Investigating Crowdsourcing Protocols for Evaluating Factual Consistency of Summarization

Xiangru Tang[1], Alexander R. Fabbri[1], Ziming Mao[1], Griffin Adams[3], Haoran Li[2], Yashar Mehdad[2], Dragomir Radev[1]

Yale University[1] Facebook AI[2] Columbia University[3]

LILY Lab

Introduction

Large language models can create fluent summaries with high coverage, yet it is now well-documented that these summaries often suffer from factual inconsistencies. Though many automatic metrics have been proposed to characterize the level of hallucinations and factual errors, manual evaluation is recognized to be needed. Manual evaluation is essential to judge progress, yet not as much work has been done to address the ideal human evaluation setup. In this paper, we crowdsource annotations for factual consistency across Likert and Ranking-based evaluation scales across the CNN-Daily Mail and XSum datasets. We find that the BWS protocol produces more reliable annotations in general. But the Likert scale is not bad to be used for factuality evaluation in some cases, contrary to the argument in prior work. Based on our quantitative experiments, we also present the prerequisites of using Likert.

Study Design

Previous work has analyzed the biggest limitations of modern pre-trained models for abstractive document summarization and found that these models are highly prone to hallucinate content that is unfaithful to the input document. Thus, we conduct studies on the faithfulness task to elicit summary judgments for analysis. In the task, we ask annotators to judge the factual consistency of the summaries. We organize our findings along three main research questions (RQ) outlined in this section.

- RQ1: Ranking (BWS) vs. Likert on factual consistency.
- RQ2: Considering different models and datasets, how is the reliability of scale methods?
- RQ3: How serious is the four limitations we defined of these scale methods?
  1. Inconsistencies in annotations by the same annotator: an annotator might assign different scores to the same item when the annotations are spread over time.
  2. Inconsistencies in annotations by different annotators: one annotator might assign a score of 7 to the word good on a 1-to-9 sentiment scale, while another annotator can assign a score of 8 to the same word.

Table 1. Overall results of baselines across the tasks of XSum.

Model	Rouge-1	Rouge-2	Rouge-L
PEGASUS	46.84[1]	24.52[1]	39.10[1]
ProphetNet	43.23[3]	19.96[3]	35.16[3]
BART	44.15[2]	21.28[2]	35.94[2]
BERTSUM	38.21[4]	16.11[4]	30.83[4]

Table 2. Overall results of baselines across the tasks of CNN/DM.

Model	Rouge-1	Rouge-2	Rouge-L
PEGASUS	44.19[1]	21.45[1]	41.08[1]
ProphetNet	42.45[3]	19.90[3]	39.31[3]
BART	44.07[2]	21.13[2]	40.89[2]
BERTSUM	41.82[4]	19.39[4]	38.67[4]

Table 3. The average Likert scores and the average rankfor all systems on XSum and CNN/DM.

Models	XSum			CNN/DN	
	BWS	Likert	Likert-10	BWS	Likert
PEGASUS	3.247[3]	3.350[1]	6.247[2]	3.230[2]	3.887[2]
ProphetNet	3.360[2]	3.293[3]	6.427[2]	3.100[3]	3.860[4]
BART	3.570[1]	3.433[2]	6.937[1]	3.593[1]	4.017[1]
BERTSUM	2.827[4]	2.790[4]	5.163[4]	3.087[4]	3.863[3]

Table 4. Instance-level reliability computed by Krip-pendorff 's alpha (α) on the CNN/DM and XSum.

krippendorff	XSum	CNN/DM
BWS	0.2477	**0.1582**
Likert	**0.2951**	0.0443

Table 5. System-level split-half reliability computedby Split-Half Reliability (SHR) on the CNN/DM andXSum.

SHR	XSum	CNN/DM
BWS	0.90314	**0.87657**
Likert	**0.92777**	0.45619

Figure 1. Score distribution of Likert for faithfulness. Each data point shows the number of times a particular score was assigned to each system.

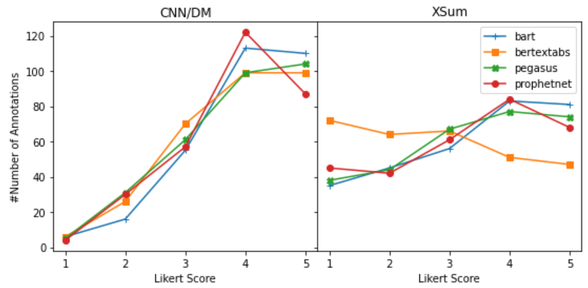


Figure 2. Score distribution of Likert scale 1-10 for faithfulness. Each data point shows the number of times a particular score was assigned to each system.

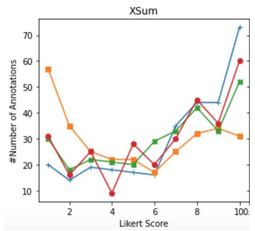
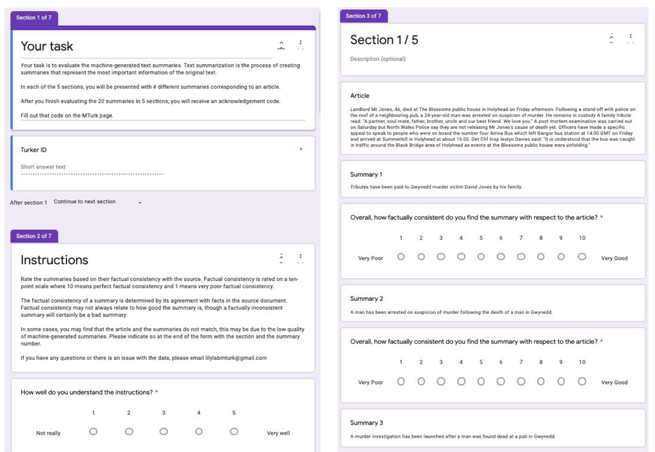


Figure 3. Mturk interface.



3. Scale region bias: annotators often have a bias towards a part of the scale, for example, preference for the middle of the scale.
4. Fixed granularity: in some cases, annotators might feel too restricted with a given rating scale and may want to place an item inbetween the two points on the scale. On the other hand, a fine-grained scale may overwhelm the respondents and lead to even more inconsistencies in annotation.

Results

We use both Likert and BWS on the output of four recent summarizers. The majority of evaluations is conducted using judgements, with the second most frequent method being rank-based annotations. Best-worst scaling (BWS) is a specific type of ranking-oriented evaluation that requires annotators to specify only the first and last rank, which is claimed to be reliable. We are using BART, ProphetNet, Pegasus, and BERTSUM models on XSUM and CNN/DM. Each study contains 100 original passages, and 400 summaries generated from four pre-trained models. We randomly sample 100 documents from the corpus with corresponding summaries from models to form the item set for all our studies. We demand 3 judgments per summary to mirror a common setup in manual evaluation, while each annotator only can annotate 5 summaries. We separated our corpus into 20 blocks of 5 documents and included all 4 generated summaries for each document in the same block, which results in 5 × 4 = 20 summaries per block.

Conclusion

- Regrading RQ1, on the general used CNN/DM, we show that ranking-style evaluations (BWS) are more reliable and cost-efficient than Likert scale.
- Regrading RQ2, We found different phenomena happened on different datasets XSum, while BWS and Likert are comparable, Likert even is a little better. Because in CNN/DM, it's easy to distinguish which one is better (we see many 4 and 5 in Figure 1), which benefits BWS. But in XSum, swinging between close scores increases the noise and bias.

Acknowledgement

Thanks to Drago and my collaborators for this project. Thanks, Alex, Ziming. This couldn't happen without their insights, efforts, and leadership.