Towards Best Experiment Design for Evaluating Dialogue System Output

Sashank Santhanam and Samira Shaikh
Computer Science
University of North Carolina at Charlotte
Charlotte, NC, USA
{ssantha1,sshaikh2}@uncc.edu

Abstract

To overcome the limitations of automated metrics (e.g. BLEU, METEOR) for evaluating dialogue systems, researchers typically use human judgments to provide convergent evidence. While it has been demonstrated that human judgments can suffer from the inconsistency of ratings, extant research has also found that the *design* of the evaluation task affects the consistency and quality of human judgments. We conduct a between-subjects study to understand the impact of four experiment conditions on human ratings of dialogue system output. In addition to discrete and continuous scale ratings, we also experiment with a novel application of Best-Worst scaling to dialogue evaluation. Through our systematic study with 40 crowdsourced workers in each task, we find that using continuous scales achieves more consistent ratings than Likert scale or ranking-based experiment design. Additionally, we find that factors such as time taken to complete the task and no prior experience of participating in similar studies of rating dialogue system output *positively* impact consistency and agreement amongst raters.

1 Introduction and Related Work

A tremendous amount of recent research has focused on approaches towards generating responses for conversations in an open-domain setting (Radford et al., 2019; Xing et al., 2018; Wolf et al., 2019). An equally challenging task for natural language generation systems is evaluating the quality of the generated responses. Evaluation of generated output is typically conducted using a combination of crowdsourced human judgments and automated metrics adopted from machine translation and text summarization (Liu et al., 2016; Novikova et al., 2017). However, studies conducted by Liu *et al.*(2016) and Novikova *et al.* (2017) show that the automated metrics have poor correlation with human judgments. Despite their shortcomings, automated metrics like BLEU, ROUGE, and METEOR are used due to a lack of alternative metrics. This puts a major imperative on obtaining high-quality crowdsourced human judgments. Previous research which employs crowdsourced judgments has focused on metrics including *ease of answering*, *information flow* and *coherence* (Li et al., 2016; Dziri et al., 2018), *naturalness* (Asghar et al., 2018), *interestingness* (Asghar et al., 2017; Santhanam and Shaikh, 2019), *fluency* or *readability* (Zhang et al., 2018), *engagement* (Venkatesh et al., 2018). While experiment designs primarily use Likert scales, Belz and Kow (2010) argue that discrete scales, such as the Likert scales, can be unintuitive and certain individuals may avoid extreme values in their judgments. Prior research has also shown that use of continuous scales is more viable for language evaluation (Novikova et al., 2018; Belz and Kow, 2011). Such evidence places more emphasis on a careful study towards obtaining reliable and consistent human ratings for dialogue evaluation.

To address this research problem, we focus on a systematic comparison of four experimental conditions by incorporating ***continuous, relative*** and ***ranking scales*** for obtaining crowdsourced human judgments. In this initial study, we evaluate the use of two metrics: ***Readability*** and ***Coherence***.

Our key findings are:

1. Use of Likert scales results in the lowest inter-rater consistency and agreement when compared to other experiment conditions
2. Use of continuous scales results in higher inter-rater consistency and agreement
3. Raters who have no prior experience in evaluating dialogue system output have greater inter-rater consistency and agreement than do those who have previously participated in such rating tasks.

Our findings have the potential to help the research community in the design of their evaluation tasks to obtain higher quality human judgments for natural language generation output.

2 Data and Models

We used the Reddit conversation corpus to train our models. The Reddit conversation corpus, made available by Dziri *et al.* (2018), consists of data extracted from 95 top-ranked subreddits that discuss various topics such as sports, news, education and politics. The corpus contains 9M training examples, 500K development dialogues and 400K dialogues as test data.[1] We trained three models on the Reddit conversational dataset described below. All the pre-trained models and supporting analysis code along with user study data are available at https://www.github.com/sashank06/INLG_eval. The models trained for this study include:

- **Seq2Seq:** Simple encoder-decoder model with attention mechanism (Bahdanau et al., 2014)
- **HRED:** ***Hierarchical Encoder-Decoder*** (Serban et al., 2016) which incorporates an utterance and intra-utterance layer to model context.
- **THRED:** ***Topic Augmented Hierarchical Encoder-Decoder*** (Dziri et al., 2018) which uses topic words along with a hierarchical encoder-decoder to produce a response.

3 Metrics

For this initial study, we focus on two metrics, readability and coherence. These metrics are among those essential to evaluate the quality of generated responses (Novikova et al., 2017; Dziri et al., 2019). We describe an automated method to compute each metric.

Readability or Fluency measures the linguistic quality of text and helps quantify the difficulty of understanding the text for a reader (Gatt and Krahmer, 2018; Novikova et al., 2017). We use the Flesch Reading Ease (FRE) (Kincaid et al., 1975) that counts the number of words, syllables and sentences in the text.[2] Higher readability scores indicate that utterance is easier to read and comprehend.

Coherence measures the ability of the dialogue system to produce responses consistent with the topic of conversation (Venkatesh et al., 2018). To calculate coherence, we use the method proposed by Dziri *et al.* (2018). This metric computes the cosine similarity on embedding vectors of generated response and target while accounting for dull and generic responses through a penalty factor.

To overcome the issue of dull and generic responses, Dziri *et al.* (2018) induce a penalty factor which takes into account

$$P = 1 + \log \frac{2 + L'}{2 + L''} \quad (1)$$

where L' indicates the length of response after dropping stop words and punctuation and L'' indicates the length of non-dull parts of the response after dropping stop words. The penalized semantic similarity (SS) score is then calculated as:

$$SS(utt_{i,j}, resp_i) = P \times (1 - cos(utt_{i,j}, resp_i) \quad (2)$$

where i represents the index of the dialogue in the dataset and j denotes index of the utterance in the conversation history.

4 Experiment Designs

In our study, we use three well-known question types of Likert Scale, Magnitude Estimation and Best-Worst Ranking. We chose these questions types to investigate as these are commonly used across various language evaluation tasks (Belz and Kow, 2011; Asghar et al., 2018; Novikova et al., 2018; Kiritchenko and Mohammad, 2017) . With the help of these three types of questions, we design four rating procedures that are explained below.

Likert Scale (LS): is typically used in experiments for crowdsourcing human evaluation of dialogue systems (Asghar et al., 2018; Lowe et al., 2017). In our experiment, we ask the raters to rate the generated responses on a 6-point scale, following Novikova *et al.* (2018) (where 1 is the lowest and 6 is the highest on the metrics of readability and coherence).

Rank-Based Magnitude Estimation (RME): Prior research by Belz and Kow (2011) demonstrates through six separate experiments that continuous scales are more viable and offer distinct advantages over discrete scales in evaluation tasks. Recently, Novikova *et al.* (2018) adopted magnitude estimation by providing the rater with a *standard value* for a reference sentence to evaluate output from goal-oriented systems. Following Novikova *et al.* (2018), we also set the value of the

[1] https://github.com/nouhadziri/THRED
[2] https://bit.ly/1IZ0FG4

standard (reference utterance) as 100 since the reference utterance was produced by humans and is considered as gold-standard. The crowd-sourced workers are asked to provide a score relative to 100 (from 0 to 999) for three system-generated outputs.

Biased Magnitude Estimation (BME): Our third experiment design is biased magnitude estimation (BME). The main difference between RME and BME method is that the standard value we provide for the reference utterance is not uniformly set to 100 for all examples, but instead calculated by automated methods (explained in Section 3). Our motivation to do so is to understand if **anchoring bias** may affect the ratings when judgments are made relative to a fixed value (100) or relative to a value calculated by automated means. Anchoring bias is the tendency to rely too heavily on one piece of information offered (the "anchor", in this case, the number 100) when making decisions (Kahneman, 2016).

Best-Worst Scaling (BWS): Our last experiment condition is best-worst scaling (BWS) in which raters are asked to rank the generated responses in order of best to worst on both metrics (readability and coherence). This approach has previously been used to estimate emotion intensity and has been demonstrated to produce high quality and consistent judgments from humans (Kiritchenko and Mohammad, 2017).

Each task includes 50 randomly sampled conversations from the test set in our corpus along with generated responses from the three models and the ground truth (reference utterance). For each task, we collected ratings from 40 workers with Master qualifications through Amazon Mechanical Turk.

5 Experiment Results

We organize our findings along five main research questions (RQs) outlined in this section. In the following section, we report on statistical significance using two-way ANOVAs on the between-subject ratings across the four experiment conditions (Tables 1– 7).

RQ1: What is the effect of experiment design on the reliability on human ratings? We use intra-class correlation (ICC) to measure the reliability across multiple raters (Shrout and Fleiss, 1979; Landis and Koch, 1977). To compare the scores obtained from magnitude estimation experiments to the ratings from the task using discrete Likert scales, we perform a normalization of the magnitude estimation scores on a logarithmic scale as suggested by Bard *et al.* (1996).

		Likert	RME	BME	BWS
ICC-C	Readability	0.75	0.95†	0.83	0.75
	Coherence	0.83	0.92	0.81	0.80
ICC-A	Readability	0.59	0.95†	0.83	0.75
	Coherence	0.77	0.92	0.81	0.80

Table 1: ICC scores on the metrics of readability and coherence for each experiment design. All values are statistically significant p-value<0.001 except those indicated by †. n=40 for all four designs.

Table 1 represents the ICC scores on consistency (ICC-C) and agreement (ICC-A) for our four experiment tasks. We observe that use of Magnitude Estimation with anchors (RME or BME) results in more reliable ratings than using Likert Scale or using Best-Worst ranking (BWS). This result is consistent with prior research by Novikova *et al.* (2018) and Belz and Kow (2011).

RQ2: Does time taken to complete the survey influence reliability of the rankings? To analyze RQ2, we calculated the total time spent by each participant from the start to the end of the experiment. We found that BME task had longest on average time to completion (43 minutes), followed by RME (42.8 minutes) and Likert scale (33 minutes; Best-Worst ranking had shortest average completion time (32.5 minutes). We then test the hypothesis that raters who spent longer than average time on the task would be more reliable in their ratings than those who completed in less than average time. Table 2 represents the ICC scores for raters who spent higher than average time for the task, while Table 3 represents scores for raters who spent less than average time. Surprisingly, we find that consistency and agreement among raters who spend less than average time is higher than those who spend more time, for the Likert, BME or BWS experiment designs. When using the RME design, raters who spend more time have higher consistency and agreement.

RQ3: Does prior experience of evaluating dialogue system output or engaging with conversational agents affect reliability of rankings? We asked each rater two additional questions at the end of the task. The questions asked raters to indicate whether or not they had prior experi-

		Likert (n=15)	RME (n=16)	BME (n=15)	BWS (n=16)
ICC-C	Readability	0.58	0.93	0.51	0.62
	Coherence	0.74	0.85	0.55	0.64
ICC-A	Readability	0.52	0.93	0.51	0.62
	Coherence	0.69	0.86	0.56	0.64

Table 2: ICC scores when participants spend **above average time**. All values in this table are statistically significant with p-value<0.001

		Likert (n=25)	RME (n=24)	BME (n=25)	BWS (n=24)
ICC-C	Readability	0.61	0.88	0.81	0.65
	Coherence	0.66	0.85	0.75	0.76
ICC-A	Readability	0.36	0.88	0.81	0.66
	Coherence	0.55	0.85	0.75	0.76

Table 3: ICC scores when participants spend **below average time**. All values in this table are statistically significant with p-value<0.001

ence taking part in studies (a) to evaluate dialogue system output; and (b) to engage with a conversational agent.

Tables 4 and 5 show how reliable the ratings from the participants based on their prior experience of taking part in studies about evaluating conversational response. We find that participants who have not taken part in prior studies are more consistent and have a higher agreement score than participant who have prior experience. These results are also validated by Tables 6 and 7 which shows that participants with no prior experience of engaging with conversational agents are more consistent and reliable.

		Likert (n=15)	RME (n=7)	BME (n=18)	BWS (n=13)
ICC-C	Readability	0.45	0.37	0.51	0.54
	Coherence	0.38	0.48	0.55	0.63
ICC-A	Readability	0.35	0.38	0.52	0.55
	Coherence	0.32	0.49	0.55	0.63

Table 4: ICC scores when participants **have** prior experience evaluating dialogue system output. All values statistically significant at p-value<0.001.

RQ4: How well do automated methods to calculate readability and coherence correlate with human ratings? We report on correlation between readability and coherence scores that are

		Likert (n=25)	RME (n=33)	BME (n=22)	BWS (n=27)
ICC-C	Readability	0.71	0.95†	0.83	0.70
	Coherence	0.82	0.92	0.76	0.72
ICC-A	Readability	0.50	0.95†	0.83	0.70
	Coherence	0.75	0.92	0.77	0.72

Table 5: ICC scores when participants **do not have** prior experience evaluating dialogue system output. All values statistically significant at p-value<0.001 except those indicated by †.

		Likert (n=18)	RME (n=11)	BME (n=23)	BWS (n=18)
ICC-C	Readability	0.46	0.69	0.60	0.57
	Coherence	0.44	0.65	0.62	0.67
ICC-A	Readability	0.37	0.69	0.61	0.57
	Coherence	0.38	0.65	0.62	0.67

Table 6: ICC scores when participants **have** prior experience engaging with conversational agents. All values statistically significant at p-value<0.001.

		Likert (n=22)	RME (n=29)	BME (n=17)	BWS (n=22)
ICC-C	Readability	0.70	0.95†	0.84	0.67
	Coherence	0.82	0.91	0.76	0.68
ICC-A	Readability	0.48	0.95†	0.84	0.67
	Coherence	0.75	0.91	0.76	0.68

Table 7: ICC scores when participants **do not have** prior experience engaging with conversational agents. All values statistically significant at p-value<0.001 except those indicated by †.

calculated using automated methods (outlined in Section 3) with the human ratings in Table 8. Readability scores were computed using the Flesh Reading Ease (Kincaid et al., 1975) and coherence scores were computed based on method proposed by Dziri *et al.* (2018). We observe that the automated metrics for Readability (Kincaid et al., 1975) and Semantic Similarity (Dziri et al., 2018) show low correlation to human judgments ratings.

	Likert	RME	BME	BWS
	Automated Metric			
Readability	0.26	-0.11	-0.12	-0.06
Coherence	-0.12	-0.13	-0.11	0.01

Table 8: Spearman correlation between the ratings obtained from the automated metrics to human ratings.

RQ5: Is there any correlation between ratings of readability and coherence for each of the four experiment conditions? To evaluate whether there is any correlation between the ratings obtained for readability and coherence through of four experimental designs, we report the Spearman correlation values in Table 9. We find that there is high correlation between the human ratings of readability and coherence obtained through RME and BME (statistically significant). One likely factor affecting correlation may be anchoring bias towards the fixed value of the standard utterance provided in RME (100) and reference value provided in BME. We aim to investigate this further in future work.

	Likert	**RME**	**BME**	**BWS**
		Readability		
Coherence	0.1	0.79***	0.77***	0.5***

Table 9: Spearman correlation between the ratings of readability and coherence obtained on four different experiment designs. *** p-value$<$0.001

6 Conclusion and Future Work

In this paper, we present our work on designing a systematic experiment with four experiment conditions to evaluate the output of dialogue systems. Different from prior work where a similar study was conducted with output from goal-oriented systems (Novikova et al., 2018), our study focuses on evaluating output in open-domain situations. Consistent with prior findings, metrics calculated using automated methods (Dziri et al., 2019) were found to have a negative correlation with human judgments (c.f. Table 8). This finding points to the need for more effective automated metrics.

We find that that use of continuous scales to obtain crowdsourced ratings provides more consistent and reliable ratings than ratings obtained through Likert scales or Best-Worst scaling. This finding is consistent with prior work conducted by Novikova *et al.* (2018). Novel in our study was the testing of the Best-Worst scaling method to evaluate responses against one another. Although the Best-Worst scaling method has been shown to be effective in obtaining crowdsourced ratings of emotions (Kiritchenko and Mohammad, 2017), we did not find it to be effective in this study. We aim to investigate further whether this finding can be reproduced in a different experiment.

Further, we were able to identify the effects of time taken to complete the task on rating reliability. We find that workers who spent less than average time on the task had higher consistency (for the Likert, BME and BWS experiment conditions) than did the workers who spent more than average time. This finding is counter-intuitive, we expect that spending more time would positively impact inter-rater consistency. Our first step in the analysis of the effects of time taken on reliability included analyzing data from workers who spent more or less than average time, which offers admittedly a limited perspective; an interesting next step would be to more thoroughly study the effects of time taken on reliability by taking into account the full distribution of the time spent data.

We also find that *lack of* prior experience of evaluating open-domain dialogue system output results in more reliable ratings. One potential explanation for this could be that workers may have pre-conceived notions based on their past experience. One limitation of our current study is that although we had output from three separate models, we conducted the study using data from one corpus. Reproducing our findings across additional corpora, additional metrics and other experiment designs would help substantiate these findings further. An analysis of the interaction effects between independent variables such as time taken and prior experience would also help strengthen the findings of our study.

By using a larger sample size (n=40), we are able to make claims about statistical significance across experiment conditions. In future work, we plan to evaluate the impact of cognitive biases such as anchoring and confirmation bias in-depth and how it affects consistency and reliability along with testing continuous scale ratings with no reference value.

Acknowledgments

This work was supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No FA8650-18-C-7881. All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of AFRL, DARPA, or the U.S. Government. We thank the anonymous reviewers for the helpful feedback.

References

Nabiha Asghar, Pascal Poupart, Jesse Hoey, Xin Jiang, and Lili Mou. 2018. Affective neural response generation. In *European Conference on Information Retrieval*, pages 154–166. Springer.

Nabiha Asghar, Pascal Poupart, Xin Jiang, and Hang Li. 2017. Deep active learning for dialogue generation. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 78–83, Vancouver, Canada. Association for Computational Linguistics.

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

Ellen Gurman Bard, Dan Robertson, and Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language*, pages 32–68.

Anja Belz and Eric Kow. 2010. Comparing rating scales and preference judgements in language evaluation. In *Proceedings of the 6th International Natural Language Generation Conference*, pages 7–15. Association for Computational Linguistics.

Anja Belz and Eric Kow. 2011. Discrete vs. continuous rating scales for language evaluation in nlp. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 230–235. Association for Computational Linguistics.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019. Evaluating coherence in dialogue systems using entailment. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3806–3812, Minneapolis, Minnesota. Association for Computational Linguistics.

Nouha Dziri, Ehsan Kamalloo, Kory W Mathewson, and Osmar Zaiane. 2018. Augmenting neural response generation with context-aware topical attention. *arXiv preprint arXiv:1811.01063*.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Daniel Kahneman. 2016. 36 heuristics and biases. *Scientists Making a Difference: One Hundred Eminent Behavioral and Brain Scientists Talk about Their Most Important Contributions*, page 171.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202. 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. 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.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. RankME: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 72–78, New Orleans, Louisiana. Association for Computational Linguistics.

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

Sashank Santhanam and Samira Shaikh. 2019. A survey of natural language generation techniques with a focus on dialogue systems-past, present and future directions. *arXiv preprint arXiv:1906.00500*.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building

end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420.

Anu Venkatesh, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Behnam Hedayatnia, Angeliki Metallinou, et al. 2018. On evaluating and comparing conversational agents. *arXiv preprint arXiv:1801.03625*.

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

Chen Xing, Yu Wu, Wei Wu, Yalou Huang, and Ming Zhou. 2018. Hierarchical recurrent attention network for response generation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

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