Best Practices for Crowd-based Evaluation of German Summarization: Comparing Crowd, Expert, and Automatic Evaluation

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

Abstract

One of the main challenges in the development of summarization tools is summarization quality evaluation. On the one hand, the human assessment of summarization quality conducted by linguistic experts is slow, expensive, and still not a standardized procedure. On the other hand, the automatic assessment metrics are reported not to correlate high enough with human quality ratings. As a solution, we propose crowdsourcing as a fast, scalable, and cost-effective alternative to expert evaluations to assess the intrinsic and extrinsic quality of summarization by comparing crowd ratings with expert ratings and automatic metrics such as ROUGE, BLEU, or BertScore on a German summarization data set. Our results provide a basis for best practices for crowd-based summarization evaluation regarding major influential factors such as the best annotation aggregation method, the influence of readability and reading effort on summarization evaluation, and the optimal number of crowd workers to achieve comparable results to experts, especially when determining factors such as overall quality, grammaticality, referential clarity, focus, structure & coherence, summary usefulness, and summary informativeness.

1 Introduction

Even though there has been an enormous increase in automatic summarization research, human evaluation of summarization is still an understudied aspect. One the one hand, there is no standard procedure for conducting human evaluation, which is leading to a high degree of variation and different results (Van Der Lee et al., 2019); on the other hand, human evaluation is usually carried out in a traditional laboratory environment by linguistic experts, which is costly and time-consuming to run and prone to subjective biases (Celikyilmaz et al., 2020). Therefore, automatic evaluation metrics such as BLEU and ROUGE have been used as substitutes for human evaluation (Papineni et al., 2002; Lin, 2004). However, they require expert summaries as references to be calculated and are often reported not to correlate with human evaluations regarding the readability, grammaticality, and content-related factors (Novikova et al., 2017).

In the other NLP domains, crowdsourcing has been proposed as an alternative to overcome these challenges, showing that crowd workers' aggregated responses could produce quality approaching those produced by experts (Snow et al., 2008; Callison-Burch, 2009; Nowak and Rüger, 2010). In the summarization evaluation, very few researchers have investigated crowdsourcing as an alternative, eventually concluding that the chosen crowd-based evaluation methods are not reliable enough to produce consistent scores (Gillick and Liu, 2010; Fabbri et al., 2020). However, the authors did not apply any pre-qualification test, did not provide information about the number of crowd workers, did not apply annotation aggregation methods, or did not analyze the effect of reading effort and readability of source texts caused by the text's structural, and formal composure. Additionally, they used the TAC and CNN/Daily Mail data set derived from high-quality English texts. So, there is a research gap regarding the best practices for crowd-based evaluation of summarization, especially for languages other than English and noisy internet data.

We address this gap in the following ways: 1) We use a German summarization data set derived from an online question-answering forum; 2) We apply pre-qualification tests and set a threshold for minimum task completion duration in crowdsourcing; 3) We collect intrinsic and extrinsic quality ratings from 24 different crowd workers per summary in order to analyze consistency; 4) We use different annotation aggregation methods on crowdsourced data; 5) We analyze the effect of annotation aggregation methods, reading effort, and the number

Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems (Eval4NLP), pages 164–175,
November 20, 2020.

of crowd workers per item on robustness, comparing results from a) expert assessment; b) crowd assessment; c) state of the art automatic assessment metrics. Especially, languages other than English can benefit from our results, since they lack easy-to-use automatic evaluation metrics in the form of simplified toolkits, and a well-executed evaluation can accelerate the research on automatic summarization (Fabbri et al., 2020).

2 Related Work

2.1 Automatic Summarization Evaluation

The automatic evaluation of summarization can be categorized into two categories: untrained automatic metrics, which do not require machine learning but are based on string overlap, or content overlap between machine-generated and expert generated summaries (ground-truth), and machine-learned metrics that are based on machine-learned models (Celikyilmaz et al., 2020).

2.1.1 Untrained Automatic Metrics

The most common untrained automatic metrics for summarization evaluation are BLEU, METEOR, and ROUGE, which rely on counting n-grams and calculating Precision, Recall, and F-measure by comparing one or several system summaries to reference summaries generated by experts (Papineni et al., 2002; Denkowski and Lavie, 2014; Lin, 2004). As Gao et al. (2019) stated, ROUGE is the most popular method to assess the summarization quality, and at least one of the ROUGE variant is used in 87% of papers on summarization in ACL conferences between 2013 and 2018. In recent years, many variations on ROUGE and other measures have been introduced in the literature (Zhou et al., 2006; Ng and Abrecht, 2015; Ganesan, 2018). However, they have been criticized because of the wide range of correlations being weak to strong with human assessment reported in the summarization literature and for being not suitable for capturing important quality aspects (Reiter and Belz, 2009; Graham, 2015; Novikova et al., 2017; Peyrard and Eckle-Kohler, 2017). Therefore, more and more researchers refrain from using automatic metrics as a primary evaluation method (Reiter, 2018). Still, Van Der Lee et al. (2019) report that 80% of the empirical papers presented at the ACL track on NLG or at the INLG conference in 2018 using automatic metrics due to the lack of alternatives and the fast and cost-effective nature.

2.1.2 Trained Automatic Metrics

Over the last few years, NLP researchers proposed new machine-learned automatic metrics trained using BERT contextual embeddings such as BertScore, BLEURT, and BLANC to evaluate the natural language generation (NLG) quality, which can also be applied to summarization evaluation (Devlin et al., 2019; Zhang et al., 2019; Sellam et al., 2020; Vasilyev et al., 2020). BertScore and BLEURT still require expert generated summaries as ground-truth and computes the similarity of two summaries as a sum of cosine similarities between their tokens' embeddings. Zhang et al. (2019) reported that BertScore correlates better than the other state of the art metrics in the domain of machine translation and image captioning tasks, Sellam et al. (2020) showing that BLEURT correlates better than BertScore with human judgments on the WMT17 Metrics Shared Task. Unlike these metrics, the BLANC score is designed not to require any reference summaries aiming for fully human-free summary quality estimation (Vasilyev et al., 2020). BLANC was shown to correlate as good as ROUGE on CNN/DailyMail data set.

2.2 Human Evaluation

Human evaluation can be conducted as pair comparison (compared to expert summaries) or using absolute scales without having a reference. One of the common human evaluation methods using pair comparison is the PYRAMID method (Nenkova and Passonneau, 2004). In the PYRAMID method, sentences in summaries are split into Summary Content Units for both system and reference summaries and compared with each other based on content. So, it measures only the summaries' relative quality and does not give a sense of the summary's absolute quality. In this paper, we focus on absolute quality measurement in which the generated summaries are demonstrated to the evaluators one at a time, and they judge summary quality individually by rating the quality along a Likert or sliding scale. Therefore, we do not use the PYRAMID method in our human evaluation and collect human ratings on two categories: intrinsic (linguistic) and extrinsic (content) evaluation (Jones and Galliers, 1995; Steinberger and Ježek, 2012).

2.2.1 Intrinsic (Linguistic) Evaluation

In intrinsic evaluation, domain experts are usually asked to evaluate the quality of the given summary, either as overall quality or along some specific

dimension without reading the source document (Celikyilmaz et al., 2020). To determine the intrinsic quality of summarization, the following five text readability (linguistic quality) scores are most commonly used: grammaticality, non-redundancy, referential clarity, focus, and structure & coherence. In the section 3, we determine these scores based on the definitions in Dang (2005).

2.2.2 Extrinsic (Content) Evaluation

In extrinsic evaluation, domain experts evaluate a system's performance on the task for which it was designed, so the evaluation of summary quality is accomplished based on the source document (Lloret et al., 2018). The most common extrinsic quality measures are: 1) "Summary usefulness" - also called content responsiveness - which determines the summary's usefulness concerning how useful the extracted summary is to satisfy the given goal; 2) "Source text usefulness" - also called relevance assessment - which examines how useful the source document is to satisfy the given goal; 3) "Summary informativeness" measuring how much information from the source document is preserved in the extracted summary (Mani, 2001; Conroy and Dang, 2008; Shapira et al., 2019).

2.3 Crowdsourcing for Summarization Evaluation

Crowdsourcing has been used as a fast and cost-effective alternative to traditional subjective evaluation with experts in summarization evaluation; however, it has not been explored as thoroughly as other NLG tasks, such as evaluating machine translation (Lloret et al., 2018). In the few papers where crowdsourcing has been used for summarization evaluation, the quality of crowdsourced data has been repeatedly questioned because of the crowd worker's inaccuracy and the complexity of summarization evaluation.

For example, Gillick and Liu (2010) found that the ratings from non-expert crowd workers do not correlate the expert ratings on the TAC summarization data set, which contains 100-word summaries of a set of 10 newswire articles about a particular topic. A similar conclusion was reached by Lloret et al. (2013), who created a corpus for abstractive image summarization with five crowd workers per item. However, besides the fact that results were obtained from other domains than the presented telecommunication domain in this work, in both works, the authors did not apply any pre-qualification test or did not provide information about crowdsourcing task details, which can also cause a rather large influencing effect. Following, Gao et al. (2018); Falke et al. (2017); Fan et al. (2018) have used crowdsourcing as the source of human evaluation to rate their automatic summarization systems. Nevertheless, they did not question the robustness of crowdsourcing for this task and compared the crowd with expert data. Also, we have shown that crowdsourcing achieves almost the same results as the laboratory studies using 7-9 crowd workers, but we did not compare the crowd with experts (Iskender et al., 2020). Fabbri et al. (2020) compared the crowd with expert evaluation on CNN/Daily Mail data set using only five crowd workers per summary. They also found that crowd and expert ratings do not correlate and emphasized the need for protocols for improving the human evaluation of summarization.

To improve the quality of crowdsourcing, researchers have developed several methods such as *filtering* and *aggregation* (Kairam and Heer, 2016). When filtering crowd workers, the first approach focuses on the pre-qualification tasks designed based on the task characteristics (Mitra et al., 2015). While aggregating crowd judgments, the majority vote is the most common technique (Chatterjee et al., 2019). Much more complex annotation aggregation methods such as probabilistic models of annotation, accounting item level effects, or clustering methods have been introduced in the recent years (Passonneau and Carpenter, 2014; Whitehill et al., 2009; Luther et al., 2015).

To provide the best practices for crowd-based summarization evaluation, we apply pre-qualification and focus on the following aggregation methods in this paper: 1) MOS: Mean Opinion Score (MOS) takes the mean of all judgments for a given item and is one of the most popular metrics for subjective quality evaluation (Streijl et al., 2016; Chatterjee et al., 2019), 2) Majority Vote: In Majority Vote, the answer with the highest votes is selected as the final aggregated value, and it is the most popular method in subjective quality evaluation with crowdsourcing (Hovy et al., 2013; Hung et al., 2013), 3) Crowdtruth: It represents the crowdsourcing system in its three main components – input media units, workers, and annotations. It is designed to capture inter-annotator disagreement in crowdsourcing and aims to collect gold standard data for training and evaluation

of cognitive computing systems using crowdsourcing (Dumitrache et al., 2018a). Dumitrache et al. (2018b) have shown that the Crowdtruth performs better than the majority vote in different domains, 4) MACE: Multi-Annotator Competence Estimation (MACE) is a probabilistic model that computes competence estimates of the individual annotators and the most likely answer to each item (Hovy et al., 2013). Paun et al. (2018) have shown that MACE performs better than the other annotation aggregation methods in evaluations against the gold standard. This model is possibly most widely applied to linguistic data (Plank et al., 2014; Sabou et al., 2014; Habernal and Gurevych, 2016).

3 Experiments

3.1 Data Set

In our experiments, we used the same German summary data set with 50 summaries as described in Iskender et al. (2020). The corpus contains queries with an average word count of 7.78, the shortest one with four words, and the longest with 17 words; posts from a customer forum of Deutsche Telekom with an average word count of 555, the shortest one with 155 words, and the longest with 1005 words; and corresponding query-based extractive summaries with an average word count of 63.32, the shortest one with 24 words, and the longest one with 147 words.

3.2 Crowdsourcing Study

We collected crowd annotations using Crowdee[1] Platform. Crowd workers were only allowed to perform the summary evaluation task after passing two qualification tests in the following order: 1) German language proficiency test provided by the Crowdee platform with a score of 0.9 and above (scale [0, 1]), 2) Summarization evaluation test containing deliberately designed bad and good examples of summaries to be recognized by the crowd. Here, a maximum of 20 points could be reached by crowd workers, and we kept crowd workers exceeding 12 points. Besides, according to our expert pre-testing, we set 90 seconds as a threshold for the minimum task completion duration and eliminated all the crowd answers under this threshold.

In the main task, a brief explanation of the summary creation process was shown first with an example of a query, forum posts, and a summary to provide background information. After reading all instructions, crowd workers evaluated nine quality factors of a single summary using a 5 point scale with the labels *very good, good, moderate, bad, very bad* in the following order: 1) overall quality, 2) grammaticality, 3) non-redundancy, 4) referential clarity, 5) focus, 6) structure & coherence, 7) summary usefulness, 8) post usefulness and 9) summary informativeness. In the first six questions, the corresponding forum posts and the query were not shown to the crowd workers (intrinsic quality); in question 7, we showed the original query; in questions 8 and 9, the original query and the corresponding forum posts. In total, 24 repetitions per item for each of these nine questions were collected, resulting in 10,800 labels (50 summaries x 9 questions x 24 repetitions). Compensation was carefully calculated to ensure the minimum wage of €9.35 per hour in Germany. Overall, 46 crowd workers (19f, 27m, $M_{age} = 43$) completed the individual sets of tasks within 20 days where they spent 249,884 seconds, ca. 69.4 hours at total.

3.3 Expert Evaluation

We used a similar approach to the Delphi method to obtain a consensus among experts in an iterative procedure (Linstone et al., 1975; Sanchan et al., 2017). In the first evaluation round, two experts, who are Masters students in linguistics, evaluated separately the same summarization data set using the same task design as crowd workers by using Crowdee Platform to avoid any user interface biases. After the first evaluation round, the inter-rater agreement calculated by Cohen's κ showed that the experts often diverted in their assessment. In order to reach an acceptable inter-rater agreement score, physical follow-up meetings with experts were arranged. In these meetings, experts discussed causes and backgrounds of their ratings for each item they disagreed, simultaneously creating a more detailed definition and evaluation criteria catalog for each score for future experiments. After the meeting, acceptable inter-rater agreement scores were achieved (see Section 4). In total, 900 ratings (50 Summary x 9 questions x 2 experts) were collected.

3.4 Automatic Evaluation

We calculated the BLEU and ROUGE scores using the sumeval library[2] for German, BertScore[3], and

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

[2] https://github.com/chakki-works/sumeval

[3] https://github.com/Tiiiger/bert_score

BLEURT[4] scores using bert-base-german-cased configuration. All of these four metrics require gold standard summaries, which were created by the two linguistic experts. The gold standard summaries have an average word count of 58.18, the shortest one with 14 words, and the longest with 112 words. In addition, we calculated the human-free summary quality estimation metric BLANC[5] using bert-base-german-cased configuration. The reason for selecting these five metrics is that they either are the baseline of automatic summarization evaluation metrics (BLEU and ROUGE) or the latest AI-based metrics (BertScore, BLEURT, BLANC) which have not been applied to a German summarization data set.

4 Results

Results are presented for the scores overall quality (OQ), the five intrinsic quality scores (including grammaticality (GR), non-redundancy (NR), referential clarity (RC), focus (FO), structure & coherence (SC)) and the three extrinsic quality scores (summary usefulness (SU), post usefulness (PU) and summary informativeness (SI)). We will refer to these labels by their abbreviations in this section. For our human-based evaluation, we analyzed 10,800 ratings from the crowdsourcing study and 900 ratings from the expert evaluation. For automatic evaluation, we analyzed the BLEU, ROUGE-1, ROUGE-2, ROUGE-L, BertScore (we use F-scores for these metrics), BLEURT by taking the mean of scores calculated using two expert summaries and the BLANC scores resulting in 350 scores (50 summaries x 7 automatic metrics).

4.1 Comparing Crowd with Expert

Before comparing expert ratings with the crowd, we calculated Cohen's κ and Krippendorff's α scores to measure the inter-rater agreement between two experts and the raw agreement scores as recommended in Van Der Lee et al. (2019) (see Table 1). Looking at the raw agreement, we see that experts gave the same ratings at least 70 % of the data for all nine measures after the second evaluation round. Further, Cohen's κ scores show that there is substantial (0.6-0.8] or almost perfect agreement (0.80-1.0] between experts for all measures except for NR, PU, and SI being weak (0.40-0.59) (Landis and Koch, 1977).

Measure	Raw Agr. in %	κ	α
OQ	82	0.637	0.820
GR	78	0.626	0.815
NR	70	0.520	0.796
RC	88	0.819	0.907
FO	80	0.685	0.777
SC	82	0.743	0.893
SU	76	0.635	0.835
PU	70	0.469	0.630
SI	76	0.565	0.764

Table 1: Raw agreement in %, Cohen's κ and Krippendorff's α of expert ratings

Also, we calculated Krippendorff's α, which is technically a measure of evaluator disagreement rather than agreement and the most common of the measures in the set NLG papers surveyed in Amidei et al. (2019). The Krippendorff's α scores for all the other measures are good [0.8-1.0] except for PO and SI measures, which are tentative [0.67-0.8) and PU measure, which should be discarded because it is 0.04 lower than the threshold 0.67 Krippendorff (1980). Because of the minimal difference of 0.04, we decided to still use the PU measure in our further analysis for interpretation. With these results, we achieved a better agreement level than the average expert agreement of summarization evaluation reported in other papers Van Der Lee et al. (2019).

We use the mean of expert ratings for all quality measures as our ground-truth for our further analysis. To test the normality of expert ratings, we carried out Anderson-Darling tests showing that the measures OQ, NR, FO, and SI were not normally distributed ($p < 0.05$). Therefore, we apply non-parametric statistics in the following sections.

4.1.1 Annotation Aggregation Methods

To investigate the effect of the annotation aggregation methods on the correlation coefficients between the crowd and expert ratings, we compared MOS with the baseline Majority Vote and two weighted-rank metrics CrowdTruth and MACE using crowdtruth-core[6] and MACE[7] libraries. Table 2 shows the Spearman's ρ correlation coefficients between crowd and experts by using these four

[4] https://github.com/google-research/bleurt

[5] https://github.com/PrimerAI/blanc

[6] https://github.com/CrowdTruth/CrowdTruth-core

[7] https://github.com/dirkhovy/MACE

Measure	MOS	Maj. Vote	Crowdtruth	MACE
OQ	**.730**	.624	.702	.654
GR	.706	.696	**.721**	.633
NR	**.581**	.523	.553	.490
RC	**.741**	.619	.726	.647
FO	**.656**	.516	.636	.516
SC	.828	.690	**.834**	.748
SU	**.688**	.60	.677	.561
PU	**.464**	NS	.435	NS
SI	**.619**	.482	.609	.523

$p < 0.05$ for all correlations
NS: Not Significant

Table 2: Spearman's ρ correlation coefficients between crowd and expert ratings for all measures by the aggregation methods MOS, Majority Vote, Crowdtruth and MACE

aggregation methods, and the bold coefficients correspond to row maxima.

For all measures, Majority Vote and MACE performed worse than MOS and Crowdtruth. For measures OQ, NR, RC, FO, SU, PU, and SI, MOS performed better than the Crowdtruth, and for GR and SC, Crowdtruth performed better than MOS by all correlation coefficients. To determine if these differences are statistically significant, we applied Zou's confidence intervals test for dependent and overlapping variables and found out that the differences between correlation coefficients were not statistically significant for all nine measures (Zou, 2007). Based on this correlation analysis, we recommend using MOS as the aggregation method for crowd-based summarization evaluation since aggregation using MOS delivers the most comparable aggregates compared to experts and easy to apply.

Analyzing the Spearman's ρ correlation coefficients between the crowd and expert ratings by MOS, we see that all correlation coefficients were statistically significant, ranging from moderate (NR, PU) to strong (OQ, GR, RC, FO, SU, SI) and very strong (SC), where SC had the highest correlation coefficient of 0.828 and PU the lowest correlation coefficient of 0.464. This result suggests that crowdsourcing can be used instead of experts when determining the structure & coherence of a summarization. For determining OQ, GR, RC, FO, SU, and SI, crowdsourcing can be preferred since the overall correlation coefficients are strong, but the results should be interpreted with some degree of caution. However, when evaluating the non-redundancy and post usefulness, experts should be used for more robust results.

To investigate the differences between the crowd and expert judgments, we conducted the Mann-Whitney U test for each pair of nine quality scores. We observed no significant difference between the median ratings of OQ, SC, SU, and SI measures. This result suggests that crowdsourcing can be used instead of experts when determining these four measures without significant deviation in absolute score rating value. Please note that the ratings' distributions allow for significant equality in estimated mean values (here as the median) even on levels where correlations did not show very strong but only strong magnitudes.

However, there were statistically significant difference between GR_{Crowd} ($M = 3.667$) and $\text{GR}_{\text{Expert}}$ ($M = 4.0$), NR_{Crowd} ($M = 3.865$) and $\text{NR}_{\text{Expert}}$ ($M = 4.0$), RC_{Crowd} ($M = 3.794$) and $\text{RC}_{\text{Expert}}$ ($M = 4.0$), FO_{Crowd} ($M = 4.048$) and $\text{FO}_{\text{Expert}}$ ($M = 4.250$), as well as PU_{Crowd} ($M = 3.856$) and $\text{PU}_{\text{Expert}}$ ($M = 4.0$), showing that the crowd workers rated these factors statistically lower than the experts. This observation might be explained by the fact that the nature of extractive summarization and inherent text quality losses - compared to naturally composed text flow - are more familiar to experts than to non-experts, so they can distinguish between the unnaturalness and the linguistic quality in more robust ways.

4.1.2 Effect of Reading Effort

In this section, we analyzed the seven measures which achieve a correlation coefficient above 0.6 with experts: OQ, GR, RC, FO, SC, SU, and SI. Because the text's structural and formal composure, among many other factors, can cause difficulty in summarization evaluation, we analyzed the quality assessment performance of crowd workers regarding two distinct factors: a) readability of the text; b) reading effort in terms of overall stimuli length by dividing our data into six groups.

As our first reading effort criteria, we used the automated readability index (ARI), a readability test designed to assess a text's understandability, where a low ARI score indicates higher readability of a text (Feng et al., 2010). We split the packaged data into two groups by the median ARI scores of source texts (ARI-Low, ARI-High) calculated using textstat[8] library. Because the amount of information to be read and understood by any crowd

[8] https://github.com/shivam5992/textstat

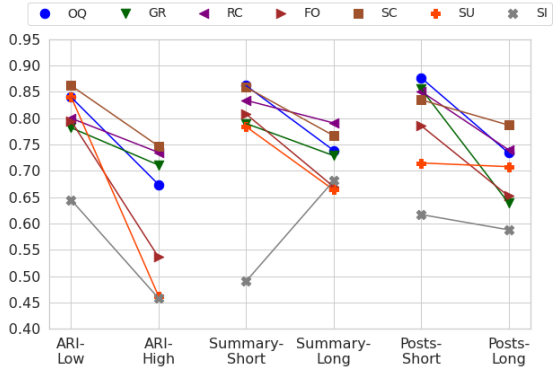

Figure 1: Spearman's ρ correlation coefficients between crowd and expert ratings for six groups

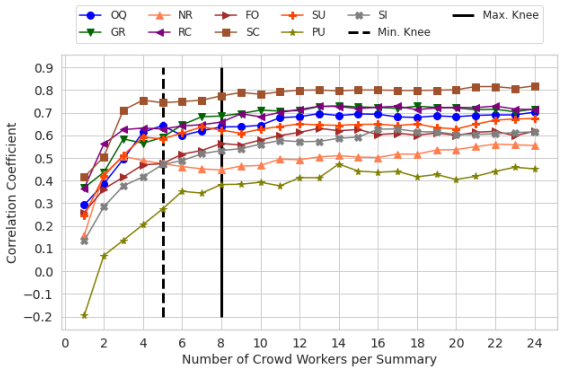

Figure 2: Spearman's ρ correlation coefficients between crowd and expert ratings by the number of crowd workers

participant may cause degrading concentration and motivation levels when the reading effort gets too long, we also split the data by the median of the word count of the summaries ($M = 56$) (Summary-Short, Summary-Long), and by the median of the forum posts ($M = 516$) (Posts-Short, Posts-Long).

Figure 1 displays all the correlation coefficients for the six groups. Here, we recognized that there was a certain pattern for all group pairs where correlation coefficients between the crowd and expert ratings were in groups "ARI-Low", "Summary-Short", and "Posts-Short" higher than the correlation coefficients in groups "ARI-High", "Summary-Long", and "Posts-Long" except for SI. The reason for the opposite trend of SI in groups divided by the summary length might be that the long summaries naturally contain more information, so it is easier for crowd workers to identify the summary informativeness. Other than this opposite trend of SI, we can derive the intuitive assumption that text understandability and reading effort have a noticeable effect on crowd judgments' robustness. Crowd workers may be used instead of experts for the evaluation of rather short summaries derived from documents with high readability.

4.1.3 Optimal Crowd Worker Number

To find out the optimal number of required crowd workers assessments per item, we plot the change of correlation coefficients between the crowd and expert ratings for all nine measures, where the x-axis shows the number of crowd workers per item in measured order, and the y-axis displays the Spearman ρ correlation coefficients between the crowd and expert ratings in Figure 2.

Looking at Figure 2, three or fewer crowd workers as annotators are not sufficient, and a study with a low number of crowd workers would not deliver a qualitative result since the correlation coefficient increase by increasing the number of crowd workers. However, this increase ends a saturation point between the number of repetitions and the resulting correlation coefficient. In order to determine the accurate optimal number of repetitions, we applied the method described in our paper Iskender et al. (2020), where multiple randomized runs are simulated in order to determine a "knee point" robustly, after which any additional repetitions no longer cause an adequate increase of overall correlation coefficients between the crowd and expert ratings. Our findings are directly in line with our findings in Iskender et al. (2020), where we applied this method to compare the crowd rating with laboratory ratings and stated that 7-9 crowd workers are the optimal number to achieve almost the same results as laboratory results in general.

We found that the knee point is 5 for RC; 7 for OQ, GR, NR, FO, SC, SU, and SI; 8 for PU. This result shows that generally, after collecting data from 5-8 different crowd workers depending on the measure, collecting one more additional crowd judgment was no longer worth the increase in correlation coefficient between the crowd and expert.

4.2 Human vs. Automatic Evaluation

As explained in section 2.1, we calculated BLEU ($\bar{x} = 0.294$), F1-Scores of ROUGE-1 ($\bar{x} = 0.459$), ROUGE-2 ($\bar{x} = 0.345$), ROUGE-L ($\bar{x} = 0.380$), and BertScore ($\bar{x} = 0.371$) as well as BLANC ($\bar{x} = 0.281$) and BLEURT ($\bar{x} = -0.492$) scores for our data set using the summaries from two experts as our gold standard.

While analyzing the Spearman's correlation co-

Measure	OQ	RC	FO	SU	SI
ROUGE-1	0.351	NS	0.323	0.395	0.420
ROUGE-2	NS	NS	NS	0.304	0.326
ROUGE-L	NS	NS	NS	0.284	0.315
BertScore	0.333	0.302	0.322	0.390	0.397

$p < 0.05$ for all correlations
NS: Not significant

Table 3: Spearman's ρ correlation coefficients between ROUGE scores, BertScore, and crowd ratings

Measure	OQ	GR	RC	SI
ROUGE-1	0.315	0.365	NS	0.377
BertScore	0.04	0.320	0.318	NS

$p < 0.05$ for all correlations
NS: Not significant

Table 4: Spearman's ρ correlation coefficients between ROUGE scores, BertScore, and expert ratings

efficients between the automatic scores and the crowd ratings, we observed that only ROUGE and BertScore scores correlated with OQ, RC, FO, SU, and SI of the crowd judgments (see Table 3). Looking at the correlation coefficients between expert ratings and automatic metrics (see Table 4), we also found that there was a significant correlation between only ROUGE and BertScore scores and OQ, GR, RC, and SI of expert ratings. Generally, we observed that overall correlations were of weak level, looking at the magnitude of any significant correlation found. Even though we used most recent metrics other than ROUGE trained on BERT, such as BertScore (Van Der Lee et al., 2019), our findings verify that automatic metrics do not correlate with linguistic quality metrics in the summarization domain.

Although Papineni et al. (2002); Lin (2004); Zhang et al. (2019) reported high correlations with humans on the content-related quality assessment in the corresponding original papers, we showed that these metrics correlate poorly with any human rating, from crowd or expert, verifying the findings of Van Der Lee et al. (2019) for our data set. The reason for this difference is that the BLEU score is developed for measuring machine translation quality and tested on a translation data set. BertScore is also not evaluated using a summarization data set in the original corresponding paper. Only, ROUGE metric is tested on summarization data sets. However, in the human evaluation part of the original paper, the evaluators assigned content coverage scores to a candidate summary compared to a manual summary, which is very similar to the way of working of ROUGE calculating the n-gram match of a candidate summary in comparison to a manual summary. In our human evaluation, we did not apply pair comparison, and the ratings were given on an absolute scale, which might be the reason for the low correlation coefficients between automatic metrics and human ratings in our study.

We also calculated BLEURT and BLANC scores, but we treat them as preliminary results since we did not apply any special pre-training to these metrics. We found that BLEURT does not correlate with any of the crowd and expert ratings significantly. Similarly, BLANC does not correlate with any of the crowd rating except for NR ($\rho = -0.342$), and surprisingly it correlates significantly and negatively with expert ratings for NR ($\rho = -0.473$) , RC ($\rho = -0.308$), and SC ($\rho = -0.347$). We can not explain the reasons for the negative correlation and speculate that this might be due to not applying pre-training.

5 Conclusion and Future Work

In this paper, we provide a basis for best practices for crowd-based summarization evaluation by comparing different annotation aggregation methods, analyzing the effect of reading effort and readability, and approaching an estimate of an optimal number of required crowd workers per item in order to as closely as possible resemble experts' assessment quality through crowdsourcing.

When determining structure & coherence, we suggest that crowdsourcing can be used as a direct substitute for experts proven by the very strong correlation coefficient. For determining overall quality, grammaticality, referential clarity, focus, summary usefulness, and summary informativeness, crowdsourcing can be preferred as the overall correlation still results strong, but the results should be interpreted carefully. However, when evaluating non-redundancy and post usefulness, experts should be used for more robust results, as correlations result moderate only.

Our experiments further recommend following best-practices when using crowdsourcing instead of experts: 1) In general 5-8 crowd workers should annotate a given summary, 2) MOS should be used as an aggregation method to achieve optimally comparable results to experts, 3) Crowdsourcing may be

used at best when readability of the source and reading effort of the task is of rather low and straightforward nature. We also confirm the findings of Dumitrache et al. (2018b) that Crowdtruth performs better than the MACE. Further, we confirm that the automatic evaluation metrics BLEU, ROUGE, and BertScore can not be used to evaluate the linguistic quality, and we show that automatic evaluation metrics correlate poorly with any content-related absolute human rating, from crowd or expert, verifying the findings of Van Der Lee et al. (2019) for our domain. Therefore, crowdsourcing should generally be the preferred evaluation method over automated scores in the summarization evaluation.

Since the vast majority of research on summarization bases on the TAC or CNN/Dailymail data sets, there is a lack of works from other languages or domains. We address this gap by using a German forum summarization data set derived from an online forum in the telecommunication domain. Contrary to the findings of Gillick and Liu (2010) and Fabbri et al. (2020), we achieve significant correlations between the crowd and expert ratings ranging from moderate to very strong magnitude, as well as no significant difference in absolute mean rating in between the crowd and expert assessment for overall quality, structure & coherence, summary usefulness, and summary informativeness. Other scales show a slight but still significant bias towards lower ratings of about less than 0.3pt absolute. These are important findings in the development of NLG tools for summarization. In particular, summarization tools developed for languages other than English for which it is harder to conduct expert evaluations and find easy-to-use automatic metrics could benefit highly from our findings.

However, this study has some limitations since we conduct our analysis using only a single data set derived from an online forum in the telecommunication domain. The level of domain knowledge of crowd workers and experts about the telecommunication service might play a role when determining content-related quality measures such as post usefulness. So, the effect of the domain knowledge should be investigated in detail in future work. Another shortcoming of this paper is that our summarization data set is derived from noisy internet data, and the summary length does not differ much. As shown in section 4.1.2, the readability of the source document and varying lengths of summaries might affect the results; therefore, the same analysis should be conducted on one more data set. Additionally, our data set was monolingual, so exploring the language-based effects will also be part of future work.

Further, this study is that we did not investigate the effect of the crowdsourcing task design and learning effect on the correlation coefficient between the crowd and expert ratings. Questions regarding the limitation to the number of assignments taken on by an evaluator (both for crowd and expert) and evaluators' behavior (becoming more lenient or strict over time) should also be analyzed in future work. Also, we did not use the pairwise comparison in our task design and only focused on absolute quality rating. For that reason, investigating the pairwise comparison using crowdsourcing and its comparison to absolute rating should be considered as an essential aspect of the crowdsourcing task design in future work.

Despite the limitations of our study, this paper is the first paper in the summarization evaluation literature that provides evidence for clear support for using crowdsourcing to evaluate summarization quality and adds to a growing corpus of research on the summarization evaluation.

References

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2019. Agreement is overrated: A plea for correlation to assess human evaluation reliability. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 344–354.

Chris Callison-Burch. 2009. Fast, cheap, and creative: evaluating translation quality using amazon's mechanical turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 286–295. Association for Computational Linguistics.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Sujoy Chatterjee, Anirban Mukhopadhyay, and Malay Bhattacharyya. 2019. A review of judgment analysis algorithms for crowdsourced opinions. *IEEE Transactions on Knowledge and Data Engineering*.

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-Volume 1*, pages 145–152. Association for Computational Linguistics.

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

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Anca Dumitrache, Oana Inel, Lora Aroyo, Benjamin Timmermans, and Chris Welty. 2018a. Crowdtruth 2.0: Quality metrics for crowdsourcing with disagreement. In *1st Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing, and Short Paper 1st Workshop on Disentangling the Relation Between Crowdsourcing and Bias Management, SAD+ CrowdBias 2018*, pages 11–18. CEUR-WS.

Anca Dumitrache, Oana Inel, Benjamin Timmermans, Carlos Ortiz, Robert-Jan Sips, Lora Aroyo, and Chris Welty. 2018b. Empirical methodology for crowdsourcing ground truth. *Semantic Web*.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2020. Summeval: Re-evaluating summarization evaluation. *arXiv preprint arXiv:2007.12626*.

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.

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.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment. In *Coling 2010: Posters*, pages 276–284.

Kavita Ganesan. 2018. Rouge 2.0: Updated and improved measures for evaluation of summarization tasks. *arXiv preprint arXiv:1803.01937*.

Yang Gao, Christian M Meyer, and Iryna Gurevych. 2018. April: Interactively learning to summarise by combining active preference learning and reinforcement learning. In *EMNLP*.

Yanjun Gao, Chen Sun, and Rebecca J Passonneau. 2019. Automated pyramid summarization evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 404–418.

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. Association for Computational Linguistics.

Yvette Graham. 2015. Re-evaluating automatic summarization with bleu and 192 shades of rouge. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 128–137.

Ivan Habernal and Iryna Gurevych. 2016. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with mace. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130.

Nguyen Quoc Viet Hung, Nguyen Thanh Tam, Lam Ngoc Tran, and Karl Aberer. 2013. An evaluation of aggregation techniques in crowdsourcing. In *International Conference on Web Information Systems Engineering*, pages 1–15. Springer.

Neslihan Iskender, Tim Polzehl, and Sebastian Möller. 2020. 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.

Karen Sparck Jones and Julia R Galliers. 1995. *Evaluating natural language processing systems: An analysis and review*, volume 1083. Springer Science & Business Media.

Sanjay Kairam and Jeffrey Heer. 2016. Parting crowds: Characterizing divergent interpretations in crowd-sourced annotation tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 1637–1648. ACM.

Klaus Krippendorff. 1980. Content analysis: An introduction to its methodology.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

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. *Language Resources and Evaluation*, 52(1):101–148.

Kurt Luther, Jari-Lee Tolentino, Wei Wu, Amy Pavel, Brian P Bailey, Maneesh Agrawala, Björn Hartmann, and Steven P Dow. 2015. Structuring, aggregating, and evaluating crowdsourced design critique. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 473–485. ACM.

Inderjeet Mani. 2001. Recent developments in text summarization. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 529–531. ACM.

Tanushree Mitra, Clayton J Hutto, and Eric Gilbert. 2015. Comparing person-and process-centric strategies for obtaining quality data on amazon mechanical turk. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1345–1354. ACM.

Ani Nenkova and Rebecca J Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the human language technology conference of the north american chapter of the association for computational linguistics: Hlt-naacl 2004*, pages 145–152.

Jun Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for rouge. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1925–1930.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252.

Stefanie Nowak and Stefan Rüger. 2010. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, pages 557–566. ACM.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Rebecca J Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2:311–326.

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. 2018. Comparing bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585.

Maxime Peyrard and Judith Eckle-Kohler. 2017. Supervised learning of automatic pyramid for optimization-based multi-document summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1084–1094.

Barbara Plank, Dirk Hovy, Ryan McDonald, and Anders Søgaard. 2014. Adapting taggers to twitter with not-so-distant supervision. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1783–1792.

Ehud Reiter. 2018. A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401.

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.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *LREC*, pages 859–866.

Nattapong Sanchan, Ahmet Aker, and Kalina Bontcheva. 2017. Gold standard online debates summaries and first experiments towards automatic summarization of online debate data. In *International Conference on Computational Linguistics and Intelligent Text Processing*, pages 495–505. Springer.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Ori Shapira, David Gabay, Yang Gao, Hadar Ronen, Ramakanth Pasunuru, Mohit Bansal, Yael Amsterdamer, and Ido Dagan. 2019. Crowdsourcing lightweight pyramids for manual summary evaluation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 682–687.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. 2008. Cheap and fast—but is it

good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics.

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 Systems*, 22(2):213–227.

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.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the blanc: Human-free quality estimation of document summaries. *arXiv preprint arXiv:2002.09836*.

Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier R Movellan, and Paul L Ruvolo. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in neural information processing systems*, pages 2035–2043.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Liang Zhou, Chin-Yew Lin, Dragos Stefan Munteanu, and Eduard Hovy. 2006. Paraeval: Using paraphrases to evaluate summaries automatically. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 447–454. Association for Computational Linguistics.

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