Towards a Reliable and Robust Methodology for Crowd-Based Subjective Quality Assessment of Query-Based Extractive Text Summarization

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

Abstract

The intrinsic and extrinsic quality evaluation is an essential part of the summary evaluation methodology usually conducted in a traditional controlled laboratory environment. However, processing large text corpora using these methods reveals expensive from both the organizational and the financial perspective. For the first time, and as a fast, scalable, and cost-effective alternative, we propose micro-task crowdsourcing to evaluate both the intrinsic and extrinsic quality of query-based extractive text summaries. To investigate the appropriateness of crowdsourcing for this task, we conduct intensive comparative crowdsourcing and laboratory experiments, evaluating nine extrinsic and intrinsic quality measures on 5-point MOS scales. Correlating results of crowd and laboratory ratings reveals high applicability of crowdsourcing for the factors *overall quality, grammaticality, non-redundancy, referential clarity, focus, structure & coherence*, *summary usefulness*, and *summary informativeness*. Further, we investigate the effect of the number of repetitions of assessments on the robustness of mean opinion score of crowd ratings, measured against the increase of correlation coefficients between crowd and laboratory. Our results suggest that the optimal number of repetitions in crowdsourcing setups, in which any additional repetitions do no longer cause an adequate increase of overall correlation coefficients, lies between seven and nine for intrinsic and extrinsic quality factors.

Keywords: micro-task crowdsourcing, optimal repetition number, subjective linguistic quality evaluation

1. Introduction

In recent years, there has been an enormous increase in the need for multi-document summarization, trying to process the ever-increasing volume of information on the world wide web. Because of the high cost and time barriers of expert summarization, alternative solutions such as machine summarization tools or crowd-based summarization have been addressed over the past few years (Lloret et al., 2018). Especially, the evaluation of summary quality created by automatic tools is crucial in improving the automatic summarization tools. On the one hand, the appropriateness of a summary based on different criteria needs to be assessed for the training of these tools, and on the other hand, the quality of summaries generated by automatic summarization tools should be measured to determine their performance. Due to the subjectivity and ambiguity of summary quality evaluation, as well as the high variety of summarization approaches, there is a set of possible measures for the summary quality evaluation which can be broadly classified into two categories: extrinsic and intrinsic evaluation which are usually carried out in a traditional laboratory environment or with the help of experts (Jones and Galliers, 1995; Steinberger and Ježek, 2012).

Most of the cost and time barriers of qualitative and quantitative laboratory studies, controlled experiments, and expert evaluations can be reduced by using micro-task crowdsourcing (Horton et al., 2011; Gadiraju, 2018). Thus micro-task crowdsourcing has been widely used for quick and easy, isolated tasks such as image tagging, or print document digitization (Kittur et al., 2011), several researchers have attempted to explore crowdsourcing for challenging and expert tasks such as programming, product design, or NLP tasks (Kittur et al., 2013; Valentine et al., 2017). Particularly, empirical examination of numerous NLP tasks such as image recognition, sentiment analysis, and assessment of the performance of machine translation systems via crowdsourcing has shown that collective responses of crowd workers may provide gold standard data sets with quality approaching those generated by experts (Snow et al., 2008; Callison-Burch, 2009; Nowak and Rüger, 2010).

Inspired by these findings, we suggest using micro-task crowdsourcing to evaluate the extrinsic and intrinsic quality of query-based extractive forum summarization to overcome these time and cost barriers of summary quality evaluation. In particular, when the naive end-users viewpoint is needed to evaluate an automatic summarization application or any summarization method, the subjective quality evaluation plays an important role. For this subjective evaluation, micro-task crowdsourcing can provide the desired diversity of the potential testers accumulating a vast unfiltered number of crowd workers from all over the world (Hossfeld et al., 2014).

To our knowledge, only prior work by the authors themselves has considered the intrinsic quality evaluation of query-based extractive summarization as an application area of micro-task crowdsourcing and found out that crowdsourcing shows high applicability for determining the intrinsic quality factors (Iskender et al., 2019; Iskender et al., 2020). However, no study has considered micro-task crowdsourcing as an application area of extrinsic quality evaluation of query-based extractive forum summarization. Besides, the promise of time and cost reduction of micro-task crowdsourcing might be jeopardized if the repetition number per item used in mean opinion score (MOS) is too high. Therefore, the optimal point, in which the additional cost to increase the robustness of MOS in crowdsourcing is no longer worth the expected benefit, should also be investigated in detail. Again, to our knowledge, no other study in this domain has considered this aspect.

To do so, we concentrate in this paper on subjective qual-

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020) , pages 245–253
Marseille, 11–16 May 2020

ity assessment of query-based extractive text summaries by conducting both laboratory and crowdsourcing experiments to investigate the appropriateness of micro-task crowdsourcing for this task. Further, we collect 24 repetitions per item in both experiments to find out the optimal repetition number per item in the crowdsourcing setup.

2. Related Work

2.1. Subjective Summary Quality Evaluation

Summary quality assessment is crucial to the enhancement of machine summary tools. On the one hand, a summary data set should be assessed depending on different criteria to train these machine summarization tools, and on the other hand, the quality of machine summaries created by different tools should be evaluated to calculate the performance of these different tools. There is a set of possible measures for the summary quality evaluation, which can be broadly classified into two categories: intrinsic and extrinsic evaluation.

2.1.1. Intrinsic Quality Evaluation

In intrinsic evaluation, the summary quality assessment is directly based on itself without considering the source document and is usually carried out by comparison with a gold standard reference summary created by experts (Jones and Galliers, 1995). The text quality evaluation (or linguistic quality evaluation) and content evaluation are the two primary strategies to evaluate the intrinsic quality (Steinberger and Ježek, 2012). The linguistic quality evaluation is typically executed manually by humans and includes the evaluation of grammaticality, non-redundancy, referential clarity, focus, and structure & coherence (Dang, 2005). In the section 3., we determine these factors based on the definitions in Dang (2005) and Lloret et al. (2018).

In contrast to linguistic quality evaluation, the content evaluation is calculated automatically and finds out how many word sequences/sentences of gold standard reference summary are included in the peer summary. One of the popular automatic quality metrics is ROUGE, which provides a set of statistics (e.g., ROUGE-1 which uses 1-grams, ROUGE-2 which uses 2-grams) by applying a sequence of recall metrics based on n-gram co-occurrence between a peer summary and a list of gold standard reference summaries which are created by experts (Lin, 2004; Torres-Moreno et al., 2010). Nonetheless, the linguistic quality factors listed above can not be measured automatically in most cases (Steinberger and Ježek, 2012). The existing automatic evaluation methods for these are limited (Lin et al., 2011; Pitler et al., 2010; Ellouze et al., 2017), usually do not take into account the complex and subjective nature of the linguistic quality factors. Therefore, we do not focus on these automatic quality measurement tools in this paper.

2.1.2. Extrinsic Quality Evaluation

In extrinsic evaluation, the evaluation of summary quality is accomplished based on the source document. For the particular case of query-based extractive forum summarization, the source document consists of two bases: forum posts and the corresponding user query. The relationship between these forum posts, query, and extracted summary, as shown in figure 1 can be investigated extrinsically by using three main measures: *summary usefulness, post usefulness,* and *summary informativeness*.

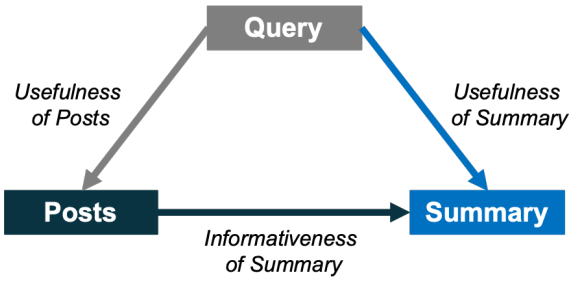


Figure 1: Extrinsic Quality Measures for Query-based Extractive Forum Summarization

The first extrinsic measure *summary usefulness*, also called content responsiveness, introduced in DUC 2003 (NIST, 2003), examines the summary's usefulness concerning external information need or goal basis (Shapira et al., 2019). In the evaluation of summary usefulness, the human evaluators give a single subjective score using a Likert scale, answering the question of how useful is the extracted summary to satisfy the given goal, in our case, to answer the given query. Following, the second factor *post usefulness*, also called relevance assessment, determines if the source document contains relevant information about the user's need by answering the question of how useful is the source document, in our case the forum posts, for answering the user need, in our case the user query (Mani, 2001; Conroy and Dang, 2008). Lastly, to measure the third factor *summary informativeness*, the human evaluators answer the question of how much information from the source document, in our case the forum posts, is preserved in the extracted summary (Mani, 2001).

2.2. Crowdsourcing for Summary Quality Evaluation

Recent studies in crowdsourcing have shown that even some complex tasks such as content writing, product design, or programming can be successfully accomplished by non-expert crowd workers with suitable task design and technological support (Kittur et al., 2013; Chatterjee et al., 2019; Chatterjee et al., 2017; Kairam and Heer, 2016). Notably, the need for scalable, low-cost corpus creation has increased the interest of researchers to use non-expert crowd workers for NLP annotation tasks, which are traditionally carried out by experts (Kairam and Heer, 2016; De Kuthy et al., 2016; Cocos et al., 2017).

In our case - the intrinsic and extrinsic summary quality evaluation - the use of micro-task crowdsourcing has not been examined as extensively as other NLP tasks, such as translation (Lloret et al., 2018). Gillick and Liu (2010) have used micro-task crowdsourcing to explore the reliability of crowd-based summary quality evaluation and revealed that non-expert crowd workers can not assess the quality of summaries as good as experts. Also, Gao et al. (2018), Falke et al. (2017), and Fan et al. (2017) have applied crowdsourcing to test the quality of their automatic

summarization tools, but not questioned the reliability of micro-task crowdsourcing for this task. Additionally, our previous work (Iskender et al., 2019; Iskender et al., 2020) has focused on crowd-based summary quality evaluation. Our results showed that micro-task crowdsourcing can be applied instead of laboratory studies to evaluate the intrinsic quality of text summaries, and crowd workers correlate moderately with experts only assessing the low-quality summaries.

However, the crowd is typically composed of people with unknown and very diverse abilities, skills, interests, personal objectives, and technological resources, which lead to several challenges related to lack of control on participants and consistency of output quality in crowdsourcing (Hossfeld et al., 2014). Therefore, outputs produced by the crowd must be checked for quality, and so the quality of crowd-based NLP annotations has been repeatedly questioned (Lloret et al., 2018).

To improve the quality of crowdsourcing annotations for NLP, researchers have developed several methods such as *filtering*, *aggregation* and *inferring bias* (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). Another approach is filtering low-quality data after task completion by adding trapping questions or using behavioral traces (Kittur et al., 2008; Rzeszotarski and Kittur, 2011). Further, the inferring bias method focuses on identifying and removing individual worker biases with a probabilistic model (Snow et al., 2008). Lastly, the crowd rating aggregation methods contain probabilistic models of annotation, accounting item level effects, and clustering methods (Passonneau and Carpenter, 2014; Whitehill et al., 2009; Luther et al., 2015), but the traditional majority voting or mean opinion score is still the most common technique (Chatterjee et al., 2019).

In this paper, we focus on the appropriateness of micro-task crowdsourcing for subjective summary quality evaluation by comparing crowdsourcing with the laboratory results. To our knowledge, there is no best practice guideline for summary quality evaluation regarding the optimal number of repetitions per item in crowdsourcing studies used in MOS. Therefore, we explore the relationship between the number of repetitions and the correlation coefficient between crowdsourcing and laboratory results to provide a best practice guideline regarding the optimal repetition number in MOS.

3. Experiments

3.1. Data Set

Our data set consists of query-based extractive forum summaries originating from the customer posts and queries of the forum *Deutsche Telekom hilft*. In this forum, the Telekom customers and support agents answer in a forum structure the customer questions regarding the Telekom products such as "where can I find my customer number?" or "My internet does not work. How can I fix this?". In our previous work (Iskender et al., 2019), the overall quality of the summary data set that we use in our experiment was already annotated in a crowdsourcing experiment using a 5-point MOS scale with three repetitions per item. Calculating the means of these ratings, the quality of these summaries was ranging from 1.667 to 5. Based on these overall quality ratings, we selected 50 summaries with ten distinct quality groups ranging from lowest to highest scores (lowest group [1.667, 2]; highest group (4.667, 5]), each included five summaries to create stratified data of widely varying qualities. The average word count of these selected summaries was 63.32, the shortest one with 24 words, and the longest one with 147 words where the average word count of the corresponding posts was 555, the shortest posts with 155 words, and the longest with 1005 words. Accordingly, customer queries had an average word count of 7.78, the shortest one with four words, and the longest with 17 words.

3.2. Crowdsourcing Study

We used the Crowdee([1]) platform for all of our crowdsourcing experiments using a simple workflow with two steps: qualification task and summary quality evaluation task. For the qualification task, we accepted only crowd workers who passed the German language proficiency screener provided by the Crowdee platform with a score of 0.9 and above (scale [0, 1]).

The qualification task started with a brief explanation of the summarization process. It was highlighted that the summaries were created by simple copy-paste from forum posts, and therefore they can appear slightly unnatural. Next, crowd workers were asked to evaluate the overall quality and the content quality of four reference summaries (two very good, two very bad). The quality of these reference summaries have already been determined by experts on a 5-point MOS scale using the labels *very good, good, moderate, bad, very bad* and these ratings have not been shown to crowd workers during the qualification task. For each exact rating match to expert rating, crowd workers gained 4 points, and for each point deviation from the expert rating, they got a point less, so deviations from the expert ratings were linearly punished. Out of 1569 screened crowd workers holding a German language score $>= 0.9$, 82 crowd workers completed this qualification task, which was online for one week. Sixty-seven out of screened crowd workers passed the qualification task with a point ratio $>= 0.625$. After two weeks, the summary quality evaluation task was published, and 46 qualified crowd workers came back to participate in this task.

In the summary quality evaluation task, we explained again that the summaries are created by copying sentences from forum posts to answer the user query without adjusting the original wording in the forum posts. After that, we presented an example of a query, forum posts, and a summary to provide some background information. Following, crowd workers rated nine quality factors of a single summary using a 5 point MOS Scale in 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

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

workers (intrinsic quality), in question 7, we showed the original query; in questions 8 and 9, the original query and the corresponding forum posts. The scoring of each aspect of a single summary was done on a separated page, which contained a short, informal definition of the particular aspect (sometimes illustrated with an example), the summary and the 5-point MOS scale (*very good, good, moderate, bad, very bad*).

Twenty-four repetitions were collected for each of these nine items for 50 summaries, resulting in 10,800 labels (50 summaries x 9 questions x 24 repetitions). We paid both for the qualification task and the summary quality evaluation task 1.2 Euros. The total cost of the crowdsourcing experiments was 1538.4 Euros (1.2 Euros x 82 participants in the qualification task, 50 summaries x 1.2 Euros x 24 repetitions in summary quality evaluation task). 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. Laboratory Study

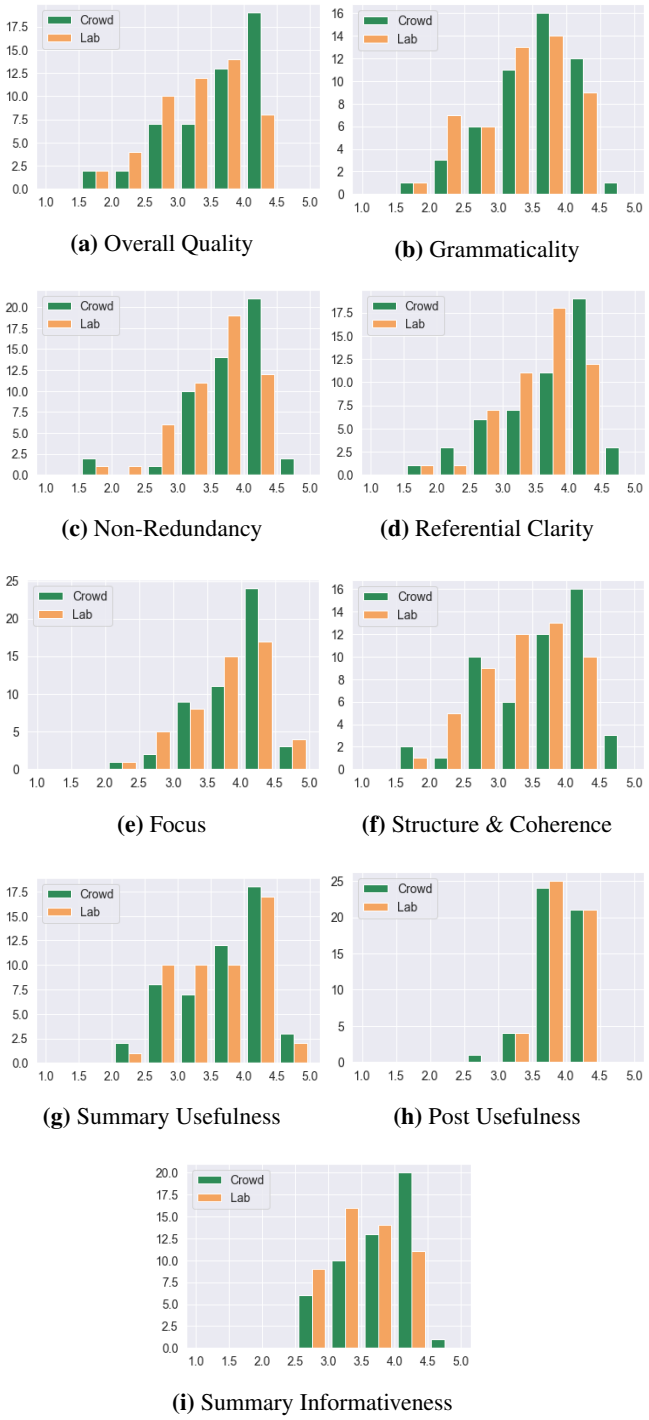
In the laboratory experiments, the participants were recruited via a local participant pool accepting German natives only. They did not perform any other qualification task. The summary quality evaluation task itself was the same as the one in the crowdsourcing study described in section 3.2..

In a controlled laboratory environment, participants used the Crowdee platform for performing the summary quality evaluation task. In contrast to the crowdsourcing study, all the participants were also instructed in a written form following the standard practice for laboratory tests. Besides, all of their questions during the written instruction and performing the task were answered immediately by the lab instructor. The experiment duration was set to one hour, and the participants were told to evaluate as many summaries as they can. Each of the 50 summaries were again rated by 24 different participants, resulting in further 10,800 labels (50 summaries x 9 questions x 24 repetitions).

In addition to the summary quality assessment, we collected participant information about age, gender, education, and knowledge about the services and products of telecommunication service *Telekom*. The laboratory experiment was completed in 51 days by 71 participants (33f, 38m, M_{age} = 29) who spent overall 295,033 seconds, ca. 82 hours for 10,800 labels. The average number of evaluated summaries in an hour was 12, and they were paid 15 Euros per hour, resulting in a total cost of 1065 Euros (15 Euros x 71 participants). Looking at the participant demographics, 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.

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.



(a) Overall Quality **(b)** Grammaticality

(c) Non-Redundancy **(d)** Referential Clarity

(e) Focus **(f)** Structure & Coherence

(g) Summary Usefulness **(h)** Post Usefulness

(i) Summary Informativeness

Figure 2: Histogram of Crowd (green) and Laboratory (orange) Ratings

Overall, we analyzed 10,800 labels (50 summaries x 9 questions x 24 repetitions) from the crowdsourcing and 10,800 labels from the laboratory study. As our aggregation method, we took the mean of 24 judgments per item and analyzed 450 labels (50 summaries x 9 questions x av-

Table 1: Median Values, Spearman Rank Order Correlation Coefficients, and Mann Whitney U Test Results of Crowd and Laboratory Ratings

Measure	OQ	GR	NR	RC	FO	SC	SU	PU	SI
Median-Lab	3.313	3.438	3.583	3.688	3.896	3.417	3.708	3.917	3.50
Median-Crowd	3.708	3.694	3.833	3.877	4.042	3.833	3.776	3.833	3.792
Spearman Corr. (Lab & Crowd)	0.925	0.873	0.807	0.844	0.852	0.919	0.858	0.637	0.808
U statistic	9850.5	1014.50	941.0	1046.50	1082.0	994.0	1162.0	1231.50	876.50
p value	0.034	0.053	0.017	0.081	0.124	0.039	0.273	0.451	0.005
Significant Diff.	**Yes**	No	**Yes**	No	No	**Yes**	No	No	**Yes**

$p < 0.001$ for all correlations

erage of 24 repetitions) from crowdsourcing and 450 labels from the laboratory study in section 4.1.. In section 4.2., we again used the aggregated 450 labels from the laboratory experiments and analyzed each of 24 repetitions crowd repetitions separately.

4.1. Comparing Crowdsourcing and Laboratory

The figure 2 displays the histograms for nine quality ratings collected in crowdsourcing and laboratory experiments. All of the distributions are slightly negatively skewed, with its peak shifted toward the upper end of its range, indicating a non-normal distribution. To test the normality of these crowdsourcing and laboratory ratings, we carried out Anderson Darling tests. The test results showed that OQ, NR, and SU from crowd ratings and NR, SU, and SI from laboratory ratings were not normally distributed ($p < 0.05$).
Because of these non-normal distributions, Spearman rank-order correlations coefficients and Mann Whitney U Tests were calculated to determine the relationship between crowd and laboratory ratings, as shown in table 1. All correlations between crowd and laboratory were statistically significant, ranging from moderate to very strong, where OQ had the highest correlation coefficient of 0.925 and PU the lowest correlation coefficient of 0.637. Following, to compare the environmental differences, Mann-Whitney U tests were conducted for each pair of nine quality measures (see table 1. The test results showed that there were no significant difference between GR, RC, FO, SU, and PU ratings with respect to the corresponding crowd and laboratory ratings. However, there were statistically signifcant difference between OQ_{Crowd} ($MdN = 3.708$) and OQ_{Lab} ($MdN = 3.313$), NR_{Crowd} ($MdN = 3.833$) and NR_{Lab} ($MdN = 3.583$), SC_{Crowd} ($MdN = 3.833$) and SC_{Lab} ($MdN = 3.417$), as well as SI_{Crowd} ($MdN = 3.792$) and SI_{Lab} ($MdN = 3.50$), showing that the crowd workers are rating these factors statistically higher than the laboratory participants.

4.2. Optimal Repetition Number: Curve Estimation and Knee Algorithm

In our data set, we have 24 different judgments for each nine quality dimensions of each summary. However, the relationship between crowd and laboratory ratings often reach a point at which the relative cost to increase the repetition number in crowdsourcing is no longer worth the corresponding correlation coefficient increase, so the need for 24 judgments should be investigated for a comprehensive cost-benefit analysis. To find out the optimal repetition number, we applied following algorithm to the each nine quality measures:

1. Let $S = \{s_1, ..., s_n\}$ be set of summaries to be evaluated and $J = \{j_1, ..., j_m\}$ set of judgments for a single quality measure e.g. overall quality. Combining these two sets J x S results in m x n matrix.

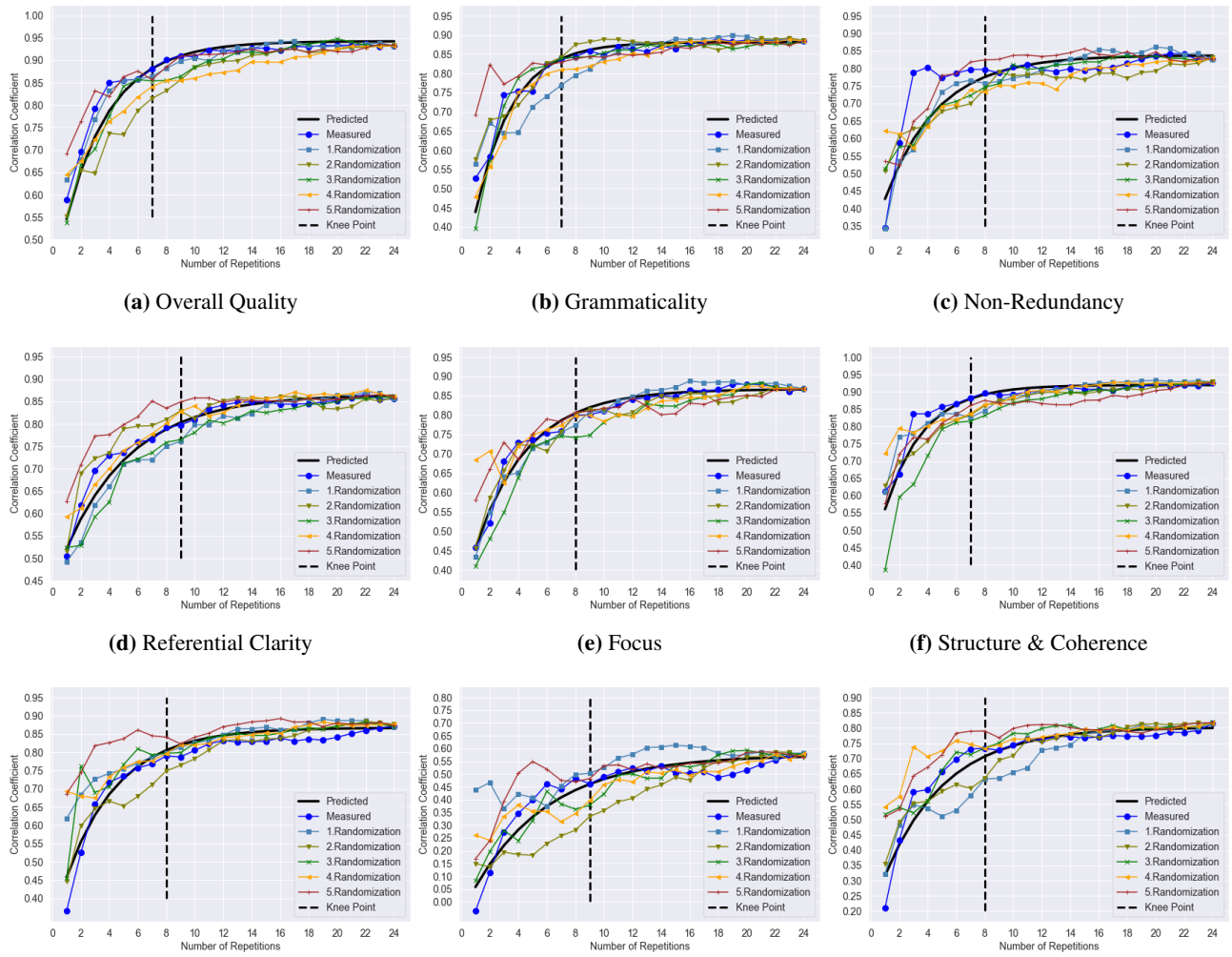
2. In each row of this matrix, the 24 different judgments for a single summary is represented. Next, MOS by repetition is applied to this matrix, meaning we get a new m x n matrix where each column shows the set of means M per item for a single summary:
$M = \{m_1, ... , m_m\}$, where
$m_1 = j_1$,
$m_2 = (j_1 + j_2) / 2, ...$,
$m_m = (j_1 + ... + j_m) / m$.

3. Following, let $L = \{l_1, ..., l_n\}$ be set of MOS in laboratory for each summary for a given item in form of n x 1 matrix.

4. Further, compute Spearman rank order correalation coefficient between the set L and the matrix J x S. Let the result $C = \{c_1, ... , c_n\}$ is set of correlation coefficients in form of m x 1 correlation matrix.

5. Lastly, we shuffle k times the order of the judgments in set J and repeat the steps 2, 3, and 4. The result is set of correlation coefficients for different randomizations $R = \{C_1, ... , C_k\}$.

In our case, we had 50 summaries in our data set and 24 repetitions for each summary quality measures, meaning $m = 24$, $n = 50$. And, we randomized the order of the judgments five times to lower the effect of lurking variable, so $k = 5$ where the number 5 was selected arbitrarily. Consequently, we got a set of correlation coefficients for unrandomized judgments $C_{measured}$, and C_1, C_2, C_3, C_4, and C_5 for five randomizations. All of these forms the set of correlation coefficients for different randomizations $R = \{C_{measured}, C_1, C_2, C_3, C_4, C_5\}$.
Looking at figure 3, it is clear that there is a saturation point



(a) Overall Quality **(b)** Grammaticality **(c)** Non-Redundancy

(d) Referential Clarity **(e)** Focus **(f)** Structure & Coherence

(g) Summary Usefulness **(h)** Post Usefulness **(i)** Summary Informativeness

Figure 3: Curve Estimation and Knee Points of Correlation Coefficients between Laboratory Ratings (All) and Crowd Ratings by Repetition for Overall Quality, Intrinsic and Extrinsic Quality Factors

between the number of repetitions and the resulting correlation coefficient where data points follow an exponential curve relationship. We aim to find out this saturation point, respectively optimal repetition number, called also *knee point*, in which the additional cost to increase the robustness of crowd judgment is no longer worth the expected correlation coefficient increase. To do so, we conduct curve fitting for each of the nine quality measures using Least Square Estimation for the following equation:

$$y = (a * (1 - e^{-b*x})) + c. \quad (1)$$

In figure 3, the correlation coefficients between laboratory (mean of all judgments) and crowd judgments by repetitions are shown for the nine quality measures. The black line plot labeled as *predicted* displays the estimated correlation coefficients using the curve fitting for the equation 1. Following, the dark blue line plot labeled as *measured* shows the correlation coefficients between aggregated laboratory ratings and the crowdsourcing ratings by repetition in observed order, the blue line plot labeled as *1. Randomization* the order after the first randomization, the olive line plot labeled as *2. Randomization* the order after the second randomization, the green line plot labeled as *3. Randomization* the order after the third randomization, the orange line plot labeled as *4. Randomization* the order after the fourth randomization, the brown line plot labeled as *5. Randomization* the order after the fifth randomization. Moreover, the dotted vertical black line plot displays the knee point.

We calculated these knee points by using the kneedle algorithm on the estimated exponential functions of each nine quality measures. Satopaa et al. (2011) have developed this algorithm based on the concept that in the data set, the points of the maximum curvature - called knee points - represents nearly the set of local maxima points in the curve when the curve is rotated θ degrees clockwise about (x_{min}, y_{min}) through the line formed by the points (x_{min}, y_{min}) and (x_{max}, y_{max}). The reason for choosing this line is to protect the overall behavior of the data set by applying the best fit line. When the curve is rotated about this line, the knee points display the points where the curve is most different from the straight line segment that links the first and last data points.

The table 2 shows all the estimated coefficients, knee points, and R^2 values for observed order and five random-

Table 2: Estimated Coefficients, Knee Points, and R^2 Values for Overall Quality, Intrinsic and Extrinsic Quality Measures

Quality Measure	Estimated Coefficients	Knee Point	R^2 Values for Measured and 5 Randomizations					
			Measured	1.Rand.	2.Rand.	3.Rand.	4.Rand.	5.Rand.
OQ	a=0.543, b=0.313, c=0.401	7	94.5%	94.0%	82.5%	95.8%	82.9%	79.5%
GR	a=0.652, b=0.386, c=0.230	7	94.7%	79.6%	88.6%	97.2%	95.3%	50.7%
NR	a=0.537, b=0.271, c=0.301	8	69.5%	94.2%	83.8%	94.4%	73.7%	86.7%
RC	a=0.425, b=0.213, c=0.440	9	96.2%	93.5%	84.5%	91.3%	95.3%	62.9%
FO	a=0.540, b=0.271, c=0.330	8	96.9%	96.7%	95.1%	86.2%	69.2%	82.6%
SC	a=0.519, b=0.370, c=0.40	7	93.0%	88.6%	90.1%	58.8%	74.1%	86.5%
SU	a=0.540, b=0.271, c=0.33	8	92.8%	78.6%	87.8%	81.1%	71.7%	36.5%
PU	a=0.624, b=0.190, c=-0.05	9	90.1%	29.5%	61.7%	92.5%	78.0%	62.4%
SI	a=0.612, b=0.232, c=0.190	8	92.2%	81.3%	93.3%	82.3%	54.3%	64.7%

ized orders of crowd judgments of overall quality, five intrinsic and three extrinsic quality measures. The overall model fit, R^2, for different orders of crowd judgments has reached at least one time 90% for all of the nine quality measures, 96.9% being the highest fit. Furthermore, it was mostly above 60% for all of the orders with a couple of exceptions in grammaticality, summary usefulness, post usefulness, and summary informativeness. Notably, in the first randomization of post usefulness, the R^2 was quite low, with a value of 29.5%. Also, in the fifth randomization of summary usefulness, R^2 was under the average with a value of 36.5%.

Looking at the optimal points, the minimum knee point was seven, the maximum nine, and the average eight for nine quality measures. This result shows that, on average, after eight repetitions, collecting one more additional crowd judgment was no longer worth the increase in correlation coefficient between crowd and laboratory ratings.

5. Discussion and Conclusion

In this paper, we have analyzed the appropriateness of micro-task crowdsourcing for the subjective task of overall quality, intrinsic quality and extrinsic quality evaluation of query-based extractive forum summarization, and also investigated the trade-off point in between the assessment repetition number, which is linked to increased study costs, and the quality gain obtained by additional repetitions with respect to overall robustness in micro-task crowdsourcing.

With the comparison of crowd and laboratory ratings (cf. 4.1.), we have shown that there is a statistically significant very strong correlation between crowd and laboratory ratings of overall quality and structure & coherence, strong correlation between crowd and laboratory ratings of grammaticality, non-redundancy, referential clarity, focus, summary usefulness and post informativeness, as well as moderate correlation between crowd and laboratory ratings of post usefulness. The reason for the comparatively lower correlation coefficient of post usefulness might be the different backgrounds of test participants, i.e., level of domain knowledge on the Telekom services and products, which makes it difficult to judge the usefulness of information given in the forum post for answering the user query.

When looking at the median values of nine quality factors in crowdsourcing and laboratory, we observed that crowd workers rated systematically higher than the laboratory participants except for post usefulness. Further, the results from the Mann Whitney U tests show that this higher rating is only for overall quality, non-redundancy, structure & coherence, and summary informativeness statistically significant. So, this higher rating tendency of crowd workers should be noted when using crowdsourcing instead of the laboratory experiments while evaluating the different subjective quality aspects of an automatic summarization tool or any summarization technique. Again, the reason for the non-fitting of post usefulness to this pattern might be the different background knowledge of participants about the Telekom services and products.

With these results, we showed as one of the main findings of this paper that the degree of control on noise, mental distraction, and continuous work does not lead to any difference in the summary quality evaluation. Although crowd workers were not equally well-instructed compared to the laboratory participants, e.g., receiving a personal introduction, a pre-written instructions sheet, and being able to verbally clarify irritations, the crowdsourcing shows high applicability for all these quality factors except for post usefulness when executing the task in the field.

Following, our results in section 4.2. reveal the first attempt of providing best practice guidelines regarding the optimal number of repetitions in crowdsourcing studies while evaluating the intrinsic and extrinsic summary quality. Based on this result, we can conclude that if the crowdsourcing study in this paper were conducted by collecting eight repetitions instead of 24 repetitions, then the cost and the time of the whole study would be reduced by 66% and we would still get very similar results. This is an important finding for future NLP research, especially when a reliable, cost-effective way of quality evaluation is needed for the comparison of large scale corpora or tool performance. Since the automatic evaluation of text summaries always requires gold standard data to calculate metrics such as ROUGE (Lin, 2004), NLP research might profit from using crowd-

sourcing for this task, especially when assessing the quality of end-user directed summarization applications.

However, the revealed optimal point is calculated based on an estimated model, and the model fit for randomized orders of 24 crowd ratings is ranging from 29.5% to 96.9% and become quite low in a couple of randomizations of summary informativeness and post usefulness. While evaluating these both quality aspects, the level of domain knowledge, here about Telekom's products and services, plays an essential role. So, this knowledge might be the lurking variable that is not included in the estimated model but can affect the correlation coefficient between aggregated laboratory ratings and aggregation of crowdsourcing ratings by repetition. In future work, the reasons for different model fits will be investigated by collecting more demographic and skill-related data of crowd workers so that the domain knowledge biases can be taken into account.

Furthermore, this work does not include any special data cleaning or annotation aggregation method other than the calculating mean values over 24 different judgments for a single item. Therefore, further analysis needs to be performed in order to find out the optimal aggregation method along with the corresponding optimal repetition number, such that comparable results to the laboratory can be obtained in a reliable and cost-effective way. A more in-depth analysis of which evaluation measures are more sensitive to varying annotation quality will also be part of future work. Lastly, we also plan to investigate the relationship between expert and crowd ratings both for intrinsic and extrinsic quality evaluation in order to more deeply understand the relationship between very high domain knowledge, very high linguistic expertise, and the process of multi-expert label convergence finding.

6. Bibliographical References

Callison-Burch, C. (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.

Chatterjee, S., Mukhopadhyay, A., and Bhattacharyya, M. (2017). Quality enhancement by weighted rank aggregation of crowd opinion. *arXiv preprint arXiv:1708.09662*.

Chatterjee, S., Mukhopadhyay, A., and Bhattacharyya, M. (2019). A review of judgment analysis algorithms for crowdsourced opinions. *IEEE Transactions on Knowledge and Data Engineering*.

Cocos, A., Qian, T., Callison-Burch, C., and Masino, A. J. (2017). Crowd control: Effectively utilizing unscreened crowd workers for biomedical data annotation. *Journal of biomedical informatics*, 69:86–92.

Conroy, J. M. and Dang, H. T. (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.

Dang, H. T. (2005). Overview of duc 2005. In *Proceedings of the document understanding conference*, volume 2005, pages 1–12.

De Kuthy, K., Ziai, R., and Meurers, D. (2016). 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)*, pages 110–119.

Ellouze, S., Jaoua, M., and Hadrich Belguith, L. (2017). Mix multiple features to evaluate the content and the linguistic quality of text summaries. *Journal of computing and information technology*, 25(2):149–166.

Falke, T., Meyer, C. M., and Gurevych, I. (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.

Fan, A., Grangier, D., and Auli, M. (2017). Controllable abstractive summarization. *arXiv preprint arXiv:1711.05217*.

Gadiraju, U. (2018). *Its Getting Crowded! Improving the Effectiveness of Microtask Crowdsourcing*. Gesellschaft für Informatik eV.

Gao, Y., Meyer, C. M., and Gurevych, I. (2018). April: Interactively learning to summarise by combining active preference learning and reinforcement learning. *arXiv preprint arXiv:1808.09658*.

Gillick, D. and Liu, Y. (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.

Horton, J. J., Rand, D. G., and Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental economics*, 14(3):399–425.

Hossfeld, T., Keimel, C., Hirth, M., Gardlo, B., Habigt, J., Diepold, K., and Tran-Gia, P. (2014). Best practices for qoe crowdtesting: Qoe assessment with crowdsourcing. *IEEE Transactions on Multimedia*, 16(2):541–558.

Iskender, N., Gabryszak, A., Polzehl, T., Hennig, L., and Möller, S. (2019). A crowdsourcing approach to evaluate the quality of query-based extractive text summaries. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE.

Iskender, N., Polzehl, T., and Möller, S. (2020). Crowdsourcing versus the laboratory: towards crowd-based online information extraction from query-based forum discussions. In *The conference for intelligent content solutions (QURATOR 2020) (Accepted)*, pages 1–16.

Jones, K. S. and Galliers, J. R. (1995). *Evaluating natural language processing systems: An analysis and review*, volume 1083. Springer Science & Business Media.

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

Kittur, A., Chi, E. H., and Suh, B. (2008). Crowdsourcing user studies with mechanical turk. In *Proceedings of*

the SIGCHI conference on human factors in computing systems, pages 453–456. ACM.

Kittur, A., Smus, B., Khamkar, S., and Kraut, R. E. (2011). Crowdforge: Crowdsourcing complex work. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 43–52. ACM.

Kittur, A., Nickerson, J. V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M., and Horton, J. (2013). The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1301–1318. ACM.

Lin, Z., Ng, H. T., and Kan, M.-Y. (2011). 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*, pages 997–1006. Association for Computational Linguistics.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Lloret, E., Plaza, L., and Aker, A. (2018). The challenging task of summary evaluation: an overview. *Language Resources and Evaluation*, 52(1):101–148.

Luther, K., Tolentino, J.-L., Wu, W., Pavel, A., Bailey, B. P., Agrawala, M., Hartmann, B., and Dow, S. P. (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.

Mani, I. (2001). Summarization evaluation: An overview.

Mitra, T., Hutto, C. J., and Gilbert, E. (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.

NIST. (2003). Duc 2003: Documents, tasks, and measures. `https://duc.nist.gov/duc2003/tasks.html`. Accessed: 2019-11-20.

Nowak, S. and Rüger, S. (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.

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

Pitler, E., Louis, A., and Nenkova, A. (2010). Automatic evaluation of linguistic quality in multi-document summarization. In *Proceedings of the 48th annual meeting of the Association for Computational Linguistics*, pages 544–554. Association for Computational Linguistics.

Rzeszotarski, J. M. and Kittur, A. (2011). Instrumenting the crowd: using implicit behavioral measures to predict task performance. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 13–22. ACM.

Satopaa, V., Albrecht, J., Irwin, D., and Raghavan, B. (2011). Finding a" kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st international conference on distributed computing systems workshops*, pages 166–171. IEEE.

Shapira, O., Gabay, D., Gao, Y., Ronen, H., Pasunuru, R., Bansal, M., Amsterdamer, Y., and Dagan, I. (2019). Crowdsourcing lightweight pyramids for manual summary evaluation. *arXiv preprint arXiv:1904.05929*.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (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.

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

Torres-Moreno, J.-M., Saggion, H., Cunha, I. d., SanJuan, E., and Velázquez-Morales, P. (2010). Summary evaluation with and without references. *Polibits*, (42):13–20.

Valentine, M. A., Retelny, D., To, A., Rahmati, N., Doshi, T., and Bernstein, M. S. (2017). Flash organizations: Crowdsourcing complex work by structuring crowds as organizations. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 3523–3537. ACM.

Whitehill, J., Wu, T.-f., Bergsma, J., Movellan, J. R., and Ruvolo, P. L. (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.