How to Evaluate a Summarizer: Study Design and Statistical Analysis for Manual Linguistic Quality Evaluation

Julius Steen Katja Markert
Department of Computational Linguistics
Heidelberg University
69120 Heidelberg, Germany
(steen|markert)@cl.uni-heidelberg.de

Abstract

Manual evaluation is essential to judge progress on automatic text summarization. However, we conduct a survey on recent summarization system papers that reveals little agreement on how to perform such evaluation studies. We conduct two evaluation experiments on two aspects of summaries' linguistic quality (coherence and repetitiveness) to compare Likert-type and ranking annotations and show that best choice of evaluation method can vary from one aspect to another. In our survey, we also find that study parameters such as the overall number of annotators and distribution of annotators to annotation items are often not fully reported and that subsequent statistical analysis ignores grouping factors arising from one annotator judging multiple summaries. Using our evaluation experiments, we show that the total number of annotators can have a strong impact on study power and that current statistical analysis methods can inflate type I error rates up to eight-fold. In addition, we highlight that for the purpose of system comparison the current practice of eliciting multiple judgements per summary leads to less powerful and reliable annotations given a fixed study budget.

1 Introduction

Current automatic metrics for summary evaluation have low correlation with human judgements on summary quality, especially for linguistic quality evaluation (Fabbri et al., 2020). As a consequence, manual evaluation is still vital to properly compare the linguistic quality of summarization systems.

While the document understanding conferences (DUC) established a standard manual evaluation procedure (Dang, 2005), we conduct a comprehensive survey of recent works in text summarization that reveals a wide array of different evaluation questions and methods in current use. Furthermore, DUC procedures were designed for a small set of expert judges, while current evaluation campaigns are often conducted by untrained crowd-workers. The design of the manual annotation, specifically *the overall number of annotators* as well as *the distribution of annotators to annotation items*, has substantial impact on power, reliability and type I errors of subsequent statistical analysis. However, most current papers (see Section 2) do not consider the interaction of annotation design and statistical analysis. We investigate the optimal annotation methods, design and statistical analysis of summary evaluation studies, making the following contributions:

1. We conduct a comprehensive survey on the current practices in manual summary evaluation in Section 2. Often, important study parameters, such as the total number of annotators, are not reported. In addition, statistical significance is either not assessed at all or with tests (t-test or one-way ANOVA) that lead to inflated type I error in the presence of grouping factors (Barr et al., 2013). In summarization evaluation, grouping factors arise whenever one annotator rates multiple summaries.

2. We carry out annotation experiments for coherence and repetition. We use both Likert- and ranking-style questions on the output of four recent summarizers and reference summaries. We show that ranking-style evaluations are more reliable and cost-efficient for coherence, similar to prior findings by Novikova et al. (2018) and Sakaguchi and Van Durme (2018). However, on repetition, where many documents do not exhibit any problems, Likert outperforms ranking.

3. Based on our annotation data, we perform

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 1861–1875
April 19 - 23, 2021.

Monte-Carlo simulations to show the risk posed by ignoring grouping factors in statistical analysis and find up to eight-fold increases in type I errors when using standard significance tests. As an alternative, we propose to either use mixed effect models (Barr et al., 2013) for analysis or to design studies in such a manner that results can be aggregated into independent samples, amenable to simpler analysis tools.

4. Finally, we show that the common practice of eliciting repeated judgements for the same summary leads to less reliable and powerful studies for system-level comparison when compared to studies with the same budget but only one judgement per summary.

Code and data for our experiments is available at https://github.com/julmaxi/summary_lq_analysis.

2 Literature Survey

We survey all summarization papers in ACL, EACL, NAACL, ConLL, EMNLP, TACL and the *Computational Linguistics* journal in the years 2017-2019. We choose this timeframe as we are interested in current practices in summarization evaluation: 2017 marks the publication of the pointer generator network (See et al., 2017), which has been highly influential for neural summarization. We focus our analysis on papers that present *a novel system* for single- or multi-document summarization and take a single or multiple full texts as input and also output text (SDS/MDS). This allows us to concentrate on recommendations for human evaluation of newly developed summarization systems.[1]

Out of the resulting **105** SDS/MDS system papers, we identify all papers that conduct at least one new comparative system evaluation with human annotators for further analysis, leading to **58** papers in the survey. The fact that this is only about half of all papers is troubling given that it has been recently demonstrated that current automatic evaluation measures such as ROUGE (Lin, 2004) are

[1]Excluded from the analysis are sentence summarization or headline generation papers, although most of the points we make hold for their evaluation campaigns as well. Summarization evaluation papers that do not present a new system but concentrate on sometimes large-scale system comparisons are discussed in the Related Work section instead. Lists of all included and excluded papers are given in Supplementary Material, which also contains exact evaluation parameters per paper in a spreadsheet.

	Category	Pa.	St.
Evaluation Questions	Overall	17	23
	Content	45	65
	Fluency	29	34
	Coherence	10	11
	Repetition	14	17
	Faithfulness	6	8
	Referential Clarity	2	2
	Other	8	9
Evaluation Method	Likert	32	43
	Pairwise	10	14
	Rank	9	9
	BWS	6	9
	QA	9	14
	Binary	4	4
	Other	2	2
Number of Documents in Evaluation	< 20	6	10
	20-34	22	41
	35-49	3	4
	50-99	14	21
	100	11	14
	> 100	4	4
	not given	1	1
Number of Systems considered	< 3	13	20
	3	17	23
	4	16	23
	5	6	10
	> 5	12	19
	w/ Reference	16	25
	w/o Reference	45	70
Number of Annotations per Summary	1	2	5
	2-3	20	30
	4-5	12	27
	6-10	3	5
	not given	23	28
Overall Number of Annotators	1-5	19	25
	6-10	3	3
	> 10	5	9
	not given	32	58
Annotator Recruitment	Crowd	25	49
	Other	35	46
Statistical Evaluation	t-test	9	16
	ANOVA	9	18
	CI	4	6
	Other/unspecified	7	8
	None	32	47

Table 1: Our survey for 58 system papers with 95 manual evaluation studies (2017-2019). We show numbers both for individual studies and per paper. As a paper may contain several studies with different parameters, counts in the paper column do not always add up.

not good at predicting summary scores for modern systems (Schluter, 2017; Kryscinski et al., 2019; Peyrard, 2019).

We assess both *what* studies ask annotators to judge, as well as *how* they elicit and analyse judgements. The survey was conducted by one of the authors: for most papers, the categories they fell into were obvious. For difficult cases (unclear specifications, papers that do not fit the normal mould) the two authors discussed the categorisations. Survey results are given in Table 1. Further details about the choices made in the survey, including category groupings/definitions and what is included under *Other*, can be found in Appendix B. As many papers conduct more than one human evaluation (for example on different corpora), we also list individual annotation studies (a total of 95).

Of the systems that do have human evaluation, many focus on *content*, including informativeness, coverage, focus, and relevance. Where linguistic quality is evaluated, most focus on general questions about fluency/readability, with a smaller number of papers evaluating coherence and repetition.

In the rest of this section we focus on the three aspects of evaluation we cover in this paper: How to elicit judgements, how these judgements are analysed statistically and how studies are designed.

2.1 Methods

The majority of evaluations is conducted using Likert-type judgements, with the second most frequent method being rank-based annotations, including pairwise comparison. Best-worst scaling (BWS) is a specific type of ranking-oriented evaluation that requires annotators to specify only the first and last rank (Kiritchenko and Mohammad, 2017). QA (Narayan et al., 2018) is used for content evaluation only. This motivates us to compare both Likert and ranking annotations in Section 4.1.

2.2 Statistical Analysis

If a significance test is conducted, most papers analyse their data either using ANOVA or a sequence of paired t-tests. Both tests are based on the assumption that judgements (or pairs of judgements, in case of paired t-test) are sampled *independently* from each other. However, in almost all studies, annotators give judgements on more than one summary from the same system. Thus the resulting judgements are only independent if we assume that all annotators behave identically. Given that prior work (Gillick and Liu, 2010; Amidei et al., 2018), as well as our own reliability analysis in Section 4.1, show that especially crowd-workers tend to disagree about judgements, this assumption does not seem warranted. As a consequence, traditional significance tests are at high risk of inflated type I error rates. This is well known in the broader field of linguistics (Barr et al., 2013), but is disregarded in summarization evaluation. We show in Section 5 that this is a substantial problem for current summarization evaluations and suggest alternative analysis methods.

2.3 Design

Most papers only report the number of documents in the evaluation and the number of judgements *per summary*. This, however, is not sufficient to describe the design of a study, lacking any indication about the overall number of annotators that made these judgements. A study with 100 summaries and 3 annotations per summary can mean 3 annotators did all judgements in one extreme, or a study with 300 distinct annotators in the other. Only 26 of the 95 studies describe their annotation design in full, almost all of which use designs in which a small number of annotators judge all summaries. Only **6** of **49** crowdsourced studies report the full design.

We show in Section 5 that a low total number of annotators aggravates type I error rates with improper statistical analysis. In Section 6 we further show that with proper analysis, a low total number of annotators leads to less powerful experiments. Almost all analysed papers choose designs with multiple judgements per summary. However, we show in Section 6.2 that this — for the purpose of system ranking — leads to loss of reliability as well as power when compared to a study with the same budget and only one annotation per summary.

3 Coherence and Repetition Annotation

To elicit summary judgements for analysis, we conduct studies on two linguistic quality tasks. In the first, we ask annotators to judge the *coherence* of the summaries, while in the second we ask for the *repetitiveness* of the summary. We select these two tasks over the more frequent *Fluency* task as we found in preliminary investigations that many recent summarization systems already produce highly fluent text, making them hard to differentiate. We do not evaluate *Overall* and *Content* as both require access to the input document, which differentiates these questions from the linguistic

quality evaluation of the summaries.

For both tasks, we conduct one study using a seven-point Likert-scale (`Likert`) and another using a ranking-based annotation method (`Rank`), where annotators rank summaries for the same document from best to worst. Screenshots of the interfaces for both approaches and full annotator instructions are given in Appendix A.

Corpus and Systems. Mirroring a common setup (see Section 2), we select four abstractive summarization systems and the reference summaries (`ref`) for analysis.

- The pointer generator summarizer (`PG`) (See et al., 2017), which is still often used as a baseline for abstractive summarization
- The abstractive sentence rewriter (`ASR`) of Gehrmann et al. (2018), which is a strong summarization system that does not rely on external pretraining for its generation step
- `Seneca` (Sharma et al., 2019), a system that combines explicit modelling of coreference information with an external coherence model
- `BART` (Lewis et al., 2020), a transformer network that achieves SotA on CNN/DM.

We randomly sample 100 documents from the popular CNN/DM corpus (Hermann et al., 2015) with corresponding summaries from all systems to form the item set for all our studies.

Study design. We ensure a sufficient total number of annotators by using a block design. We separated our corpus into 20 blocks of 5 documents and included all 5 summaries for each document in the same block, which results in $5 \times 5 = 25$ summaries per block.

All items in a block were judged by the same set of three annotators. No annotator was allowed to judge more than one block. This results in a total of $3 \times 20 = 60$ annotators and 1500 judgements per task. Figure 1 shows a schematic overview of our design, which balances the need for a large enough annotator pool with a sufficient task size to be worthwhile to annotators.

We recruited native English speakers from the crowdsourcing platform Prolific[2] and carefully adjusted the reward to be no lower than £7.50 per hour based on pilot studies. Summaries (or sets of summaries for `Rank`) within a block were presented in random order.

[2] `prolific.com`

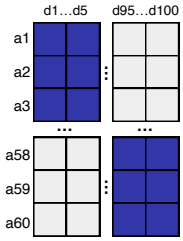


Figure 1: Schematic representation of our study design. Rows represent annotators, columns documents. Each blue square corresponds to a judgement of the summaries of all five systems for a document. Every rectangular group of blue squares forms one block.

4 Ranking vs. Likert

Table 2 shows the average Likert scores and the average rank for all systems, tasks and annotation methods. We use mixed-effect ordinal regression to identify significant score differences (see Section 5 for details). Both annotation methods provide compatible system rankings for the two tasks, though for the repetition task both methods struggle to differentiate between systems. If we were interested in the true ranking, we could conduct a power analysis given some effect size of interest and elicit additional judgements to improve the ranking. However, as we are concerned with the *process* of system evaluation and not the evaluation itself, we do not conduct any further analysis.

In the remainder of this section, we focus on the reliability of the two methods as well as their cost-effectiveness.

4.1 Reliability

Traditionally, reliability is computed by chance-adjusted agreement on individual instances. However, for NLG evaluation, Amidei et al. (2018) argue that a low agreement often reflects variability in language perception. Additionally, we are not interested in individual *document scores*, but in whether independent runs of the same study would result in consistent *system scores*. In Table 3 we thus report split-half reliability (SHR) in addition to Krippendorffs α (Krippendorff, 1980). To compute SHR, we randomly divide judgements into two groups that share neither annotators nor documents, i.e. two independent runs of the study. We

System	Likert (Coh)	Rank (Coh)	Likert (Rep)	Rank (Rep)
BART	$5.25^{(1)}$	$1.73^{(1)}$	$5.85^{(2/3)}$	$2.88^{(2/3/4)}$
ref	$4.33^{(3/4)}$	$3.31^{(3/4)}$	$6.14^{(1/2)}$	$2.41^{(1/2)}$
ASR	$4.17^{(3/4)}$	$3.17^{(3/4)}$	$4.88^{(4/5)}$	$3.51^{(4/5)}$
PG	$4.81^{(2)}$	$2.68^{(2)}$	$5.63^{(3)}$	$2.92^{(3/4)}$
seneca	$3.52^{(5)}$	$4.11^{(5)}$	$5.16^{(4/5)}$	$3.27^{(3/4/5)}$

Table 2: Results of our annotation experiment. Numbers in brackets indicate rank for a system for a given annotation method. Multiple ranks in the brackets indicate systems at these ranks are not statistically significantly different ($p \geq 0.05$, mixed-effects ordinal regression).

System	α	SHR
Coh: Likert	0.22	0.96
Coh: Rank	0.43	0.98
Rep: Likert	0.27	0.95
Rep: Rank	0.18	0.91

Table 3: Krippendorffs α with ordinal level of measurement and Split-Half-Reliability for both annotation methods on the two tasks.

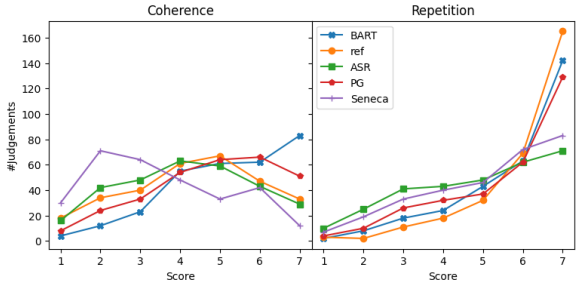


Figure 2: Score distribution of Likert for both tasks. Each data point shows the number of times a particular score was assigned to each system.

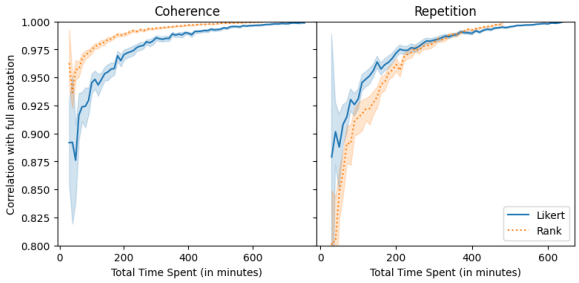


Figure 3: Time spent on annotation (in minutes) vs. correlation with the full-sized score. We gather annotation times in buckets with a width of ten minutes and show the 95% confidence interval for each bucket.

then compute the correlation[3] between the system scores in both halves. The final score is the average correlation after 1000 trials.

Though agreement on individual summaries is relatively low for all annotation methods, studies still arrive at consistent system scores when we average over many annotators as demonstrated by the SHR. This reflects similar observations made by Gillick and Liu (2010).

We find that on coherence, Rank is more reliable than Likert, though not on repetition. An investigation of the Likert score distributions for both tasks in Figure 2 shows that coherence scores are relatively well differentiated whereas a majority of repetition judgements give the highest score of 7, indicating no repetition at all in most summaries. We speculate overall agreement suffers, because ranking summaries with similarly low level of repetition (and not allowing ties) is potentially arbitrary.[4]

4.2 Cost-efficiency

While more reliable annotation methods allow for fewer annotations, the cost of a study is ultimately determined by the work-time that needs to be invested to achieve a reliable result. To investigate this, we randomly sample between 2 and 19 blocks from our annotations and compute the total time annotators spent to complete each sample. We also compute the Pearson correlation of the system scores in each sample with the scores on the full annotation set. We relate time spent to similarity between sample and full score in Figure 3.

For coherence, Rank is more efficient than Likert. On repetition, the lower reliability of Rank also results in lower efficiency. However, with additional annotation effort, reliability becomes on-par with Likert. This is a consequence of the overall lower annotator workload for Rank.

[3] We use the Pearson correlation implementation of scipy (Virtanen et al., 2020).

[4] This is supported by feedback we received from annotators that the summaries were difficult to rank as they mostly avoided repetition well.

5 Statistical Analysis and Type I Errors

The two most common significance tests in summarization studies, ANOVA and t-test (see Table 1), both assume judgements (or pairs of judgements, in the case of t-test) are independent. This is, however, not true for most study setups as a single annotator typically judges multiple summaries and multiple summaries are generated from the same input document. Both documents and annotators are thus grouping factors in a study that must be taken into account by the statistical analysis. Generalized mixed effect models (Barr et al., 2013) offer a solution but have, to the best of our knowledge, not been used in summarization evaluation at all. We choose a mixed effect ordered logit model to analyse our Likert data for both tasks.[5] We will show that traditional analysis methods have a substantially elevated risk of type I errors, i.e. differences between systems found in manual analysis might be overstated.

Method. The ordered logit model we employ can be described as follows:

$$logit(P(Y \leq c)) = \mu_c - (X\beta + Z_a u_a + Z_d u_d)$$

where $P(Y \leq c)$ is the probability that the score of a summary is at most c. μ_c is the threshold coefficient for level c, β is the vector of fixed effects and u_a, u_d are the vectors of annotator- and document-level random effects respectively, where u_a, u_d are both drawn from normal distributions with mean 0. Finally, X, Z_a, Z_d are design matrices for fixed and random effects. As the only fixed effect, we use a dummy-coded variable indicating the system that has produced the summary, with `ref` as the reference level. We estimate both random intercepts and slopes for both documents and annotators following advice of Barr et al. (2013) to always specify the maximal random effect structure. In practical terms this means that we allow annotators to both differ in how harsh or generous they are in their assessment, as well as in which system they prefer. Similarly, we allow system performance to vary per-document, leading to both generally higher or lower scores, as well as different system rankings per document.

[5]We do not include Rank data as the ordinal regression model does not generate ranks.

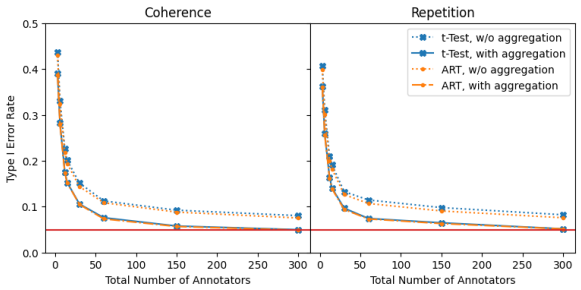


Figure 4: Relation of type I error rates at $p < 0.05$ to the total number of annotators for different designs, all with 100 documents and 3 judgements per summary. We conduct the experiment with both the t-test and approximate randomization test (ART). We show results both with averaging results per document and without any aggregation. We run 2000 trials per design. The red line marks the nominal error rate of 0.05.

We fit all models using the `ordinal` R-package (Christensen, 2019) and compute pairwise contrasts between the parameters estimated for each system using the `emmeans`-package (Lenth et al., 2018) with Tukey-adjustment.

To demonstrate the problem of ignoring the grouping factors, we can now sample artificial data from the model distribution and try to analyse it with inappropriate tests. This Monte-Carlo simulation is similar to the more general analysis of Barr et al. (2013).

We set β to 0 so all systems perform equally well on the population level and only keep the (zero-mean) document and annotator effects in the model. The false-positive rate of statistical tests on this artificial data should thus be no higher than the significance level. We then repeatedly apply both the t-test and the approximate randomization test (ART) (Noreen, 1989), a non-parametric test, to samples drawn from the model and determine the type I error rate at $p < 0.05$. We set the number of documents to 100 and demand 3 judgements per summary to mirror a common setup in manual evaluation. We then vary the total number of annotators between 3 and 300 by changing how many summaries a single annotator judges.

Results. We report results given the model estimated for `Likert` in Figure 4. Ignoring the dependencies between samples leads to inflated type I error rates, whether using the t-test or the ART. This is especially severe when only few annotators judge the whole corpus. In the extreme case with only three annotators in total, the null-hypothesis is rejected in about 40% of trials at a significance

level of 0.05 in both tasks. Even our original design with 60 annotators still sees an increase of the type I error rate by about 3%. Only if every annotator judges a single document and annotations are averaged per document, samples are independent and thus the real error is at the nominal 0.05 level. This design, however, is unrealistic given that annotators must be recruited and instructed.

We suggest two solutions to this problem: Either use mixed effect models or aggregate the judgements so samples become independent. This allows the assumptions of simpler tools such as ART to be met. In our study, we could average judgements in every block to receive independent samples. This is only possible, however, if the design of the study considers this problem in advance: a crowd-sourcing study that allows annotators to judge as many samples as they like is unlikely to result in such a design.

6 Study Design and Study Power

When conducting studies for system comparison, we are interested in maximizing their power to detect differences between systems. For traditional analysis, the power is ultimately determined by the number of documents (or judgements, when no aggregation takes place) in the study. However, when analysis takes into account individual annotators, power becomes additionally dependent on the total number of annotators and how evenly they participated in the study. This gives additional importance to the design of evaluation studies. In this section, we thus focus on how to optimize studies for power and reliability.

We first show that for well-powered experiments, we need to ensure that a sufficient total number of annotators participates in a study. In the second part of this section, we will then demonstrate studies can improve their power by not eliciting multiple judgements per summary.

6.1 Overall Number of Annotators

To demonstrate the difference in power caused by varying the total number of annotators in a study, we determine the power for a design with the same total number of documents and judgements per document but different total numbers of annotators.

We run the experiment both with regression and ART with proper aggregation of dependent samples as described in Section 5. We refer to the latter as `ARTagg` to differentiate it from normal ART.

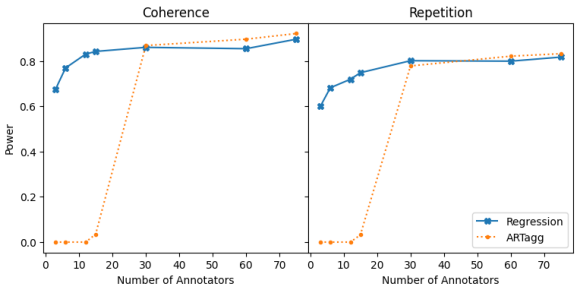


Figure 5: Power for 100 documents and 3 judgements per summary with different number of total annotators.

For each design we repeatedly sample artificial data from the `Likert` model and apply both tests to the data. The process is the same as in Section 5 except we do not set β to zero and count acceptances of the null-hypothesis.[6]

We again set the number of documents to 100 and the number of repeat judgements to 3 and vary the total number of annotators between 3 and 75 by varying the number of blocks between 1 and 25. We test for power at a significance level of 0.05.

Figure 5 shows how power drops sharply when only few annotators take part in the study. This is in line with the theoretical analysis of Judd et al. (2017) that shows that the number of participants is crucial for power when analysing studies with mixed effect models. The drop is worse for `ARTagg` as fewer annotators mean fewer independent blocks and thus a lack of datapoints for the analysis.

6.2 Annotator Distribution

Most studies elicit multiple judgements per summary, following best practices in NLP for corpus design (Carletta, 1996). While this leads to better judgements per *document*, the goal of many summarization evaluations is a per *system* judgement.

For this kind of study, Judd et al. (2017) show that for mixed models that include both annotator and target (in our case, input document) effects, a design where targets are *nested* within annotators, i.e. every annotator has its own set of documents, is always more powerful than one where they are (partially) *crossed* with annotators, i.e. a study with multiple annotations per summary, *given the same total number of judgements*. In fact, power could be maximized by having each annotator judge the sum-

[6] As this is an observed power analysis it probably overestimates the power of our analysis for the true effect. The analysis is thus only useful to compare designs under our best estimate of actual effect sizes.

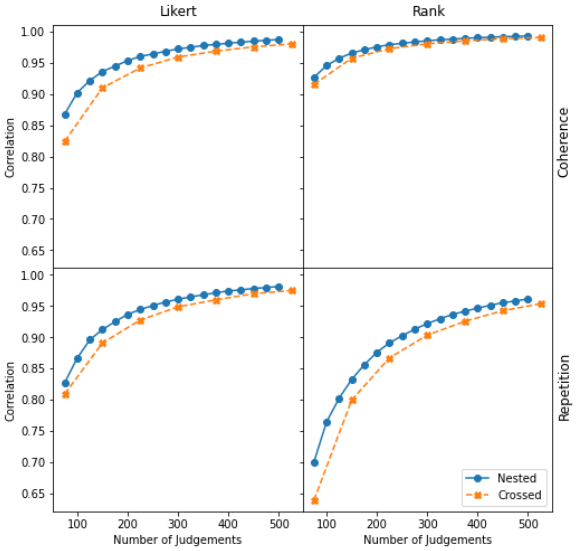


Figure 6: Reliabilities of nested vs. crossed designs for `Rank` and `Likert` for both tasks.

maries for only a single, unique document. However, this is usually not realistic due to the fixed costs of annotator recruitment and instruction. We demonstrate on our dataset how both reliability and power are affected by nested vs. crossed design.

To compare reliability, we randomly sample both nested and crossed designs from our full study and then compute the Pearson correlation of the system scores given by this smaller annotation set with the system scores given by the full study. As shown in Figure 6, nested samples are always at least as good and mostly better at approximating the results of the full annotation compared to a crossed sample with the same annotation effort.

We also conduct a power analysis for regression and `ARTagg` comparing nested and crossed designs. We again turn to Monte-Carlo simulation on the `Likert` models and sample nested and crossed designs with the same total number of judgements (i.e. the same cost). We keep the block size constant at 5 and vary the number of annotators between 3 and 60. For nested designs, we drop the document-level random effects from the ordinal regression, as document is no longer a grouping factor in nested designs.

Figure 7 shows that nested designs always have a power advantage over crossed designs, especially when few judgements are elicited. We also find that ART can be used to analyse data without loss of power when there are enough independent blocks. This might be attractive as ART is less computationally expensive than ordinal regression.

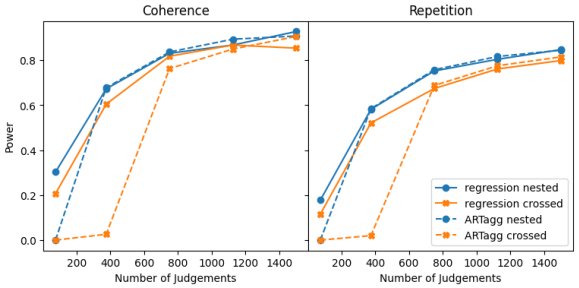


Figure 7: Power for $p < 0.05$ of nested and crossed designs for `ARTagg` and regression. X-axis shows the number of judgements elicited, Y-axis the power-level.

7 Related Work

Human evaluation has a long history in summarization research. This includes work on the correlation of automatic metrics with human judgements (Lin, 2004; Liu and Liu, 2008; Graham, 2015; Peyrard and Eckle-Kohler, 2017; Gao et al., 2019; Sun and Nenkova, 2019; Xenouleas et al., 2019; Zhao et al., 2019; Fabbri et al., 2020; Gao et al., 2020) and improving the efficiency of the annotation process (Nenkova and Passonneau, 2004; Hardy et al., 2019; Shapira et al., 2019). The impact of annotator inconsistency on system ranking has been studied both by Owczarzak et al. (2012) and Gillick and Liu (2010). To the best of our knowledge, we are the first to investigate the implications of annotator variance on the statistical analysis and the design in summarization system comparison studies.

For general NLG evaluation, van der Lee et al. (2019) establish best practices for evaluation studies. We extend on their advice by conducting experimental studies specifically for summary evaluation. In addition, we show the importance of study design and consideration of annotator-effects in analysis on real world data. The advice of Mathur et al. (2017) regarding annotation sequence effects should be taken into account in addition to our suggestions.

Method Comparison. Ranking has been shown to be effective in multiple NLP-tasks (Kiritchenko and Mohammad, 2017; Zopf, 2018), including NLG quality evaluation (Novikova et al., 2018). In this work we confirm this for coherence evaluation, although we find evidence that ranking is less efficient on repetition, where many documents do not exhibit any problems. We also add the dimension of annotator workload as a primary determinant of cost to the analysis of the comparison.

Multiple methods have been suggested to reduce study cost by sample selection (Sakaguchi et al., 2014; Novikova et al., 2018; Sakaguchi and Van Durme, 2018; Liang et al., 2020) or integration with automatic metrics (Chaganty et al., 2018). These efforts complement ours, as care still needs to be taken in analysis and study design.

Recently, rank-based magnitude estimation has been shown to be a promising method for eliciting judgements in NLG tasks and offers a combination of ranking and rating approaches (Novikova et al., 2018; Santhanam and Shaikh, 2019). However, it has not yet found widespread use in the summarization community. While magnitude estimation has been shown to reduce annotator variance, our advice regarding experimental design and grouping factors in statistical analysis applies to this method as well, as annotators can still systematically differ in which systems they prefer.

Statistical analysis. With regard to statistical analysis of experimental results, Dror et al. (2018) give advice for hypothesis testing in NLP. However, they do not touch on the problem of dependent samples. Rankel et al. (2011) analyse TAC data and show the importance of accounting for input documents in statistical analysis of summarizer performance and suggest the use of the Wilcoxon signed rank test for analysis. Sadeqi Azer et al. (2020) argue that p-values are often not well understood and advocate bayesian methods as an alternative. While the analysis in our paper is frequentist, the mixed effect model approach can also be integrated into a bayesian framework. Kulikov et al. (2019) model annotator bias in such a framework but do not account for differences in annotator preferences. In work conducted in parallel to ours, Card et al. (2020) show that many human experiments in NLP underreport their experimental parameters and are underpowered, including Likert-type judgements. Their simulation approach to power analysis is very similar to our experiments. In addition to their analysis, we show that ignoring grouping factors in statistical analysis of human annotations leads to inflated type I error rates. We also show that power can be increased by choosing nested over crossed designs with the same budget. The problem of underpowered studies has also been tackled outside of NLP by Brysbaert (2019).

For psycholinguistics, Barr et al. (2013) demonstrate how generalizability of results is negatively impacted by ignoring grouping factors in the analysis. Mixed effect models have found use in NLP before (Green et al., 2014; Cagan et al., 2017; Karimova et al., 2018; Kreutzer et al., 2020), but to the best of our knowledge they have not been used in summary evaluation.

8 Conclusion

We surveyed the current state of the art in manual summary quality evaluation and investigated methods, statistical analysis and design of these studies. We distill our findings into the following guidelines for manual summary quality evaluation:

Method. Both ranking and Likert-type annotations are valid choices for quality judgements. However, we present preliminary evidence that the optimal choice of method is dependent on task characteristics: If many summaries are similar for a given aspect, Likert may be the better option.

Analysis. Analysis of elicited data should take into account variance in annotator preferences to avoid inflated type I error rates. We suggest the use of mixed effect models for analysis that can explicitly take into account grouping factors in studies. Alternatively, traditional tests can be used with proper study design and aggregation.

Study Design. Study designers should control the number of annotators and how many summaries each individual annotator judges to ensure sufficient study power. Additionally, to ensure reliability of results, studies should report the design and the total number of annotators in addition to the number of documents and repeat judgements. Studies with repeat judgements on the same summary do not provide any advantage for system comparison and are less reliable and powerful than nested studies of the same size.

We hope that these findings will help researchers plan their own evaluation studies by allowing them to allocate their budget better. We also hope that our findings will encourage researchers to take more care in the statistical analysis of results. This prevents misleading conclusions due to ignoring the effect of differences in annotator behaviour.

Acknowledgements

We would like to thank Stefan Riezler for many fruitful discussions about the applications of mixed effect models.

References

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018. Rethinking the agreement in human evaluation tasks. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3318–3329, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278.

Marc Brysbaert. 2019. How many participants do we have to include in properly powered experiments? a tutorial of power analysis with reference tables. *Journal of Cognition*, 2(1).

Tomer Cagan, Stefan L. Frank, and Reut Tsarfaty. 2017. Data-driven broad-coverage grammars for opinionated natural language generation (ONLG). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1331–1341, Vancouver, Canada. Association for Computational Linguistics.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274, Online. Association for Computational Linguistics.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Arun Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evalaution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Melbourne, Australia. Association for Computational Linguistics.

R. H. B. Christensen. 2019. ordinal—regression models for ordinal data. R package version 2019.12-10. https://CRAN.R-project.org/package=ordinal.

Hoa Trang Dang. 2005. Overview of duc 2005. In *In Proceedings of the Document Understanding Conf. Wksp. 2005 (DUC 2005) at the Human Language Technology Conf./Conf. on Empirical Methods in Natural Language Processing (HLT/EMNLP*.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

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

Yang Gao, Wei Zhao, and Steffen Eger. 2020. SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354, Online. Association for Computational Linguistics.

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, Hong Kong, China. Association for Computational Linguistics.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

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

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, Lisbon, Portugal. Association for Computational Linguistics.

Spence Green, Sida I. Wang, Jason Chuang, Jeffrey Heer, Sebastian Schuster, and Christopher D. Manning. 2014. Human effort and machine learnability in computer aided translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1225–1236, Doha, Qatar. Association for Computational Linguistics.

Hardy Hardy, Shashi Narayan, and Andreas Vlachos. 2019. HighRES: Highlight-based reference-less evaluation of summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3381–3392, Florence, Italy. Association for Computational Linguistics.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1693–1701, Cambridge, MA, USA. MIT Press.

Charles M Judd, Jacob Westfall, and David A Kenny. 2017. Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology*, 68:601–625.

Sariya Karimova, Patrick Simianer, and Stefan Riezler. 2018. A user-study on online adaptation of neural machine translation to human post-edits. *Machine Translation*, 32(4):309–324.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.

Julia Kreutzer, Nathaniel Berger, and Stefan Riezler. 2020. Correct me if you can: Learning from error corrections and markings. *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*. Sage, Beverly Hills, CA.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China. Association for Computational Linguistics.

Ilia Kulikov, Alexander Miller, Kyunghyun Cho, and Jason Weston. 2019. Importance of search and evaluation strategies in neural dialogue modeling. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 76–87, Tokyo, Japan. Association for Computational Linguistics.

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

Russell Lenth, Henrik Singmann, Jonathon Love, Paul Buerkner, and Maxime Herve. 2018. Emmeans: Estimated marginal means, aka least-squares means. *R package version*, 1(1):3.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Weixin Liang, James Zou, and Zhou Yu. 2020. Beyond user self-reported Likert scale ratings: A comparison model for automatic dialog evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1363–1374, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Feifan Liu and Yang Liu. 2008. Correlation between ROUGE and human evaluation of extractive meeting summaries. In *Proceedings of ACL-08: HLT, Short Papers*, pages 201–204, Columbus, Ohio. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2017. Sequence effects in crowdsourced annotations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2860–2865, Copenhagen, Denmark. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Ranking sentences for extractive summarization with reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1747–1759, New Orleans, Louisiana. Association for Computational Linguistics.

Ani Nenkova and Rebecca 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, Boston, Massachusetts, USA. Association for Computational Linguistics.

Eric W Noreen. 1989. *Computer-intensive methods for testing hypotheses*. Wiley New York.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. RankME: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 72–78, New Orleans, Louisiana. Association for Computational Linguistics.

Karolina Owczarzak, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. 2012. An assessment of the accuracy of automatic evaluation in summarization. In *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, pages 1–9, Montréal, Canada. Association for Computational Linguistics.

Maxime Peyrard. 2019. Studying summarization evaluation metrics in the appropriate scoring range. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5093–5100, Florence, Italy. Association for Computational Linguistics.

Maxime Peyrard and Judith Eckle-Kohler. 2017. A principled framework for evaluating summarizers: Comparing models of summary quality against human judgments. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 26–31, Vancouver, Canada. Association for Computational Linguistics.

Peter Rankel, John Conroy, Eric Slud, and Dianne O'Leary. 2011. Ranking human and machine summarization systems. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 467–473, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Erfan Sadeqi Azer, Daniel Khashabi, Ashish Sabharwal, and Dan Roth. 2020. Not all claims are created equal: Choosing the right statistical approach to assess hypotheses. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5715–5725, Online. Association for Computational Linguistics.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2014. Efficient elicitation of annotations for human evaluation of machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 1–11, Baltimore, Maryland, USA. Association for Computational Linguistics.

Keisuke Sakaguchi and Benjamin Van Durme. 2018. Efficient online scalar annotation with bounded support. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 208–218, Melbourne, Australia. Association for Computational Linguistics.

Sashank Santhanam and Samira Shaikh. 2019. Towards best experiment design for evaluating dialogue system output. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 88–94, Tokyo, Japan. Association for Computational Linguistics.

Natalie Schluter. 2017. The limits of automatic summarisation according to ROUGE. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 41–45, Valencia, Spain. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Eva Sharma, Luyang Huang, Zhe Hu, and Lu Wang. 2019. An entity-driven framework for abstractive summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3280–3291, Hong Kong, China. Association for Computational Linguistics.

Simeng Sun and Ani Nenkova. 2019. The feasibility of embedding based automatic evaluation for single document summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1216–1221, Hong Kong, China. Association for Computational Linguistics.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake Vand erPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1. 0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Stratos Xenouleas, Prodromos Malakasiotis, Marianna Apidianaki, and Ion Androutsopoulos. 2019. SUM-QE: a BERT-based summary quality estimation model. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6005–6011, Hong Kong, China. Association for Computational Linguistics.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in*

Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Markus Zopf. 2018. Estimating summary quality with pairwise preferences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1687–1696, New Orleans, Louisiana. Association for Computational Linguistics.

A Interface Screenshots

We show screenshots of the instructions for both annotation methods and tasks in Figure 8 and interfaces in Figure 9.

B Survey

B.1 Categories

While most categories are self-explanatory, we elaborate on some of the decisions we made during the survey in this section.

Evaluation Questions. We allow a single study to include multiple evaluation questions, as long as all questions are answered by the same annotators and use the same method. We make no distinction between informativeness, coverage, focus and relevance and summarize them under *Content*. Similarly, we summarize fluency, grammaticality and readability under *Fluency*. *Other* includes:

- One study with a specialized set of evaluation questions evaluating the usefulness of a generated related work summary
- One study of *polarity* in a sentiment summarization context
- One study where annotators were asked to identify the aspect a summary covers in the context of review summarization
- Two studies evaluating formality and *meaning similarity* of reference and system summary
- One study evaluating diversity
- One study conducting a Turing test
- One study asking paper authors whether they would consider a sentence part of a summary of their own paper.
- One study evaluating structure and topic diversity

Evaluation Method. *Binary* includes any task with a yes/no style decision, while *pairwise* includes any method in which two systems are ranked against each other. *Other* includes

- The aspect identification task mentioned above
- One study in which participants selected a single best summary out of a set of summaries

Annotator Recruitment. *Other* includes any recruitment strategy that does not rely on crowdsourcing. This includes cases in which the recruitment was not specified, students, experts, the authors themselves and various kinds of volunteers.

Statistical Evaluation. *Other/unspecified* includes

- Four studies which reported statistical significance without reporting the test used
- Two studies using the approximate randomization test
- One study using the chi-square test
- One study using a Tukey test without prior ANOVA.

B.2 Survey Files

All papers we considered for the survey are listed in the supplementary material in the file `all_papers.yaml` by their id in the ACL anthology bib-file. The 58 SDS/MDS system papers that contain new human evaluation studies and are thus included in the survey are listed in the category `with_human_eval`.

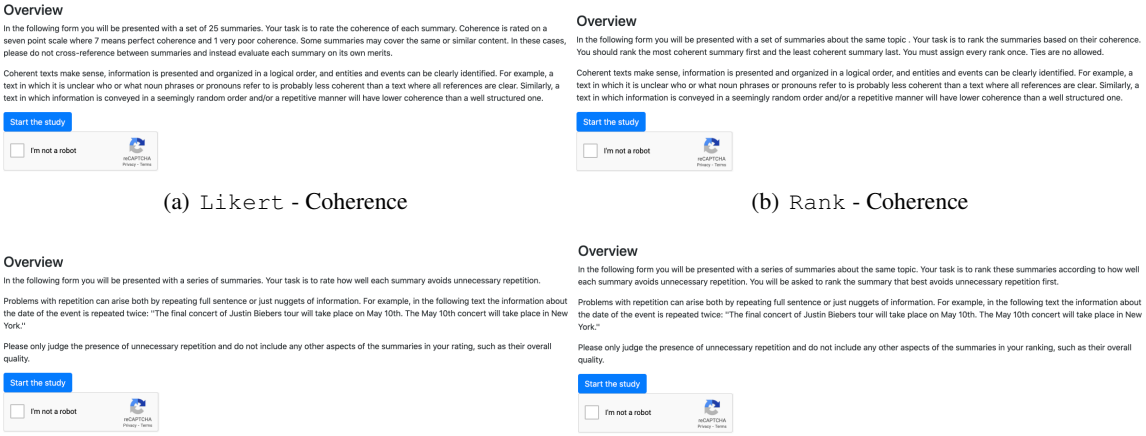
For the sake of completeness, we further list summarization papers we did not include in our survey. We separate them into the following categories:

no_human_eval 47 SDS/MDS system papers without human evaluation

sentsum 27 Sentence summarization and headline generation papers

non_system 34 summarization papers that do not introduce new systems, like surveys, opinion pieces and evaluation studies

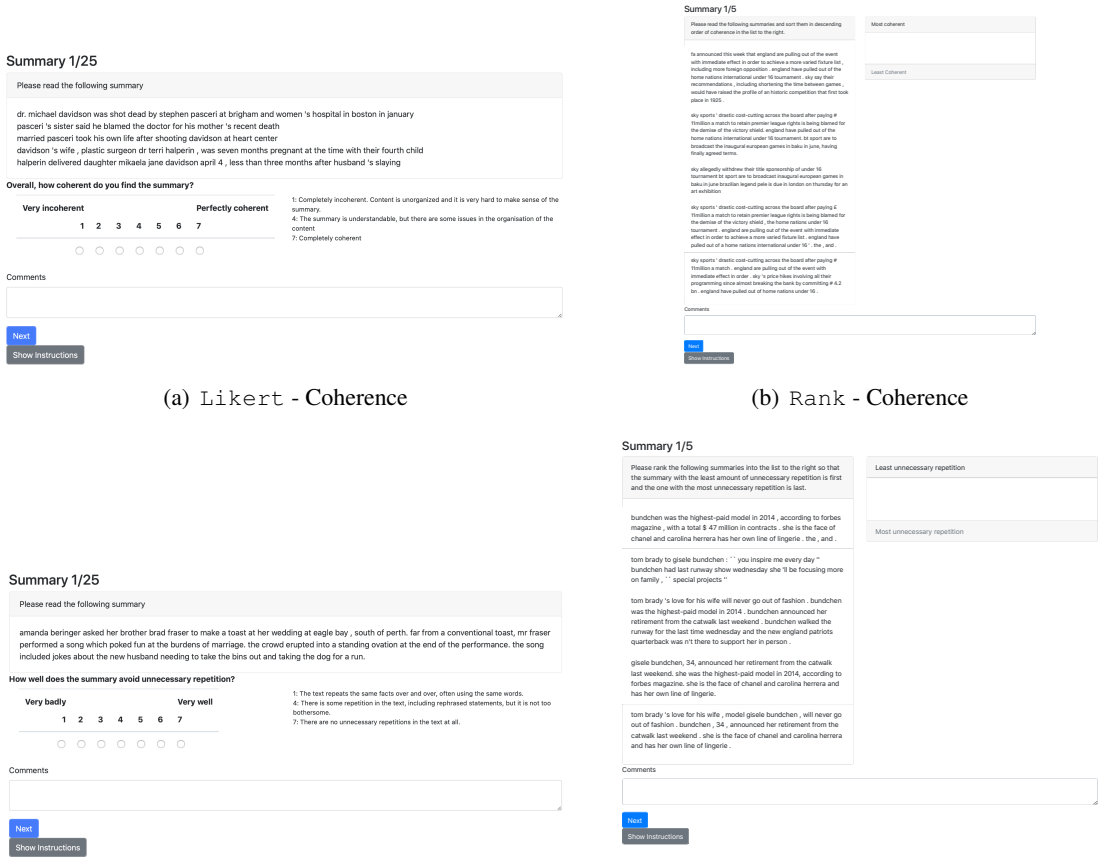
other 10 Papers that conduct summarization with either non-textual input or non-textual output



(a) `Likert` - Coherence (b) `Rank` - Coherence

(c) `Likert` - Repetition (d) `Rank` - Repetition

Figure 8: Screenshots of the Annotator Instructions.



(a) `Likert` - Coherence (b) `Rank` - Coherence

(c) `Likert` - Repetition (d) `Rank` - Repetition

Figure 9: Screenshots of the Annotation Interfaces.

We give a full list of the survey results for all papers with human evaluation studies in the file `survey_details.csv`. The file has the following columns:

paper Id of the paper in the ACL anthology

eval_id Id of the evaluation study to differentiate them in papers with multiple studies

task Summarization task of the paper: SDS vs. MDS

genre Genre of the summarized documents

#docs Number of documents in the evaluation

#systems Number of systems in the evaluation

includes_reference Whether the reference summary is included in the human evaluation

#ann_total Total number of annotators in the study

#ann_item Number of annotators per summary

content, fluency, repetition, coherence, referential_clarity, other, overall Binary columns indicating evaluation questions in the paper

measure Annotation method used in the study

anntype Annotator recruitment strategy

stattest Statistical test used

design_specified Indicates whether it is possible to determine the full design from the information given about the study in the paper

comments Comment column. This column describes the use of *other* where present.