistants to entertainers for recreation. While several automated dialogue agents such as Siri, Alexa, Cortana, and Google Assistant have been built and deployed, there is no good automatic evaluation metric to measure the quality of their conversations. Researchers have usually adopted n-gram based metrics (Papineni et al., 2002; Banerjee and Lavie, 2005; Lin, 2004) or embedding based metrics (Forgues et al., 2014; Rus and Lintean, 2012; Zhang et al., 2020a) to compare the model's response with a *single* reference. These metrics assume that a valid response should be semantically or lexically similar to the reference without taking the context of the conversation into consideration. However, in open domain conversations, a given context can have a wide range of possible responses that may be lexically and semantically very different from each other. For example, the context, ''I like dancing and swimming, what about you?'' can be responded to with ''I paint in my free time'' or ''I do not have time for hobbies right now'', both of which are valid responses. As a result, n-gram and word embedding based metrics, which rely on lexical and/or semantic match, correlate very weakly with human judgments for dialogue evaluation (Liu et al., 2016).

*The first two authors worked equally towards the project.

[1]Dataset: https://iitmnlp.github.io/DailyDialog-plusplus/.

[2]Code: https://github.com/iitmnlp/Dialogue-Evaluation-with-BERT.

Transactions of the Association for Computational Linguistics, vol. 8, pp. 810–827, 2020. https://doi.org/10.1162/tacl_a_00347
Action Editor: Xiaojun Wan. Submission batch: 6/2020; Revision batch: 8/2020; Published 12/2020.

Given the shortcomings of context-agnostic n-gram and embedding based metrics, the focus has now shifted to building neural network-based, trainable dialogue evaluation models (Lowe et al., 2017; Tao et al., 2018; Shimanaka et al., 2019; Ghazarian et al., 2019). Such models are trained to identify whether a given response can be considered as a valid continuation of the given context or not. In other words, the model should (i) assign a high score to *all* relevant responses no matter how diverse they are and (ii) assign a low score to all irrelevant responses, preferably with a clear margin of separation from relevant responses. Although there exist several open-domain dialogue datasets (Forsythand and Martell, 2007; Tiedemann, 2012; Ritter et al., 2010; Li et al., 2017b) that are used for training dialogue response generation systems, they are not suitable for training and testing such evaluation models. This is because these datasets have only *a single* relevant response and no irrelevant responses. Irrelevant responses can of course be generated by sampling random utterances from other contexts, but such examples typically do not have any overlap with the context and hence are easier for the model to distinguish from relevant responses (as we will show in our results later). We refer to the randomly sampled responses as *random* negatives.

Some efforts have been made to build dialog datasets with multiple relevant responses (i.e., multiple references), but these datasets are either very small (1000 contexts) (Moghe et al., 2018; Gupta et al., 2019) or automatically constructed from Reddit conversations, hence, potentially noisy (Gao et al., 2019). Further, these datasets do not have any carefully crafted *adversarial* irrelevant responses. We define an *adversarial* irrelevant response as one that has a significant word overlap with the context but is still an irrelevant response (hence harder to identify than randomly selected irrelevant examples, which may not have any relation to the context). To overcome this limitation of existing datasets, we propose a large scale multi-reference dataset, *DailyDialog++*, which is an extension of the DailyDialog dataset. In particular, for each of the 19K contexts derived from DailyDialog, we collect an additional 5 reference responses with the help of human annotators. Further, for $\sim$11K contexts in DailyDialog, we also ask human annotators to carefully craft irrelevant responses that have a significant word overlap with the context. This dataset will be made publicly available and help towards better training and more robust evaluation of dialogue evaluation metrics.

Using this dataset, we extensively evaluate a wide range of n-gram-based and embedding-based metrics. In particular, we compute (i) the correlation of these metrics with binary human judgments and (ii) the accuracy obtained by using the scores assigned by the metrics to classify relevant/irrelevant responses. The performance of these metrics improves when presented with multiple references as opposed to a single reference, but they still leave a lot to be desired. On the other hand, most model-based evaluation metrics, when trained and evaluated using multiple relevant and random negative responses, perform significantly better than the n-gram-based and embedding-based methods. However, their performance drops substantially on the adversarial examples in our dataset.

Lastly, one could argue that dialog evaluation metrics could be improved by pretraining on large amounts of data. To check if this is indeed the case, we propose a new BERT-based evaluation metric called DEB (Dialog Evaluation using BERT), which is pretrained on 727M Reddit conversations. Indeed, this model performs significantly better on random negatives with an accuracy of 88.27% in distinguishing the positive and random negative responses. It also correlates well with human judgments on responses generated by five dialog generation systems (Serban et al., 2016, 2017; Park et al., 2018; Zhang et al., 2020b). In particular, the Spearman rank correlation between human scores and DEB scores is 0.52 at the response level scores and 0.70 at the system level scores, calculated by aggregating the scores on all responses by each system. However, once again, when evaluated on adversarial examples from our dataset, its performance drops substantially, underscoring that even large-scale pretrained models are not robust to adversarial examples.

2 Proposed Dataset

Our goal was to build a dataset with manually created multiple relevant and adversarial irrelevant responses. For this, we wanted to start with an existing base dataset that already has one

relevant response for every context, and then extend it to include multiple responses. For the base dataset, we considered several popular datasets such as Twitter (Ritter et al., 2010), Reddit (Henderson et al., 2019), Open Subtitles (Tiedemann, 2012), NPS Chat (Forsythand and Martell, 2007), PersonaChat (Zhang et al., 2018), and DailyDialog (Li et al., 2017b). Of these, Twitter and Reddit are generally considered noisy, so we chose not to use either of them as the base dataset. Similarly, Open Subtitles and NPS Chat did not have speaker-aligned utterances, and hence were not suitable for our purposes. We found that the DailyDiaog dataset was clean, human-written, readily available, and covered a diverse set of generic topics such as *ordinary life, school life, tourism, attitude & emotion, relationship, health, work, politics, culture & education*, and *finance*. It contains a total of 13K conversations with an average of 8 turns between exactly 2 speakers. Alternatively, we could have also chosen PersonaChat, which is of a similar size and also contains chit-chat style conversations, but we chose the antecedent DailyDialog dataset.

For shorter conversations in DailyDialog (having less than 8 turns) we collected multiple relevant responses only for the last utterance. For longer conversations (having 8 turns or more), we divided the conversation into two or more smaller chunks and collected multiple relevant responses for the last utterance in every chunk. In this way, from the 13K conversations[3] in DailyDialog, we were able to create 19K sub-conversations with multiple relevant responses for the last utterance in each sub-conversation or context. The responses were created by in-house annotators. Each context was shown to 2–3 annotators, and each of them was asked to generate 1–3 alternative responses for the last utterance, capping the total number of alternative responses to 5 (in addition to the one response already available in DailyDialog). The annotators were strictly instructed to avoid short generic responses (''Okay'', ''Thank you'', ''Sure'', etc.), and write longer meaningful responses containing at least 8–10 words. These responses were then verified (and if needed, corrected and re-validated) by a different set of annotators.

2.1 Adversarial Irrelevant Responses

In addition to collecting multiple relevant responses for each context, we also wanted to collect irrelevant responses for each context. Most of the models that are trained for the task of dialogue evaluation (and dialogue generation) (Tao et al., 2018; Ghazarian et al., 2019; Li et al., 2017a) procure irrelevant responses by randomly sampling responses from other contexts. Such random negatives are often entirely out of context (unrelated) and hence are too easy for the model to distinguish. To allow for a more critical or adversarial examination of dialogue evaluation systems, we propose creating adversarially crafted irrelevant responses that have lexical or semantic overlap with the context but are still unacceptable as valid responses.

For obtaining such tricky negative responses, the annotators were asked to choose some words from the context and use them directly or indirectly while writing the responses. Indirect usage here refers to using words closely related to the context words. For example, using synonyms, antonyms, homonyms, subwords, or other words that are known to frequently co-occur with the words in the context (e.g., the words ''flexibility'' and ''injuries'' co-occur with ''acrobatics''). Once again, each context was shown to 2–3 annotators, and each of them was asked to generate 1–3 adversarially crafted responses for the last utterance, capping the total number of alternative responses to 5. Each response was then validated by two different annotators. The validating annotators were instructed to either eliminate or modify the responses that were not negative or were borderline. A final check was made by one more evaluator to ensure that the responses were adversarially crafted, irrelevant, and grammatically correct. We collected 5 such responses for 11,429 contexts. Table 1 shows examples of relevant and irrelevant responses in our dataset and Table 2 shows some statistics about our dataset.

We acknowledge that, in practice, a given context can have a large number of relevant responses ($>> 5$). However, exhaustively collecting all such responses is prohibitively expensive and time-consuming. Although it is desirable to have even more than 5 responses for

[3]Out of the 13K conversations released in DailyDialog, we found that a good number of contexts were repeated, either with slightly different spellings or through some subtle differences such as representing numbers using digits versus using words. We filtered out the repetitions and worked with the remaining $\sim$11K contexts.

Context	Valid responses	Invalid, adversarial responses
FS: Can you do push-ups ? SS: Of course I can . It's a piece of cake ! Believe it or not , I can do 30 push-ups a minute. FS: Really ? I think that's impossible ! SS: You mean 30 push-ups ? FS: Yeah !	SS: You don't believe me, do you? SS: Start your timer, here we go. SS: Watch me do it. SS: That's because you can't do it. SS: You don't know that I am a fitness trainer, do you ?	SS: Push up the window and look out for a minute SS: Would you like to eat a piece of cake before gym? SS: I like watching the Ripley's Believe it or Not show where they discuss nearly impossible feats and gymnastics SS: I have enough time for my treadmill exercises SS: Are you asking me to do 40 squats?

Table 1: Examples from DailyDialog++ dataset with the context consisting of 2 speakers [annotated as FS (First Speaker) and SS (Second Speaker)], and multiple reference responses and adversarial negative responses. The underlined, purple colored words in the adversarial responses are those that overlap or are closely related to the theme or words in the context.

Total # of contexts	19,071
Avg. # of turns per context	3.31
Avg. # of words per context	45.32
Avg. # of words per utterance	13.55
# of contexts with 5 relevant responses	19,071
# of contexts with 5 adv. irrelevant responses	11,429
Avg. # of words per relevant response	10.13
Avg. # of words per irrelevant response	13.8

Table 2: DailyDialog++ dataset statistics.

every context, we believe that having at least 5 is a good starting point given the dearth of such multi-reference conversation datasets. The proposed dataset thus serves as a pragmatic substitute for an ideal dataset that would have contained a large number of responses per context. Having said that, we would also like to point out that the value of the proposed dataset goes beyond having multiple relevant references as it is also the first dataset containing adversarial irrelevant responses for given contexts.

3 Existing Metrics

In this section, we present a brief overview of the existing automatic metrics used for dialogue evaluation. The existing metrics can be broadly classified into two categories, namely, (i) Untrained metrics, and (ii) Trained metrics. Untrained evaluation metrics, usually adopted from the NLG literature, use a predefined formula to compare the candidate response with a reference without taking the context into account. On the other hand, trained metrics are usually trained specifically for the task of dialogue response evaluation to identify valid and invalid responses for a given context.

3.1 Untrained Metrics

Untrained metrics can be further sub-classified into (i) n-gram based, (ii) word embedding based, and (iii) contextualized embedding based metrics.

***N*-gram Based:** N-gram based metrics score a candidate response based on the amount of n-gram overlap it has with a given reference. BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) are among the most commonly adopted n-gram based metrics to evaluate dialogue systems. BLEU is calculated using n-gram precision scores between the candidate response and the reference. ROUGE-L (Lin, 2004) is based on the F-measure of the longest common subsequence between the candidate and reference responses. METEOR (Banerjee and Lavie, 2005) relaxes the exact match criteria by including word stems, synonyms, and paraphrases. More recently, Galley et al. (2015) proposed deltaBLEU, which takes in multiple references and rewards n-gram matches with positive references and penalizes the matches with the negative references.

Word Embedding Based: These methods use word embeddings to compute the similarity between the candidate response and the reference response. The most commonly used word embedding based metrics are Embedding Average (Wieting et al., 2016), Vector Extrema (Forgues et al., 2014), and Greedy Matching (Rus and Lintean, 2012). Embedding Average defines a sentence embedding as the average word embedding of the constituent words. The final score is calculated using the cosine similarity of candidate and reference sentence embeddings. Vector Extrema (Forgues et al., 2014) instead computes the sentence embedding by taking the most extreme value for each dimension. In other words,

the value of the i-th dimension of the sentence embedding is computed by taking a maximum over the i-th dimension of all words in the sentence. Greedy Matching (Rus and Lintean, 2012) first computes the maximum cosine similarity that every word in the candidate response has with any word in the reference response. Similarly, the highest cosine similarity for each of the reference words with any of the candidate response words is calculated. The similarity between the candidate response and reference response is then computed by taking an average of the maximum cosine similarities computed above.

BERTScore: Recently, Zhang et al. (2020a) proposed BERTScore, which uses contextualized word embeddings of the candidate and reference sentences to compute the score. BERTScore is similar to greedy matching but uses contextualized embeddings from BERT instead of static word embeddings.

3.2 Trained Metrics

ADEM: Automatic Dialogue Evaluation Model (ADEM) (Lowe et al., 2017) uses pretrained vector representations of the the dialogue context $\mathbf{c}$, reference response $\mathbf{r}$, and proposed response $\hat{\mathbf{r}}$ to compute the evaluation score as follows:

$$\text{Score}(\mathbf{c}, \mathbf{r}, \hat{\mathbf{r}}) = (\mathbf{c}^T \mathbf{M} \hat{\mathbf{r}} + \mathbf{r}^T \mathbf{N} \hat{\mathbf{r}} - \alpha)/\beta \quad (1)$$

where $\mathbf{M}$, $\mathbf{N} \in \mathbb{R}^{n \times n}$ are learned matrices, and α, β are scalar constants used to re-scale scores in the range $[1, 5]$. The context, proposed response and reference response are encoded using a Hierarchical RNN (H-RNN) encoder consisting of utterance-level and context-level RNNs. The H-RNN encoder is pretrained on a Twitter dataset (Dhingra et al., 2016) in a generative setup using the latent variable hierarchical recurrent encoder decoder (VHRED) model (Serban et al., 2017). The weight matrices, $\mathbf{M}, \mathbf{N}$, are later finetuned for the task of dialogue response evaluation.

RUBER: (Tao et al., 2018) introduced an unreferenced evaluation model consisting of GRU encoders (Chung et al., 2014) to measure the relatedness between the dialogue context and a given response. The authors train the model on Chinese dialogue data with the hinge loss objective.

BERT Regressor[4]**:** Shimanaka et al. (2019) propose a BERT based evaluation model to score a candidate sentence based on a reference. Unlike BERTScore, the BERT model is finetuned to predict human judgement scores from the concatenated reference and candidate sentence.

BERT+DNN[5]**:** Ghazarian et al. (2019) use contextualized embeddings to compute a relatedness score between the dialogue context and response. The best performing model of Ghazarian et al. (2019) consists of a multilayer perceptron that takes the concatenation of contextualized representations of the context and response as input. The contextualized representations are obtained by max-pooling the respective BERT embeddings for each token. Note that the BERT embeddings are not finetuned.

4 Dialogue Evaluation using BERT

In the last two years, considerable success in NLP has been driven by large pretrained transformer-based models (Radford et al., 2019; Devlin et al., 2019; Zhang et al., 2019). These models are typically trained with a language model objective and leverage large amounts of unlabeled data. However, none of the trained metrics discussed in the previous section leverage pretraining on large-scale dialogue corpora. With the hope that such pretraining should help dialog evaluation models also, we introduce DEB (Dialog Evaluation using BERT) which is trained using a masked language model objective (similar to BERT) and a modified next response prediction objective.

We set up the the task of next response prediction as one of identifying whether the given response is a valid next response for the given context. Formally, given a context $\mathbf{C} = \{w_1^c, \ldots, w_n^c\}$ and a response $\mathbf{R} = \{w_1^r, \ldots, w_m^r\}$, we first pass the concatenated sequence $\mathbf{U} = \{[\text{CLS}], w_1^c, \ldots, w_n^c, [\text{SEP}], w_1^r, \ldots, w_m^r\}$ through the BERT transformer and obtain $\mathbf{H}_{cls} \in \mathbb{R}^H$, the last-layer activations corresponding to the special [CLS] token. We then make our final next response predictions as follows: $\hat{y} = \text{softmax}(\mathbf{W}\mathbf{H}_{cls})$, where $\mathbf{W} \in \mathbb{R}^{2 \times H}$ is a learnable matrix. We

[4]Because we couldn't find an exact name for the evaluator model by Shimanaka et al. (2019) , we adopt the name 'BERT regressor' from their paper's title.

[5]Due to the lack of a specific name for the models in Ghazarian et al. (2019), we refer to the model adopted from their work as 'BERT+DNN'

use cross entropy loss with binary targets for the next-response prediction. In addition, we use the standard masked language model objective by randomly masking 15% of the words in $\mathbf{C}$ and $\mathbf{R}$.

Note that the proposed model is a straightforward extension of the standard BERT model used for language modeling. We do not claim any novelty on this front. The key contribution here is to assess if pretraining on large-scale dialogue corpora improves the performance of dialogue evaluation metrics. Existing BERT-based evaluation metrics (Shimanaka et al., 2019; Ghazarian et al., 2019) do not use such pretraining on any large-scale, domain-related corpora. In other words, they do not leverage the more successful recipe of (i) pretraining with a masked language modeling objective and (ii) finetuning with a task-specific objective (dialog evaluation in this case). The idea behind DEB is to check if this successful recipe can be replicated for dialog evaluation, making use of the dialogues in the large-scale Reddit corpus.

4.1 Training Details

For pretraining, we use a massive open-domain dialogue dataset of Reddit comments from 2005 to 2019 consisting of 256M threads with a total of 3.68B comments. From this dataset, we extracted a total of 727M $\{$*context*, *positive response*$\}$ pairs with 654M for training and 73M for testing following the method described in Henderson et al. (2019). We used an equal number of negative responses by randomly sampling responses from other contexts. We use the BERT base model with 110M parameters consisting of 12 layers, 768 dimensional hidden space, and 12 attention heads per layer in all our experiments. We finetune the pretrained DEB model on our DailyDialog++ dataset for 1 epoch (we did not see any advantage of finetuning beyond 1 epoch). Note that during finetuning we only use the next response prediction objective.

5 Experimental Setup

Our goal is to check if the adversarial responses in our dataset, which are specifically crafted to target context-dependent model-based metrics (such as ADEM, RUBER, BERT+DNN, and DEB), indeed affect the performance of such models. To do so, we first need to benchmark the models' performance on random negatives and then check if the performance drops when evaluated on adversarial examples. Hence, in this section, we describe (i) the process of creating and validating such random negatives and (ii) the process used for training model-based metrics.

We randomly divide our dataset into train (80% contexts), validation (10% contexts), and test (10% contexts) splits. Note that adversarial negatives are not used for training or finetuning the models unless explicitly specified.

5.1 Creating & Validating Random Negatives

For every context in our dataset, which has 5 relevant responses, we also sample 5 random negatives. While sampling random negatives, we avoid short responses that may be generic and relevant for any context. To verify whether the sampled random negatives were indeed irrelevant, we asked human annotators to manually check 500 such sampled responses. More specifically, we showed them the original context and the sampled random negative response and asked them if it was a relevant or irrelevant response. In 95% of the cases, the annotators confirmed that the random negative response was irrelevant, thereby confirming that a random sampling strategy indeed results in irrelevant responses (although they may not be as hard as our adversarial negative examples as shown later).

5.2 Pretraining & Finetuning Trained Metrics

We describe the pretraining and finetuning procedure for the various models used in our analysis below.

ADEM: As previously mentioned in Section 3, ADEM was pretrained on Twitter corpus using the VHRED setup and then finetuned for dialogue response evaluation. We take this publicly available model and finetune it further using our DailyDialog++ dataset with a target of 5 for positive responses and 1 for random negatives. The reference response could be any of the other four relevant responses. Note that ADEM produces a score on a scale of 1 to 5 whereas the other models produce a score on a scale of 0 to 1. For easier comparison, we scale the output of ADEM so that it lies in the range of 0 to 1.

BERT regressor: We finetune the publicly available pretrained BERT base model (110M

parameters) on our DailyDialog++ dataset. We train the model with a label of 1 for positive responses and 0 for random negative responses using any one of the other four positive responses as the reference. We train the model using cross-entropy loss and follow the same set of hyperparameters as used by Shimanaka et al. (2019) during finetuning.

BERT+DNN: We use the best performing model from Ghazarian et al. (2019), which consists of a three-layered feed-forward neural network and uses pretrained BERT embeddings as input. We train the model on our DailyDialog++ dataset with random negatives using cross entropy loss.

RUBER and RUBER-Large: We experiment with two variants of Tao et al.'s (2018) models with different sizes, namely (i) RUBER (34M parameters), which consists of single-layer GRUs with a hidden size of 1024, and (ii) RUBER-Large (236M parameters), which consists of two layered GRUs with a hidden size of 2048. As shown in Vaswani et al. (2017), the training time for RNN based architectures is very high when compared with the transformer models that allow much greater parallelization. We observed an estimated time of over 200 days to train the RUBER-Large model on the 727M Reddit corpus on a 1080ti GPU, thereby making it practically infeasible to train such models on large-scale datasets. Taking the computational costs into consideration, we pretrained RUBER and RUBER-Large on a sample of 20M contexts with relevant and random irrelevant responses from Reddit. We then finetuned these models on our proposed dataset with random negatives.[6]

DEB: We pretrained DEB on the entire 727M Reddit corpus using the masked language model and the modified next response prediction objective. Pretraining DEB took 4 days on a single Google Cloud TPUv2. We achieved a test accuracy of 90% on the next response prediction task and a perplexity of 15.47 (58% accuracy) on the masked language modeling task in the pretraining corpus. We then finetuned DEB on our dataset with random negatives.

[6]We agree that this may not be a fair comparison but we were constrained by the inherent limitations of such RNN-based, sequential models that make large-scale pretraining prohibitively expensive and time-consuming.

5.3 Untrained Metrics with Multiple References

Untrained metrics like METEOR, Greedy Matching, and so forth, usually work with a single reference response but can also be adapted to work with multiple reference responses. For example, for a given candidate response c and a set of reference responses $r_1, r_2, r_3, ..., r_k$, we can compute the multi-reference METEOR score as:

$$METEOR_{multi} = max_{i=1}^{k} METEOR(c, r_i)$$

Instead of the *max* function we can also use the average function. We use a similar formula for all the untrained metrics.

A few metrics like BLEU, deltaBLEU, and ROUGE-L have their own standard formula to incorporate multiple references. BLEU calculates the number of matches for each n-gram based on the maximum number of times the n-gram occurs in common with any one of the references. deltaBLEU further extends the same idea to incorporate a score for each reference. We follow the implementation from Galley et al. (2015) to compute the deltaBLEU scores. For ROUGE-L, we follow the strategy in Sharma et al. (2017), where the score is an F-measure of the maximum precision and maximum recall over all the references. In addition to the average and maximum aggregations, we also report these standard multi-reference scores for BLEU, deltaBLEU, and ROUGE-L.

6 Results

In this section, we compare the performance of different dialog evaluation metrics in separating relevant references from (i) random negatives, (ii) synthetically crafted adversarial irrelevant responses (explained below), and (iii) manually crafted adversarial irrelevant responses (as in our DailyDialog++ dataset).

6.1 Performance on Random Negatives

For every context in our test split, we obtain the scores assigned by a given metric to the 5 positive and 5 random negative responses. In particular, we treat each of the 5 relevant and 5 random irrelevant responses as a candidate response. For all untrained metrics other than deltaBLEU, we consider the remaining 4 relevant responses as reference responses. For deltaBLEU, we consider the remaining 4 relevant responses as references

Metric	Point Biserial Correlation (p-value)				Accuracy in percentage			
	Single	Multiple			Single	Multiple		
		Avg	Max	Standard		Avg	Max	Standard
BLEU-1	0.26 (<1e-9)	0.42 (<1e-9)	0.41 (<1e-9)	0.41 (<1e-9)	61.26	68.60	68.75	70.36
BLEU-2	0.22 (<1e-9)	0.39 (<1e-9)	0.36 (<1e-9)	0.40 (<1e-9)	58.09	68.26	68.37	68.66
BLEU-3	0.14 (<1e-9)	0.26 (<1e-9)	0.24 (<1e-9)	0.28 (<1e-9)	53.11	58.85	58.90	58.89
BLEU-4	0.08 (<1e-9)	0.17 (<1e-9)	0.15 (<1e-9)	0.18 (<1e-9)	51.16	53.56	53.56	53.50
METEOR	0.23 (<1e-9)	0.40 (<1e-9)	0.41 (<1e-9)	–	59.77	68.51	68.01	–
ROUGE-L	0.23 (<1e-9)	0.41 (<1e-9)	0.40 (<1e-9)	0.37 (<1e-9)	59.47	67.89	68.25	68.43
deltaBLEU (Galley et al., 2015)	–	–	–	0.29 (<1e-9)	–	–	–	64.89
Embed Avg	0.23 (<1e-9)	0.25 (<1e-9)	0.23 (<1e-9)	–	61.27	61.56	62.67	–
Vec Extr (Forgues et al., 2014)	0.24 (<1e-9)	0.35 (<1e-9)	0.33 (<1e-9)	–	59.22	63.70	63.90	–
GreedyMatch (Rus and Lintean, 2012)	0.24 (<1e-9)	0.36 (<1e-9)	0.32 (<1e-9)	–	60.02	63.99	65.56	–
BERTScore (Zhang et al., 2020a)	0.29 (<1e-9)	0.39 (<1e-9)	0.39 (<1e-9)	–	63.71	69.05	68.59	–
ADEM (Lowe et al., 2017)	0.40 (<1e-9)				64.74			
BERT regressor (Shimanaka et al., 2019)	0.52 (<1e-9)				73.40			
BERT+DNN (Ghazarian et al., 2019)	0.57 (<1e-9)				74.67			
RUBER (Tao et al., 2018)	0.64 (<1e-9)				78.18			
RUBER-Large (Tao et al., 2018)	0.69 (<1e-9)				82.36			
DEB (ours)	**0.79*** (<1e-9)				**88.27***			

Table 3: Automatic evaluation metrics performance on random negatives (PBC refers to point-biserial correlation. Column subheading 'Single' refers to experiments using single reference response and 'Avg' and 'Max' are the average and maximum aggregation strategies when using multiple reference responses. 'Standard' is applicable when the metric aggregates multiple references differently. * indicates statistical significance in performance over all other metrics (with p-values <1e-9) on William's test for comparing correlations and Chi-squared test for accuracies. p-values for individual correlations are in parenthesis.

with a score of 1 and the remaining 4 irrelevant responses as references with a score of -1. We expect a good evaluation metric to provide high scores on relevant responses and low scores on the irrelevant responses. We then quantify the performance of all metrics using two measures. First, we compute the Point Biserial correlation (PBC) between the scores assigned by a metric and the binary target i.e., a score of 1 for positive responses and 0 for random negative responses.[7] Second, we compute the classification accuracy of the metric by using a threshold and marking all responses having a score above this threshold as positive and others as negative. We use a threshold of 0.5 for the trained metrics. For all the untrained metrics, we perform a search from 0 to 1 with step size of 0.01 and select the threshold that minimizes the error rate on the validation set.[8] Later in Section 6.1.1, we shall observe that if we use 0.5 as the threshold, the performance of most untrained metrics would be abysmally poor. Note that for the trained metrics we found that the scores were spread evenly in the range of 0 to 1 and there was no benefit of doing a grid search to find the threshold—a threshold of 0.5 was adequate.

In Table 3, we report PBC and accuracy of the different untrained metrics with both single and multiple references, and the trained metrics. When evaluating using single references, we use any one of the 5 relevant responses as a reference response (other than the one being used as a candidate). We observe that with a single reference, all the untrained metrics are poor at distinguishing between the positive and random negative responses as inferred from the low accuracy and correlation values. When we use multiple responses, we observe a relatively better performance. We notice that the performance is largely similar across the aggregation techniques: average, maximum, and standard (when applicable). Metrics such as BLEU-1, METEOR, ROUGE-L, and BERTScore with multiple references are able to achieve modest correlations with the binary target. Interestingly, we observe that all the word embedding based methods, even in the presence of multiple references, perform badly in scoring the positive and random negative responses. In contrast, trained metrics such as

[7]Note that it can be shown that PBC is equivalent to the Pearson correlation when one of the variables is binary, as is the case above.

[8]With this approach of setting a threshold, we want to be lenient with the untrained metrics and investigate how best they can be adopted. One might also think of using the median of all the scores assigned by a metric as its threshold, however, such an approach is error-prone and has several boundary conditions that would fail the purpose. We hence estimate the threshold by minimizing the risk.

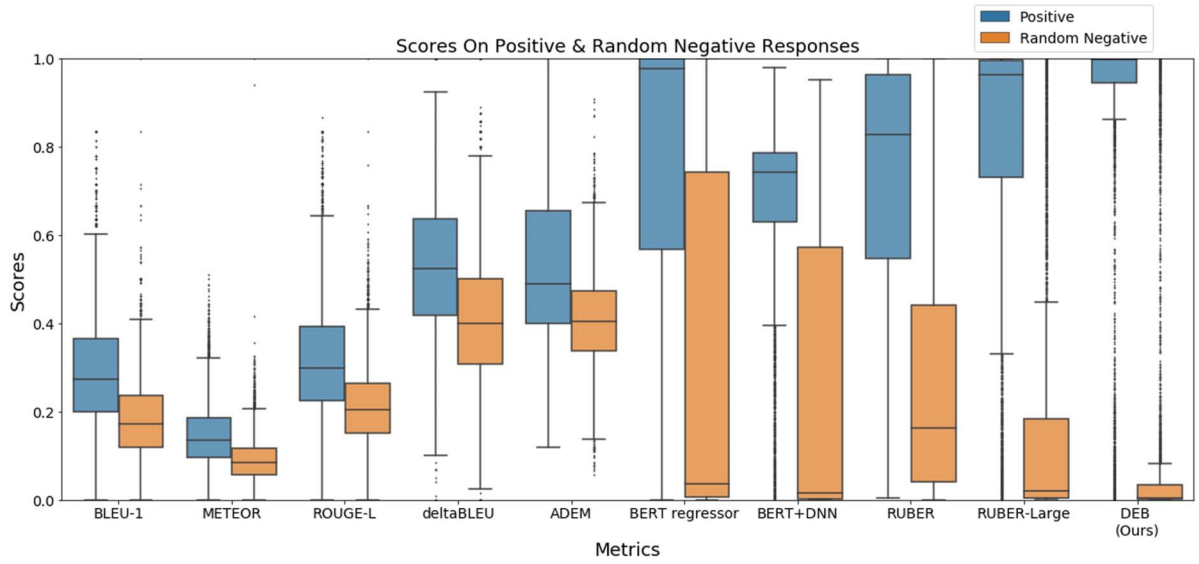


Figure 1: Box plots of the scores given by various metrics to the positive and random negative responses.

BERT regressor, RUBER, BERT+DNN, and DEB perform substantially better than the untrained metrics. Our proposed DEB model achieves state-of-the-art performance, with an accuracy of 88.27% and a strong correlation of 0.79.

6.1.1 Analysis using Box Plots

We now visualize the box plots of the scores given by the various metrics to the positive and random negative responses. Figure 1 shows these box plots for the multi-reference untrained metrics (max aggregation) and the trained metrics. We observe several shortcomings of the untrained metrics. Firstly, all the untrained metrics have a significant overlap in the interquartile range of the positive and random negative scores, implying that there is a high degree of intermixing of scores given to the positive and random negative responses. The overlap is even higher for word embedding based metrics, which obtain low point biserial correlations. Secondly, we note that the score distributions of the untrained metrics are highly skewed. For instance, the scores of BERTScore are almost always greater than 0.75 even though it scores responses in the range [0,1]. Therefore, it is difficult to tell at what value of the metric a response can be safely considered relevant. These observations suggest that untrained metrics even with multiple references cannot be reliably used to score dialogue responses.

For the ADEM evaluation model, we observe that it outputs scores close to the mean score of 0.5 with little spread in their values. Sai et al. (2019) also made similar observation about the clustering of the scores around the mean in ADEM, which they explain using linear system theory. In BERT regressor, there is a high overlap in the scores given to positives and random negatives. We further observe that the RUBER and BERT+DNN are able to better distinguish the positive and random negative responses. Although there is separation in the interquartile range for the two classes in RUBER and BERT+DNN scores, there is a greater spread within each class and a lot of points of the two classes substantially overlap. RUBER-Large is able to reduce the overlap, while DEB further achieves better performance by pushing the scores for positive responses close to 1 and the scores for random negatives to 0 with high accuracy. We shall show in Section 7.3 that DEB achieves this by pushing the $\mathbf{H}_{cls}$ embeddings for the positive and random negative responses farther apart in space.

6.2 Performance on Synthetically Crafted Adversarial Responses

Due to space constraints, in the remainder of this section we present results only for the best performing evaluation metrics from Table 3, namely, BERT+DNN, RUBER, RUBER-Large, and DEB. Before evaluating them using the adversarial examples in our dataset, we first investigate the performance of the models with synthetically crafted adversarial attacks, similar to Sai et al. (2019). In particular, we perform simple transformations on relevant responses by

Modification	DEB	RUBER-Large	RUBER	BERT+DNN
	% classified as positive			
Unmodified positives	87.9%	81.7%	77.5%	93.5%
Reverse word order	60.0%	70.3%	71.3%	80.4%
Jumble word order	69.3%	71.2%	72.3%	77.4%
Retain only nouns	60.1%	27.9%	27.8%	0.0%
Remove punctuation	86.4%	72.9%	72.4%	88.5%
Remove stopwords	85.8%	73.6%	69.6%	29.3%
Replace with synonyms	81.2%	70.8%	65.6%	91.1%
	Pearson Correlation with human scores			
Remove stopwords	0.58 (<1e-9)	0.56 (<1e-9)	0.52 (<1e-9)	0.056 (0.26)
Replace with synonyms	0.68 (<1e-9)	0.57 (<1e-9)	0.54 (<1e-9)	−0.017 (0.67)

Table 4: Fraction of responses classified as positives with synthetic modifications. Unmodified positives are presented in the 1st row for reference (p-values for individual correlations in brackets).

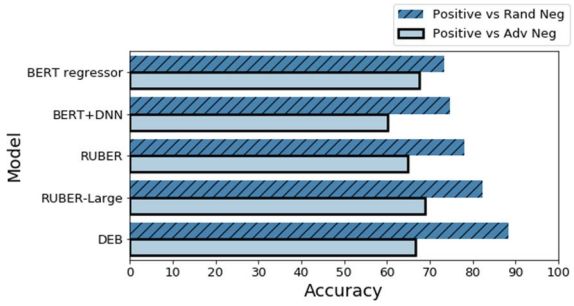


Figure 2: Accuracy of different models in identifying adversarial and random negatives versus positive responses.

(i) jumbling words in the sequence, (ii) reversing the sequence, (iii) dropping all words except nouns, (iv) dropping all stop words, (v) dropping punctuation, and (vi) replacing words with synonyms. These results are presented in Table 4.

The modifications of reversing and jumbling the word order in a relevant response make it irrelevant (grammatically wrong) and hence we expect to see more of the original true positives get classified as negatives. BERT+DNN classifies a majority of these responses as positives. One possible reason for this is that their model only uses a max pooled aggregation on BERT embeddings and does not explicitly model the sequential order of words. On the other hand, DEB fares better than the other models as seen by the drop in fraction of responses identified as positives. However, RUBER variants and BERT+DNN do better than DEB when retaining only nouns in a response. On removing punctuation, we expect that most of the positive responses without punctuation would remain positive and hence the percentage of responses marked positive should remain about the same. In this case, both DEB and BERT+DNN perform better than the RUBER models. For the modifications of removing stopwords and replacing words with synonyms, it is hard to generalize the trend that is observed. Hence, we perform human evaluations by presenting in-house annotators with contexts and modified responses. We ask them to provide scores in the range 0 to 3, with higher scores meaning better responses. We obtain human scores on 400 samples for this task and compute the Pearson correlation of the model predictions with the human judgements. In this case, we find DEB is better correlated with human judgements on both the modifications.

6.3 Performance of Model-Based Metrics on Manually Crafted Adversarial Responses

So far we have established that (i) untrained metrics perform poorly compared to trained metrics even for separating random negatives from positives (ii) trained models like RUBER, BERT+DNN, RUBER-Large and DEB perform remarkably well in distinguishing relevant responses from random responses (iii) RUBER variants and DEB perform well on most synthetically mutated responses whereas BERT+DNN performs poorly against certain mutations. However, we still need to check if the trained models are robust to adversarial examples which are specifically crafted to fool such *context-dependent*, model-based metrics. Note that none of the untrained metrics are context dependent as they directly compute the similarity between the reference and candidate response without considering the context.

We consider the 5 relevant and the 5 adversarial irrelevant responses in our dataset and just as before compute the scores assigned by the different metrics to each of these responses. We then compute the accuracy of a metric using the target label as 0 for irrelevant responses and 1 for relevant responses. As expected, the accuracy of all the models drops, as seen in Figure 2. In particular, we observe that the models wrongly classify most of the irrelevant responses as positive/relevant responses. This can be seen from the confusion matrices in Table 5, where it is clear that the number of false positives is very high.

TP FN FP TN	Positive vs Random negatives		Positive vs Adversarial negatives	
BERT+DNN	5337	373	5337	373
	2520	3190	4179	1531
BERT regressor	3442	1126	3442	1126
	1304	3264	1837	2731
RUBER	4420	1280	4420	1280
	1207	4493	2714	2986
RUBER-Large	4659	1041	4659	1041
	970	4730	2500	3200
DEB	5011	689	5011	689
	646	5054	3101	2599

Table 5: Confusion matrix showing changes in the performance of different models on DailyDialog++ with random and adversarial negatives.

Model	Pos vs Rand Neg	Pos vs Adv Neg
BERT original	72.65	58.10
DEB pretrained on Reddit	84.16	59.82
Pretrained DEB finetuned on rand neg	88.29	66.75

Table 6: Ablation studies on DEB.

7 Discussion

In this section, we do further analysis of DEB.

7.1 Ablation Studies on DEB

There are different stages of training our DEB model. First, the underlying BERT model is already pretrained on English Wikipedia and the BooksCorpus. We then pretrain it further for our task using Reddit corpus and finally finetune it on the DailyDialog++ dataset. We now evaluate the contributions of each of these stages of training (see Table 6). First, we find that the original BERT model when adopted directly for the task of dialog evaluation gives an accuracy of 72.65% and 58.10% on random and adversarial negatives respectively. On further analysis, we find that it has a high false positive rate, with more than 52% of the adversarial negatives getting classified as positives. After pretraining it with Reddit data, it achieves an accuracy of 84.16% on DailyDialog++ even though it has not seen any training instances from this dataset.

Model	Training/ Finetuning Data	Pos vs Rand Neg	Pos vs Adv Neg
BERT regressor	Rand neg	73.40	67.57
	Adv neg	69.89	75.92
	Rand + Adv neg	72.77	74.55
BERT+DNN	Rand neg	74.67	60.14
	Adv neg	60.49	87.67
	Rand + Adv neg	73.87	86.61
RUBER (Pretrained)	Rand neg	78.18	64.96
	Adv neg	70.82	76.50
	Rand + Adv neg	75.11	83.88
RUBER-Large (Pretrained)	Rand neg	82.35	68.94
	Adv neg	63.99	90.49
	Rand + Adv neg	79.91	86.54
DEB (Pretrained)	Rand neg	88.29	66.75
	Adv neg	86.24	82.04
	Rand + Adv neg	**88.67**	**92.65**

Table 7: Accuracy in classifying Pos vs Rand Neg and Pos vs Adv Neg responses for various model variants trained/finetuned on DailyDialog++.

However, there is only a marginal improvement on adversarial negatives. Finally, finetuning BERT on DailyDialog++ using only random negatives further improves the accuracy to 88.29% and 66.75%, respectively.

7.2 Training with Adversarial Examples

We examine whether the evaluation models can learn to distinguish the adversarial negatives when specifically finetuned for that task. By training on DailyDialog++ with adversarial negatives rather than random negatives, we find that all models give an improved performance in identifying adversarial negatives (see Table 7). However, with such training, every model's performance drops when evaluated on DailyDialog++ with random negatives, with BERT+DNN dropping substantially to 60.49%. The best overall performance is seen when the models are finetuned with both random and adversarial negatives, with DEB achieving the highest accuracies on both test sets. While such improvement is expected given the capacity of the models, obtaining such adversarial examples for training is not always feasible.

Effect of the Number of Adversarial Negatives Added to Training: Because of the difficulty in manually creating adversarial examples, we study the effect of the number of the adversarial examples added to the training set. Our findings are presented in Figure 3, where we progressively increase the percentage of adversarial negative examples added as input to the DEB model during training with random negatives. As expected,

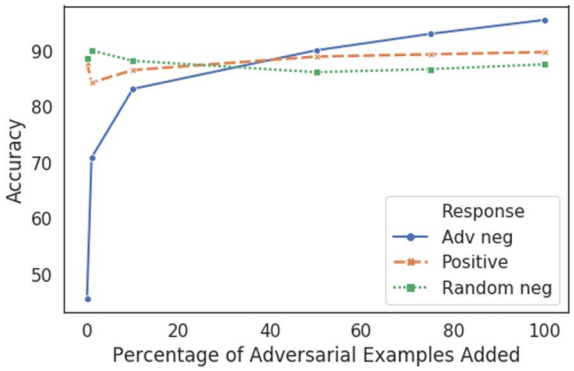


Figure 3: Effect of varying the amount of adversarial negatives added to the training set.

the accuracy in identifying adversarial negatives improves as the model is exposed to more data points of the same type, where we specifically note the considerable improvement from 45.6% to 70.85% after adding just 1% of adversarial negatives from our dataset (i.e., 100 contexts with 5 adversarial examples each). With the addition of more adversarial negatives, we find a small drop in the accuracy of identifying random negatives. There is also a slight decrease in the performance on the positives responses when the number of adversarial examples are small. We note that the adversarial negatives are hard negatives close to the positive responses in the embedding space, as we elaborate in Section 7.3, thereby confusing the model.

7.3 Conicity Analysis on DEB

We analyze the embeddings from the final embeddings projection space, that is, the one used by softmax layer for next response prediction. We check for the spread of the embeddings of the positive and negative responses. Specifically, let P, R, and A be the set of embeddings of all positive responses, random negative responses, and adversarial negative responses respectively for a given context. We want that if we consider the set P then the spread of this set should be low in the projected space (all positive responses embedded close to each other). At the same time, if we consider the union of the sets P, R, and A then the spread of this set should be high (positive responses separated from negative responses). We measure this spread using conicity analysis (Chandrahas et al., 2018). Conicity on a set of vectors V is defined as the average of the cosine similarity of the vectors with their mean vector, $\bar{v}$. The lower the conicity, the higher the spread.

For each utterance in DailyDialog++, we first construct the sets P, R, and A using the pretrained DEB model. We find that the average conicity of the set P is 0.89 (averaged over all utterances), indicating that the positive responses get mapped very close to each other. The average conicity of the set $P \cup R$ is 0.59, indicating that the positive responses are well separated from the random negatives. However, the average conicity of the set $P \cup A$ is 0.74, indicating that the positive responses are not well separated from the adversarial negative responses. We illustrate this in Figure 4a by representing the mean vector of each of the sets along a corresponding highlighted region where the vectors of the set lie on average.[9] We then finetune the DEB model on the DailyDialog++ dataset. Once again, for every utterance we construct the sets P, R, and A using this finetuned model. We now observe that the average conicity of the sets P, $P \cup R$, and $P \cup A$ are 0.86, 0.37, and 0.35 respectively. Thus, after finetuning, the model is able to achieve a clear separation between positive responses and random or adversarial negative responses. Furthermore, the positive responses are still close to each other (illustrated in Figure 4b).

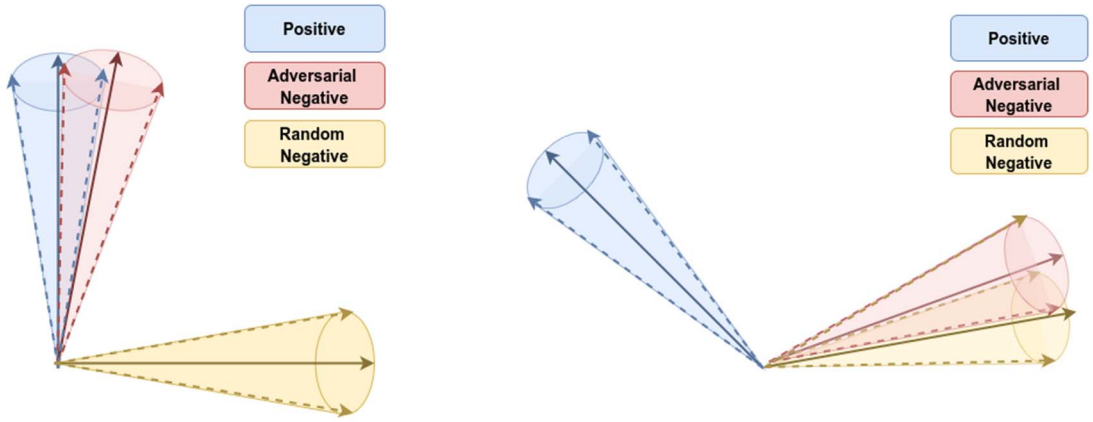
8 Generalization to Other Datasets

In this section, we investigate how well the different model-based metrics trained on DailyDialog++ generalize to other datasets that are not seen during training. We evaluate the 3 unreferenced models, BERT+DNN, RUBER, and DEB, which require only context and candidate response as inputs on these 3 datasets.

Twitter: Microsoft Research Social Media Conversation Corpus (Sordoni et al., 2015) contains a curated list of 3-turn Twitter conversations, all of which are human-verified as good responses.

PersonaChat: The dialogues in PersonaChat (Zhang et al., 2018) are associated with well-defined personalities of the speakers involved. We consider the verified human-human chat logs, released by See et al. (2019), as positive examples.

[9]Note that separation of cones in the figure does not indicate complete separation of all the vectors between the sets, rather separation on average, as there could be some overlap or outliers, as evident from the model's performance in various experiments.



(a) Before finetuning on DailyDialog++ (b) After finetuning on DailyDialog++

Figure 4: Illustration of the spread of the positive and negative response embeddings by DEB (not to scale).

Model	Persona	Twitter	Holl-E
BERT+DNN	71.01	48.71	54.60
RUBER	61.17	71.18	54.83
RUBER-Large	62.32	77.18	55.94
DEB	**78.55**	**82.71**	**62.74**

Table 8: Transferability to other datasets.

Holl-E: This dataset (Moghe et al., 2018) contains conversations about movies, where each response is generated by copying and modifying content from a relevant background document. We use the multi-reference test set of Holl-E containing 4 positive responses for each context.

For all the 3 datasets, we consider the reference responses as positive responses and obtain negative examples by randomly sampling responses from other contexts. We reiterate that we do not train the models on these datasets but simply evaluate the models trained on DailyDialog++ on these datasets. Table 8 shows that DEB outperforms the other unreferenced models on all the 3 datasets. With Holl-E dataset being specific to conversations about movies rather than generic topics, we find the scores are relatively lower on it for all the models. The other evaluation models and metrics cannot be compared on PersonaChat and Twitter without additional reference responses, since the available single reference in these datasets is being evaluated. On the multi-reference test set of Holl-E, however, we find that their performance is lower than the three unreferenced models.

9 Correlations with Human Judgments on System Generated Responses

Lastly, we wanted to check if DEB scores correlate well with scores assigned by humans on responses generated by dialogue systems (as opposed to humans). To do so, we collected responses generated by the following five dialogue response generation models:

HRED: Hierarchical Recurrent Encoder Decoder (HRED) (Serban et al., 2016) extends the traditional seq2seq model by adding an additional utterance-level RNN.

VHRED: Latent Variable HRED (VHRED) (Serban et al., 2017) includes a latent variable at the decoder, and is trained by maximizing a variational lower-bound on the log-likelihood.

VHCR: Variational Hierarchical Conversation RNN (VHCR) (Park et al., 2018) further extends VHRED by drawing a prior encoding for each conversation.

DialoGPT small: Zhang et al. (2020b) pretrained GPT-2-like (Radford et al., 2019) transformer models on 147M conversations extracted from Reddit comments. The small version contains 12 layers and 768 hidden dimensions.

DialoGPT medium: The medium version of DialogGPT contains 24 layers and 1024 hidden dimensions.

For the RNN-based models (HRED, VHRED, VHCR), we use a single-layer bidirectional encoder and single-layer decoder each with a hidden size of 1024. We pretrain the RNN-based

Model	Pearson	Spearman	Kendall tau
Response level			
BERT+DNN	0.016 (0.73)	0.009 (0.89)	0.007 (0.88)
RUBER	0.111 (2.5e-2)	0.126 (1.1e-2)	0.090 (8.9e-2)
RUBER-Large	0.265 (<1e-7)	0.256 (<1e-6)	0.173 (<1e-6)
DEB w/o Reddit	0.356 (<1e-9)	0.295 (<1e-9)	0.202 (<1e-9)
DEB w/o DD++	0.274 (<1e-9)	0.337 (<1e-9)	0.232 (<1e-9)
DEB	**0.440*** (<1e-9)	**0.523*** (<1e-9)	**0.374*** (<1e-9)
System level			
BERT+DNN	0.050 (0.89)	-0.100 (0.87)	0.000 (1.1)
RUBER	0.221 (0.72)	0.300 (0.62)	0.200 (0.81)
RUBER-Large	0.679 (0.20)	0.499 (0.39)	0.399 (0.483)
DEB w/o Reddit	0.784 (0.12)	0.600 (0.28)	0.400 (0.48)
DEB w/o DD++	0.855 (0.06)	0.600 (0.28)	0.400 (0.48)
DEB	**0.973** (5.2e-3)	**0.700** (0.18)	**0.600** (0.23)

Table 9: Human correlations on DailyDialog++ data with different models. (Individual p-values in parenthesis.) * indicates statistical significance in performance over other models, with p-values <1e-6 on the William's test.

models on the casual conversation subset of the Reddit dataset, consisting of 10M conversation exchanges. We finetune all the models on the DailyDialog++ dataset.

We conducted human evaluations to compare the extent to which the model-based metrics agree with human judgements. We randomly sampled 100 contexts from the test set of the DailyDialog++ dataset and obtained the responses generated by each of the above models. Annotators were shown a context-response pair and were asked to rate how human-like the response is with respect to the context, on a scale of 0–3. The annotators were asked to check for both fluency and coherence. A total of 15 in-house annotators participated in the human evaluation study. The annotators were Computer Science graduates competent in English. Each context-response pair was rated by 5 annotators and the final score was obtained by averaging the 5 scores. We also obtained scores at the system level by aggregating the scores for each model. In Table 9, we report the correlations of human judgments with the model scores at the response level and system level. We observe that the BERT+DNN model, which only has a feed-forward neural network that is learnable, does not have any significant correlation with human judgments. On the other hand, RUBER, consisting of pretrained GRUs, obtains low to moderate correlations. RUBER-Large further obtains improved correlations, indicating that using large-scale pretrained models helps. This trend is also observed in the comparisons of DEB with its ablated versions (without Reddit pretraining and without finetuning on DailyDialog++), indicating the contribution of these steps in training the final model. Our proposed DEB model obtains significantly higher correlations at response level. We checked for significance using William's test to compare DEB with all other models and found p-values to be $< 1e^{-6}$. This establishes the effectiveness of DEB in scoring model generated responses. At the system level, we find that DEB correlates substantially higher than other models, with the human rankings of the models. However, the p-values in this case are not significant due to the limited number of systems. In hindsight, we realize that reporting system level correlations is not very informative as the number of samples are very small (as many as the number of systems). Hence, these numbers are not very reliable. However, following Lowe et al. (2017), we still report the system-level correlations (along with the p-values) for the sake of completeness.

10 Related Work

We point the reader to Serban et al. (2018) for an excellent survey of existing datasets containing single reference responses. Recently, there has been some effort to create datasets containing multiple references but these datasets are either too small (around 1,000 contexts) (Moghe et al., 2018; Gupta et al., 2019) or noisy (Gao et al., 2019).

We have already reviewed all the existing dialog metrics in Section 3 and hence we do not discuss them again here. Instead, we quickly mention existing works which critically examine dialog evaluation metrics. For example, Liu et al. (2016) show that existing n-gram based metrics do not correlate well with human judgements for dialog evaluation. We report similar results but additionally show that the correlation improves in the presence of multiple references. Similarly, Sai et al. (2019) have critically examined ADEM and shown that in most cases it produces a score close to 2.5 (on a scale of 1 to 5) and hence does not clearly separate relevant and irrelevant responses.

Lastly, we also mention a very recent work, Zhang et al. (2020b), which has pretrained a large scale transformer on Reddit corpus for building conversation systems. However, their focus is on dialog generation and not on evaluation metrics.

11 Conclusions

We propose a multi-reference open-domain dialogue dataset with multiple relevant responses and adversarial irrelevant responses. We perform an extensive study of the existing dialogue evaluation metrics using this dataset and also propose a new transformer-based evaluator pretrained on large-scale dialogue datasets. We identify the strengths and weaknesses of such a model through studies of its performance on untrained and synthetically modified data. We find DEB to be easily adaptable to other open-domain dialogue datasets. We also present the scope of the adversarial responses in our dataset towards bringing out better evaluation metrics, since all the current models do not perform well on those unless explicitly trained.

Acknowledgments

We thank the Department of Computer Science and Engineering, IIT Madras and the Robert Bosch Center for Data Science and Artificial Intelligence, IIT Madras (RBC-DSAI) for providing us resources required to carry out this research. We are grateful to Google for the TFRC credits that supported our usage of TPUs for several experiments in this paper. We also thank Google for supporting Ananya Sai through their Google India Ph.D. Fellowship Program. We thank the action editor, Xiaojun Wan, and all the anonymous reviewers for their very helpful comments in enhancing the work. We thank the in-house human annotators and evaluators for helping us create the dataset.

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*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Chandrahas, Aditya Sharma, and Partha P. Talukdar. 2018. Towards understanding the geometry of knowledge graph embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 122–131. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/P18-1012

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NeurIPS 2014 Workshop on Deep Learning, December 2014*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Bhuwan Dhingra, Zhong Zhou, Dylan Fitzpatrick, Michael Muehl, and William W. Cohen. 2016. Tweet2vec: Character-based distributed representations for social media. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*. The Association for Computer Linguistics. **DOI:** https://doi.org/10.18653/v1/P16-2044

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*, volume 2.

Eric N. Forsythand and Craig H. Martell. 2007. Lexical and discourse analysis of online chat dialog. In *Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007), September 17-19, 2007, Irvine, California, USA*, pages 19–26. IEEE Computer Society. **DOI:** https://doi.org/10.1109/ICSC.2007.55

Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris

Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. 2015. deltaBLEU: A discriminative metric for generation tasks with intrinsically diverse targets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 445–450, Beijing, China. Association for Computational Linguistics. **DOI:** https://doi.org/10.3115/v1/P15-2073

Xiang Gao, Sungjin Lee, Yizhe Zhang, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2019. Jointly optimizing diversity and relevance in neural response generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1229–1238. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/N19-1125

Sarik Ghazarian, Johnny Wei, Aram Galstyan, and Nanyun Peng. 2019. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 82–89. Association for Computational Linguistics, Minneapolis, Minnesota. **DOI:** https://doi.org/10.18653/v1/W19-2310

Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskénazi, and Jeffrey P. Bigham. 2019. Investigating evaluation of open-domain dialogue systems with human generated multiple references. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, SIGdial 2019, Stockholm, Sweden, September 11-13, 2019*, pages 379–391. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/W19-5944, **PMCID:** PMC6813692

Matthew Henderson, Paweł Budzianowski, Iñigo Casanueva, Sam Coope, Daniela Gerz, Girish Kumar, Nikola Mrkšić, Georgios Spithourakis, Pei-Hao Su, Ivan Vulic, and Tsung-Hsien Wen. 2019. A repository of conversational datasets. In *Proceedings of the Workshop on NLP for Conversational AI*. Data available at github.com/PolyAI-LDN/conversational-datasets. **DOI:** https://doi.org/10.18653/v1/W19-4101

Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017a. Adversarial learning for neural dialogue generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2157–2169, Copenhagen, Denmark. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017b. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 986–995. Asian Federation of Natural Language Processing.

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

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1116–1126. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/P17-1103

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M. Khapra. 2018. Towards exploiting background knowledge for building conversation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2322–2332. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/D18-1255

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics. **DOI:** https://doi.org/10.3115/1073083.1073135

Yookoon Park, Jaemin Cho, and Gunhee Kim. 2018. A hierarchical latent structure for variational conversation modeling. 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 1792–1801, New Orleans, Louisiana. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/N18-1162

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

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of Twitter conversations. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2-4, 2010, Los Angeles, California, USA*, pages 172–180. The Association for Computational Linguistics.

Vasile Rus and Mihai C. Lintean. 2012. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, BEA@NAACL-HLT 2012, June 7, 2012, Montréal, Canada*, pages 157–162. The Association for Computer Linguistics.

Ananya B. Sai, Mithun Das Gupta, Mitesh M. Khapra, and Mukundhan Srinivasan. 2019. Re-evaluating ADEM: A deeper look at scoring dialogue responses. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27-February 1, 2019*, pages 6220–6227. AAAI Press. **DOI:** https://doi.org/10.1609/aaai.v33i01.33016220

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? How controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1702–1723. Association for Computational Linguistics.

Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2018. A survey of available corpora for building data-driven dialogue systems: The journal version. *Dialogue Discourse*, 9(1):1–49.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 3776–3784. AAAI Press. **DOI:** https://doi.org/10.5087/dad.2018.101

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3295–3301. AAAI Press.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *CoRR*, abs/1706.09799.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2019. Machine translation evaluation with BERT regressor. *ArXiv*, abs/1907.12679.

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. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 196–(205. The Association for Computational Linguistics. **DOI:** https://doi.org/10.3115/v1/N15-1020

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. RUBER: an unsupervised method for automatic evaluation of open-domain dialog systems. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 722–729. AAAI Press.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 2214–2218. European Language Resources Association (ELRA).

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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Towards universal paraphrastic sentence embeddings. Yoshua Bengio and Yann LeCun, editors, In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

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, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2204–2213. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/P18-1205

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. BERTScore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. DialoGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, pages 270–278. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/2020.acl-demos.30

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1441–1451. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/P19-1139