Studying the Effects of Cognitive Biases in Evaluation of Conversational Agents

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

Abstract

We conducted a between-subjects study with 77 crowdsourced workers to understand the role of cognitive biases, specifically anchoring bias, when humans are asked to evaluate the output of conversational agents. We find increased consistency in ratings across two experimental conditions may be a result of anchoring bias. We also determine that external factors such as time in similar tasks have effects on inter-rater consistency. Our results provide insight into how best to evaluate conversational agents.[1]

1 Introduction

Novikova *et al.* (2018) have shown that continuous scales help improve the consistency and reliability of human ratings across several language evaluation tasks as opposed to Likert scales. In their experiments, Novikova *et al.* (2018) found that consistency of crowd-sourced workers improved when workers were asked to rate the conversational agent output by comparing it against a given (gold) standard. But what if this increased consistency is a result of the very presence of the predetermined gold standard, possibly because humans evaluators are anchored on that standard value of 100? *Anchoring bias*, which is the tendency of people to focus on the first piece of information presented; also defined as "*inability of the people to make sufficient adjustments starting from the initial value (anchor) to yield the final answer*" (Kahneman, 2003). To investigate the effects of cognitive biases, specifically anchoring bias, on decision-making around evaluating chatbot output, we designed a $2X2$ experiment with 77 crowdsourced workers.

We find systematic effects of anchoring in the **magnitude** of participants' ratings: participants who are presented with an anchor will provide a rating that is closer to the anchor value than those who are not presented with an anchor.

[1]This paper has been accepted at CHI 2020, Hawaii, USA

2 Data and Models

We used the Reddit Conversational Corpus to train our models made available by Dziri *et al.* (2018). The corpus contains 9M training examples, 500K development dialogues and 400K dialogues as test data. We trained three models: 1. Seq2seq, 2. Hierarchical Encoder Decoder (HRED) 3. Topic Augmented Hierarchical Encoder-Decoder.

3 Experiment Design

We built an interface to allow participants to evaluate the generated responses. We initially focus on 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 use magnitude estimation (ME) questions to obtain ratings from crowdsourced workers. We design four experiment conditions, namely **Anchor**: With or Without Anchor and **Presentation Order**: Both Questions (Readability and Coherence together) or Single Question (Readability and Coherence on separate screens). We had 40 participants in Setup 1 (Both Questions) and 22 were in the anchoring condition (anchor value$=100$) and 18 in no anchor condition. We had 37 participants in setup 2 (one question per screen) and 18 in no anchor condition and 19 in anchoring condition.

4 Results

RQ1:What is the effect of anchors and type of setup on the magnitude of ratings? Figure 1 presents ratings for the metrics of readability and

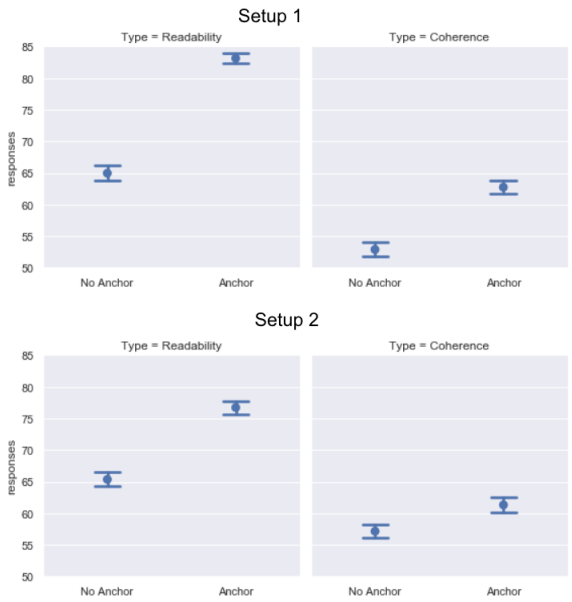

Figure 1: Mean of the responses bootstrapped with 95% confidence intervals across Setups 1 and 2 on the metrics of Readability and Coherence.

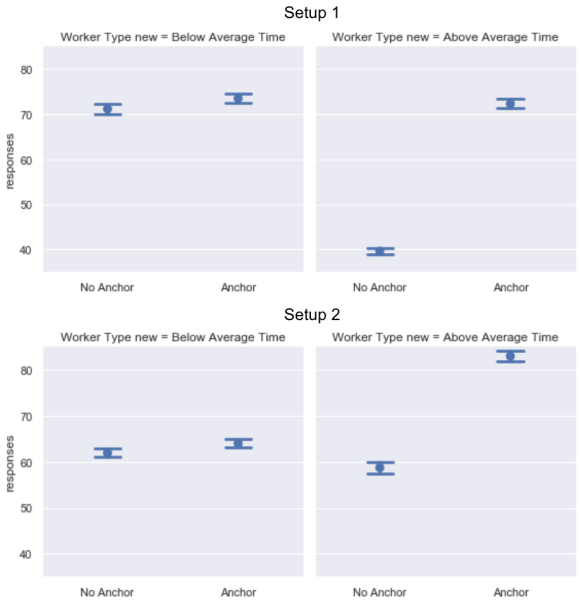

Figure 2: Mean of the responses bootstrapped with 95% confidence intervals across Setups 1 and 2 based on amount of time spent on study

coherence separately. We find that across both setups, the difference between anchor and no anchor conditions to be larger for the metrics of readability than coherence (statistically significant with $p<0.001$). We find that in Setup 1, readability values have a mean of 83.13 in the anchor condition and in no anchor condition the mean of the responses drop down to 64.97. Also in Setup 1, we find that for coherence metric, the mean of responses in the anchoring condition is M=62.74 and without anchor M=52.89. We find similar trends in the responses provided in Setup 2 for both metrics of readability and coherence.

RQ2:What is the effect of time taken to complete the task on the magnitude of the ratings? We find that participants who are presented with anchors spend *more time* on average taking the study than participants in no anchor conditions across both setups. From the total of 77 participants, the mean time taken to complete the study was 57.17 minutes. In Setup 1, we find that participants took an average of 66 minutes in the with anchor condition and average of 54.83 minutes in the without anchor conditions. Similarly, in Setup 2 we find participants took an average of 54.94 minutes with anchor condition and 50.94 with no anchor condition. We grouped the participants based on the amount of time spent into two categories: (1) **Below Average** - when participants spend less than mean time; (2) **Above Average** - when participants spend more than mean time. Across both setups, we find that people in the above average group show significant differences in their responses. In Setup 1, in the above average group, the mean of responses in no anchor condition was 39.65 and mean of the responses in anchor condition was 72.35. We find similar evidence in Setup 2 with people in anchor condition provide higher values (83) close to the numerical anchor (100).

5 Conclusion

Our findings are a step towards understanding the impact of experiment design and the possible role of cognitive bias such as anchoring bias towards dialogue evaluation. We find the effect of anchoring is more pronounced in instances when participants are asked to provide ratings on two metrics at the same time (Both Questions/Setup 1) and the effect of anchoring is slightly less pronounced when participants are asked to provide ratings for a single metric on a single screen (Single Question/Setup 2).

References

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. 2003. A perspective on judgment and choice: mapping bounded rationality. *American psychologist*, 58(9):697.

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.

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*.