Crowdsourcing versus the laboratory: Towards crowd-based linguistic text quality assessment of query-based extractive summarization

Neslihan Iskender[1], Tim Polzehl[1], and Sebastian Möller[1]

[1]Quality and Usability Lab, TU Berlin, Berlin, Germany
neslihan.iskender@tu-berlin.de, tim.polzehl1@tu-berlin.de,
sebastian.moeller@tu-berlin.de

Abstract. Curating text manually in order to improve the quality of automatic natural language processing tools can become very time consuming and expensive. Especially, in the case of query-based extractive online forum summarization, curating complex information spread along multiple posts from multiple forum members to create a short meta-summary that answers a given query is a very challenging task. To overcome this challenge, we explore the applicability of microtask crowdsourcing as a fast and cheap alternative for query-based extractive text summarization of online forum discussions. We measure the linguistic quality of crowd-based forum summarizations, which is usually conducted in a traditional laboratory environment with the help of experts, via comparative crowdsourcing and laboratory experiments. To our knowledge, no other study considered query-based extractive text summarization and summary quality evaluation as an application area of the microtask crowdsourcing. By conducting experiments both in crowdsourcing and laboratory environments, and comparing the results of linguistic quality judgments, we found out that microtask crowdsourcing shows high applicability for determining the factors *overall quality, grammaticality, non-redundancy, referential clarity, focus,* and *structure & coherence.* Further, our comparison of these findings with a preliminary and initial set of expert annotations suggest that the crowd assessments can reach comparable results to experts specifically when determining factors such as overall quality and structure & coherence mean values. Eventually, preliminary analyses reveal a high correlation between the crowd and expert ratings when assessing low-quality summaries.

Keywords: digitally curated text · microtask crowdsourcing · linguistic summary quality evaluation

1 Introduction

With the widespread usage of the world wide web, crowdsourcing has become one of the main resources to work at so-called "micro-tasks" that require human intelligence to annotate text or solve tasks that computers cannot yet solve and connect to external knowledge and expertise. In this way, a fast and relatively inexpensive mechanism is provided so that the cost and time barriers of qualitative and quantitative laboratory studies and controlled experiments can be mostly overcome [15, 12].

Although microtask crowdsourcing has been primarily used for simple, independent tasks such as image labeling or digitizing the print documents [20], few researchers have begun to investigate the crowdsourcing for complex and expert tasks such as writing, product design, or Natural Language Processing (NLP) tasks [19, 34]. Especially, empirical investigation of different NLP tasks such as sentiment analysis and assessment of translation quality in crowdsourcing has shown that aggregated responses of crowd workers can produce gold-standard data sets with quality approaching those produced by experts [31, 1, 28]. Inspired by these results, we propose using microtask crowdsourcing as a fast and cheap way of curating complex information spread along multiple posts from multiple forum members to create a short meta-summary that answers a given query along with the quality evaluation of these summaries. To our knowledge, only prior work by the authors themselves has considered the query-based extractive forum summarization and linguistic quality evaluation of query-based extractive summarization as an application area of micro-task crowdsourcing, without finalizing the decision on crowdsourcing's appropriateness for this task [16]. To fill this research gap, we focus on the subjective linguistic quality evaluation of a curation task containing the compilation of online discussion forum summarization by conducting comparative crowdsourcing and laboratory experiments. Given such a meta-summary encompassing all forum posts aggregated and summarized towards a certain query can reach high quality results, both human and automated search as well as summary embedding in different contexts can become a much more valuable mean to raise efficiency and accuracy in information retrieval applications.

In the remainder of this paper, we answer the research question "Can crowd successfully create query-based extractive summaries and asses the overall and linguistic quality of these summaries?" by conducting both laboratory and crowdsourcing experiments and comparing the results. Following, as preliminary results, we compare laboratory and crowd assessments with an initial data set of expert annotations to determine the reliability of non-expert annotations for summary quality evaluation.

2 Related Work

2.1 Evaluation of Summary Quality

The evaluation of summary quality is crucial for determining the success of any summarization method, improving the quality of both human and automatic summarization tools and as well for their commercialization. Due to the subjectivity and ambiguity of summary quality evaluation, as well as the high variety of summarization approaches, the possible measures for the summary quality evaluation can be broadly classified into two categories: extrinsic and intrinsic evaluation [17, 32].

In extrinsic evaluation, the evaluation of summary quality is accomplished on two bases: content responsiveness which examines the summary's usefulness with respect to external information need or goal basis, and relevance assessment which determines if the source document contains relevant information about the user‘s need or query [26, 5]. The extrinsic measures are usually assessed manually with the help of experts or non-expert crowdworkers. In intrinsic evaluation, the evaluation of the summary is directly based on itself and is often done by comparison with a reference summary [17]. Two main approaches to measure the intrinsic quality are the linguistic quality evaluation (or readability evaluation) and content evaluation [32]. The content evaluation is often performed automatically and determines how many word sequences of reference summary are included in the peer summary [21]. In contrast, linguistic quality evaluation contains the assessment of grammaticality, non-redundancy, referential clarity, focus, structure and coherence [6].

Hence, there are widely used automatic quality metrics developed to measure the quality of a summary such as ROUGE which offers a set of statistics (e.g. ROUGE-2 which uses 2-grams) by executing a series of recall measures based on n-gram co-occurrence between a peer summary and a list of reference summaries [21, 33]. These scores can only provide content-based similarity based on a gold standard summary created by an expert. However, linguistic summary quality features such as grammaticality, non-redundancy, referential clarity, focus, and structure & and coherence, can not be assessed automatically in most cases [32]. The existing automatic evaluation methods for linguistic quality evaluation are rare [22, 29, 9], often do not consider the complexity of the quality dimensions, and can require language-dependent adaptation for the recognition of very complex linguistic features. Therefore, linguistic quality features should be assessed manually by experts, or not evaluated at all due to the time and cost efforts. So, there are still new manual quality assessment methods needed in the research of summary quality evaluation.

In this paper, we focus on intrinsic evaluation measures, especially on the linguistic quality evaluation as defined in [6], performed under both crowd-working and laboratory conditions.

2.2 Crowdsourcing for Summary Creation and Evaluation

In recent years, researchers have found that even some complex and expert tasks such as writing, product design, or NLP tasks may be successfully completed by non-expert crowd workers with appropriate process design and technological support [19, 3, 2, 18]. Especially, using non-expert crowd workers for NLP tasks which are usually conducted by experts has become one of the research interests due to the organizational and financial benefits of microtask crowdsourcing [18, 7, 4].

Although crowdsourcing services provide quality control mechanisms, the quality of crowdsourced corpus generation has been repeatedly questioned because of the crowd worker's inaccuracy and the complexity of text summarization. Gillick and Lui [14] have shown that non-expert crowdworkers can not produce summaries with the same linguistic quality results as the experts. Besides, Lloret et. al. [23] have conducted a crowdsourcing study for corpus generation of abstractive image summarization and their results suggest that non-expert crowdworkers perform poorly due to the complexity of summarization task and non-motivation of crowdworkers. However, El-Haj et.al. [8] have shown that Amazon's Mechanical Turk (AMT)[1] is appropriate for carrying out summary creation task of human-generated single-document summaries from Wikipedia and newspaper article in Arabic. One reason for different results regarding the appropriateness of crowdsourcing for summary evaluation may be that these studies have concentrated on different kinds of summarization tasks such as abstractive image summarization or extractive generic summarization which leads to varying levels of task complexity.

Focusing on summarization quality evaluation, the application of crowdsourcing has not been explored as thoroughly as for other NLP tasks, such as translation [24]. However, the subjective quality assessment is needed to determine the quality of automatic or human-generated summaries [30]. Using crowdsourcing for subjective summary quality evaluation provides a fast and cheap alternative to the traditional subjective testing with experts but the crowdsourced annotations must be checked for quality since they are produced by workers with unknown or varied skills and motivations [27, 25].

Only, Gillick and Lui [14] have conducted various crowdsourcing experiments to investigate the quality of crowdsourced summary quality evaluation and showed that non-expert crowdworkers can not evaluate the summaries as good as experts. Also, Iskender et. al.[16] have carried out a crowdsourcing study to evaluate the summary quality, showing that experts and crowd workers correlate only assessing the low quality summaries. Gao et.al. [13], Falke et.al. [10], and Fan et. al. [11] have used crowdsourcing as a source of human evaluation to evaluate their automatic summarization systems, but not questioned the robustness of crowdsourcing for this task.

Therefore, more empirical studies should be conducted to find out which kind of summarization evaluation tasks are appropriate for crowdsourcing, how to design crowdsourcing tasks appropriate to crowd workers, as well as how to assure the quality of crowdsourcing for summarization evaluation.

3 Experimental Setup

3.1 Data Set

We used a German data set of crowdsourced summaries created as the query-based extractive summarization of forum queries and posts. These forum queries and posts originate from the forum *Deutsche Telekom hilft* where Telekom customers ask questions about the company's products and services and the questions are answered by other customers or the support agents of the company.

[1] http://www.mturk.com.

This summary data set was already annotated using a 5-point MOS in a previous crowdsourcing experiment. After aggregating the three different judgments per summary with a majority voting in this experiment, the quality of these summaries was ranging from 1.667 to 5. Based on these annotations, we allocated 50 summaries within 10 distinct quality groups ranging from lowest to highest scores (lowest group [1.667, 2]; highest group (4.667, 5]) each represented by five summaries to generate stratified data of widely varying qualities. The average word count of these summaries was 63.32, the shortest one with 24 words, and the longest one with 147 words. The corresponding posts had an average word count of 555, the shortest posts with 155 words, and the longest with 1005 words. Accordingly, the average length of customer queries was 7.78, the shortest one with 4 words, and the longest with 17 words.

3.2 Crowdsourcing Study

All the crowdsourcing tasks were completed using Crowdee platform[2]. For the crowd worker selection, we used two different tasks: German language proficiency screener provided by the Crowdee platform and a task-specific qualification job developed by the author team. We admitted only crowd workers who passed the German language test with a score of 0.9 and above (scale [0, 1]) to participate in the qualification job.

In this qualification task, we gave explanations about the process of extractive summary crafting and asked the crowd workers to rate the overall linguistic and content quality of four reference summaries (two very good, two very bad) whose quality were already annotated by experts on a 5-point MOS scale using the labels *very good, good, moderate, bad, very bad.* Expert scores were not shown to the participants. For each rating matching the experts' rating crowd workers earned 4 points. For each MOS-scale point deviation from the expert rating, crowd workers earned a point less, so increasing deviations from the experts' ratings were linearly punished.

We paid 1.2 Euros for the qualification task and its average compilation duration was 417 seconds, ca. 7 minutes. The qualification task was online for one week. Out of 1569 screened crowd workers of Crowdee platform holding a German language score >= 0.9, 82 crowd workers participated in the qualification task, 67 out of them passed the test with a point ratio >= 0.625, and 46 qualified crowd workers returned in order to perform the summary quality assessment task when they were published after 2 weeks time.

In the summary quality assessment task, crowd workers were presented with a brief explanation of how the summaries are created. It was highlighted that the summaries were constructed by the simple act of copying sentences from forum posts, potentially incurring a certain degree of unnatural or isolated composure. After that, an example of a query, forum posts, and a summary were shown. Next, crowd workers answered 9 questions regarding the quality of a single summary in the following order: 1) overall quality, 2) grammaticality, 3) non-redundancy,

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

4) referential clarity, 5) focus, 6) structure & coherence, 7) summary usefulness, 8) post usefulness and 9) summary informativeness.

The overall quality was asked first to avoid the influence of more detailed aspects on the overall quality judgment. The scoring of each aspect of a single summary was done on a separated page, which contained a short, informal definition of the respective aspect (sometimes illustrated with an example), the summary and the 5-point MOS scale (*very good, good, moderate, bad, very bad*). Additionally, in question 7 we showed the original query; in questions 8 and 9, the original query and the corresponding forum posts were displayed.

Each of the 50 summaries was rated by 24 different crowd workers, resulting in 10,800 labels (50 summaries x 9 questions x 24 repetitions). The estimated work duration for completing this task was 7 minutes. Accordingly, the payment was calculated based on the minimum hourly wage floor (9,19 Euros in Germany), so the crowd workers were paid 1.2 Euros. 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. With an average of 55.543 answers accepted at Crowdee platform in total, the crowd workers were relatively experienced users of the platform.

3.3 Laboratory study

The summary quality evaluation task design itself was identical to the crowdsourcing study. Participants also used the Crowdee platform to answer the questions, however, this time in a controlled laboratory environment. The experiment duration was set to one hour and the participants were instructed to evaluate as many summaries as they can. Following common practice for laboratory tests, all the participants were instructed in a written form before the experiment with the general task description and all their questions regarding the experiment rules or general questions were answered immediately. Each of the 50 summaries was again rated by 24 different participants, resulting in further 10,800 labels (50 summaries x 9 questions x 24 repetitions).

Participants were recruited using a local participant pool admitting German natives only. Before conducting the laboratory experiment, we collected participant information about age, gender, education and knowledge about the services and products of telecommunication service *Telekom* from which the queries and posts originate. Overall, 71 participants (33f, 38m, $M_{age} = 29$) completed the laboratory study in 51 days, spending 295,033 seconds, ca. 82 hours at total. The average number of evaluated summaries in an hour was 12 and they were paid 15 Euros per hour. Attained education was distributed over the complete range with 46% having completed high school, 7% college, 24% a Bachelor's degree and 23% Master's degree or higher. The question about knowledge on telecommunication service *Telekom* resulted in self-assessments of a 10% *very bad*, 24% *bad*, 39% *average*, 25% *good* and 1% *very good* answer distribution.

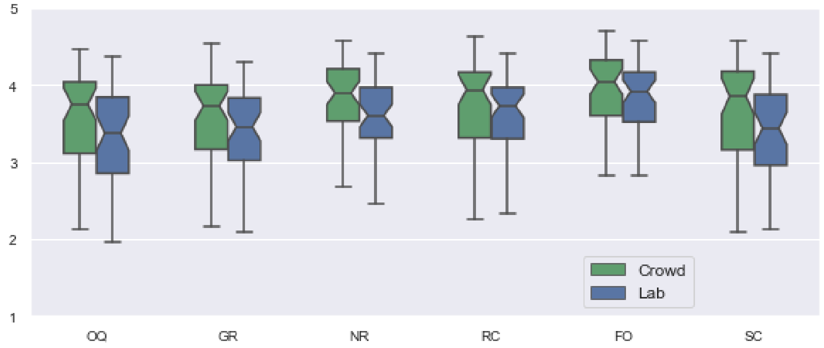


Fig. 1: Boxplots of Mean Ratings from Laboratory (blue) and Crowdsourcing (green) Assessment for OQ, GR, NR, RC, FO, SC

4 Results

Results are presented for the scores overall quality (OQ) and the five linguistic quality scores (including grammaticality (GR), non-redundancy (NR), referential clarity (RC), focus (FO), structure & coherence (SC)), and we will refer to these labels by their abbreviations in this section. The evaluation of the extrinsic factors, summary usefulness (SU), post usefulness (PU) and summary informativeness (SI), is future work.

Overall, we analyzed 7,200 labels (50 summaries x 6 questions x 24 repetitions) from the crowdsourcing study collected with 46 crowd workers and 7,200 labels from laboratory study collected with 71 participants. We used majority voting as the aggregation method which leads to 300 labels (50 summaries x 6 questions x average of 24 repetitions) from crowdsourcing and 300 labels from laboratory study. After data cleaning and basic outlier removal, we analyzed 288 labels, $N = 48$ per question, (48 summaries x 6 questions x average of 24 repetitions) from crowdsourcing and 288, $N = 48$ per question, from laboratory study.

4.1 Evaluation of Crowdsourcing Ratings

Anderson Darling tests for normality check were conducted to test the distribution of crowd ratings for OQ, GR, NR, RC, FO, and SC, indicating that all items are normally distributed with $p > 0.05$. Figure 1 shows the boxplots of each crowd rated item.

To determine the relationship between OQ and GR, NR, RC, FO, SC, Pearson correlations were computed (cf. table 1). With each of these linguistic quality items, OQ obtained a significant high correlation coefficient $r_p > .84$ and $p < .001$ which indicates a very strong linear relationship between the individual

Table 1: Pearson Correlation Coefficients of Crowd Ratings

Measure	**OQ**	**GR**	**NR**	**RC**	**FO**
GR	.894				
NR	.847	.711			
RC	.974	.881	.807		
FO	.970	.860	.806	.971	
SC	.960	.856	.857	.956	.945

$p < 0.001$ for all correlations

linguistic quality items and OQ, with the correlation between OQ and RC being the strongest ($r_p = .97$). In addition, linguistic quality items inter-correlate with each other significantly with $p < .001$ and $r_p > .71$ while the correlation coefficient between GR and NR being the weakest ($r_p = .711$), and correlation between RC and FO results the strongest ($r_p = .971$).

Before conducting a one-way ANOVA test to compare the means of OQ and the 5 linguistic quality scores for significant differences, Levene's test to check the homogeneity of variances was carried out with respective assumptions met. There were statistically significant differences between group means revealed by the one-way ANOVA ($p < .05$). Post hoc test applying Tukey criterion revealed that the mean of FO ($M = 3.937$) was significantly higher than the mean of OQ ($M = 3.588$, $p < 0.05$) and the mean of GR ($M = 3.565$, $p < 0.05$). No other significant differences were found.

4.2 Evaluation of Laboratory Ratings

Anderson Darling tests for normality check were conducted to test the distribution of laboratory ratings for OQ, GR, NR, RC, FO, and SC, indicating that all items are normally distributed ($p > 0.05$). Figure 1 shows the boxplots of each in laboratory rated item.

To determine the relationship between OQ and the five linguistic quality scores, Pearson correlations were computed (cf. table 2). With each of these linguistic quality items, OQ obtained a significant correlation coefficient $r_p > .79$ and $p < .001$ indicating a strong linear relationship between linguistic quality items and OQ, showing the correlation between OQ and SC as strongest ($r_p = .946$). In addition, linguistic quality items inter-correlate with each other significantly with $p < .001$ and $r_p >= .58$, where the correlation coefficient between GR and NR results the weakest ($r_p = .58$), and the correlation between RC and FO results the strongest ($r_p = .929$).

Before conducting a one-way ANOVA test to compare the OQ and the five linguistic quality scores with each other, Levene's test to check the homogeneity of variances was carried out resulting respective assumptions to be met. There were statistically significant differences between group means determined by one-way ANOVA ($p < .001$). Post hoc test (Tukey criterion) revealed that the mean

Table 2: Pearson Correlation Coefficients of Laboratory Ratings

Measure	OQ	GR	NR	RC	FO
GR	.836				
NR	.792	.580			
RC	.918	.733	.739		
FO	.920	.720	.789	.929	
SC	.946	.768	.828	.903	.903

p < 0.001 for all correlations

of FO (M = 3.827) was statistically higher than the means of the OQ (M = 3.359, $p < 0.01$), GR (M = 3.354, $p < 0.01$) as well as SC (M = 3.406, $p < 0.001$). Again, no other significant differences were found.

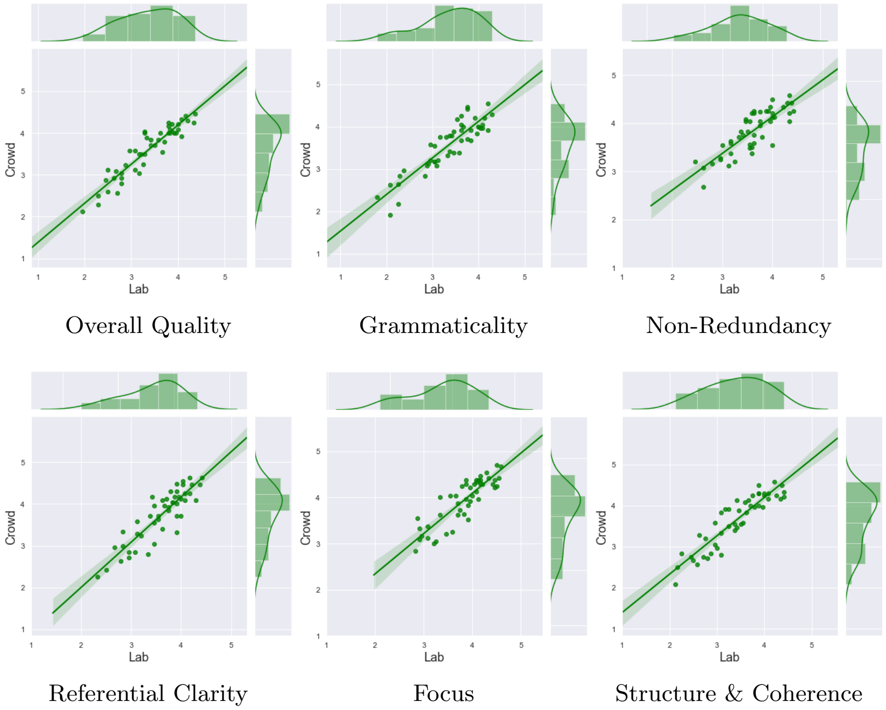
4.3 Comparing Crowdsourcing and Laboratory

When calculating Pearson correlation coefficients between MOS from crowd assessments and MOS from laboratory assessments, results of overall quality (r_p = .935), GR (r_p = .90), NR (r_p = .833), RC (r_p = .881), FO (r_p = .869) and SC (r_p = .911) with $p < .001$ for all correlations reveal overall very strong significant linear relationship between crowd and laboratory assessments. Figure 3 shows the dependency of crowd and laboratory ratings in scatter plots.

To compare OQ and the five linguistic quality scores from crowdsourcing with their respective items from laboratory ratings, T-tests assuming independent sample distributions were conducted. Before that, Anderson Darling tests for normality check and Levene's test to check the homogeneity of variances were carried out, with respective assumptions met. T-Test results revealed that there was no significant difference between OQ, GR, RC, FO and SC ratings with respect to the corresponding crowd and laboratory ratings. Only between-1, there was a significant difference, revealing that the mean of NR in laboratory (M = 3.60) was rated significantly lower than the mean of NR in crowdsourcing (M = 3.831, $p < .05$).

4.4 Preliminary Results: Towards Comparing Experts with Crowdsourcing and Laboratory

To explore the relationship between expert and non-expert ratings, we created a reference group with randomly selected 24 summaries (N = 24) from our data set to test the congruence of non-expert judgments by comparing them to an initial data set of expert judgments, with two experts rated OQ, GR, NR, RC, FO and SC of 24 summaries. We again used majority voting as our aggregation method when analysing the preliminary results of comparison expert judgments with non-expert judgments.



Overall Quality

Grammaticality

Non-Redundancy

Referential Clarity

Focus

Structure & Coherence

Fig. 3: Scatter Plot of Mean Ratings from Laboratory and Crowdsourcing Assessments

At first, we divided our data into three groups, all of which contain crowd, laboratory and expert judgments. The first group "All" includes all the summaries acting as the reference group ($N = 24$). Next, we split the data into subsets of high- and low-rated summaries by the median of crowd OQ ratings ($Mdn = 3.646$) and created the groups "Low" ($N = 12$) and "High" ($N = 12$) for crowdsourcing, laboratory and expert ratings respectively. Because of the resulting non-normal distribution in the groups, we calculated the Spearman rank-order correlation coefficients for all three groups between expert ratings and crowd ratings on the one side, and expert ratings and laboratory ratings on the other side. Table 3 shows the correlation coefficients for all groups.

Comparing correlation in between expert and crowd ratings as shown in the left plane of table 3, in the second column of the group "Low", a strong correlation can be found between OQ ratings, as well as between FO ratings. Also, the overall magnitude between experts and crowd increase in the group "Low" in comparison to the group "All". Comparing correlation in between expert and laboratory ratings as shown in the right plane of table 3, in the fifth column of the group "Low", we observe that the correlation coefficients for OQ, GR, and SC generally increase compared to group "All". However, the

Table 3: Spearman Rank-order Correlation Coefficients of Expert and Crowd Ratings, as well as Expert and Laboratory Ratings for Groups "All", "Low" and "High". Bold figures correspond to row maxima.

	Expert and Crowd			**Expert and Laboratory**		
	All	**Low**	**High**	**All**	**Low**	**High**
Measure	Corr coeff.	Corr coeff.	Corr coeff.	Corr coeff.	Corr coeff.	Corr coeff.
OQ	.441*	**.842****	.553	.422*	**.628***	.526
GR	.606**	**.731****	.411	.776***	**.854*****	.60*
NR	.279	.546	.162	.505*	.565	**.706***
RC	.553**	**.641***	.531	**.493***	.432	.497
FO	.626**	**.842****	.581*	.574**	.577*	**.646***
SC	.518**	**.614***	.462	.473*	**.756****	.242

***$p < 0.001$, **$p < 0.01$, *$p < 0.05$

correlation coefficients of NR and FO between expert and laboratory ratings increase comparing high-quality summaries to group "All".

In order to find out if labels assessed by crowd worker, laboratory participants or experts show significant differences, we compare the individual means of crowdsourcing, laboratory and experts assessments with respect to OQ and the five linguistic quality scores applying one-way ANOVA or in case of non-normal distribution Kruskal-Wallis tests. Results show significant differences in between crowdsourcing, laboratory and experts assessments with respect to means of OQ ratings ($p < .05$), GR ratings ($p < .001$), NR ratings ($p < .001$), RC ratings ($p < .001$) as well as FO ratings ($p < .001$). Eventually, ratings of SC ratings did not show significant difference. In a final analysis, we compare the absolute magnitudes and offset of ratings applying post hoc tests (Tukey criterion in case of normal distribution, and Dunn's criterion in case of non-normal distribution). Results revealed that the experts rated OQ (M = 3.771) significantly higher than the laboratory participants (M = 3.266, $p < .05$). Moreover, experts rated GR (M = 4.25), NR (M = 4.354), RC (M = 4.396) and FO (M = 4.333) significantly higher than the crowd workers (M_{GR} = 3.642, M_{NR} = 3.818, M_{RC} = 3.709, M_{FO} = 3.90, and $p < .05$ for all) and the laboratory participants (M_{GR} = 3.399, M_{NR} = 3.521, M_{RC} = 3.507, M_{FO} = 3.741, and $p < .05$ for all).

However, due to the small sample size of these three groups, these results need to be interpreted with caution and treated as preliminary results.

5 Discussion

In this paper, we have analyzed the appropriateness of micro-task crowdsourcing for the complex task of query-based extractive text summary creation by means of linguistic quality assessment.

The analysis of crowd ratings (cf. 4.1) and the analysis of laboratory ratings (cf. 4.2) have shown that there is a significant strong or very strong intercorrelation between overall quality ratings and the five linguistic scores in both environments, suggesting that non-experts associate the overall quality strongly with the linguistic quality. Additionally, an analysis of one-way ANOVA for both environments has revealed that the means of the individual scores do not differ from each other significantly, except the mean of focus. Interestingly, significantly higher overall ratings with respect to focus scores in both experiments indicate that the query-based summaries can well be assessed as highly focused on a specific topic on the one side, while at the same time they can be showing lower grammaticality or structure & coherence assessments. Potentially, this can be connected to the nature of query-based summaries from individual posts, the latter being likely to be focused on a given query.

With the comparison of crowd and laboratory ratings (cf. 4.3), we have shown that there is a statistically significant very strong correlation between overall quality ratings and the five linguistic quality scores, although crowd workers are not equally well-instructed compared the laboratory participants, e.g. receiving a personal introduction, a pre-written instructions sheet and being able to verbally clarify irritations. As the main finding of this paper, we showed that the degree of control on noise, mental distraction, and continuous work does not lead to any difference in the overall quality. Also, the presented results from the independent-samples T-tests support these findings, except for non-redundancy, which needs to be analyzed in more detailed work in the future. These findings highlight that crowdsourcing can be used instead of laboratory studies to determine the subjective overall quality and the linguistic quality of text summaries.

Additionally, as the preliminary results in section 4.4 reveal, crowd workers may even be preferred compared to experts in certain cases such as identifying overall quality and focus of low-quality summaries or determining the mean overall quality and structure & coherence of a summary data set. Following, laboratory participants may be preferred to experts in such cases like assessing the grammaticality and referential clarity of low quality summaries. Again, laboratory studies might be used to determine the mean structure & coherence of a summarization data set. Especially, using crowdsourcing to eliminate the bad quality summaries from the data set might be quite beneficial when training an automatic summarization tool with not annotated noisy data or when to decide on the application of experts or crowds in order to procure cost-efficient high quality text summaries at scale.

Further, since the automatic evaluation of text summaries always requires gold standard data to calculate metrics such as ROUGE [21], NLP research might profit from using crowdsourcing to determine the mean overall quality of an automatic summarization tool. In this way, the performance of different automatic summarization tools can be compared with each other without having a gold standard data which is costly to create. Especially, when assessing summaries to prepare training data for end-user directed summarization application, a naive assessment by non-expert crowd workers may even reflect a more realistic assessment with respect to non-expert level understanding and comprehension in the end user group in comparison to expert evaluation.

For all other items e.g. grammaticality, non-redundancy, referential clarity and focus, experts rate significantly higher than the crowd workers and laboratory participants. This observation might be explainable 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, hence their quality degradation may be more distinguishable and accessible to experts.

6 Conclusion and Future Work

In this paper, we execute a first step to answer the question “Can crowd successfully be applied to create query-based extractive summaries?”. Although crowd workers are not equally well-instructed compared to laboratory participants, crowdsourcing can be used instead of traditional laboratory assessments to evaluate the overall and linguistic quality of text summaries as shown above. This finding highlights that crowdsourcing facilitates a prospective of large-scale subjective overall and linguistic quality annotation of text summaries in a fast and cheap way, especially when naive end-users viewpoint is needed to evaluate an automatic summarization application or any kind of summarization method.

However, preliminary results also suggest that if expert annotations are needed, crowdsourcing and laboratory assessments can be used instead of experts only in certain cases such as identifying summaries with the low overall quality, grammaticality, focus and structure & coherence and also determining the mean overall quality and structure & coherence of a summary data set. So, if there is a not annotated summary data set which needs expert annotation, then the crowd workers or laboratory participants can not replace the experts since the correlation coefficients for mixed quality summaries are generally moderate. Additionally, the correlation coefficients of experts with non-experts vary in a range from weak to very strong in between groups. Currently, we cannot precisely explain these different correlation magnitudes of the experts and non-experts. The reasons for the dissent can be of multiple nature, for example, a different understanding of the guidelines, varying weighting of the objected summary parts or the lack of expertise. Right now, based on the comparative results of laboratory and crowd ratings, we can only exclude the online working characteristics of crowdsourcing such as unmotivated and unconcentrated crowd workers.

In future work, the reasons for these different correlation magnitudes will be investigated by collecting more expert data, as the expert ratings are collected for half of our data set. Also, qualitative interviews will be conducted with crowd workers to find out how well the guidelines are understood by them. Furthermore, this work does not include any special data cleaning or annotation aggregation method of 24 different judgments for a single item. Therefore, further analysis needs to be performed to answer the question of how many repetitions are enough for crowdsourcing and laboratory assessments so comparable results to experts can be obtained. Lastly, already collected extrinsic quality data will be analyzed to explore the relationship between overall, intrinsic and extrinsic quality factors. A deeper analysis of which evaluation measures are more sensitive to varying annotation quality will be also part of future work, in order to analyze more elaborately dependencies, requirements and applicability of a general application of crowd-based summary creation to help both humans and automated tools curating large online texts.

References

1. Callison-Burch, C.: 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. pp. 286–295. Association for Computational Linguistics (2009)
2. Chatterjee, S., Mukhopadhyay, A., Bhattacharyya, M.: Quality enhancement by weighted rank aggregation of crowd opinion. arXiv preprint arXiv:1708.09662 (2017)
3. Chatterjee, S., Mukhopadhyay, A., Bhattacharyya, M.: A review of judgment analysis algorithms for crowdsourced opinions. IEEE Transactions on Knowledge and Data Engineering (2019)
4. Cocos, A., Qian, T., Callison-Burch, C., Masino, A.J.: Crowd control: Effectively utilizing unscreened crowd workers for biomedical data annotation. Journal of biomedical informatics **69**, 86–92 (2017)
5. Conroy, J.M., Dang, H.T.: 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. pp. 145–152. Association for Computational Linguistics (2008)
6. Dang, H.T.: Overview of duc 2005. In: Proceedings of the document understanding conference. vol. 2005, pp. 1–12 (2005)
7. De Kuthy, K., Ziai, R., Meurers, D.: Focus annotation of task-based data: Establishing the quality of crowd annotation. In: Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016). pp. 110–119 (2016)
8. El-Haj, M., Kruschwitz, U., Fox, C.: Using mechanical turk to create a corpus of arabic summaries (2010)
9. Ellouze, S., Jaoua, M., Hadrich Belguith, L.: Mix multiple features to evaluate the content and the linguistic quality of text summaries. Journal of computing and information technology **25**(2), 149–166 (2017)
10. Falke, T., Meyer, C.M., Gurevych, I.: 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). pp. 801–811 (2017)
11. Fan, A., Grangier, D., Auli, M.: Controllable abstractive summarization. arXiv preprint arXiv:1711.05217 (2017)
12. Gadiraju, U.: Its Getting Crowded! Improving the Effectiveness of Microtask Crowdsourcing. Gesellschaft für Informatik eV (2018)

13. Gao, Y., Meyer, C.M., Gurevych, I.: April: Interactively learning to summarise by combining active preference learning and reinforcement learning. arXiv preprint arXiv:1808.09658 (2018)
14. Gillick, D., Liu, Y.: 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. pp. 148–151. Association for Computational Linguistics (2010)
15. Horton, J.J., Rand, D.G., Zeckhauser, R.J.: The online laboratory: Conducting experiments in a real labor market. Experimental economics **14**(3), 399–425 (2011)
16. Iskender, N., Gabryszak, A., Polzehl, T., Hennig, L., Möller, S.: A crowdsourcing approach to evaluate the quality of query-based extractive text summaries. In: 2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX). pp. 1–3. IEEE (2019)
17. Jones, K.S., Galliers, J.R.: Evaluating natural language processing systems: An analysis and review, vol. 1083. Springer Science & Business Media (1995)
18. Kairam, S., Heer, J.: Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing. pp. 1637–1648. ACM (2016)
19. Kittur, A., Nickerson, J.V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M., Horton, J.: The future of crowd work. In: Proceedings of the 2013 conference on Computer supported cooperative work. pp. 1301–1318. ACM (2013)
20. Kittur, A., Smus, B., Khamkar, S., Kraut, R.E.: Crowdforge: Crowdsourcing complex work. In: Proceedings of the 24th annual ACM symposium on User interface software and technology. pp. 43–52. ACM (2011)
21. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. Text Summarization Branches Out (2004)
22. Lin, Z., Ng, H.T., Kan, M.Y.: Automatically evaluating text coherence using discourse relations. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. pp. 997–1006. Association for Computational Linguistics (2011)
23. Lloret, E., Plaza, L., Aker, A.: Analyzing the capabilities of crowdsourcing services for text summarization. Language resources and evaluation **47**(2), 337–369 (2013)
24. Lloret, E., Plaza, L., Aker, A.: The challenging task of summary evaluation: an overview. Language Resources and Evaluation **52**(1), 101–148 (2018)
25. Malone, T.W., Laubacher, R., Dellarocas, C.: The collective intelligence genome. MIT Sloan Management Review **51**(3), 21 (2010)
26. Mani, I.: Summarization evaluation: An overview (2001)
27. Minder, P., Bernstein, A.: Crowdlang: A programming language for the systematic exploration of human computation systems. In: International Conference on Social Informatics. pp. 124–137. Springer (2012)
28. Nowak, S., Rüger, S.: 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. pp. 557–566. ACM (2010)
29. Pitler, E., Louis, A., Nenkova, A.: Automatic evaluation of linguistic quality in multi-document summarization. In: Proceedings of the 48th annual meeting of the Association for Computational Linguistics. pp. 544–554. Association for Computational Linguistics (2010)

30. Shapira, O., Gabay, D., Gao, Y., Ronen, H., Pasunuru, R., Bansal, M., Amsterdamer, Y., Dagan, I.: Crowdsourcing lightweight pyramids for manual summary evaluation. arXiv preprint arXiv:1904.05929 (2019)
31. Snow, R., O'Connor, B., Jurafsky, D., Ng, A.Y.: 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. pp. 254–263. Association for Computational Linguistics (2008)
32. Steinberger, J., Ježek, K.: Evaluation measures for text summarization. Computing and Informatics **28**(2), 251–275 (2012)
33. Torres-Moreno, J.M., Saggion, H., Cunha, I.d., SanJuan, E., Velázquez-Morales, P.: Summary evaluation with and without references. Polibits (42), 13–20 (2010)
34. Valentine, M.A., Retelny, D., To, A., Rahmati, N., Doshi, T., Bernstein, M.S.: Flash organizations: Crowdsourcing complex work by structuring crowds as organizations. In: Proceedings of the 2017 CHI conference on human factors in computing systems. pp. 3523–3537. ACM (2017)