UNDER REVIEW

RESCIENCE C

Replication / ML Reproducibility Challenge 2022

[Re] DialSummEval - Evaluation of automatic summarization evaluation metrics

Anonymous[1, ID]

[1] Anonymous Institution

Edited by
(Editor)

Reviewed by
(Reviewer 1)
(Reviewer 2)

Received
03 February 2023

Published
–

DOI
–

Reproducibility Summary

Scope of Reproducibility – In this paper, we perform a reproduction study of the original work of Gao and Wan [1] on the evaluation of automatic dialogue summarization metrics and models. They concluded that (1) few metrics are efficient across dimensions, (2) metrics perform differently in the dialogue summarization task than when evaluating conventional summarization, (3) models tailored for dialogue summarization capture *coherence* and *fluency* better than *consistency* and *relevance*.

Methodology – Three annotators evaluated the outputs of 13 summarization models and their human reference summaries, following the guidelines of the original paper. This took on average 20 hours. A new annotation tool was developed to address the limitations of the Excel interface. An ablation study was conducted with a subset of data annotated with the original process. Finally, we implemented modified parts of the author's code to apply the metrics over the summaries and compare their scores with our human judgments. All experiments were run on CPU.

Results – The original paper's main claims were reproduced. While not all original authors' arguments were replicated (e.g. ROUGE scoring higher for *relevance*), the correlation between metrics and human judgments showed similar tendencies as in [1]. The annotations correlated with the original at a Pearson score of 0.6, sufficient for reproducing main claims.

What was easy – The reproducibility strengths of the original paper lie primarily in its profound methodological description. The rich and detailed incorporation of tables made the comparison with our reproduced results fairly easy.

What was difficult – The reimplementation of the original paper's code was relatively complex to navigate and required a fair amount of debugging when running the metrics. Certain deficiencies in the annotation guidelines also resulted in rather time-consuming decision-making for the annotators. Finally, the methodological description of the post-processing of the annotations was relatively unclear and the code calculating the inter-annotator agreement was missing.

Communication with original authors – We contacted the paper's first author, twice, to request the annotation guidelines, the missing code parts, and clarifications regarding the annotation post-processing. Their responses were prompt and helpful.

Copyright © 2023 Anonymous, released under a Creative Commons Attribution 4.0 International license.
The authors have declared that no competing interests exists.
Code is available at https://anonymous.4open.science/r/Reproducing_DialSummEval-7371/.

1 Introduction

As noted by Belz et al. [2] in their survey of reproducibility research in Natural Language Processing (NLP), increased attention has been directed towards the reproduction of results in the field following a "reproducibility crisis" in science [3]. A great deal of NLP reproduction studies in recent years have focused on metric scores [4]. However, even though human evaluations have been a central part of Natural Language Generation (NLG) research for years, little is known about their reproducibility. Some effort has been made to investigate the topic [5]. For example, Iskender, Polzehl, and Möller [6] investigate the reliability of human evaluations focusing on factors such as annotators' demographics and the design of the task.

Dialogue summarization is a task in NLP that involves generating a summary of a conversation or dialogue. This task is important for applications such as chatbots, where a summary of a conversation can help users quickly understand its content and make informed decisions. The quality of dialogue summaries is typically evaluated using automatic evaluation metrics such as ROUGE [7], BLEU [8] or BARTScore [9]. However, these metrics seem to not accurately or sufficiently assess the performance of dialogue summarization models, as they do not consider the multifaceted nature of the task and its specific challenges. In other words, it appears that each of these metrics fails to capture one or more aspects determining the quality of a summary [10].

The work of Gao and Wan [1] investigates these shortcomings by re-evaluating a range of automatic evaluation metrics and correlating them with human evaluation to identify the strengths and weaknesses of current evaluation methods of dialogue summarization. Creating the DialSummEval dataset is a significant contribution of the paper, providing a valuable resource for evaluating dialogue summarization models.

In this paper, we reproduce the methodology and main findings of Gao and Wan [1], which are more extensively introduced in Section 2. In addition, we reflect on the reproducibility process with respect to its determining factors and main challenges.

2 Scope of reproducibility

The present study builds on the work of the original paper, "DialSummEval: Revisiting Summarization Evaluation for Dialogues" [1], which proposed a re-evaluation of automatic evaluation metrics for dialogue summarization. Gao and Wan [1] observed that current methods for evaluating the quality of summaries in dialogue summarization, such as relying on the SAMSum dataset [11] and ROUGE [7], are flawed and may not accurately assess the performance of dialogue summarization models. The paper re-evaluates a range of automatic evaluation metrics in terms of *coherence*, *consistency*, *fluency*, and *relevance*. Additionally, they conducted a human evaluation of various summarization models based on the same four quality aspects. The human evaluation was the primary focus of the reproduction study, as the resulting dataset was used in the subsequent experiments to evaluate different evaluation metrics.

The original paper presents the following main claims:

1. Few automatic evaluation metrics perform well in all dimensions of dialogue summarization. Recently proposed metrics, such as BARTScore and QA-based, are the best performers.

2. There are some trends in the performance of different evaluation metrics for dialogue summarization that differ from those observed in conventional summarization tasks.

3. Models specifically designed for dialogue summarization perform well in terms of *coherence* and *fluency* but still have shortcomings in terms of *consistency* and *relevance*.

3 Methodology

The original paper selected dialogues and their corresponding human reference summaries from the SAMSum dataset and collected automatic summaries for each dialogue stemming from 13 summarization models (see Section 3.2.1). Six models already had SAMSum summaries included in their publication. Other models did not come with these specific summaries, so Gao and Wan [1] generated them by applying the models to the dialogues in question. Once all dialogues and the 14 summaries for each dialogue (13 model outputs and one human summary) were collected, they were annotated by three human annotators. The annotators evaluated each summary on four dimensions. Aside from the human annotation, each summary was evaluated using 32 automatic metrics (see Section 3.2.2). The score of each metric was compared to the respective human annotation score for each quality dimension using Pearson's R correlation on both summary and system level, to assess the relatedness between automatic evaluation and human judgment.
For the present reproduction, the same random selection of dialogues from the SAMSum dataset was used. The original authors provide all model-generated summaries that they used and the results of all automated metrics. We decided not to generate new model outputs or rerun the automatic evaluation metrics, as the main claims and contributions of Gao and Wan [1] pertain to the outcome of the human annotation process. While the third claim does concern the performance of the models, it references only systems designed specifically for dialogue summarization which the authors of the original paper did not re-train themselves. Furthermore, reproducing only the human annotation part of Gao and Wan [1] allows us to control for the variation within the model-produced summaries and metric scores. However, since the human evaluation was redone, the final correlation scores of the automated metrics and the human scores were recalculated.

3.1 Datasets

The SAMSum dataset is a collection of 16k chat dialogues written by linguists fluent in English and contains one summary per dialogue written by a language expert. The dataset consists of training, validation, and test sets of 14732, 818, and 819 dialogues, respectively. Gao and Wan [1] sampled random 100 dialogues from the test portion of the SAMSum dataset [11].

3.2 Model and metric descriptions

Model descriptions — Each of the 100 dialogues was summarized using 13 models. Two models were extractive (LEAD-3 and LONGEST-3), two were neural summarization models (PGN [12] and Transformer [13]), and three were generic pre-trained generative models (BART [14], PEGASUS [15] and UniLM [16]). The original paper retrained all these models to acquire their outputs. The present paper uses these generated outputs. The remaining six models were designed for dialogue summarization, and their summaries for the SAMSum dataset were already available. These models were CODS [17], ConvoSumm [18], MV-BART [19], PLM-BART [20], Ctrl-DiaSumm [21], and S-BART [22].

Metrics — Gao and Wan [1] evaluated 32 automatic evaluation metrics, which fall into five categories. Examples of metrics in each category include (1) n-gram overlap (ROUGE [7], BLEU [8], and METEOR [23]), (2) pre-trained language models (BERTScore [24], MoverScore [25], and BARTScore [9]), (3) word embeddings (SMS [26], Embedding average [27], and Vector extrema [28]), (4) question-answering (FEQA [29], SummaQA [30], and QuestEval [31]), and (5) entailment classification (FactCC [32], DAE [33]).

3.3 Annotation

The focus of this project was on the reproducibility of the original annotations and their correlation with automatic evaluations. Three of the present paper's authors annotated all of the summaries, which is the same number of annotators as employed in Gao and Wan [1]. All three have a background in Linguistics and Natural Language Processing and were thus deemed adequate annotators. No annotation guidelines were present in the original paper, but the authors provided these upon request. These guidelines can be found in the Appendix in Section 7.1. In line with Gao and Wan [1], all annotators were asked to annotate all data (i.e., 100 dialogues x 14 model outputs = 1400 instances) to maintain consistency within the annotations. The dialogue was first presented to the annotators; subsequently, the model outputs were shown one by one. The annotators were asked to score each summary on a Likert scale from 1 to 5 in four dimensions: *consistency*, *coherence*, *fluency*, and *relevance*. The explanation of these dimensions can be found in the annotation guidelines (see Appendix 7.1). As in the original paper, the order of the model outputs was the same for every dialogue. The total time required per annotator to evaluate all summaries was between two and three full working days (16-24 hours).

We deviated from the original annotation procedure and guidelines on two minor points. Firstly, Gao and Wan [1] employed an Excel sheet for annotating the dialogue summaries. For faster and easier annotation, we developed an annotation tool (see Appendix 7.2) that can be used in a Jupyter Notebook. The code can be found on our Github[1]. An ablation study was conducted to investigate the possible influence of the tool, compared to using Excel (see Section 4.2.2).

The second deviation concerns the scoring of the summaries on the category of *coherence*, which pertained to the summaries that were composed of a non-complex sentence. These were challenging to score on *coherence*, as this metric assesses the quality of all sentences in relation to one another. The original guidelines did not address this issue. An amendment was made to give these summaries a default *coherence* score of 4.

3.4 Processing annotations

We followed the annotation processing procedure adopted by [1]. Before calculating the inter-annotator agreement using Krippendorf's Alpha, Gao and Wan [1] clean the noise. Noise is defined as the outlier score when two of the three annotators agree. The noise then no longer influences the agreement negatively. Subsequently, the resulting annotations are aggregated into one set of four scores per summary (one per dimension). Thus, the majority vote was taken as the gold standard, and when none of the annotators agree the average score was used. All further experiments were performed on the cleaned annotations. The results of calculating the inter-annotator agreement are discussed in Section 4.2.

3.5 Experimental setup and code

For details on the annotation set-up and tool see Section 3.3. We used the original paper's code to run the correlation calculations between our human evaluation scores and the automatic metrics. The only necessary adjustment was to adhere to the different file formats in which the human evaluations were stored. The full code, including the annotation tool, Inter-annotator agreement calculations, and the correlation experiment can be found on Github[1]. All experiments could be run on a machine without a dedicated graphics card.

[1] https://anonymous.4open.science/r/Reproducing_DialSummEval-7371/

4 Results

The results of our reproduction of each of the main claims (see Section ??) are given below. Additional experiments were conducted to further investigate the validity of the human annotations and the influence of our annotation tool. These results are discussed in Section 4.2.

4.1 Results reproducing original paper

Result 1 — The results in this section concern the first main claim made in the original paper, namely (1) there are few automatic evaluation metrics that perform well in all dimensions of dialogue summarization and (2) recently proposed metrics such as BARTScore and QA-based metrics perform the best.
The first part of this claim is supported by the findings of the current reproduction study. The authors of the original paper state that a metric can be seen as a good performer when it shows significant strength in all four dimensions. As is visible in Table 1, there is indeed no metric that has a significantly high correlation with the human judgments in all dimensions. Some metrics, such as BERTScore-f1 and BARTScore-r-h perform moderately well in three dimensions, but fail in one (*consistency*).
The second part of the above-mentioned claim, that BARTScore and QA-based metrics outperform the other metrics, is supported by the findings of this study. It is difficult to define a threshold for when a model outperforms other models. Using the five highest correlating metrics on a system level as an indication, it is evident that a large share of these come from BARTScore or a QA-based metric, though to a lesser extent than in the original paper. Table 1 shows that on the system level, two out of five highest correlating metrics on *coherence* are either a BARTScore or a QA-based metric, four out of five for *consistency*, two out of five for *fluency*, and three out of five on *relevance*. Therefore, the results of this reproduction support the first main claim of the original paper.

Result 2 — This result pertains to the second claim outlined in Section 2, which states that the trends observed for the automatic evaluation metrics in the results of Gao and Wan [1], i.e. for dialogue summarization, differ from the patterns observed for conventional text summarization in previous studies. Gao and Wan [1] base the assertion mostly on the results of ROUGE. More specifically, they note the following:

- *Increasing the size of n in ROUGE-n did not lead to improvement on almost all dimensions, contrasting the findings of Rankel et al. [34] and Fabbri et al. [18].* This result has been replicated in our reproduction. In fact, as can be seen in Table 1, increasing the size of n led to lower results for every single ROUGE metric on all dimensions, both on system and summary level.

- *The scores obtained by ROUGE for the dimension of relevance were not as high as could be expected, given its commonly-believed ability to reflect content selection.* This result was **not** replicated in the current study. In fact, on the system level, out of the four dimensions, all ROUGE metrics obtained the best results for *relevance*. Furthermore, also on the system level, ROUGE-1 and ROUGE-l obtained the second and the fourth highest score, respectively, out of all 32 metrics.

- *Metrics based on n-gram overlap such as ROUGE and CHRF obtained lower scores on dialogue summarization than they do on conventional text summarization in Fabbri et al. [18], while metrics that make use of source documents such as BLANC performed better.* This result was **not** replicated in the current study. In the current reproduction, ROUGE has obtained higher scores than Fabbri et al. [18] for *relevance* and *coherence* (all sub-metrics), and *fluency* (some sub-metrics). Fabbri et al. [18] observes better performance of ROUGE only for *consistency*. Moreover, CHRF has scored higher

than in Fabbri et al. [18] on two metrics of *coherence* and *relevance*. Finally, BLANC has obtained lower scores than those observed in Fabbri et al. [18].

While not all individual results were replicated, the second main claim made by Gao and Wan [1] can still be supported, as the observed results show differences in the performance of the automatic evaluation metrics on dialogue versus conventional summarization.

	Coherence		Consistency		Fluency		Relevance	
Metrics	sys	sum	sys	sum	sys	sum	sys	sum
ROUGE-1	0.67	**0.49**	0.38	0.25	**0.66****	**0.46**	**0.73****	**0.42**
ROUGE-2	0.52	0.4	0.36	0.24	0.49	0.36	0.66**	0.4
ROUGE-3	0.41	0.32	0.34	0.23	0.38	0.28	0.60*	0.37
ROUGE-4	0.35	0.28	0.31	0.22	0.31	0.23	0.56*	0.34
ROUGE-L	0.66**	**0.49**	0.35	0.23	0.64*	0.45	**0.70****	**0.41**
BERTScore-p	**0.84****	**0.66**	0.09	0.07	**0.78****	**0.58**	0.53	0.32
BERTScore-r	0.45	0.36	0.4	0.28	0.44	0.33	0.68**	**0.43**
BERTScore-f1	**0.69****	**0.55**	0.24	0.17	0.65*	**0.49**	0.63*	0.39
MoverScore	0.57*	0.45	0.34	0.24	0.55*	0.41	0.67**	**0.41**
SMS	0.33	0.24	0.32	0.22	0.30*	0.21	0.56	0.34
BARTScore-s-h+	-0.44	-0.36	**0.65***	**0.44**	-0.26	-0.21	0.26	0.17
BARTScore-h-r	0.15	0.09	-0.08	-0.06	0.23	0.17	-0.04	-0.04
BARTScore-h-r	0.48	0.36	0.5	**0.32**	0.48	0.34	**0.76****	**0.45**
BARTScore-r-h	**0.85****	**0.6**	0.29	0.21	**0.85****	**0.59**	**0.69****	0.38
BLANC-help+	-0.79**	-0.56	**0.52**	**0.36**	-0.64*	-0.43	0.11	0.11
BLANC-tune+	-0.82**	-0.59	0.48	**0.3**	-0.68**	-0.46	0.06	0.07
FEQA+	**0.89****	0.41	0.37	0.15	**0.92****	0.4	0.63**	0.22
QuestEval+	0.1	0.04	**0.89****	**0.3**	0.31	0.08	**0.72****	0.24
SummaQA-conf+	-0.66*	-0.34	0.64*	**0.33**	-0.52	-0.26	0.23	0.13
SummaQA-fscore+	-0.74**	-0.32	**0.55***	0.24	-0.59*	-0.24	0.16	0.07
PPL-	0.17	0.06	-0.48	-0.25	-0.02	-0.08	-0.35	-0.17
CHRF	0.43	0.34	0.41	0.28	0.42	0.32	0.68**	**0.42**
BLEU-1	0.41	0.33	0.29	0.2	0.37	0.28	0.57*	0.35
BLEU-2	0.34	0.28	0.29	0.21	0.29	0.23	0.54*	0.35
BLEU-3	0.3	0.25	0.27	0.2	0.25	0.2	0.51	0.33
BLEU-4	0.25	0.21	0.27	0.19	0.2	0.16	0.49	0.32
METEOR	0.38	0.31	0.36	0.26	0.35	0.28	0.62*	0.4
EmbeddingAverage	**0.78****	**0.53**	0.12	0.09	**0.77****	**0.51**	0.59*	0.31
VectorExtrema	0.56*	0.43	0.3	0.2	0.53*	0.39	0.64*	0.38
GreedyMatching	0.56*	0.42	0.3	0.2	0.54*	0.4	0.65*	0.37
FactCC+	-0.83**	-0.5	0.45	0.21	-0.73**	-0.41	0	-0.04
DAE+	-0.79**	-0.53	**0.51**	0.27	-0.67**	-0.43	0.04	0

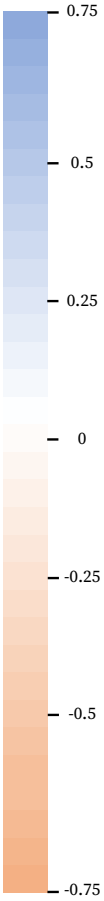

Table 1. Best viewed in color. Orange values are scores below what the original paper reports and blue values are higher. The differences are shown numerically in Appendix 7.3. The following is taken from Gao and Wan [1] due to identical table layout: The correlation (Pearson's R) of annotations computed on system level and summary level along four quality dimensions between automatic metrics and human judgments. For evaluation, all metrics require at least the summaries to be evaluated as input. Metrics with + indicate that the source dialogues are used, metrics with - mean no other input is required, others need to use the reference summaries. The five most-correlated metrics in each column are bolded (For system level, **=significant for $p \leq 0.01$, *=significant for $p \leq 0.05$). Suffixes are added to distinguish the different variants of metrics. For BARTScore, h, r, and s are abbreviations of hypotheses, references, and source dialogues respectively. BARTScore-s-h measures the probability to generate hypotheses using source dialogues as inputs, while BARTScore-h measures the probability to generate hypotheses without other inputs, and so on. For BLANC, BLANC-tune refers to the way of fine-tuning on a generated summary and then conducting nature language understanding tasks on source dialogues, while BLANC-help refers to the way of inferring with a generated summary concatenated together. For SummaQA, SummaQA-fscore measures the average overlap between predictions and ground truth answers, and SummaQA-conf corresponds to the confidence of the predictions.

Result 3 — This final result concerns the third of the claims stated in Section 2, namely that models created specifically for dialogue summarization (i.e., CODS, ConvoSumm, MV-BART, PLM-BART, Ctrl-DiaSumm, and S-BART) obtain scores comparable to reference summaries on the dimensions of *coherence* and *fluency* but perform worse on *consistency* and *relevance*. This result was replicated. Table 2 shows that while the reference summary and the models obtained higher *consistency* scores in the reproduction than in the original paper, in both studies, the best-performing model on that dimension still obtained a result at least 0.433 lower than the reference summary (0.433 for Gao and Wan [1]; 0.434 for the current reproduction).

Models	Coherence	Consistency	Fluency	Relevance	R-1	R-2	R-L
reference	**4.76**	4.917	4.88	**4.287**	1.000	1.000	1.000
LONGEST-3	1.423	**4.967**	2.98	3.117	0.304	0.092	0.267
LEAD-3	2.42	**4.937**	3.103	2.82	0.309	0.092	0.296
PGN	3.487	2.253	3.523	1.707	0.356	0.126	0.357
Transformer	3.28	1.423	3.667	1.447	0.329	0.098	0.319
BART	4.54	4.29	**4.99**	3.363	0.533	0.299	0.521
PEGASUS	4.53	4.427	4.95	3.203	0.508	0.254	0.476
UniLM	4.293	3.827	4.72	3.06	0.489	0.232	0.470
CODS	4.37	4.277	4.91	3.287	0.523	0.278	0.509
ConvoSumm	4.623	4.483	4.98	3.377	0.532	0.268	0.498
MV-BART	**4.747**	4.45	**4.99**	**3.76**	0.539	0.289	0.513
PLM-BART	4.627	4.34	4.93	3.487	0.533	0.284	0.506
Ctrl-DiaSumm	4.73	4.4	4.96	3.663	**0.564**	**0.312**	**0.549**
S-BART	4.517	3.873	4.93	3.213	0.497	0.244	0.472

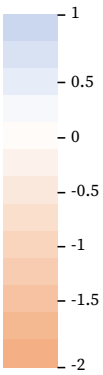

Table 2. Best viewed in color. Average human ratings across the four dimensions for each model output summary. Additional ROUGE-1,2,l scores were calculated using the present sampling data. The two best-performing summaries for each dimension are highlighted in bold. The blue values are scores higher than those reported in the original paper, and the orange scores are lower. All ROUGE scores are identical to the original paper. The differences are shown numerically in Appendix 7.3

4.2 Results beyond original paper

Additional Result 1 — We perform a statistical comparison between the annotations performed by Gao and Wan [1] and our annotations to examine their deviation. To estimate the impact of the noise removal, we first compare the inter-annotator agreement on uncleaned annotations, measured with Krippendorff's Alpha score. Our annotations have higher agreement across all dimensions with *consistency* superseding the other three (see Table 3). Given that a usually required Krippendorff's Alpha is around 0.80 and the lowest acceptable score is 0.67 [35], it is clear that the uncleaned version of the original paper's annotations fails to meet this threshold. Following the noise removal, we observe that our reproduction is left with a higher number of annotations throughout. This indicates that we had fewer cases in which all three annotators assigned different scores. Additionally, our cleaned annotations display a higher Krippendorff's Alpha with three of the dimensions scoring slightly below the recommended threshold (0.80) and *consistency* scoring well above (0.92). At the same time, the removal of noise increased the agreement results in the original paper, especially in the case of *fluency*. However, even with the cleaning effort, the agreement on *relevance* did not surpass the lowest threshold (0.67).

We further calculated the Pearson's R correlation between annotations for each dimension, see Table 4. We observe moderate correlation for*coherence*(0.42) and *fluency*(0.55), *consistency* and *relevance* show a higher uniformity.

	Coherence	Consistency	Fluency	Relevance
Total	4200	4200	4200	4200
Krippendorff's α (total)	0.38 →0.61	0.49 →0.79	0.13 →0.52	0.39 →0.52
Cleaned	3161 →3607	3360 →3754	3050 →3625	3439→3394
Krippendorff's α (cleaned)	0.76 →0.78	0.67 →0.92	0.68 →0.78	0.56 →0.72

Table 3. Annotations and agreement: "original paper →reproduction results". The first row represents the total amount of annotations, and the second represents the IAA on the total. The third row is the number of annotations left after cleaning, and the fourth row shows the IAA on the cleaned annotations.

Additional Result 2 — We conducted an ablation study to examine the impact of the annotation tool on the annotation procedure. 140 summaries (14 summaries per 10 randomly selected dialogues) were annotated by the same three annotators as in our main annotation process. They used the same method as in Gao and Wan [1], working in Excel where each model's summaries were displayed on separate sheets. Table 4 reveals a strong correlation between the results obtained through the tool and the original annotation process, supporting the use of the tool.

	Coherence	Consistency	Fluency	Relevance
Reproduction-Original	0.42	0.77	0.55	0.69
Full Reproduction-Ablation	0.71	0.66	0.77	0.51

Table 4. Row 1: Pearson's R correlation between the reproduction annotations and Gao and Wan [1]. Row 2: Pearson's R correlation between the full reproduction annotations and the ablation results.

5 Discussion

Discussion of the results — The results of the current study exhibit the same tendencies as those observed by Gao and Wan [1], thus effectively replicating the paper's main claims outlined in Section 2. However, it can be observed that in our reproduction, for all four dimensions, the reference summaries were given higher scores than in Gao and Wan [1]. The original authors note that the reference summaries often lack important information [1], which is a statement that our annotators agree with. This is reflected in the relatively low score on the dimension of *relevance* in both studies. Nevertheless, since despite the differences across the other dimensions we can observe the same general tendencies as the original authors, we can attribute this deviation to the subjectivity of the annotators and/or the ambiguity of the annotation guidelines. Another note to be made is that the annotators could become aware of 'good' or 'bad' models, due to the fixed order, and this may have resulted in them scoring the summaries differently than if the order had been randomized.

New annotation tool — While we have deviated from the exact approach of Gao and Wan [1] by utilizing a new annotation tool, the results of our ablation study show that this had no significant impact on our results. Thus, we recommend its implementation in future studies to increase the ease of annotation. Finally, reporting the exact code used to create such a tool contributes to ensuring reproducibility.

Annotation Noise Removal — Our final concern regarding the human evaluation process pertains to the authors' decision to treat the disagreement in the annotations as noise and remove it. We found this approach rather counterintuitive, as low agreement is

usually interpreted as a sign that refinement of the annotation guidelines is needed. Although the original paper does not reference the motivation underlying this approach, when contacted, the authors cited Bhandari et al. [36] as the inspiration. However, we found that this work displays some differences from the current study. In particular, the annotation process involves a binary label and four annotators, as opposed to a 5-point Likert scale and three annotators employed in our case. Thus, we believe that the suitability of the approach for this study design remains an open question.

What was easy — The reproducibility strengths of the original paper lie primarily in its profound methodological description. The rich and detailed incorporation of tables made the comparison with our reproduced results fairly easy.

What was difficult — The reimplementation of the original paper's code was relatively complex to navigate and required a fair amount of debugging when running the metrics. Certain deficiencies in the annotation guidelines also resulted in rather time-consuming decision-making. Finally, the methodological description of the post-processing of the annotations was relatively unclear and the code calculating the inter-annotator agreement was missing.

Recommendations for reproducibility — When it comes to the original paper's final conclusions, it can be argued that some claims were rather vaguely expressed and, therefore, it was challenging to judge whether they were successfully reproduced. For instance, Gao and Wan [1] concluded that very few metrics perform well across all dimensions. Regardless of its truthfulness, this argument requires a more fine-grained definition of efficient metric performance. By quantifying the latter and delimiting it inside a certain threshold, it would be substantially easier to compare our reproduction with the original results and make more confident conclusions.
Additionally, based on the annotators' reflections and the results in Section 4.2, most notably the low correlation obtained for the dimension of *coherence*, we believe that the annotation guidelines could benefit from a greater level of detail. Specifically, more fine-grained definitions and a section with examples of how to score ambiguous and borderline cases could increase the reproducibility of the task.

6 Communication with original authors

We contacted the first author of the original paper via email. They provided us with the exact annotation guidelines, the raw inter-annotator agreement scores before cleaning for our comparison in Section 4.2.1, and the missing code for conducting the noise removal. On all emails, we received swift replies and we would like to thank the authors for the correspondence.

References

1. M. Gao and X. Wan. "DialSummEval: Revisiting Summarization Evaluation for Dialogues." In: **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. 2022, pp. 5693–5709.
2. A. Belz, S. Agarwal, A. Shimorina, and E. Reiter. **A Systematic Review of Reproducibility Research in Natural Language Processing**. 2021.
3. M. Baker. "Reproducibility crisis." In: **Nature** 533.26 (2016), pp. 353–66.
4. A. Belz, A. Shimorina, S. Agarwal, and E. Reiter. "The ReproGen Shared Task on Reproducibility of Human Evaluations in NLG: Overview and Results." In: **Proceedings of the 14th International Conference on Natural Language Generation**. Aberdeen, Scotland, UK: Association for Computational Linguistics, Aug. 2021, pp. 249–258.

5. S. Gehrmann, E. Clark, and T. Sellam. "Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text." In: **arXiv preprint arXiv:2202.06935** (2022).
6. N. Iskender, T. Polzehl, and S. Möller. "Reliability of Human Evaluation for Text Summarization: Lessons Learned and Challenges Ahead." In: **Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)**. Online: Association for Computational Linguistics, Apr. 2021, pp. 86–96.
7. C.-Y. Lin and E. Hovy. "ROUGE: A Package for Automatic Evaluation of Summaries." In: **Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)**. 2004, pp. 74–81.
8. K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. "BLEU: a Method for Automatic Evaluation of Machine Translation." In: **Proceedings of the 40th annual meeting on association for computational linguistics**. 2002, pp. 311–318.
9. F. Yuan, Y. Liu, K. Toutanova, and Z. Chen. "BARTScore: Better Evaluation for Text Generation Models." In: **Transactions of the Association for Computational Linguistics** 9 (2021), pp. 243–257.
10. D. Deutsch and D. Roth. "Understanding the Extent to which Content Quality Metrics Measure the Information Quality of Summaries." In: **Proceedings of the 25th Conference on Computational Natural Language Learning**. Online: Association for Computational Linguistics, Nov. 2021, pp. 300–309.
11. D. Gliwa and H. Schutze. "SAMSum: Sentence-Aligned Multilingual Summarization Evaluation." In: **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**. 2019, pp. 3778–3788.
12. A. See, P. J. Liu, and C. D. Manning. "Get to the point: Summarization with pointer-generator networks." In: **arXiv preprint arXiv:1704.04368** (2017).
13. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is all you need." In: **Advances in neural information processing systems** 30 (2017).
14. M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension." In: **arXiv preprint arXiv:1910.13461** (2019).
15. J. Zhang, Y. Zhao, M. Saleh, and P. Liu. "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization." In: **International Conference on Machine Learning**. PMLR. 2020, pp. 11328–11339.
16. L. Dong, N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, and H.-W. Hon. "Unified language model pre-training for natural language understanding and generation." In: **Advances in Neural Information Processing Systems** 32 (2019).
17. C.-S. Wu, L. Liu, W. Liu, P. Stenetorp, and C. Xiong. "Controllable abstractive dialogue summarization with sketch supervision." In: **arXiv preprint arXiv:2105.14064** (2021).
18. A. R. Fabbri, F. Rahman, I. Rizvi, B. Wang, H. Li, Y. Mehdad, and D. Radev. "ConvoSumm: Conversation summarization benchmark and improved abstractive summarization with argument mining." In: **arXiv preprint arXiv:2106.00829** (2021).
19. J. Chen and D. Yang. "Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization." In: **arXiv preprint arXiv:2010.01672** (2020).
20. X. Feng, X. Feng, L. Qin, B. Qin, and T. Liu. "Language model as an annotator: Exploring DialoGPT for dialogue summarization." In: **arXiv preprint arXiv:2105.12544** (2021).
21. Z. Liu and N. F. Chen. "Controllable neural dialogue summarization with personal named entity planning." In: **arXiv preprint arXiv:2109.13070** (2021).
22. J. Chen and D. Yang. "Structure-aware abstractive conversation summarization via discourse and action graphs." In: **arXiv preprint arXiv:2104.08400** (2021).
23. S. Banerjee and A. Lavie. "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments." In: **Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)**. Ann Arbor, Michigan: Association for Computational Linguistics, June 2005, pp. 24–31.
24. J. Zhang, Y. Liu, J. Gao, K. Toutanova, and Z. Chen. "BERTScore: Evaluating Text Generation with BERT." In: **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**. Online: Association for Computational Linguistics, July 2020, pp. 4920–4926.
25. K. Zhao, Y. Wu, M. Eskenazi, and Y. Tian. "MoverScore: Text Generation Evaluation as Empirical Learning-to-Rank." In: **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3071–3081.
26. K. Clark, U. Khandelwal, O. Levy, and C. D. Manning. "Sentence-Mover: A Distance Metric for Sentence Embeddings." In: **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 1765–1775.
27. T. K. Landauer and S. T. Dumais. "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." In: **Psychological review** 104.2 (1997), p. 211.

28. G. Forgues, T. Lavergne, and J. Allan. "Automatic evaluation of summaries using vector extrema." In: **Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. 2014, pp. 36–46.
29. E. Durmus, V. Van Asch, A. Gatt, K. Georgila, and M.-F. Moens. "FEQA: Fine-Grained Factual Consistency Evaluation for Text Generation." In: **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**. 2020, pp. 6644–6653.
30. T. Scialom, L. Delabrouille, and T. Lavergne. "SummaQA: Answering Questions on Summaries using Pre-Trained Models." In: **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**. 2019, pp. 1687–1697.
31. T. Scialom, T. Lavergne, and L. Delabrouille. "QuestEval: A Simple QA-based Metric for Referenceless Evaluation of Text Generation." In: **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short and Student Papers)**. 2021, pp. 620–628.
32. D. Kryscinski, A. Burchardt, and L. Specia. "FactCC: Evaluating Factual Consistency of Generated Text." In: **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**. 2020, pp. 6654–6665.
33. A. Goyal and G. Durrett. "DAE: Dependency-Based Factual Consistency Evaluation for Text Generation." In: **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing**. 2020, pp. 6801–6812.
34. P. A. Rankel, J. M. Conroy, H. T. Dang, and A. Nenkova. "A Decade of Automatic Content Evaluation of News Summaries: Reassessing the State of the Art." In: **Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 131–136.
35. K. Krippendorff. **Content analysis: An introduction to its methodology (2 nd Thousand Oaks**. 2004.
36. M. Bhandari, P. N. Gour, A. Ashfaq, P. Liu, and G. Neubig. "Re-evaluating Evaluation in Text Summarization." In: **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Online: Association for Computational Linguistics, Nov. 2020, pp. 9347–9359.

7 Appendix

7.1 Annotation Guidelines

Guide for Dialogue Summarization Quality Evaluation
Guidelines by: M. Gao and X. Wan. "DialSummEval: Revisiting Summarization Evaluation for Dialogues." Alterations made for the present paper are preceded by "Addition:"

1. Instructions
In this task you will evaluate the quality of summaries written for a dialogue from daily life.
To correctly solve this task, follow these steps:
1. Carefully read this dialogue, be aware of the information it contains.
2. Read the proposed summaries A-N (14 in total).
3. Rate each summary on a scale from 1(worst) to 5(best) by its consistency, relevance, fluency, coherence.

2. Definitions

1. Relevance (1-5): The rating measures how well the summary captures the key points of the dialogue. Consider whether all and only the important aspects are contained in the summary.

2. Consistency (1-5): The rating measures whether the facts in the summary are consistent with the facts in the dialogue. Consider whether the summary does reproduce all facts accurately and does not make up untrue information.

3. Fluency (1-5): The rating measures the quality of individual sentences, are they well-written and grammatically correct. Consider the quality of individual sentences.

4. **Coherence (1-5)**: The rating measures the quality of all sentences collectively, to the fit together and sound naturally. Consider the quality of the summary as a whole. *Addition: If the summary consists of one sentence, it receives a coherence score of 4*

3. Error Examples

Extrinsic Hallucination: Content in summaries is not mentioned by dialogues. That is, the fact in the summary is neither supported nor contradicted by the source dialogue.

Example1: Dialogue
Luke: Hey sis, send me the pic of the parrot you painted yesterday?
Gina: file_photo
Gina: If you want better quality I need to send you PDF file.
Luke: It's ok. This parrot looks fantastic!!! I can't believe you've discovered your talent so late!
Gina: Haha thanks? fil_other Catch a PDF.
Luke: Thanks!

Summary
Luke sent his sister , Gina , a photo of a parrot she painted yesterday .

Example2:
Dialogue
Dave: Hey, is Nicky still at your place? Her phone is off
Sam: She just left
Dave: Thanks!

Summary
Nicky just left her phone at Dave ' s place .

Wrong References: Summaries contain information that is not faithful to the original dialogue, and associate one's actions/locations with a wrong speaker.

Example1:

Dialogue
Liam: will pick you up at 8
Liam: be ready
Kane: cool man
Liam: just don't get us late please

Summary
Liam will pick Liam up at 8 .

Example2:

Dialogue
Ola: Hey running late
Ola: I should be free by 8
Kurt: Sure no prob, call me

Summary
Ola will be late. Kurt will call him by 8.

Incorrect Reasoning: Summaries reasoned relations in dialogues incorrectly, thus came to wrong conclusions.

Example1:
Dialogue
Sean: Hey, I won't be able to take the car to the carwash
Sean: They want me to finish report first :(
Alice: shoot, but it's crazily dirty
Alice: Will we have tomorrow?
Sean: file_gif
Sean: We can leave a bit earlier or get it washed somewhere on the road
Alice: it might be good idea, let's do it tomorrow then
Sean: great!

Summary
Sean won't be able to take the car to the carwash tomorrow.

Example2:
Dialogue
Sophie: When r u going to Poanań?
Murphy: On Tuesday.
Sophie: And you're coming back the same day?
Murphy: Yes, in the afternoon, but I don't know the exact hour.

Summary
Murphy is going to Poanań on Tuesday and coming back the same day in the afternoon.

Grammatical Error: The grammar of the sentence is so wrong that it becomes meaningless.
Example: The Ebola vaccine accepted have already started.

4. A potential scoring criterion
Using consistency as an example
5 points: The facts in the summary are **all** consistent with the facts in the dialogue.
4 points: **A small part** of the facts in the summary are not consistent with the facts in the dialogue.
3 points: **Some of** the facts in the summary are consistent with the facts in the dialogue, while some are not.
2 points: **Most of the facts** in the summary are not consistent with the facts in the dialogue.
1 point: **None of the facts** in the summary are consistent with the facts in the dialogue.

7.2 Annotation Tool

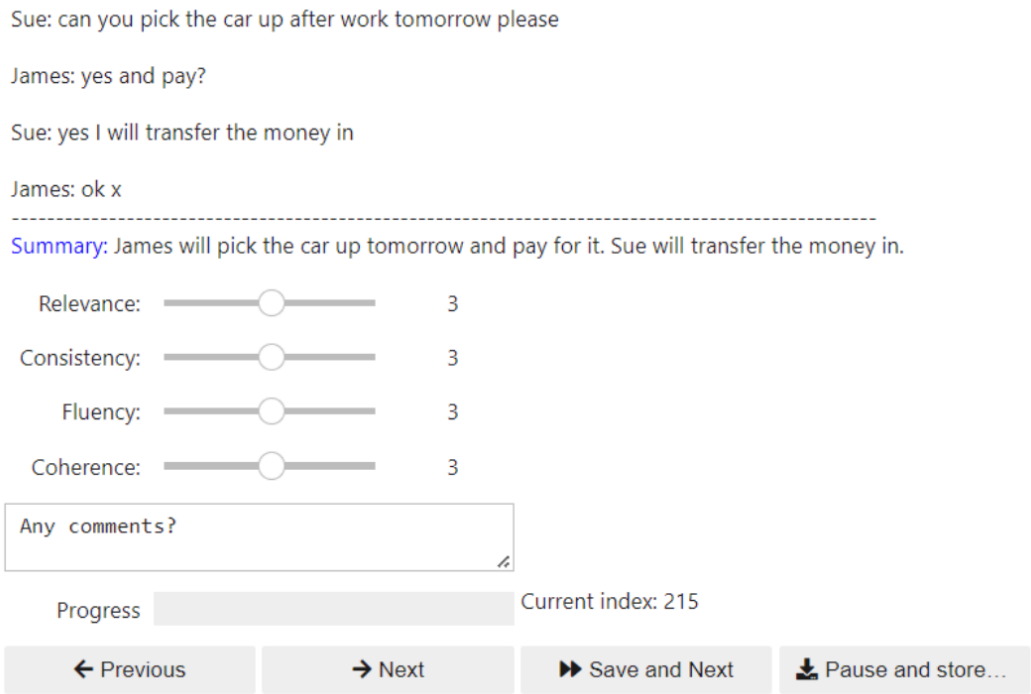

Figure 1. Example of the annotation tool developed for the current reproduction study: The dialogue is presented at the top, followed by the current model or reference summary in blue. The sliding cursors can be adjusted for each dimension on a scale from 1 to 5. Inside the box below them, the annotators can write comments to facilitate decision tracking. Directly below that, the *Current index* displays the number of data points annotated so far. The annotators can navigate the dataset by clicking on ←*Previous* and →*Next*, but without saving the scores of the current summary. To achieve the latter they have to click on *Save and Next*, while *Pause and store* can be used for taking a break, while ensuring that their progress is saved.

7.3 Numerical differences

	Coherence		Consistency		Fluency		Relevance	
Metrics	sys	sum	sys	sum	sys	sum	sys	sum
ROUGE-1	0.08	0.19	-0.04	-0.08	0.08	0.19	0.33	0.12
ROUGE-2	0.05	0.14	-0.05	-0.08	0.06	0.14	0.25	0.10
ROUGE-3	0.02	0.10	-0.05	-0.07	0.05	0.11	0.20	0.07
ROUGE-4	0.01	0.08	-0.06	-0.05	0.04	0.09	0.18	0.06
ROUGE-L	0.09	0.17	-0.04	-0.07	0.10	0.18	0.33	0.14
BERTScore-p	0.27	0.29	-0.02	-0.03	0.28	0.27	0.45	0.26
BERTScore-r	0.02	0.15	-0.05	-0.10	0.02	0.13	0.22	0.04
BERTScore-f1	0.16	0.24	-0.04	-0.07	0.17	0.22	0.36	0.17
MoverScore	0.07	0.17	-0.05	-0.08	0.09	0.16	0.29	0.10
SMS	0.00	0.06	-0.06	-0.06	0.03	0.07	0.16	0.05
BARTScore-s-h+	-0.53	-0.44	0.03	0.00	-0.50	-0.36	-0.34	-0.25
BARTScore-h	0.07	0.04	0.01	0.03	0.07	0.04	0.14	0.08
BARTScore-h-r	-0.02	0.15	-0.05	-0.14	-0.03	0.13	0.20	-0.01
BARTScore-r-h	0.18	0.18	-0.02	-0.02	0.18	0.19	0.43	0.21
BLANC-help+	-0.47	-0.35	-0.02	-0.09	-0.51	-0.35	-0.49	-0.39
BLANC-tune+	-0.45	-0.36	-0.02	-0.08	-0.50	-0.36	-0.50	-0.36
FEQA+	0.07	0.14	0.05	-0.01	0.08	0.14	0.38	0.12
QuestEval+	-0.40	-0.11	0.04	-0.09	-0.44	-0.12	-0.11	-0.13
SummaQA-conf+	-0.58	-0.31	-0.01	-0.06	-0.55	-0.25	-0.44	-0.26
SummaQA-fscore+	-0.48	-0.21	-0.03	-0.02	-0.53	-0.18	-0.46	-0.22
PPL-	0.30	0.07	0.01	0.05	0.32	0.07	0.08	0.13
CHRF	0.01	0.14	-0.05	-0.10	0.01	0.12	0.21	0.03
BLEU-1	0.06	0.18	-0.05	-0.09	0.07	0.15	0.21	0.05
BLEU-2	0.03	0.12	-0.06	-0.08	0.04	0.11	0.17	0.05
BLEU-3	0.02	0.10	-0.06	-0.07	0.04	0.09	0.15	0.05
BLEU-4	0.00	0.07	-0.06	-0.06	0.03	0.07	0.13	0.04
METEOR	0.01	0.12	-0.06	-0.09	0.02	0.11	0.19	0.05
EmbeddingAverage	0.35	0.36	-0.05	-0.11	0.25	0.29	0.44	0.12
VectorExtrema	0.09	0.21	-0.05	-0.08	0.10	0.18	0.29	0.12
GreedyMatching	0.13	0.21	-0.05	-0.11	0.11	0.19	0.29	0.07
FactCC+	-0.54	-0.41	-0.01	0.02	-0.50	-0.32	-0.49	-0.23
DAE+	-0.55	-0.46	0.01	-0.02	-0.52	-0.41	-0.50	-0.28

Table 5. Differences between the present paper and Gao and Wan [1] for the correlation between the metrics and human evaluations, represented numerically

Models	Coherence	Consistency	Fluency	Relevance	R-1	R-2	R-L
reference	0.26	0.55	0.32	0.08	0	0	0
LONGEST-3	-1.81	0.57	-1.12	-1.25	0	0	0
LEAD-3	-1.95	0.84	-1.10	-1.02	0	0	0
PGN	-0.08	0.15	-0.13	-0.59	0	0	0
Transformer	-0.12	-0.15	-0.01	-0.20	0	0	0
BART	0.06	0.62	0.32	-0.14	0	0	0
PEGASUS	-0.06	0.70	0.31	-0.21	0	0	0
UniLM	-0.01	0.51	0.20	-0.23	0	0	0
CODS	0.10	0.64	0.34	-0.11	0	0	0
ConvoSumm	0.12	0.74	0.34	-0.06	0	0	0
MV-BART	0.43	0.51	0.33	0.01	0	0	0
PLM-BART	0.27	0.62	0.25	-0.01	0	0	0
Ctrl-DiaSumm	0.41	0.51	0.31	-0.01	0	0	0
S-BART	0.29	0.57	0.41	-0.12	0	0	0

Table 6. Numerical difference between average human rating between Gao and Wan [1] and the present paper