Investigating Evaluation of Open-Domain Dialogue Systems With Human Generated Multiple References

Prakhar Gupta[1], Shikib Mehri[1], Tiancheng Zhao[1],
Amy Pavel[2], Maxine Eskenazi[1], and Jeffrey P. Bigham[1,2]

[1]Language Technologies Institute, Carnegie Mellon University,
[2]Human-Computer Interaction Institute, Carnegie Mellon University,
{prakharg,amehri,tianchez,apavel,max+,jbigham}@cs.cmu.edu

Abstract

The aim of this paper is to mitigate the shortcomings of automatic evaluation of open-domain dialog systems through multi-reference evaluation. Existing metrics have been shown to correlate poorly with human judgement, particularly in open-domain dialog. One alternative is to collect human annotations for evaluation, which can be expensive and time consuming. To demonstrate the effectiveness of multi-reference evaluation, we augment the test set of DailyDialog with multiple references. A series of experiments show that the use of multiple references results in improved correlation between several automatic metrics and human judgement for both the quality and the diversity of system output.

1 Introduction

Dialog agents trained end-to-end to hold open-domain conversations have recently progressed rapidly, generating substantial interest (Ghazvininejad et al., 2018; Serban et al., 2017, 2016a; Sordoni et al., 2015; Vinyals and Le, 2015). Development of these systems is driven by available data and benchmarks based on only a single ground truth reference response for a given context. However, such single-reference evaluation does not account for all the plausible responses for any given conversational context (Table 1). This is known as the *one-to-many* response problem (Zhao et al., 2017a). Computing word-overlap metrics against a single-reference response may penalize perfectly valid responses (Deriu et al., 2019) (e.g., "Was anything stolen?", "Is anyone hurt") that deviate from the particular target response ("When was the break-in?"). Unlike human evaluation, automatic evaluation with a single-reference may also disproportionately benefit models that produce generic responses with more probable words (e.g., "I don't know") which is known as the dull-response problem (Li et al., 2016c). As a result, single-reference evaluations correlate weakly with human judgments of quality (Liu et al., 2016).

Dialog Context:
Person A: 911 emergency. What is the problem?
Person B: I would like to report a break-in.
single-reference Response:
When was this break-in?
Other Valid Responses:
Was anything stolen?
Is anyone hurt or injured?
Is the perpetrator still inside the house?
I will send someone right away.

Table 1: Example of a dialog context where appropriate responses do not share words and meaning with a single-reference response.

To address these problems, this paper proposes to carry out automatic evaluation using multiple reference responses instead of a single-reference. Multiple reference evaluation is attractive for several reasons. First, the additional information in the multiple reference response can be used to provide more robust quality evaluation under the one-to-many condition. Second, we can use the multiple references to better measure the diversity of the model, which is a widely studied topic in open-domain response generation (Kulikov et al., 2018; Li et al., 2016a; Zhang et al., 2018; Li et al., 2016b; Zhao et al., 2017a; Gao et al., 2019).

Prior explorations in this area either rely on synthetically created or small scale reference sets (Galley et al., 2015; Qin and Specia, 2015), or perform experiments only on a small set of metrics focused on only response quality (Sugiyama et al., 2019). Our investigations for using multiple references for automatic evaluation covers the

Proceedings of the SIGDial 2019 Conference, pages 379–391
Stockholm, Sweden, 11-13 September 2019.

following aspects - 1) We propose methodology for evaluating both the quality and the diversity of generated responses using multiple references. 2) The proposed evaluation framework is metric-agnostic and the experiments cover a large spectrum of existing metrics, and 3) We augmented the exiting test set of DailyDialog dataset (Li et al., 2017) with multiple references and perform human judgment correlation studies with human-generated references. Our extensive experimental results show that using multiple test references leads to significantly better correlation of automated metrics with human judgment in terms of both response quality and diversity. This suggests that the use of multiple references serves to make automatic metrics more reliable mechanisms for evaluating open-domain dialog systems. Moreover, follow up studies are conducted to better understand the nature of the multi-reference evaluation, such as the number of reference responses needed to achieve high correlation.

The contributions of this paper are:

1. We show that multi-reference evaluation achieves better correlation with human judgments both in quality and in diversity.
2. We analyze the effect of varying the number of reference responses on the correlation with human quality judgements.
3. We construct and release an open-domain multi-reference test dataset[1].

2 Related work

The need for reliable and consistent automatic evaluation methodologies has lead to increasing interest in dialog system evaluation in recent years. In domains such as machine translation and captioning, n-gram overlap metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) and METEOR (Lavie and Agarwal, 2007) correlate well with human judgement. Several embedding-based metrics have been proposed as well, including Greedy Matching (Rus and Lintean, 2012) and Vector Extrema (Forgues et al., 2014). These automatic metrics, however, do not generalize well to open-domain dialog due to the wide spectrum of correct responses, commonly known as the one-to-many problem (Zhao et al., 2017b). Recent work has proposed several trainable evaluation metrics to address this issue. RUBER (Tao et al., 2018) evaluates generated responses based on their similarity with the reference responses and their relatedness to the dialog contexts. Lowe et al. (2017) trained a hierarchical neural network model called ADEM to predict the appropriateness score of responses. However, ADEM requires human quality annotation for training, which is costly. Sai et al. (2019) recently showed that trainable metrics are prone to gamification through adversarial attacks. While past work has focused on inventing new metrics, this paper instead aims to demonstrate that the correlation of existing metrics can be improved through the use of multiple references for evaluation in open-domain settings.

Prior attempts leveraged multiple references to improve evaluation in the context of text generation. Qin and Specia (2015) proposed variants of BLEU for machine translation based on n-gram weighting. In the dialog domain, Galley et al. (2015) proposed Discriminative BLEU, which leverages several synthetically created references obtained with a retrieval model from Twitter corpus. Sordoni et al. (2015) also followed a similar retrieval procedure for multiple-reference evaluation. Since both of them created their reference sets through retrieval followed by a rating step, their multi-reference sets do not reflect the natural variability in responses possible for a context. Sugiyama et al. (2019) proposed a regression-based evaluation metric based on multiple references. The small set of metrics and few test sentences shows promise, but also the need for further exploration. We go further with a comparison of single and multiple references for response quality evaluation and an examination of multiple references for diversity evaluation. This paper is the first, to our knowledge, to create a large test set of several human-generated references for each context. We believe that it is also the first to perform human correlation studies on a variety of automatic metrics for both quality and diversity.

Evaluating diversity in dialog model responses has been studied recently. The most commonly used metric is Distinct (Li et al., 2016a), which calculates the ratios of unique n-grams in generated responses. Distinct is, however, computed across contexts and does not measure if a model can generate multiple valid responses for a context. Xu et al. (2018) proposed Mean Diversity Score (MDS) and Probabilistic Diversity Score (PDS) metrics for diversity evaluation over groups

[1] https://github.com/prakharguptaz/multirefeval

of multiple references over a set of retrieved references. Hashimoto et al. (2019) proposed a metric for a unified evaluation of quality and diversity of outputs, which however depends on human judgements. Zhao et al. (2017a) proposed precision/recall metrics calculated using multiple hypotheses and references as an indicator of appropriateness and coverage. In this paper we leverage their recall-based metrics in our multi-reference based evaluation of diversity.

3 Methodology

We evaluated the performance of dialog response generation models from two aspects: **quality** and **diversity**. Quality tests the appropriateness of the generated response with respect to the context, and diversity tests the semantic diversity of the appropriate responses generated by the model.

We first describe the evaluation procedures used for the conventional single-reference setting. Then we present the proposed multi-reference evaluation. We define a generalized metric to be $\mathrm{d}(y, r)$ which takes a produced output y and a reference output r, and produces a matching score that measure the level of similarity between y and r. We discuss options for d in Table 2.

3.1 Baseline: Single-reference Evaluation

3.1.1 Quality

During single-reference evaluation, there is only one reference response r. As such, for a given metric d, the single-reference score will be $\mathrm{d}(y, r)$.

3.1.2 Unreferenced Diversity

Most prior work concentrates on unreferenced diversity evaluation since referenced diversity evaluation requires a multi-reference dataset. Unreferenced evaluation refers to diversity evaluation methods which ignore the reference responses, and instead compute diversity as a function only of the generated responses. The Distinct (Li et al., 2016a) metric calculates diversity by calculating the number of distinct n-grams in generated responses as a fraction of the total generated tokens. This score is calculated at the system level - over the set of responses generated for all the contexts in test set. Given a set of system responses for the same context, Self-BLEU (Zhu et al., 2018) sequentially treats each one of the generated responses as the hypothesis and the others as references. This score is computed for every context and then averaged over all contexts. A lower Self-BLEU implies greater diversity since system outputs are not similar to one another.

3.2 Proposed: Multi-Reference Evaluation

3.2.1 Quality

In multi-reference evaluation, a given context has multiple valid responses $R = \{r_1, r_2, ..., r_n\}$. As such, for a given metric d, the multi-reference score can be computed as:

$$\mathrm{score}(y, R) = \max_{r \in R} \mathrm{d}(y, r) \quad (1)$$

We score the system output against only the closest reference response because there are multiple diverse and valid responses for a given context.

3.2.2 Referenced Diversity

A multi-reference test set also allows referenced diversity evaluation. For a given context c, we are given multiple reference responses $R = \{r_1, r_2, ..., r_n\}$ and multiple system outputs $Y = \{y_1, y_2, ..., r_m\}$. For a given metric, d, we compute recall (Zhao et al., 2017a), or *coverage*, as follows:

$$\mathrm{recall(c)} = \frac{\sum_{j=1}^{M} \max_{i \in [1,N]} \mathrm{d}\left(y_i, r_j\right))}{M} \quad (2)$$

For each of the multiple reference responses, we consider the highest-scoring system output, then average these scores across the reference responses. A system that generates outputs covering a large portion of the reference responses thus receives a higher recall score.

3.3 Metrics

We consider several metrics for quality and diversity evaluation including (1) word-overlap metrics, and (2) embedding-based metrics. We describe the metrics in Table 2. Each metric represents an instantiation of the generalized scoring function d.

3.4 Compared Models

Our experiments are conducted using four models: a retrieval model and three different generative models. We treat human generated responses as an additional model.

Human: To represent ideal model performance for a particular context, we use a human-generated response for that context.

Dual Encoder: A strong baseline for dialog retrieval is the Dual Encoder (DE) architecture

Metric	Reference	Description
Word-overlap based metrics		
BLEU	Papineni et al. (2002)	BLEU is based on n-gram overlap between the candidate and reference sentences. It includes a brevity penalty to penalize short candidates.
METEOR	Lavie and Agarwal (2007)	The harmonic mean of precision and recall between the candidate and reference based on a set of alignments between the two.
ROUGE-L	Lin (2004)	An F-measure based on the Longest Common Subsequence (LCS) between the candidate and reference utterances.
Embedding based metrics		
Embedding Average	Wieting et al. (2015), others	Computes a sentence-level embedding of r and c by averaging the embeddings of the tokens composing the sentences.
Vector Extrema	Forgues et al. (2014)	Computes a sentence-level embedding by taking the most extreme value of the embeddings of tokens of the sentence for each dimension of the embedding.
Greedy Matching	Rus and Lintean (2012)	Each word in the candidate sentence is greedily matched to a word in the reference sentence based on the cosine similarity of their embeddings. The score is then averaged for each word in the candidate sentence.
Skip-Thought	Kiros et al. (2015)	Uses a recurrent network to encode a given sentence into a sentence level embedding. We use the pre-trained vectors and implementation provided by (Sharma et al., 2017).
GenSen	Subramanian et al. (2018)	Generates a sentence level embedding through a sequence-to-sequence model trained on a variety of supervised and unsupervised objectives in a multi-task framework.

Table 2: Metrics used for both quality and diversity evaluation.

(Lowe et al., 2015a). The model first encodes a given dialog context and response using an LSTM encoder. It then takes the dot-product of the two latent representations to output the likelihood of the response. The Dual Encoder is trained to differentiate between correct responses, and uniformly sampled negative responses. During inference, however, it chooses a correct response for a given context out of all the responses that occur in the training set.

Seq2Seq: Sequence-to-sequence (Seq2Seq) networks (Sutskever et al., 2014) are a typical baseline for dialog systems (Vinyals and Le, 2015). Our model consists of an LSTM encoder, an LSTM decoder and an attention mechanism (Bahdanau et al., 2014).

HRED: Hierarchical Recurrent Encoder Decoder networks (HRED) (Serban et al., 2016b) are a modification of Seq2Seq networks. Rather than encoding the context as a sequence of words, the encoding of the context is done in a two-step process. First, all the utterances of a context are independently encoded by an LSTM utterance encoder. Second, given the latent representations of each utterance, a context encoder encodes the dialog context. The attention mechanism of the decoder attends over the timesteps of context encoder.

CVAE: The Conditional Variational Autoencoder (CVAE) model (Zhao et al., 2017a). CVAE models incorporate discourse-level latent variables in HRED, in which the latent variables represent the discourse-level intentions of the system. Specifically, we reproduce the CVAE network from (Zhao et al., 2017a), where the latent variables follow a multivariate Gaussian distribution with a diagonal covariance matrix. The dimension of the latent variable is 256. To have a fair comparison, the rest of the structure is the same as the HRED with bidirectional LSTM utterance encoders and LSTM context encoder and response decoder. To alleviate the posterior collapse issue for training text CVAEs (Bowman et al., 2016), we use bag-of-words auxiliary loss (Zhao et al., 2017a) and KL-annealing (Bowman et al., 2016).

4 Multi-Reference Data Collection

We used the following procedure to prepare the DailyDialog test set for the multi-reference test set collection. A dialog *D* in the test set consists of utterances $\{u_1, u_1, ..., u_n\}$. Here, u_i denotes the utterance at the ith turn. For generating dialog contexts, we truncate the dialog at each possible utterance, except the last one. The response following each context is treated as the reference response. As an illustration, for the Dialog shown in Table 1, we would generate the following context-reference pairs: *Context 1:* "911 emergency. What is the problem?", *Reference 1:* "I would like to report a break-in.". *Context 2:* "911 emergency

Reference	Very Appropriate	Appropriate	Neutral	Not Appropriate	Not Appropriate at all
From original dataset	41%	54%	2%	3%	0%
Sampled from multi-reference collected	40%	52%	3%	5%	0%

Table 3: Results from dataset quality experiment

... report a break-in.", *Reference 2:* "'When was this break-in?'. In our multi-reference dataset, we expand each single-reference to a set of multiple references.

4.1 Data collection Procedure

We designed an interface for multi-reference data collection using Amazon Mechanical Turk (AMT). For every HIT, we asked an AMT worker to generate 4 diverse follow-up responses for a conversation. A snapshot of the data collection interface is shown in Figure 3 (Appendix). We provided instructions and examples to further clarify the task. To maintain quality post data collection, we filter out responses collected from workers who either generated very short responses or entered the responses in very short amount of time consistently.

4.2 Data Quality

Using the method described above, we collected 4 diverse responses for the 1000 dialogs in the test set, which consists of *6740 contexts*. To validate the quality of the collected dataset, an experiment on AMT is carried out for 100 contexts sampled randomly from the dataset. Workers are shown a dialog context followed by 3 responses shuffled in a random order - 1) the original response from the dataset 2) a random response from the collected multi-references, and 3) a distractor response, irrelevant to the dialog context. We use distractor responses to filter out poor annotations where the annotator gave high ratings to the distractor response. We ask the workers to rate each of the 3 responses for a dialog context on a scale of 1-5 for appropriateness, where 1 indicates *Not Appropriate at all* and 5 indicates *Very Appropriate*. We present the ratings from the experiment in Table 3 for the original responses from the dataset, and the responses from the multi-reference set. We observe that 92% sampled responses from the multi-reference set are marked Appropriate or Very Appropriate. Moreover, only 8% of the responses are marked Not Appropriate or lower, compared to 5% for the original reference set. This indicates that the collected reference set is close to the original reference set in quality. Furthermore, the responses are generated specifically for each context, they are coherent with the context.

5 Experiments

This section describes the experiments we conducted to explore the effectiveness of multi-reference evaluation.

5.1 Correlation Analysis for Quality

This analysis aims to compute the correlation between human quality judgments and two forms of automatic evaluation, both single-reference and multi-reference.

5.1.1 Human Annotations

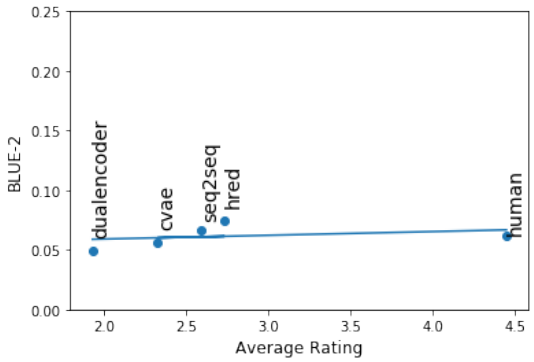
A collection of 100 dialog contexts are randomly selected from the dataset. For a particular dialog context, each of the four models produces a response. In addition, we collect a human response using Amazon Mechanical Turk (AMT), making it total of five responses for each dialog context. Given these context-response pairs, each response is rated in terms of appropriateness (from 1-5) by 5 different AMT workers. The ratings are removed for workers with a Cohen's Kappa κ (Cohen, 1968) inter-annotator agreement score of less than 0.2. The remaining workers had a mean κ score of 0.43, indicating moderate agreement.

5.1.2 Results

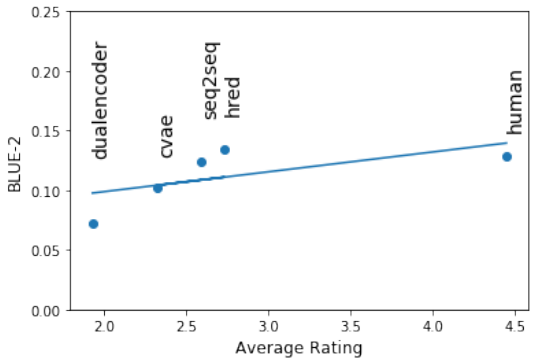
Utterance level correlation: The results of the correlation study conducted for 5 model responses for 100 contexts are shown in Table 4. Pearson correlation is computed to estimate linear correlation, and Spearman correlation to estimate monotonic correlation. The correlations with human quality judgments are computed for both single-reference and multi-reference evaluation. The multi-reference test set consists of both the original reference and the four new collected reference responses. For single-reference evaluation, except for METEOR and Vector Extrema metrics, the correlation is either small or statistically

	Single Reference				Multiple Reference			
Metrics	**Spearman**	**p-value**	**Pearson**	**p-value**	**Spearman**	**p-value**	**Pearson**	**p-value**
BLEU-1	0.0241	0.591	0.1183	0.008	0.1572	0.000	0.2190	0.000
BLEU-2	0.0250	0.577	0.1803	0.000	0.2077	0.000	0.2910	0.000
BLEU-3	0.0608	0.175	0.1269	0.005	0.2520	0.000	0.2086	0.000
BLEU-4	0.0345	0.441	0.1380	0.002	0.2202	0.000	0.2333	0.000
METEOR	0.1064	0.017	0.1871	0.000	0.2247	0.000	0.2855	0.000
ROUGE-L	0.0715	0.110	0.1408	0.002	0.2203	0.000	0.2798	0.000
Embedding Average	0.0301	0.502	-0.0067	0.880	0.1248	0.005	0.0636	0.156
Vector Extrema	0.1919	0.000	0.2114	0.000	0.2785	0.000	0.2946	0.000
Greedy Matching	0.1306	0.003	0.1150	0.010	0.2367	0.000	0.2352	0.000
Skip-Thought	-0.0029	0.949	-0.1463	0.001	0.1049	0.019	-0.0716	0.109
GenSen	0.0731	0.103	0.1110	0.013	0.1832	0.000	0.2389	0.000

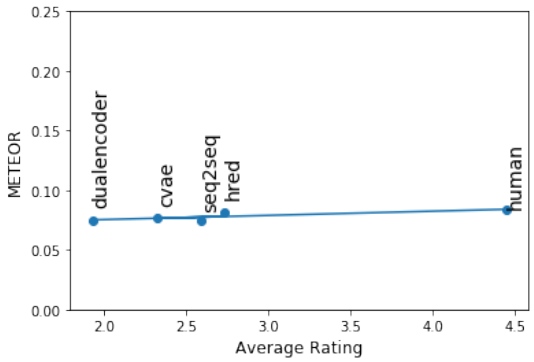
Table 4: Correlation of various metrics when evaluated using single-reference and multi-reference test sets. Evaluation using Multiple References leads to better correlation across all metrics.



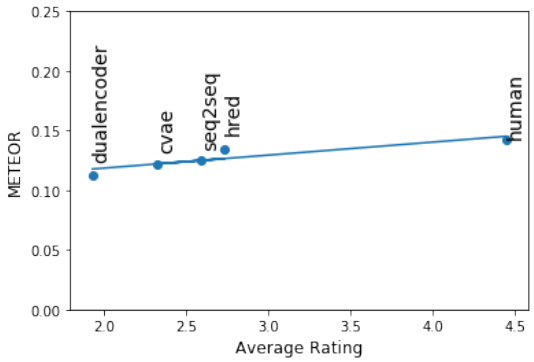
(a) BLEU-2-human ratings - single-references



(b) BLEU-2-human ratings - multiple references



(c) METEOR-human ratings - single-references



(d) METEOR-human ratings - multiple references

Figure 1: System level correlations for BLEU-2 and METEOR metrics. Multi-reference evaluation shows higher correlation with more clear differentiation in model performance.

less significant. On the other hand, every metric shows higher and significant correlation for multi-reference evaluation, with METEOR, ROUGE-L and Vector Extrema achieving the highest correlation values. These results indicate that multi-reference evaluation correlates significantly better with human judgment than single-reference, across all the metrics. This reaffirms the hypothesis that multi-reference evaluation better captures the one-to-many nature of open-domain dialog.

System level correlation: For each model used in the correlation study, the average human rating and average metric scores for 100 contexts are used to calculate system-level correlations. We show system-level correlations for metrics BLEU-2 and METEOR metrics in Figure 1. Each point in the scatter plots represents the average scores for a dialog model. Average human scores are shown on the horizontal axis, with average metric scores on the vertical axis. Humans ratings are low for responses from the retrieval model, and higher for human responses and responses from HRED

model. It is clear that the difference in scores for models when evaluated using single-references is not significant enough to compare the models, as the average metric scores have near zero or very weak correlation with average human ratings. This renders them insufficient for dialog evaluation. However, with multi-reference evaluation, the correlation is higher and significant, which differentiates the models clearly. Thus, multi-reference based evaluation correlates well with humans both at utterance level and at the system level.

5.2 Correlation Analysis for Diversity

This section aims to demonstrate that referenced diversity evaluation methods better correlate with *human judgements of diversity*, than previously used unreferenced diversity metrics. While unreferenced metrics simply reward lexical differences amongst generated outputs, referenced methods (e.g., the recall metric) aims to calculate the *coverage* of the responses. The correlation of human diversity scores is calculated with both unreferenced and referenced measures of diversity.

5.2.1 Human Annotations

Multiple hypotheses were generated from all the models. For CVAE, multiple responses are sampled from the latent space with greedy word-level decoding. For rest of the generation models, five responses were obtained using sampled decoding. For retrieval models, the top five retrieved responses were used. Human annotations of these multiple hypotheses were collected as follows: (1) Workers mark the responses which they find to be appropriate for the conversational context, (2) They then provide a score for the diversity of the responses based on how different they are in *meaning*. This two-stage annotation process captures a desired form of system diversity: generated outputs should be varied, but also appropriate. The scores are averaged across the three workers' annotations. We filtered out ratings from workers with low inter-annotator agreement as described in section 5.1.1. The final mean κ score of 0.41, which indicates moderate agreement.

5.2.2 Results

The results for the diversity correlation analysis are shown in Table 5 for a selected set of metrics[2]. The unreferenced metrics, Distinct and Self-BLEU, correlate poorly with human judgment. This is probably because these metrics evaluate lexical diversity, while humans evaluate diversity of meaning. Furthermore, unreferenced metrics do not consider the reference response and reward diverse outputs without considering appropriateness. With referenced diversity evaluation, using the recall method, BLEU-2 and Vector Extrema show the highest correlation. While metrics like Self-BLEU and Distinct can be "gamed" by producing meaningless albeit very diverse responses, the referenced recall metrics require both appropriate and diverse outputs. As such, referenced evaluation correlates significantly better with human notions of diversity. Thus, the construction of a multi-reference dataset allows for improved diversity metrics.

Metric	Spearman	p-value	Pearson	p-value
Distinct-1	0.0204	0.647	0.0465	0.299
Distinct-2	-0.1282	0.004	-0.0568	0.205
Distinct-3	-0.1316	0.003	-0.0184	0.681
Self BLEU-2	-0.1534	0.001	-0.1251	0.005
Self BLEU-4	-0.0836	0.061	-0.0304	0.497
Recall BLEU-2	0.2052	0.000	0.2469	0.000
Recall BLEU-4	0.1713	0.000	0.1231	0.005
Recall METEOR	0.1993	0.000	0.2165	0.000
Recall ROUGE-L	0.1862	0.000	0.2234	0.000
Recall Vector Extrema	0.2063	0.000	0.2314	0.000
Recall Greedy Matching	0.0797	0.075	0.1204	0.007

Table 5: Correlation scores for diversity metrics

5.3 Automatic Evaluation of Models

We use our multi-reference evaluation methodology to compare the models and the human generated responses on the whole test dataset. For the human model, we use one reference from the multi-reference set as the hypothesis. Human responses are generally more interesting and diverse than model responses, which are known to suffer from the dull response problem (Li et al., 2016c). Because of this reason, we would expect the human generated responses to get higher scores than the dialog models. However, the results presented in Table 6 show that single-reference automatic evaluation ranks few models higher than the hu-

[2]For Self-BLEU we calculate correlation with values substracted from 1 as Self-BLEU is inversely related to diversity

	Single Reference					Multiple reference				
Metric	**Dual Encoder**	**Seq2Seq**	**HRED**	**CVAE**	**Human**	**Dual Encoder**	**Seq2Seq**	**HRED**	**CVAE**	**Human**
BLEU-2	0.0399	0.0521	0.0604	0.0656	0.0513	0.0625	0.0981	0.1061	0.1033	0.1637
BLEU-4	0.0168	0.0252	0.0301	0.0291	0.0245	0.0241	0.0445	0.0497	0.0429	0.0791
METEOR	0.0653	0.0544	0.0607	0.0724	0.0592	0.1000	0.0970	0.1036	0.1120	0.1456
ROUGE-L	0.1522	0.1847	0.1998	0.2088	0.1682	0.2216	0.2927	0.3044	0.2997	0.3502
Vector Extrema	0.4005	0.5124	0.5002	0.4893	0.4823	0.4713	0.6191	0.5975	0.5722	0.6134
Greedy Matching	0.6257	0.7167	0.7104	0.7078	0.6799	0.6991	0.7649	0.7551	0.7457	0.7562
Recall BLEU-2	0.0662	0.0544	0.0766	0.1077	0.0898	0.0436	0.0377	0.0556	0.0679	0.0984
Recall Vector Extrema	0.4945	0.5127	0.5397	0.5586	0.5651	0.4934	0.5334	0.5476	0.5653	0.5881

Table 6: Model evaluation with automatic metrics on Single and Multiple references. Multiple reference evaluation is able to correctly rank human responses higher than model responses.

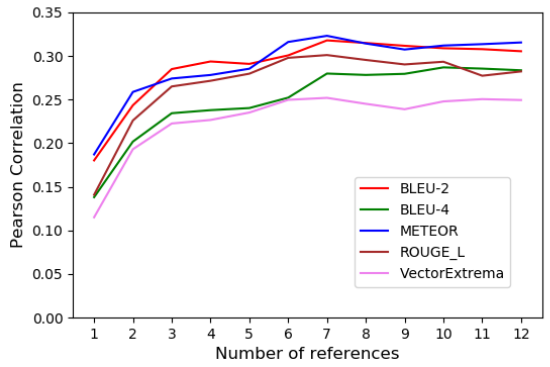


Figure 2: Change in correlation with varying number of references. Trend stablizes after 4-5 references

Dialog Context:
Person A: excuse me . check please .
Generated Response
sure , i 'll grab it and be right with you .
Single-reference Response:
ok , how was everything ?
Multi-reference Responses:
i 'll get it right away .
here is the check .
no problem , let me get your server .
i 'll be right back with it .
Average Human Rating: 5

Metric	Single reference	Multiple reference
BLEU-2	0.0275	0.3257
METEOR	0.0539	0.3425
Vector Extrema	0.5523	0.8680

Table 7: Example of difference in metric scoring for single versus multiple reference evaluation.

mans model. With multi-reference evaluation, human performance is significantly higher than model performance. We further present scores for diversity metrics on multiple hypothesis generated for 100 contexts in the last two rows of the table. The use of multi-reference evaluation covers a wider array of valid responses, which strongly rewards the diverse human responses compared to single-reference evaluation.

5.4 Effect of number of references

The correlation of automated evaluation with human judgment is calculated at various numbers of reference responses. The results shown in Figure 2 demonstrate that the Pearson correlation with human judgment generally increases sharply up to 3-5 references. It further increases slowly up to about 7 references and then seems to plateau at around eight references. This suggests that four to eight references give sufficient coverage of the response space, and collecting additional references does not provide much value in terms of mitigating the issues of the one-to-many problem.

6 Discussion and Conclusion

This work proposes a more reliable methodology for automatic evaluation of open-domain dialogues with the use of multiple references. We augment the test set of DailyDialog dataset with multiple references and show that multiple references lead to better correlation with human judgments of quality and diversity of responses. Single-reference based evaluation can unfairly penalize diverse and interesting responses which are appropriate, but do not match a particular reference in the dataset. However, multiple references can cover the possible semantic space of replies for a context better than a single reference. Thus using multi-reference test sets can improve

the way open-ended dialogue systems are currently evaluated. Our experiments also show that human-generated responses perform worse than models across most metrics when using single-reference evaluation, but multiple reference evaluation consistently ranks human responses higher than model-generated responses. Furthermore, we show how varying the number of references effects human judgement correlation. This methodology could easily be extended to other open domain datasets if the community can make similar multi-reference test sets publicly available.

We illustrate the strength of multi-reference evaluation through scores calculated for some metrics using both single and multiple references for an example context in Table 7. Multiple reference-based evaluation is often good at assigning higher scores when there is more scope for diversity in the responses as illustrated by the example. It should be noted that multiple reference evaluation generally increases the scale of metrics for all responses, and this includes dull responses.

The multi-reference data collection procedure in this paper collects the same number of responses for all contexts. However, different dialogue contexts might possess different levels of "open-endedness". For e.g., a context like "Would you like to dance?" would generally have fewer possible variations in responses than a more open-ended context like "What did you do yesterday?". Therefore, the number of references to collect for a context could be based on the expected variability in responses for the context. Such a procedure would capture more variability over the dataset for a fixed budget.

An important direction in dialog system research is to build models that have more engaging and meaningful conversations with a human. With the recent push towards models which can generate more diverse and interesting responses, appropriate evaluation methodologies are an important and urgent need for the community. Human level evaluation of generation and diversity is challenging to do in a completely automatic way, however, compared to evaluating with a single response, we show that the proposed evaluation methodology is more reliable and will facilitate progress in this direction. In this work we have chose one dataset for extensive experimentation, but in the future studies, it will be worth collecting more datasets and repeating the correlation experiments.

7 Acknowledgements

This work was funded by the Defense Advanced Research Planning Agency (DARPA) under DARPA Grant N6600198-18908, and the National Science Foundation under Award #IIS-1816012. We thank the workers on Amazon Mechanical Turk for making our research possible.

References

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

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

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

Gabriel Forgues, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay. 2014. Bootstrapping dialog systems with word embeddings. In *Nips, modern machine learning and natural language processing workshop*, volume 2.

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.

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. *arXiv preprint arXiv:1902.11205*.

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

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

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.

Ilya Kulikov, Alexander H Miller, Kyunghyun Cho, and Jason Weston. 2018. Importance of a search strategy in neural dialogue modelling. *arXiv preprint arXiv:1811.00907*.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 228–231, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016c. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*.

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

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike 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*, pages 2122–2132, Austin, Texas. 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 (Volume 1: Long Papers)*, pages 1116–1126, Vancouver, Canada. Association for Computational Linguistics.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015a. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *SIGDIAL Conference*, pages 285–294. The Association for Computer Linguistics.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015b. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ying Qin and Lucia Specia. 2015. Truly exploring multiple references for machine translation evaluation. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 113–120, Antalya, Turkey.

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

Vasile Rus and Mihai 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*, pages 157–162. Association for Computational Linguistics.

Ananya B Sai, Mithun Das Gupta, Mitesh M Khapra, and Mukundhan Srinivasan. 2019. Re-evaluating adem: A deeper look at scoring dialogue responses. *arXiv preprint arXiv:1902.08832*.

Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016a. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 3776–3783. AAAI Press.

Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016b. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 3776–3783. AAAI Press.

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

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

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

Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J Pal. 2018. Learning general purpose distributed sentence representations via large scale multi-task learning. In *International Conference on Learning Representations*.

Hiroaki Sugiyama, Toyomi Meguro, and Ryuichiro Higashinaka. 2019. *Automatic Evaluation of Chat-Oriented Dialogue Systems Using Large-Scale Multi-references*, pages 15–25. Springer International Publishing, Cham.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

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

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*.

Zhen Xu, Nan Jiang, Bingquan Liu, Wenge Rong, Bowen Wu, Baoxun Wang, Zhuoran Wang, and Xiaolong Wang. 2018. LSDSCC: a large scale domain-specific conversational corpus for response generation with diversity oriented evaluation metrics. 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 2070–2080, New Orleans, Louisiana. Association for Computational Linguistics.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating informative and diverse conversational responses via adversarial information maximization. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 1815–1825, USA. Curran Associates Inc.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017a. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664, Vancouver, Canada. Association for Computational Linguistics.

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

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 1097–1100, New York, NY, USA. ACM.

A Further Notes on Data Collection Experiments

The interface designed for multi-reference data collection is shown in Figure 3. The final design of the interface incorporates improvements based on multiple rounds of experiments and interviews on a small set of users. The workers were shown a modal box with instructions and several good and bad examples before they start the task. Then they are shown 5 contexts for a HIT, one by one. For each context, they are asked to write 4 diverse responses in the Textbox provided. Workers can enter multi-line responses and submit a response by pressing enter or clicking on a button. They are shown the number of remaining responses they need to enter for the conversation. We also record the timestamps for click and enter presses in the interface. We prevent workers from entering replies shorter than 2 characters, the exact same reply more than 1 time and show them a warning prompt if enter their response too quickly consistently.

Data Collection modes - For the collection of 4 responses per context, we have the following options - A) 4R1W- Collect 4 responses from a single worker B) 2R2W- Collect 2 responses each from 2 separate workers, and C) 1R4W - Collect 1 response each from 4 separate workers. In order to decide between these collection modes, we designed an experiment where, for 100 random contexts, we collected 4 responses using all three styles A), B) and C). In order to decide the best option, we measured lexical diversity across the 4 responses using self-BLEU (Zhu et al., 2018) and Distinct (Li et al., 2016a) metrics, and the collected responses' relevance through the average BLEU score of the multi-reference responses with the ground truth (Gt-BLEU) in the dataset. The results are reported in Table 8.

Metric	4R1W	2R2W	1R4W
SelfBLEU-1	0.3809	0.3662	0.4403
SelfBLEU-2	0.1778	0.1618	0.2657
SelfBLEU-3	0.0955	0.0851	0.2045
SelfBLEU-4	0.0548	0.0449	0.1748
Distinct-1	0.7266	0.7522	0.7082
Distinct-2	0.9240	0.9346	0.8782
Distinct-3	0.9621	0.9692	0.9092
Gt-BLEU-1	0.1213	0.1165	0.1296
Gt-BLEU-2	0.0258	0.0259	0.0352
Gt-BLEU-3	0.0091	0.0111	0.0136
Gt-BLEU-4	0.0033	0.0032	0.0033

Table 8: Diversity and relevance for different modes of data collection.

To calculate Self-BLEU, we calculate the BLEU score for every response by treating the response as a hypothesis and the others as the references, and we define the average BLEU scores calculated this way to be the Self-BLEU of the response set. A higher Self-BLEU score implies less diversity in the set. We observe that 4R1W and 2R2W achieve higher lexical diversity than 1R4W. This is because when a worker is asked to write multiple responses, they can make their responses more diverse conditioned on their previous responses. Relevance metrics Gt-BLEU-1,2,3,4 indicate that 1R4W achieve higher lexical similarity with the ground truth response in the dataset, followed by 4R1W. We chose the 4R1W mode, that is, a collection of 4 responses from 1 worker, to balance the diversity and relevance metrics.

Instructions for annotation collection for Diversity Study

We provided following instructions to the workers for collecting diversity ratings- "Please read the following conversation between two persons. Then read some possible follow-up responses for the conversation. You will be shown 5 sets of responses, with 5 responses in each set. For each response set, first select the responses you think are appropriate responses for the conversation. Then use the sliders to rate the diversity of the response set, that is, how many of the appropriate responses in the response set had different meanings or were different replies. Please provide the diversity score only for the appropriate responses you have marked. The diversity score should not be more than the number of appropriate responses in that set." These instructions were followed by an example to make the task clear.

B Choice of dataset

There are only a few open-domain multi-reference datasets and they have been collected artificially either by retrieval (Xu et al., 2018; Galley et al., 2015) or are very small in scale (Sugiyama et al., 2019). Therefore we augmented the original test set of the DailyDialog dataset (Li et al., 2017), which has a sufficiently large test set. Conversa-

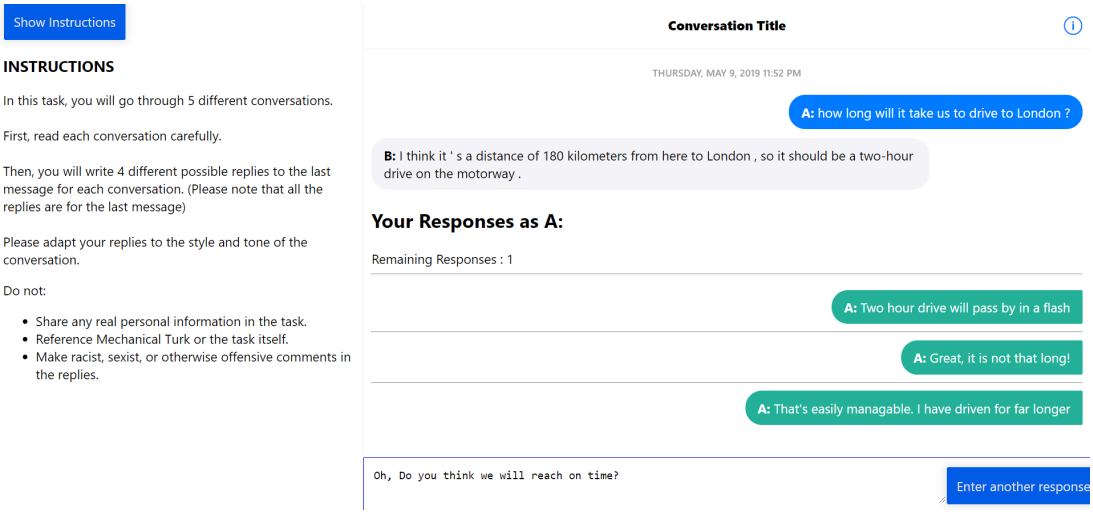


Figure 3: Interface used for multi-reference data collection.

Reference	Original	Multi-reference
Unique 1-gram	17.55	23.62
Unique 2-gram	27.88	58.69
Unique 3-gram	21.79	50.34

Table 9: Comparison of number of unique n-grams in original versus multiple references.

tions in DailyDialog cover 10 different topics on daily life. We chose to augment the DailyDialog dataset due to the following reasons- 1) The dialogs in this dataset are about daily conversation topics and thus it is easier to augment them using crowdsourcing.2) The dialogs in this dataset are generally more formal than datasets such as the Twitter Dialog Corpus (Ritter et al., 2011) and Ubuntu Corpus (Lowe et al., 2015b) which contain noise such as typos and slangs. 3) The dialogs generally have a reasonable number of turns, which makes it easier for a person to understand the context and generate a reply. Therefore, given the size of the original DailyDialog test set and the above-mentioned properties of the dataset, we chose to augment the test set of DailyDialog.

Dataset quality continued

We present the average number of unique 1, 2 and 3 grams in the original ground truth and the set of collected multi-reference ground truth in Table 9. The higher number of unique ngrams in the multi-reference ground truth indicates that the new ground truth captures more variation in the set of possible responses.