Understanding the Impact of 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

Evaluation of output from natural language generation (NLG) systems is typically conducted via crowdsourced human judgments. To understand the impact of how experiment design might affect the quality and consistency of such human judgments, we designed a between-subjects study with four experiment conditions. 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 no prior experience of participating in similar studies of rating dialogue system output *positively* impact consistency and agreement amongst raters.[1]

1 Introduction

There is a major imperative on obtaining high-quality crowdsourced human judgments of NLG output, since these are the key evidence that certain models perform better than others. Experiment designs to obtain such human judgments primarily use Likert scales. Belz and Kow (2010) argue that discrete scales, such as Likert scales, can be unintuitive and people may avoid extreme values in their judgments. We focus on a systematic comparison of four experimental conditions by incorporating ***continuous, relative*** and ***ranking scales*** for obtaining crowdsourced human judgments. Our key findings are:

1. Use of continuous scales results in higher inter-rater consistency and agreement
2. 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.

2 Data and Models

We used the Reddit Conversational Corpus made available by Dziri *et al.* (2018) to train our models. The corpus contains 9M training examples, 500K development dialogues and 400K dialogues as test data. 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 Experiment Design

We ask human raters to evaluate which model produces the better output, on the basis of two metrics: ***Readability***: which "measures the linguistic quality of text and helps quantify the difficulty of understanding the text for a reader" (Gatt and Krahmer, 2018) and ***Coherence***: "ability of the dialogue system to produce responses consistent with the topic of conversation (Venkatesh et al., 2018)". We constructed three distinct surveys (i.e. experiment conditions), each of which used one of the well-known question types of Likert Scale, Magnitude Estimation and Best-Worst Ranking. Our experiment conditions are:

[1] This work has already been published at INLG 2019, Tokyo, Japan

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 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. 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.

4 Results

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). 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). Table 1 represents the ICC scores on consistency (ICC-C). 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).

		Likert	**RME**	**BME**	**BWS**
ICC-C	Readability	0.75	0.95†	0.83	0.75
	Coherence	0.83	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.

		Likert	**RME**	**BME**	**BWS**
ICC-C Prior Exp	Readability	0.45	0.37	0.51	0.54
	Coherence	0.38	0.48	0.55	0.63
ICC-C No Prior Exp	Readability	0.71	0.95†	0.83	0.70
	Coherence	0.82	0.92	0.76	0.72

Table 2: ICC scores when participants prior experience evaluating dialogue system output. Top half represents participants with prior experience and bottom half with no prior experience. All values statistically significant at p-value$<$0.001 except those indicated by †.

RQ2: Does prior experience of evaluating dialogue system output affect reliability of rankings?

We asked each rater additional questions at the end of the task. The questions asked raters to indicate whether or not they had prior experience taking part in studies involving evaluation dialogue system output. Table 2 shows 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.

5 Conclusion

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. 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. 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 preconceived notions based on their past experience. Our findings have implications on how to best design the survey to obtain human judgments of NLG output.

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.

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 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.

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, July. 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, July. 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, June. Association for Computational Linguistics.

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*.