Abstract

Current pre-trained models applied for summarization are prone to factual inconsistencies that misrepresent the source text. Evaluating the factual consistency of summaries is thus necessary to develop better models. However, the human evaluation setup for evaluating factual consistency has not been standardized. To determine the factors that affect the reliability of the human evaluation, we crowdsource evaluations for factual consistency across state-of-the-art models on two news summarization datasets using the rating-based Likert Scale and ranking-based Best-Worst Scaling. Our analysis reveals that the ranking-based Best-Worst Scaling offers a more reliable measure of summary quality across datasets and that the reliability of Likert ratings highly depends on the target dataset and the evaluation design. To improve crowdsourcing reliability, we extend the scale of the Likert rating and present a scoring algorithm for Best-Worst Scaling that we call value learning. Our crowdsourcing guidelines will be publicly available to facilitate future work on factual consistency in summarization.

1 Introduction

Pre-trained language models have achieved promising results in abstractive text summarization (Edunov et al., 2019; Dong et al., 2019; Song et al., 2019; Zhang et al., 2019, 2020). A serious limitation of these models, however, is their tendency to produce text that is factually inconsistent with the input. Thus, evaluating the factual consistency of the generated summaries with respect to the source is an important task (Falke et al., 2019; Cao et al., 2020; Gabriel et al., 2021; Durmus et al., 2020; Huang et al., 2021; Pagnoni et al., 2021).

Recently, metrics have been proposed for evaluating factual consistency, including applying natural language inference (Falke et al., 2019; Kryscinski et al., 2020) and question-answering models (Eyal et al., 2019; Scialom et al., 2019; Durmus et al., 2020; Wang et al., 2020). However, current metrics still do not correlate highly with human judgments on factual consistency (Koto et al., 2020; Pagnoni et al., 2021). To overcome the inherent limitation of automatic metrics, researchers typically crowdsource human evaluations using platforms such as Amazon’s Mechanical Turk (MTurk) (Gillick and Liu, 2010; Sabou et al., 2012; Lloret et al., 2013). However, papers often differ in their preferred evaluation protocols (Louis and Nenkova, 2013; Hardy et al., 2019). These differences in the evaluation task design affect the quality of the resulting human judgments and system comparisons (Santhanam and Shaikh, 2019).

Two of the primary paradigms of crowdsourced evaluations are ranking-based and rating-based. Best-Worst Scaling (Louviere and Woodworth, 1991) is a ranking-based method by which the annotator selects the best and worst example out of a set of examples. Prior research has claimed that Best-Worst Scaling produces higher-quality evaluations than rating scales such as the Likert Scale for tasks such as sentiment analysis (Kiritchenko and Mohammad, 2017). In the context of summarization, Steen and Markert (2021) find that, compared to the Likert Scale, ranking-based protocols are more reliable for measuring summary coherence but less so for repetition. However, previous studies have not analyzed annotation reliability in the context of factual consistency for summarization.

Our contributions are the following: 1) We are, to the best of our knowledge, the first to study the reliability of human evaluation for summarization factual consistency. 2) We study rating and ranking-based protocols across two summarization datasets and four state-of-the-art abstractive models. We determine the factors affecting human evaluation reliability and present a novel ranking-based protocol with the highest reliability. 3) We will release our evaluation guidelines and annotations to promote future work on factual consistency evaluation.
2 Study Design

Each study consists of 100 input documents randomly sampled from each dataset, and four associated model-generated summaries.

2.1 Datasets and Models

Datasets: We conduct our study on two benchmark summarization datasets. CNN/DailyMail (Hermann et al., 2015; Nallapati et al., 2016) consists of 311,672 pairs of online articles and bullet-point summaries, typically three sentences. XSum (Narayan et al., 2018) consists of 227K online articles and single-sentence summaries.

Models: The following abstractive summarization models are chosen due to their strong cross-dataset performance: BART (Lewis et al., 2020), a denoising autoencoder for pretraining sequence to sequence and natural language understanding tasks; ProphetNet (Qi et al., 2020), a pre-trained encoder-decoder model that performs n-gram language modeling; PEGASUS (Zhang et al., 2020), a model pre-trained with a summarization-specific objective function; and BERTSUM (Liu and Lapata, 2019), a two-stage fine-tuning approach. Table 1 shows the models’ ROUGE scores (Lin, 2004).

2.2 Reliability

We follow Steen and Markert (2021) and report Krippendorff’s alpha and Split-Half Reliability as measures of the reliability of crowdsourced annotations. Krippendorff’s alpha (α) is a reliability coefficient developed to measure the agreement among multiple annotators (Krippendorff, 2011). This measures instance-level reliability, especially how reliable judgments are over individual summary instances. For system-level rankings, to measure the reliability of the rankings of summarization models, we compute Split-Half Reliability (SHR). To compute SHR, annotations are split into two independent groups, and Pearson correlations are calculated between the groups.

We follow a similar block-design described in Steen and Markert (2021). We note that we include the input document as the context of the summaries as opposed to the coherence and repetition dimensions studied in that work, which do not require reading the input article. We divided our corpus into 20 blocks of 5 documents. We include all generated summaries for each document in the same block, resulting in 5 × 4 = 20 summaries per block. We require 3 annotators per block as in Steen and Markert (2021), and each annotator is limited to annotating at most two blocks total across all tasks. A further study of the effect of the number of annotators or block design is left for future work. Crowdsourcing is done via MTurk.

2.3 Protocols

The Likert Scale (LS) is a common rating-based evaluation protocol (Asghar et al., 2018). Likert Scales applied to summarization typically range from 1-5 (Steen and Markert, 2021). Best-Worst Scaling (BWS) is a type of ranking-oriented evaluation that requires annotators to specify only the best and the worst example in a set of summaries (Hollis and Westbury, 2018; Kiritchenko and Mohammad, 2017). For BWS, the annotator labels the most factually consistent summary and the least factually consistent summary. Another type of ranking-based protocol is pairwise comparison, where each example is compared to every other example. However, this protocol is very expensive: given N items to annotate, N^2 total annotations must be collected as opposed to BWS which requires a constant factor of N total annotators. Due to this exorbitant cost as any reasonable scale, we restrict our study of ranking-based protocols to BWS, and we refer the reader to Kiritchenko and Mohammad (2017) for an in-depth discussion of the cost comparison for the task of sentiment analysis.

<table>
<thead>
<tr>
<th>Models</th>
<th>CNN/DM R-1</th>
<th>CNN/DM R-2</th>
<th>CNN/DM R-L</th>
<th>XSum R-1</th>
<th>XSum R-2</th>
<th>XSum R-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>PEGASUS</td>
<td>44.19</td>
<td>21.45</td>
<td>41.08</td>
<td>36.84</td>
<td>21.52</td>
<td>39.10</td>
</tr>
<tr>
<td>ProphetNet</td>
<td>42.45</td>
<td>19.90</td>
<td>39.31</td>
<td>43.23</td>
<td>19.96</td>
<td>35.16</td>
</tr>
<tr>
<td>BART</td>
<td>44.67</td>
<td>21.13</td>
<td>40.89</td>
<td>44.15</td>
<td>21.28</td>
<td>35.94</td>
</tr>
<tr>
<td>BERTSUM</td>
<td>41.82</td>
<td>19.92</td>
<td>38.67</td>
<td>38.21</td>
<td>16.11</td>
<td>30.85</td>
</tr>
</tbody>
</table>

Table 1: ROUGE-1/2/L scores for model reproduction on CNN/DM and XSum datasets. We apply models directly when they are already fine-tuned and otherwise re-trained them. Pegasus and BART generally obtain the highest ROUGE scores, with ProphetNet comparable in both cases and BERTSUM notably worse on XSum.

<table>
<thead>
<tr>
<th>Models</th>
<th>LS</th>
<th>LS_{10}</th>
</tr>
</thead>
<tbody>
<tr>
<td>PEGASUS</td>
<td>3.887</td>
<td>7.410</td>
</tr>
<tr>
<td>ProphetNet</td>
<td>3.860</td>
<td>7.250</td>
</tr>
<tr>
<td>BART</td>
<td>4.017</td>
<td>7.727</td>
</tr>
<tr>
<td>BERTSUM</td>
<td>3.863</td>
<td>7.453</td>
</tr>
</tbody>
</table>

Table 2: Average model rank and average rating scores across LS (5-point scale) and LS_{10} (10-point scale).
Table 3: Instance and system-level reliability computed by Krippendorff’s alpha (α) and split-half reliability (SHR) on the CNN/DM and XSum datasets.

<table>
<thead>
<tr>
<th>Protocols</th>
<th>CNN/DM</th>
<th>XSum</th>
</tr>
</thead>
<tbody>
<tr>
<td>LS</td>
<td>α</td>
<td>SHR</td>
</tr>
<tr>
<td>BWS</td>
<td>4.43</td>
<td>45.61</td>
</tr>
<tr>
<td>Ours</td>
<td>12.87</td>
<td>51.36</td>
</tr>
</tbody>
</table>

2.4 Research Questions

We study three main research questions (RQ):

- **RQ1**: Ranking (BWS) vs. LS?  We aim to determine the more reliable evaluation protocol.

- **RQ2**: What affects reliability?  We aim to determine the factors that affect the reliability of the human evaluation.

- **RQ3**: What are the protocols’ limitations and how to improve them?  Based on the analysis, we propose two protocols to improve the reliability.

3 Analysis

We show the average ratings across LS scales, including a modified LS scale we will later introduce, in Table 2. Despite the consistently higher ROUGE scores, Pegasus was not always ranked highest, which aligns with previous work suggesting that ROUGE score does not correlate with factual consistency (Durmus et al., 2020). The primary results for reliability evaluation are found in Table 3.

**RQ1**: BWS outperforms LS on CNN/DM.  We see on the left-hand side of the first two rows of Table 3 that BWS outperforms LS by a large margin on both instance-level (α) and system-level (SHR) reliability. As seen in the distribution of the LS ratings in Figures 1, many models are rated as factually consistent with scores of 4 or 5. This coincides with previous investigations on CNN/DM which conclude that recent summarization systems produce fluent texts with relatively few factual errors (Fabbri et al., 2021). We hypothesize that the greater reliability of BWS on CNN/DM data may result from the ranking task forcing the annotator to choose the best summary and distinguish these close summaries rather than allowing e.g. the annotator to give both a score of 5. This result suggests that BWS is preferable in cases where the summaries analyzed have similar factual consistency, such as CNN/DM.

Though agreement on individual summaries (α) is relatively low for all annotation methods, these numbers are comparable to those obtained in (Steen and Markert, 2021). Furthermore, we look at the relative difference between (α) of BWS and LS, and we find that studies still arrive at consistent system scores as demonstrated by the SHR. This reflects similar observations made by Gillick and Liu (2010). System-level ranks such as SHR, are also more important for evaluation purposes as the goal is generally to rank models to determine the best performing (or most factually consistent) system as opposed to examining individual examples as Krippendorff’s alpha measures.

**RQ2**: Dataset Characteristics Affect Reliability.  We extend our experiments to the XSum dataset to see whether the reliability of the protocols changes as the characteristics of the dataset change. XSum-trained models are known to suffer from factual inconsistencies because of the high compression ratio and high level of abstraction of the reference summaries (Maynez et al., 2020). As seen on the right-hand side of the first two rows of Table 3, BWS and LS both perform well, with LS slightly outperforming BWS according to SHR. As seen in Figure 1, the model scores are more spread out along the scale. This coincides with the large range of ROUGE scores and larger differences between models, as seen in Table 1, which likely explains why annotators can differentiate the model outputs better. Thus, we believe that LS is a viable option when the corpus contains a diverse quality of summaries, like XSum.
Figure 2: Score distribution of $LS_{10}$ across CNN/DM and XSum. Each data point shows the number of times a score was assigned to each system.

RQ3: Improvements and Current Limitations.

We propose two modified protocols to improve reliability and then study the presence of common limitations for evaluation protocols. Prior work has noted the effect of scale granularity (Kiritchenko and Mohammad, 2017), so for LS, we extend the scale from 5 to 10 and call it LS−10. Table 3 shows that that LS−10 is more reliable than LS A finer-grained scale may capture more nuanced differences in data points with more choices. Scores tend to move towards the extremes when we use a finer-grained scale (10 vs 5), as seen in the difference in distributions in Figures 1 and 2. Thus, for LS−10, a larger range and being less biased towards a specific region, promoting better reliability. Previous work suggests that Best-Worst Scaling fails to yield an unbiased estimate of the true quality value (Hollis, 2018). Thus, for BWS, we incorporate information about the quality of competing examples or value learning into a BS$^S_{value}$ protocol. The annotator is asked to give a score (3-point scale) for the difference between the best and the worst summary. The final ranking uses a weighted sum. The results at the bottom of Table 3 also confirm the effectiveness of this protocol.

To verify the limitations of evaluation protocols noted by Kiritchenko and Mohammad (2017), we conduct the following studies. We first analyze (a) the inconsistencies in annotations by different annotators, measured by the percentage of summaries that receive different ratings or rankings from different annotators, which we call change rate. As shown in Table 4, annotators are more likely to agree on the same ranking in BWS as opposed to the same rating for LS. We further test (b) inconsistencies by the same annotator, in particular whether annotations done by the same worker are consistent over time. We ask workers who have previously annotated XSum and CNN/DM samples to re-do their annotations one week after their initial annotations. We notified the workers to re-annotate only one week after they finished, instead of at the beginning, as we do not want to introduce design bias. In total, 43 workers redid 860 annotations. For LS, the average change in the rating of the two annotations one week apart by the same worker was 0.92.

Additionally, we examine whether LS suffers from (c) scale region bias, where different annotators are often biased towards different parts of the rating scale. For a given block and two annotators, we calculate the rating range given by each annotator. Then we calculate the overlap length between those two ranges divided by the length of the overall range from both annotators. We call this the percentage scale overlap and average over all pairs of annotators and blocks. For LS, the percentage scale overlap is (0.67, 0.88) for (CNN/DM, XSum), respectively, and (0.61, 0.82) for LS−10. The difference in scale region bias between LS and LS−10 is small, but the bias difference between CNN/DM and XSum is notable. Greater diversity in summary quality as in XSum may force the annotators to expand their use of the scale and mitigate region bias, which may explain why LS is better than BWS on XSum as opposed to CNN/DM. Future work may investigate further what exactly constitutes too wide of a scaling range.

<table>
<thead>
<tr>
<th>Change Rate (%)</th>
<th>Scale Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN/DM</td>
<td>XSum</td>
</tr>
<tr>
<td>BWS LS $LS_{10}$</td>
<td>BWS LS $LS_{10}$</td>
</tr>
<tr>
<td>74.71 87.75 96.00</td>
<td>70.25 92.25 96.25</td>
</tr>
</tbody>
</table>

Table 4: Change Rate, or percentage of summaries given different ranks or ratings by different annotators (lower is better). Scale Overlap, or average overlap of the range of rating scores between annotators (higher is better).

4 Conclusion

In this paper, we conduct studies to understand and improve the reliability of ranking and rating-based human evaluations of summarization factual consistency. We find that Best-Worst Scaling is largely reliable, and the Likert scale also has merits, but the proper scaling and dataset characteristics must be carefully studied to ensure its reliability. We improve these two protocols based on our findings and believe that our studies advance the understanding of both models and metrics as we aim to facilitate factually consistent text generation.
5 Ethical Considerations

Intellectual Properties and Privacy Rights All of the datasets (CNN/DM and XSum) used in our study are publicly available. Regarding privacy rights, the authors of the paper completed IRB human subject protection training for conducting this study. We will release the annotations, but rather than releasing the MTurk ID of the worker, we will completely anonymize this ID.

Compensation for Annotators Workers were compensated $5 per block, calibrated to equal a $15/hour payrate. We first annotated examples in-house to determine the required annotation speed. A summary block usually takes around 20 minutes.

Steps Taken to Avoid Potential Problems Annotations were completed in the form of a survey on a Google Form. We provided space for the Turkers to provide feedback. We manually uploaded the data points (articles and summaries) used in this study to avoid any offensive content.

The Number of Examples We sampled 100 examples from each dataset that did not contain exactly matching summaries. Both Likert and BWS follow the same block design, which includes the same number of examples per block. With the exception that the BWS annotation asks for the most and least factually consistent summary and the Likert asks for ratings for each individual summary. Due to space requirements, we included further details, images of the interface, in the supplementary material. We pay the same amount per block of annotations.

Qualifications of MTurk workers We use the following qualifications to recruit in total 350 MTurk workers with good track records: HIT approval rate greater than or equal to 98%, number of HITs approved greater than or equal to 500, and located in one of the following English native-speaking countries: Australia, Canada, New Zealand, United Kingdom, United States.

References


A Appendix

Besides the average model rank and average rating scores across BWS, LS-5, and LS-10 evaluations, we also provide standard deviations in Table 5.

To demonstrate our annotation template and facilitate future research, we show the interface for BWS annotations in Figures 3 and 4 and the interface for Likert annotations in Figures 5 and 6. We made use of the survey feature in Amazon Mechanical Turk (MTurk) to link to these Google Forms in Figure 7.
<table>
<thead>
<tr>
<th>Models</th>
<th>CNN/DM</th>
<th>XSum</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BWS</td>
<td>LS-10</td>
<td>BWS</td>
<td>LS</td>
</tr>
<tr>
<td>PEGASUS</td>
<td>3.230/1.150</td>
<td>3.887/1.051</td>
<td>7.410/</td>
<td>2.160</td>
</tr>
<tr>
<td>ProphetNet</td>
<td>3.300/1.026</td>
<td>3.860/1.092</td>
<td>7.250/</td>
<td>2.252</td>
</tr>
<tr>
<td>BART</td>
<td>3.593/1.113</td>
<td>4.017/1.073</td>
<td>7.727/</td>
<td>2.090</td>
</tr>
<tr>
<td>BERTSUM</td>
<td>3.087/0.984</td>
<td>3.863/1.037</td>
<td>7.453/</td>
<td>2.309</td>
</tr>
</tbody>
</table>

Table 5: Average model rank, rating, and standard deviation across BWS, LS and $L_{S_{10}}$ evaluations.

Figure 3: Screenshot of the instruction page for BWS annotation.
Figure 4: Screenshot of the evaluation page for BWS annotation.
Figure 5: Screenshot of the instruction page we used for Likert Scale annotation.
Figure 6: Screenshot of the evaluation page for Likert Scale annotation.
We are conducting an experiment about the faithfulness of text summarization. You will be presented with 20 summaries (4 articles * 5 summaries per article).

Your task is to rate the faithfulness of each summary (either through scale or ranking). Detailed instructions will be in the Google form.

For the accuracy of the experiment, you will only be allowed to do one HIT/form of this batch.

Acknowledgment code can be found after you submit the form.

Make sure to leave this window open as you complete the form. When you are finished, you will return to this page to paste the code into the box.

Provide the acknowledgment code here:

Submit