Revisiting the Gold Standard: Grounding Summarization Evaluation with Robust Human Evaluation

Yixin Liu*, Alexander R. Fabbri*, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, Dragomir Radev

Introduction

Human evaluation is the foundation upon which the evaluation of both summarization systems and automatic metrics rests. However, existing human evaluation protocols and benchmarks for summarization either exhibit low inter-annotator agreement or lack the scale needed to draw statistically significant conclusions, and an in-depth analysis of human evaluation is lacking. In this work, we address the shortcomings of existing summarization evaluation along the following axes: 1) We propose a modified summarization salience protocol, Atomic Content Units (ACUs), which relies on fine-grained semantic units and allows for high inter-annotator agreement. 2) We curate the Robust Summarization Evaluation (RoSE) benchmark, a large human evaluation dataset consisting of over 22k summary-level annotations over state-of-the-art systems on three datasets. 3) We compare our ACU protocol with three other human evaluation protocols, underscoring potential confounding factors in evaluation setups. 4) We evaluate existing automatic metrics using the collected human annotations across evaluation protocols and demonstrate how our benchmark leads to more statistically stable and significant results. Furthermore, our findings have important implications for evaluating large language models (LLMs), as we show that LLMs adjusted by human feedback (e.g., GPT-3.5) may overfit unconstrained human evaluation, which is affected by the annotators' prior, input-agnostic preferences, calling for more robust, targeted evaluation methods.

Protocol	w/ Doc	w/ Ref	Fine-grained
Prior	✗	✗	✗
Ref-free	✓	✗	✗
Ref-based	✗	✓	✗
ACU	✗	✓	✓

Statistical Power §4.1	– High statistical power is difficult to reach for human evaluation of similar-performing systems. – Increasing the sample size of human evaluation effectively raises statistical power.
Summary Length §4.2	– Summaries from different summarization systems show a large difference in average length. – Difference in summary length is not well-reflected by automatic evaluation metrics.
Evaluation Protocol Comparison §5.2	– Reference-free and reference-based human evaluation results have a near-zero correlation. – Reference-free human evaluation strongly correlates with input-agnostic, annotator preference. – Annotator's input-agnostic preference has a strong positive correlation with summary lengths. – Annotator's input-agnostic preference does not favor reference summaries. – Compared to smaller, fine-tuned models, zero-shot large language models (e.g. GPT-3) perform better under reference-free evaluation, but worse under reference-based evaluation.
Evaluating Automatic Metrics §6.2 & §6.3	– A higher-powered human evaluation dataset can lead to a more robust automatic metric evaluation, as shown by a tighter confidence interval and higher statistical power of metric evaluation. – Automatic metric performance differs greatly under different human evaluation protocols. – Automatic metrics show relatively strong system-level correlation and moderate summary-level correlation with our robust human evaluation protocol.

(a) Reference Summary: Chelsea weren't awarded a penalty for David Ospina's clash with Oscar. Arsenal goalkeeper clattered Oscar inside the box. Brazilian was taken off at half-time, with Didier Drogba replacing him.

(b) System Summary (BRIO, (Liu et al., 2022)): Oscar collided with Arsenal goalkeeper David Ospina in the 16th minute of the London derby . The Brazilian was substituted at half-time and Jose Mourinho said he suffered 'possible concussion' . Oscar was knocked back by the goalkeeper but Michael Oliver didn't award Chelsea a penalty .

(c) System Summary (GPT-3, (Brown et al., 2020)): Oscar was forced to leave the match against Arsenal after sustaining a possible concussion from a collision with the opposing goalkeeper. The referee did not award Chelsea a penalty, despite the collision appearing to warrant one. Sky Sports pundits agreed that the collision should have been penalized, with some suggesting it could have even warranted a red card.

(d) ACUs with corresponding evaluations:

- Chelsea weren't awarded a penalty. ✓ ✓
- David Ospina clashed with Oscar. ✓ ✓
- David Ospina clattered Oscar. ✓ ✓
- David Ospina plays for Arsenal. ✓ ✗
- David Ospina is a goalkeeper. ✓ ✗
- The clash occurred inside the box. ✗ ✗
- Oscar is Brazilian. ✓ ✗
- Oscar was taken off at half time. ✓ ✗
- Didier Drogba replaced Oscar. ✗ ✗

Table 2: Example of a reference summary, system summaries and corresponding ACU annotations on CNNDM. The presence or absence of the ACUs for BRIO (in blue) and GPT-3 (in green) are marked by (✓) and (✗).

	Prior	Ref-free	Ref-based	*n*ACU
Prior	-	**0.926**	-0.061	0.048
Ref-free	0.926	-	**-0.247**	-0.093
Ref-based	-0.061	-0.247	-	0.762
*n*ACU	0.048	-0.093	0.762	-
Len.	0.833	0.875	-0.550	-0.296

	Prior	Ref-free	Ref-based	ACU	Len.
BART	3.58	3.52	2.93	0.367	69.5
BRIO	3.51	3.49	**3.07**	**0.429**	66.4
T0	3.33	3.24	2.84	0.295	61.6
GPT-3	**3.72**	**3.76**	2.74	0.268	69.5
Ref.	2.85	2.94	-	-	54.9

Protocol	Prior	Ref-free	Ref-based	*n*ACU
ROUGE1	-0.061	-0.212	0.840	0.636
ROUGE2	0.000	-0.151	0.595	0.636
ROUGEl	-0.061	-0.212	0.779	0.636
METEOR	0.394	0.242	0.382	0.485
CHRF	0.576	0.424	0.199	0.485
BERTScore	-0.091	-0.182	0.779	0.485
BARTScore	-0.091	-0.182	0.656	0.364
QAEval	0.485	0.515	-0.076	0.151
SummaQA	0.515	0.424	0.260	0.303
Lite^{2}Pyramid	0.576	0.667	-0.168	0.121

Conclusion

We introduce RoSE, a benchmark whose underlying protocol and scale allow for more robust summarization evaluation across three datasets and encompassing two domains. Applying our benchmark, we re-evaluate the current state of human evaluation and its implications for both summarization system and automatic metric development. We hope that this work can be a valuable resource for future research and encourage the research community to extend our insights and help strengthen the foundation of summarization evaluation.