Reliability of human evaluation for text summarization: Lessons learned and challenges ahead

Neslihan Iskender, Tim Polzehl, Sebastian Möller
Technische Universität Berlin, Quality and Usability Lab
{neslihan.iskender, tim.polzehl1, sebastian.moeller}@tu-berlin.de

Abstract

Only a small portion of research papers with human evaluation for text summarization provide information about the participant demographics, task design, and experiment protocol. Additionally, many researchers use human evaluation as gold standard without questioning the reliability or investigating the factors that might affect the reliability of the human evaluation. As a result, there is a lack of best practices for reliable human summarization evaluation grounded by empirical evidence. To investigate human evaluation reliability, we conduct a series of human evaluation experiments, provide an overview of participant demographics, task design, experimental set-up and compare the results from different experiments. Based on our empirical analysis, we provide guidelines to ensure the reliability of expert and non-expert evaluations, and we determine the factors that might affect the reliability of the human evaluation.

1 Introduction

Evaluation of summarization quality plays a crucial role in the development of summarization tools since a well-executed evaluation can help to determine whether the system has adequately outperformed the existing tools in terms of quality and speed or whether the designed properties work as intended (van der Lee et al., 2018; Lloret et al., 2018). The human evaluation has been the most trusted evaluation method and used as gold standard for summarization evaluation (Gatt and Krahmer, 2018; Celikyilmaz et al., 2020). However, in recent years, some researchers have provided an extensive overview of papers with human evaluation and pointed out that there is a lack of standardized procedures leading to mostly non-comparable and non-reproducible results (van der Lee et al., 2019; Belz et al., 2020; Howcroft et al., 2020; van der Lee et al., 2021).

Howcroft et al. (2020) have reported based on the analysis 165 papers with human evaluation published in INLG and ENLG that more than 200 different terms have been used for human evaluation, which results in lack of clarity in reports and extreme diversity in approaches. van der Lee et al. (2021) have analyzed 304 research papers published in INLG and ACL conferences and reported that only 3% of 304 analyzed papers described the demographics, 6% provided the details about task design, 19% reported any inter-rater agreement score, 23% conducted a statistical analysis for human evaluation, and 32% reported the number of different evaluators per item, where 92% of the reported cases only one rating is used.

In this paper, we aim to contribute the human evaluation research as follows: 1) we conduct series of human evaluation with experts, crowd, and laboratory participants on two different data sets, 2) we report on the participant demographics, task design, and evaluation criteria 3) we demonstrate a comprehensive statistical analysis of human experiments, and 4) we provide guidelines to ensure the reliability of experts and non-experts and determine the factors affecting the human reliability grounded by the empirical evidence from our experiments. Data associated with this work is available at https://github.com/nesliskender/reliability_humeval_summarization.

2 Related Work

Human evaluation of text summarization can be conducted either by linguistic experts or non-experts such as laboratory participants or crowd workers. However, expert evaluation has been established as the gold standard in the summarization evaluation and the reliability of non-experts has been repeatedly questioned (Lloret et al., 2018).

Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval), pages 86–96
Online, April 19, 2021. ©2021 Association for Computational Linguistics

Gillick and Liu (2010) have conducted a crowdsourcing experiment for summarization evaluation for the first time and concluded that crowd workers can not evaluate summary quality because of the non-correlation with experts. However, they did not report the number of crowd workers per summary. Fabbri et al. (2020) have compared the crowd ratings with expert ratings using five crowd workers per item. They have also reported that crowd ratings do not correlate with experts and emphasized the need for protocols for improving the human evaluation of summarization. Further, Gao et al. (2018); Falke et al. (2017); Fan et al. (2018) have used crowd workers to evaluate the quality of their automatic summarization systems without questioning the reliability of crowd workers.

When we look at the approaches used for human summarization evaluation, they can be broadly classified into two categories: intrinsic and extrinsic evaluation (Jones and Galliers, 1996; Belz and Reiter, 2006; Steinberger and Ježek, 2012). In intrinsic evaluation, the summarization output's quality is measured based on the summary itself without considering the source text. Generally, it has been carried out as a pair comparison (compared to expert summaries) or using absolute scales without showing a reference summary (Jones and Galliers, 1996). However, the extrinsic evaluation, called also *task-based evaluation*, aims to measure the summary's impact on the completion of some task based on the source document (Mani, 2001). Reiter and Belz (2009) have argued that the extrinsic evaluation is more useful than intrinsic because the summarization systems are developed to satisfy the information need from the source text in a condensed way, but van der Lee et al. (2021) have reported that only 3% of the papers presented an extrinsic evaluation.

Further, the quality criteria used in the human evaluation and the terminology used for describing these criteria had a high degree of variation, 200+ variations in terminology (Howcroft et al., 2020). Researchers have used either the same terminology but evaluated something different or used different terminology but measured the same thing (Belz et al., 2020). In most cases, they did not define the quality criteria they investigated or cite a reference for it, making it difficult to compare the results and draw conclusions across the papers. The scales for evaluation have also varied often, such as Likert (3, 4, 5, 6, 7, 10, 11-point), categorical choice (Yes or No), or rank-based scale (van der Lee et al., 2021).

So, human evaluation lacks structured, reliable evaluation practices, and the current way of reporting human evaluation in research papers generates non-comparable and non-reproducible results. We aim to contribute to human evaluation research for text summarization by determining the intrinsic and extrinsic quality in a reliable and reproducible way with our experiments in section 3.

3 Experiments

As our source documents, we used the 67 unique post-query pairs from a telecommunication company's customer service forum in German, where customers ask questions about the company's products and services such as "*Where can I find my customer number*" or "*My internet is not working*". Each query had 6-10 corresponding forum posts, including the answers from other customers to provide a solution or at least some help to the customer problem. The average word count of the posts was 571.2, the shortest one with 150 words, and the longest one with 1006 words, where the average word count of the corresponding queries was 9.1, the shortest query with three, and the longest with 23 words.

We conducted series of human experiments on this data set shown in Table 1 in chronological order. In experiment 1, crowd workers created extractive summaries for 67 post-query pairs. In experiment 2, different crowd workers evaluated the quality of crowd-generated summaries, the output from experiment 1. Because of the high cost of human evaluation, we limited our evaluation data set for further experiments based on the overall quality ratings from experiment 2. From those, we selected 50 summaries within ten distinct quality groups ranging from lowest to highest scores (lowest group [1.667, 2]; highest group (4.667, 5]), each represented by five summaries. We generated a stratified sample of the data set consisting of summaries with low, medium, and high quality. These summaries originated from 27 post-query pairs.

This new data set, 27 post-query pairs with 50 summaries in varying qualities, has been evaluated by experts in experiment 3, by crowd workers in experiment 4, and by laboratory participants in experiment 5. In these experiments, the task design and the summaries were exactly the same to compare the effect of expertise (expert vs. non-expert) and environment (lab vs. crowd) on the

Exp. No	Type	Human	Items	#Evaluator per Item	#Total Evaluator	Average Age	Gender	Payment
1	Creation	Crowd	67 post-query pair	4	76	39.43	41m, 35f	1.2 € per task
2	Evaluation	Crowd	256 summaries (output from 1.exp)	3	86	38.8	49m, 37f	1.2 € per task
3	Evaluation	Expert	Selected 50 summ. from 1.exp output	2	2	26.5	2f	30 € per hour
4	Evaluation	Crowd	Same as in 3.exp	24	46	42.47	27m, 19f	1.2 € per task
5	Evaluation	Lab	Same as in 3.exp	24	71	29.30	38m, 33f	15 € per hour
6	Creation	Expert	27 post-query pair	2	2	26.5	2f	30 € per hour
7	Evaluation	Expert	TextRank summ. of 27 post-query pair	2	2	26.5	2f	30 € per hour
8	Evaluation	Crowd	Same as in 7.exp	10	21	28.4	15m, 6f	1.2 € per task

Table 1: Overview of all human experiments

quality assessment. Further, we created machine summaries for the same 27 post-query pairs using the sumy[1] library to investigate the effect of summary generation method (human vs. machine) on the quality assessment. We applied TextRank algorithm (Mihalcea and Tarau, 2004) for machine summarization since it is one of the limited open-source German summarization algorithm and the most used unsupervised baseline in text summarization (Allahyari et al., 2017). Experts have evaluated these machine summaries in experiment 7, crowd workers evaluated the summaries in experiment 8. Here, we did not ask laboratory participants to evaluate the machine summaries' quality since the comparisons of experiments 3, 4, and 5 revealed the insights regarding the environment's effect on the quality assessment. The experts also created the gold standard summaries for these 27 post-query pairs in experiment 6.

In human evaluation experiments, we applied both intrinsic and extrinsic approaches. As the literature reveals a high degree of variation in quality criteria used in human experiments (Belz et al., 2020; Howcroft et al., 2020; van der Lee et al., 2021), we limited the intrinsic factors to six and the extrinsic factors to three. As the limitation criteria, we narrowed the scope of human evaluation from NLG to text summarization and adopted the commonly used quality metrics. Especially, we applied the criteria from the Document Understanding Conferences (DUC[2]), which have been the forum for researchers in text summarization to compare methods and results. Additionally, we used a measure for overall quality to assess the summaries' total quality. While limiting the extrinsic quality factors, we focused on quality metrics for usefulness for the task and information need because these are the most commonly used criteria in NLG as reported in (Howcroft et al., 2020).

So, we determined intrinsic quality using six different quality criteria: overall quality, defined as "responsiveness evaluation" in Louis and Nenkova (2013), and the five readability (linguistic) measures (grammaticality, non-redundancy, referential clarity, focus, and structure & coherence) defined as in Dang (2005). We evaluated the extrinsic quality using following three measures: summary usefulness defined as "content responsiveness" in Conroy and Dang (2008), source usefulness (in our case post usefulness, because our source documents are forum posts) defined as "relevance assessment" in Mani et al. (2002), and summary informativeness defined as "informativeness" in Mani et al. (2002). We conducted all our evaluations using a continuous scale, 5-point Mean Opinion Score (MOS) with the labels *very good, good, moderate, bad, very bad*, which is one of the most applied scales in subjective quality assessment (Streijl et al., 2016).

3.1 Crowdsourcing Experiments

We conducted all of the crowdsourcing experiments using Crowdee[3] platform. Before each of our crowdsourcing experiment, we had test runs with the student workers who have acted like crowd workers and gave us feedback regarding the task design and understandability. For each new crowdsourcing experiment, we did at least ten or more alterations based on the students' feedback. Further, we payed the minimum hourly wage in Germany and determined payment based on our crowdsourcing experiments' estimated work duration.

3.1.1 Crowd Worker Selection

For crowd worker selection, we developed a two-step qualification process for both crowd creation and evaluation. In the first step, crowd workers needed to pass the German language proficiency test provided by the Crowdee platform with a score

[1] https://github.com/miso-belica/sumy

[2] https://duc.nist.gov/

[3] https://www.crowdee.com/

of 0.9 and above (scale [0, 1]). In the second step, crowd workers needed to pass a semantic task-specific pre-qualification test.

In the pre-qualification test for summary creation, at first, we presented the summary creation guidelines: 1) Summary should be non-redundant, fluent, informative, and grammatically correct, 2) Summary should be readable and understandable, 3) Summary should be created by copy-pasting 3-5 sentences from forum posts, 4) Any alternation of the sentences and also writing new sentences were not allowed. We also presented an example of a good and bad summary generated for the same post-question pair. 103 out of 144 crowd workers were approved for the summary creation task. The criterion for approval was the ROUGE score of crowd workers' summaries, calculated with summaries created by linguists of the authors' team. Further, we manually evaluated the crowd worker's summaries with a low ROUGE score (ROUGE-1 < 0.4), and if the summary quality was still acceptable, their authors were approved.

In the pre-qualification test for summary evaluation, we gave a brief explanation of the summarization process, highlighting that the summaries were created by simple cutting-out sentences from forum multiple posts, and therefore may appear slightly unnatural. Crowd workers were then asked to evaluate the overall quality of four summaries (two very good, two very bad). The quality of these summaries have already been determined by the linguists of the authors' team on a 5-point MOS scale. For each exact rating match, crowd workers got 4 points, and for each point deviation, they got a point less, so deviations were linearly punished. 98 out of 150 crowd workers passed this qualification test with a point ratio $>= 0.625$.

3.1.2 Crowd Creation

In experiment 1, we instructed the crowd workers to create one extractive, 3-5 sentences long summary for each post-query pair using the same summary creation guidelines as in the pre-qualification test. To illustrate the guidelines, we presented crowd workers an example of a post-query pair and corresponding one good and one bad summary. Additionally, forum posts were shown as an itemized list of sentences in the creation process, so that each crowd worker only had to select and copy the specified sentences into a summary. Overall 76 unique crowd workers (41m, 35f, $M_{age} = 39.43$) participated in the experiment 1. Four different crowd workers per post-query pair created 256 summaries for 67 post-query pairs after eliminating cheaters. The average work duration was 458.8 seconds, and total tasks (67 x 4) were completed in 46 hours.

3.1.3 Crowd Evaluation

In experiment 2, the crowd workers evaluated the quality of 256 crowd summaries generated in experiment 1. First, a brief explanation of the summary creation process was shown with an example of a query, forum posts, and a summary to provide background information. Next, the crowd workers were asked to evaluate two summaries regarding the overall quality and the five intrinsic quality measures in the following order: 1) overall quality, 2) grammaticality, 3) referential clarity, 4) non-redundancy, 5) focus and 6) structure & coherence. Three different crowd workers evaluated each summary, and a single crowdsourcing task included the evaluation of two summaries.

The overall quality was rated first to avoid influencing it by more detailed aspects. The evaluation of each aspect was done on a separated page, which contained a definition of the particular aspect (illustrated with an example), a summary, and a 5-point MOS scale (*very good, good, moderate, bad, very bad*) as radio buttons. To have an intrinsic (summary-focused) evaluation, crowd workers did not see the corresponding original post-query pair. Overall 86 crowd worker (49m, 37f, $M_{age} = 38.8$) completed the summary evaluation task with an average work duration of 356.36 seconds within 12 days. We noticed that conducting a crowdsourcing experiment at Christmas time has slowed the total task completion duration. Further, crowd workers had the chance to give some feedback at the end of the task, and multiple crowd workers commented about the summary content, such as "I don't find the summary very informative overall, so the overall rating was worse than the individual ratings.".

Therefore, we added questions regarding the summary's content quality to experiment 4. We used the same instructions and task description as in experiment 2 and added three extrinsic quality measures showing the original corresponding post-query pair to evaluate the summary's content quality. Also, we increased the number of unique crowd workers to 24 for each summary following the recommendations of Naderi et al. (2018) for a robust crowdsourcing study. Since reading the summary and all the source text increases the reading effort, we asked crowd workers to rate the quality of one summary in one task.

After answering the same six questions explained in the above paragraphs, we asked crowd workers to evaluate the following extrinsic quality measures: 7) summary usefulness, 8) post usefulness, 9) summary informativeness. Again, the evaluation of each aspect was done on a separate page, which contained the definition of the particular aspect with an example, the post-query pair, the summary, and the answer options as the 5-point MOS scale. Overall, 46 crowd workers (19f, 27m, $M_{age} = 43$) completed the evaluation of selected 50 summary with an average work duration of 249.88 seconds. The total of 1200 tasks (50 summary x 24 crowd worker) was published in batches, and each batch was completed within one day.

In our last crowdsourcing experiment, experiment 8, we asked crowd workers to evaluate the quality of 27 TextRank summaries using the same task design as in experiment 4. Overall, 21 crowd workers (15m, 6f, $M_{age} = 26.3$) participated in experiment 8 with an average task completion duration of 287.92 seconds, completing total tasks within three days. Our analysis from experiments 3 and 4 has shown that 8-10 crowd workers per summary delivers results corresponding to laboratory experiments. Therefore, we collected evaluations from 10 different crowd workers per summary.

3.2 Laboratory Experiment

In experiment 5, we recruited participants via a local participant pool for the summary quality evaluation experiment in a controlled laboratory environment. We accepted only the native German speakers and did not perform any other pre-qualification. The experiment design and the summaries were exactly the same as in experiment 4, where 24 different laboratory participants evaluated the nine different quality aspects of 50 summaries. They also completed the task using Crowdee platform to avoid any user interface biases.

In addition to instructions of experiment 4, all the participants were also instructed in written form before the experiment start and all of the participant's questions regarding the task's understandability were answered immediately by the lab instructor. As expected, the participants were also physically present in a controlled laboratory environment during the task. The experiment duration was set to one hour, and the participants were asked to evaluate as many summaries as they can in an hour. Overall, 71 participants (38m, 33f, $M_{age} = 29.3$) completed the experiment 5, evaluating 12 summaries per hour on average within 51 days.

3.3 Expert Experiments

In experiment 3, two experts who are Masters students in linguistics evaluated the same selected 50 summaries with the same task design as in experiment 4. At first, they evaluated the summaries separately using Crowdee platform. After the first separate evaluation round, the inter-rater agreement scores, Cohen's κ, showed that the experts often diverted in their assessment. To reach consensus among experts, we followed an iterative approach similar to the Delphi method (Linstone et al., 1975) and arranged physical follow-up meetings with experts which we refer as mediation meetings.

In these meetings, experts discussed the reasons and backgrounds of their ratings for each summary in case of disagreement and eventually aligned in case of consensus. Eventually, acceptable inter-rater agreement scores were reached for nine quality measures. One should keep in mind that elaborated follow-up meetings principally lead to the increasing convergence of expert ratings. We did not test for a saturation effect with this observation, but the effort allocated in this step clearly influences the expert rating values.

In experiment 6, the same experts created gold standard summaries for the corresponding source post-query pairs of 27 TextRank summaries using the same task design as in experiment 1. Lastly, in experiment 7, the same experts evaluated the quality of 27 TextRank summaries following the same iterative approach and same task design as in experiment 3.

4 Results

Results are presented for the mean opinion scores (MOS) of overall quality (OQ), grammaticality (GR), non-redundancy (NR), referential clarity (RC), focus (FO), structure & coherence (SC), summary usefulness (SU), post usefulness (PU) and summary informativeness (SI) collected in experiments 2, 3, 4, 5, 7, and 8 (see table 1). We will refer to these measurements by their abbreviations in this section. Further, we use non-parametric statistics in our analysis because of the non-normal distribution of some measurements in these experiments.

	Before Mediation				After Mediation			
	Crowd Summ.		TextRank Summ.		Crowd Summ.		TextRank Summ.	
	Agr. in %	κ	Agr. in %	κ	Agr. in %	κ	Agr. in %	κ
OQ	54	0.228	22.2	-0.040	82	0.637	85.2	0.717
GR	42	0.078	18.5	0.086	78	0.626	88.9	0.809
NR	34	-0.012	11.1	-0.084	70	0.520	85.2	0.797
RC	56	0.381	29.6	0.013	88	0.819	92.6	0.882
FO	52	0.249	88.9	0.779	80	0.685	96.3	0.922
SC	42	0.212	22.2	0.070	82	0.743	85.2	0.783
SU	44	0.220	37	0.093	76	0.635	88.9	0.839
PU	38	0.005	48.1	0.169	70	0.469	92.6	0.856
SI	34	-0.038	40.7	0.234	78	0.565	92.6	0.886

Table 2: Raw agreement in % and Cohen's κ scores between two experts for the evaluation of crowd summaries and TextRank summaries before mediation and after mediation

4.1 Reliability of Human Evaluation

4.1.1 Expert Evaluation

In this section, we compare the results from experiment 3 with experiment 7 to analyze expert reliability. Following the recommendations of van der Lee et al. (2019), we calculated the raw agreement in percentage and Cohen's κ as inter-rater agreement scores.

Looking at Table 2, we observe that the mediation meetings increased the agreement scores enormously both for the evaluation of crowd and TextRank summaries. Only after the mediation meetings, acceptable Cohen's κ scores between experts could be achieved with all measures having substantial (0.6-0.8] or almost perfect agreement (0.80-1.0] for all measures except for NR, PU, and SI being weak in crowd summary evaluation (0.40-0.60] (Landis and Koch, 1977).

For TextRank summaries, the increase is considerably higher than the crowd summaries. Since the same experts evaluated the TextRank summaries under the same experimental conditions as in experiment 3, we can conclude that the characteristics of machine-generated summaries such as unnaturalness or non-fluency constitute a challenge even for experts before mediation. Further, the TextRank summaries included usually same kind of mistakes which made it easier for experts to agree on a specific evaluation scheme for each evaluation criteria during mediation sessions, leading to higeher agreement in comparison to crowd summaries.

The effect of mediation on the inter-rater agreement scores shows clearly that the mediation meetings are necessary for reliable expert evaluation, especially when evaluating machine-generated summaries. We plan to use the specific evaluation criteria shaped during expert mediation sessions to improve the task design in future work.

4.1.2 Crowd Evaluation

This section compares the results from experiment 2 with experiment 4 to measure the re-test reliability of crowd experiments. To do so, we calculated the Spearman correlations between the crowd evaluations from experiment 2 (3 crowd workers per item) and experiment 4 (24 crowd workers per item) for the six intrinsic measures. To have the same number of crowd workers per summary as in experiment 2, we selected the first three evaluations per summary from experiment 4. The black circles in Figure 1a show the correlation between these first three crowd evaluations from experiment 4 and crowd evaluation from experiment 2. The correlation coefficients range from 0.497 to 0.587 for all six measures, indicating a moderate re-test reliability of crowd evaluation.

However, choosing the first 3 out of 24 crowd raters for correlation analysis is neither a conscious nor reliable choice. Would we still get the same correlations if some of the remaining 21 crowd workers would have completed the task before the first three considered above? To investigate this, we randomized 100 times the order of 24 crowd evaluations and selected the first three evaluations to correlate them with the evaluation from experiment 2. Figure 1a shows the scatter plots for these correlations, ranging from weak to strong for all six measures. We see a noticeable difference between the initial correlations (black circles in Figure 1a) and randomizations. Here, we observed that the correlations ranged from 0.2 to 0.75, showing that

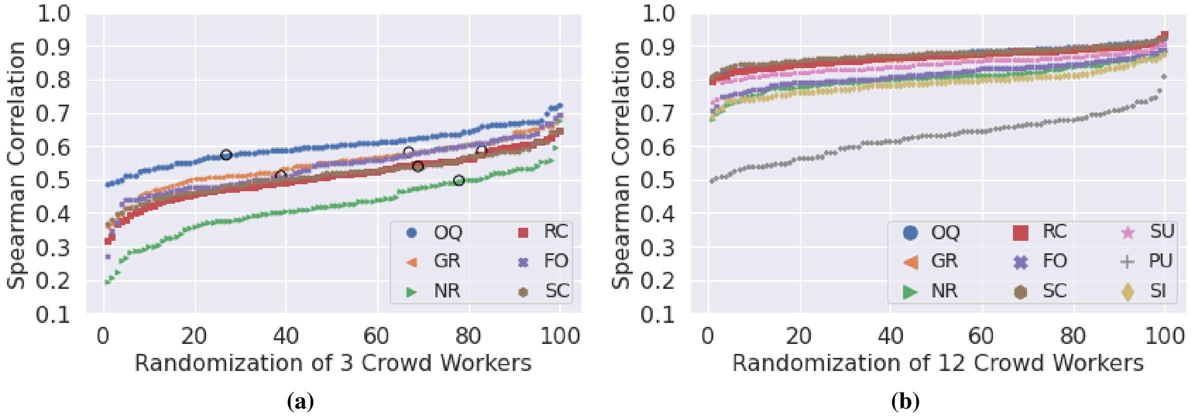

Figure 1: Spearman correlations of crowd evaluations from experiment 4 as 100 randomized groups of 3 crowd workers with crowd evaluations from experiment 2 (a) and Spearman correlations of crowd evaluations from experiment 4 as 100 randomized groups of 12 crowd workers with the remaining 12 crowd workers (b)

the crowdsourcing experiments with three crowd workers per summary still include high degree of unpredictability and can only be moderately reliable.

If we increase the number of crowd workers per item, can we overcome this unpredictability? To investigate this, we divided the existing data from experiment 4 into two random groups, two groups each with 12 crowd workers per item, and calculated Spearman correlations between them. Figure 1b shows the correlation between these two randomized groups for the nine quality measures. In comparison to Figure 1a, the slope of randomized correlations in Figure 1b is lower and the mean correlation of randomizations is very strong except for PU and SI which are strong ($\rho_{OQ} = 0.874$, $\rho_{GR} = 0.858$, $\rho_{NR} = 0.799$, $\rho_{RC} = 0.857$, $\rho_{FO} = 0.815$, $\rho_{SC} = 0.874$, $\rho_{SU} = 0.848$, $\rho_{PU} = 0.626$, $\rho_{SI} = 0.793$).

This result proves that the reliability of crowdsourcing experiments depends on the number of crowd workers per item and reliable crowdsourcing results cannot be achieved with three crowd workers per item.

4.2 Effect of Expertise and Environment

To investigate the effect of expertise and environment on the human summarization evaluation, we compare the results from experts (experiment 3), crowdsourcing (experiment 4), and laboratory (experiment 5) experiments, which are conducted on the same data set with the same task design.

Figure 2 shows the boxplots of expert, crowd, and laboratory ratings for nine quality measures. Here, we see that the experts used the upper end of the scale more often than the non-experts and gave higher ratings on average. Further, the non-expert evaluations are slightly negatively skewed using a smaller portion of the scale.

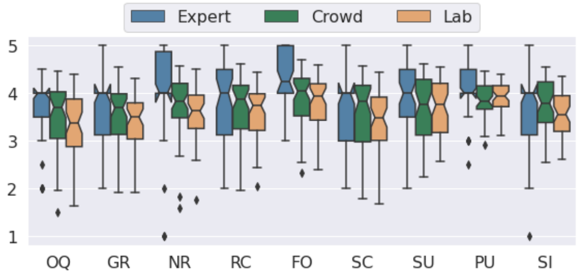

Figure 2: Boxplots of expert evaluations (blue), crowd evaluations (green) and laboratory evaluations (orange) for crowd summaries

To explore if these differences statistically significant, we calculated the non-parametric ANOVA, Kruskal-Wallis Test, between expert, crowd, and laboratory ratings. The test results revealed no significant difference between the expert and crowd evaluations except for PU and between the crowd and laboratory except for SI. However, the expert evaluations differed significantly from laboratory evaluations. Experts gave significantly higher ratings than the laboratory participants for all measures except for SU and SI. Here, we observe that significant differences exist only between the intrinsic evaluations indicating that the intrinsic evaluations require more expertise than the extrinsic evaluation.

Additionally, we calculated the Spearman correlations of expert evaluations with crowd and laboratory for all nine measures as shown in Figure 3. We found that the correlation magnitudes between ex-

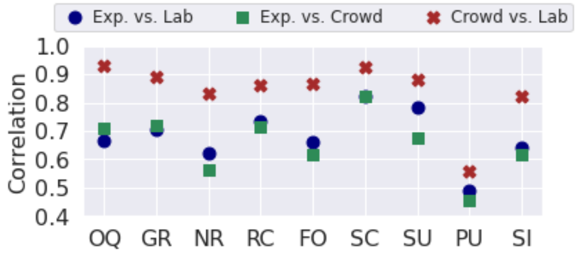

Figure 3: Spearman correlations between expert and laboratory, expert and crowd, and crowd and laboratory for the nine quality measures

pert and laboratory and between expert and crowd were very similar, ranging from moderate to very strong. However, the correlations between crowd and lab were very strong except for PU and remarkably higher than the correlations with experts. These results show that the environment does not have a significant effect on human evaluation, but the level of expertise affects the human evaluation.

4.3 Effect of Data Quality

To analyze the effect of the data quality itself on human evaluation, we compare the correlations between expert (experiment 3) and crowd (experiment 4) for crowd-generated summaries with the correlation between expert (experiment 7) and crowd (experiment 8) for TextRank-generated summaries. On average, the correlations for TextRank summaries for nine quality measures were 0.12 points lower than the crowd summaries. To determine if this is a significant difference, we applied Zou's confidence intervals test for independent variables (Zou, 2007) and found out that the differences were not statistically significant except for SC.

Further, we calculated non-parametric T-test, the Mann-Whitney U test, between crowd and expert ratings for TextRank summaries. The results revealed that the crowd workers rated OQ, RC, FO, SU, and PU of TextRank summaries significantly lower than the experts. In contrast, when evaluating crowd summaries, crowd ratings did not differ significantly from experts except for PU. This result indicates that crowd workers tend to give lower ratings than the experts for machine-generated summaries. However, the summary generation method does not affect the rank-order of their ratings, and the correlation between crowd and expert do not differ from each other significantly both for human- and machine-generated summaries.

4.4 Goodness of Automatic Metrics: With whom to compare?

The goodness of automatic summarization evaluation metrics is generally measured by their correlation to human evaluations, usually expert evaluations (Bhandari et al., 2020). In this section, we compare the correlations of commonly used automatic metrics ROUGE (Lin, 2004) and BERTScore (Zhang* et al., 2020) with expert and crowd evaluations for TextRank summaries to find out if the crowd workers can be used instead of experts.

	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
OQ$_{\text{ex}}$	0.636	0.680	0.675	0.582
OQ$_{\text{cr}}$	0.576	0.526	0.499	0.552
SU$_{\text{ex}}$	0.467	NS	0.397	NS
SU$_{\text{cr}}$	0.657	0.586	0.592	0.614
SI$_{\text{ex}}$	0.542	0.546	0.527	0.501
SI$_{\text{cr}}$	0.421	0.506	0.504	0.424

NS: Not significant

Table 3: Spearman correlations of ROUGE-1, ROUGE-2, ROUGE-L and BERTScore with expert and crowd evaluations for TextRank summaries

As human evaluation measures, we only considered the OQ, SU, and SI because the automatic metrics are content-based metrics and should rather be compared to content-based human evaluations (Lloret et al., 2013). Table 3 shows the correlations of ROUGE and BERTScore with OQ, SU, and SI measured by experts and crowd. To determine if these correlation differences are significant, we applied Zou's confidence intervals test for overlapping dependent variables and found out that there is no significant difference between any correlation. This result indicates that crowd workers can be used instead of experts to determine the goodness of automatic metrics.

5 Conclusion and Future Work

In this paper, we report a comparative analysis of series of human evaluation experiments with crowd workers, laboratory participants, and experts on two different data sets to determine the reliability of human evaluation for text summarization.

However, the research papers with expert evaluations for summarization have not reported any mediation meetings, let alone only 19 % reported the inter-rater agreement scores in the range of 0.3-0.5 (van der Lee et al., 2021). This raises the question of expert reliability, and to avoid that, we recommend having mediation meetings with experts for reliable expert evaluation based on our results in section 4.1.1. With our analysis, we showed that mediation meetings are elementary to assure the reliability of expert evaluations for all quality measures.

Further, we found out that the number of crowd workers per item determines the crowd evaluation's reliability. van der Lee et al. (2021) showed only 57 % of papers specified number of evaluators and the median was 3 among the papers which have reported the evaluator number. But our analysis in Section 4.1.2 showed that when using crowdsourcing, three crowd workers per item can only deliver moderately reliable results and around ten or more different crowd workers should evaluate each summary. This result is also inline with our previous findings in Iskender et al. (2020b,a).

While the environment (crowd vs. lab) does not affect the human evaluations, the level of expertise might have affected the human evaluation. Although there are mostly strong correlations between the experts and non-experts, their evaluations do not match 100%. Depending on the evaluation aim or the end-user group of the summarization system, the evaluator's expertise should be determined, e.g., summarization systems developed for naive end-users should be evaluated by the naive end-users rather than the experts, and expert systems should be evaluated by linguistic experts.

Additionally, the summary generation method (human vs. machine) might cause a bias in crowd assessments. Because of machine summaries' unnaturalness, the crowd workers tended to rate machine summaries lower than the experts. The feedback that the summaries were very "unnatural" and "robotic" from the crowd workers in experiment 8 also confirms this finding. But still, crowd workers can be used as a direct substitute for experts to determine the goodness of automatic evaluation metrics developed for machine summaries.

However, this paper has some limitations regarding the data set and task design. We used one task design with a single rating scale (5-point MOS scale) and the same set of definitions and explanations for our evaluation criteria in all our experiments, which were conducted on small sized data sets. In future work, we plan to include different human evaluation criteria, compare different rating scales with each other, conduct A/B testing with a second task design, which includes improved definitions of evaluation criteria based on the expert mediation sessions, and expand the data set size to increase the statistical power of our analysis. Additionally, we plan to conduct virtual mediation sessions between two or three crowd workers to find out if we can reach similar results to experts with a small number of crowd workers.

Despite the limitations of our paper, we believe that this paper makes a significant contribution to human evaluation research of text summarization. As Table 1 demonstrates, the time and organizational efforts and the cost of human experiments can be enormous. Especially, conducting laboratory and expert experiments required high organizational effort, and these experiments were completed in months while crowdsourcing experiments usually were finished in a couple of days. This shows how burdensome and time-consuming conducting human evaluation can be, which is a great challenge in a fast-moving field like summarization. Therefore, finding reliable ways of using crowdsourcing can be a promising solution and we hope to see more research in this field.

References

Mehdi Allahyari, Seyed Amin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut. 2017. Text summarization techniques: A brief survey. *CoRR*, abs/1707.02268.

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.

Anya Belz, Simon Mille, and David M. Howcroft. 2020. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 183–194, Dublin, Ireland. Association for Computational Linguistics.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey.

John M. Conroy and Hoa Trang Dang. 2008. Mind the gap: Dangers of divorcing evaluations of summary content from linguistic quality. In *Proceedings* *of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 145–152, Manchester, UK. Coling 2008 Organizing Committee.

Hoa Trang Dang. 2005. Overview of duc 2005. In *Proceedings of the document understanding conference*, volume 2005, pages 1–12.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2020. Summeval: Re-evaluating summarization evaluation.

Tobias Falke, Christian M. Meyer, and Iryna Gurevych. 2017. Concept-map-based multi-document summarization using concept coreference resolution and global importance optimization. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 801–811, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Angela Fan, David Grangier, and Michael Auli. 2018. Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 45–54, Melbourne, Australia. Association for Computational Linguistics.

Yang Gao, Christian M. Meyer, and Iryna Gurevych. 2018. APRIL: Interactively learning to summarise by combining active preference learning and reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4120–4130, Brussels, Belgium. Association for Computational Linguistics.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *J. Artif. Int. Res.*, 61(1):65–170.

Dan Gillick and Yang Liu. 2010. Non-expert evaluation of summarization systems is risky. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 148–151, Los Angeles. Association for Computational Linguistics.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Neslihan Iskender, Tim Polzehl, and Sebastian Möller. 2020a. Best practices for crowd-based evaluation of German summarization: Comparing crowd, expert and automatic evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 164–175, Online. Association for Computational Linguistics.

Neslihan Iskender, Tim Polzehl, and Sebastian Möller. 2020b. Towards a reliable and robust methodology for crowd-based subjective quality assessment of query-based extractive text summarization. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 245–253, Marseille, France. European Language Resources Association.

Karen Sparck Jones and Julia R. Galliers. 1996. *Evaluating Natural Language Processing Systems: An Analysis and Review*. Springer-Verlag, Berlin, Heidelberg.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

Chris van der Lee, Bart Verduijn, Emiel Krahmer, and Sander Wubben. 2018. Evaluating the text quality, human likeness and tailoring component of PASS: A Dutch data-to-text system for soccer. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 962–972, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. pages 74–81.

Harold A Linstone, Murray Turoff, et al. 1975. *The delphi method*. Addison-Wesley Reading, MA.

Elena Lloret, Laura Plaza, and Ahmet Aker. 2013. Analyzing the capabilities of crowdsourcing services for text summarization. *Language Resources and Evaluation*, 47(2):337–369.

Elena Lloret, Laura Plaza, and Ahmet Aker. 2018. The challenging task of summary evaluation: An overview. *Lang. Resour. Eval.*, 52(1):101–148.

Annie Louis and Ani Nenkova. 2013. Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300.

Inderjeet Mani. 2001. Summarization evaluation: An overview.

Inderjeet Mani, Gary Klein, David House, Lynette Hirschman, Therese Firmin, and Beth Sundheim. 2002. Summac: a text summarization evaluation. *Natural Language Engineering*, 8(1):43–68.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Babak Naderi, Sebastian Möller, Tobias Hossfeld, and Matthias Hirth. 2018. P.808 subjective evaluation of speech quality with a crowdsourcing approach. ITU-T Recommendation P.808, International Telecommunication Union, Geneva.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Josef Steinberger and Karel Ježek. 2012. Evaluation measures for text summarization. *Computing and Informatics*, 28(2):251–275.

Robert C. Streijl, Stefan Winkler, and David S. Hands. 2016. Mean opinion score (mos) revisited: Methods and applications, limitations and alternatives. *Multimedia Syst.*, 22(2):213–227.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech Language*, 67:101151.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Guang Yong Zou. 2007. Toward using confidence intervals to compare correlations. *Psychological methods*, 12(4):399.