Human Experts' Perceptions of Auto-Generated Summarization Quality

Maryam Lotfigolian
Department of Computer Science, Oslo Metropolitan University, 0130 Oslo, Norway
lotfigolian.m@gmail.com

Christos Papanikolaou
Department of Computer Science, Oslo Metropolitan University, 0130 Oslo, Norway
c.papanikolaou@windowslive.com

Samaneh Taghizadeh
Department of Computer Science, Oslo Metropolitan University, 0130 Oslo, Norway
samaneh71917@gmail.com

Frode Eika Sandnes
Department of Computer Science, Oslo Metropolitan University, 0130 Oslo, Norway
frodes@oslomet.no

ABSTRACT

In this study we addressed automatic summarizations generated using modern artificial intelligence techniques. Several mathematical methods for evaluating the performance of automatic summarization exist. Such methods are commonly used as they allow many test cases to be assessed with little human effort as manual assessments are challenging and time consuming. One question is whether the output of such measures matches human perception of summarization quality. In this study we document a study involving the human evaluation of the automatic summarization of 22 academic texts. The unique aspect of this study is that our participants had strong familiarity with the texts as they had studied these texts in depth. The results are quite varied but do not give the impression of unanimous agreement that automatic summarizations are of high quality and are trusted.

CCS CONCEPTS

• **Information systems** → Information retrieval; Retrieval tasks and goals; Summarization.

KEYWORDS

Automatic summarization, User perception, Quality, Evaluation, Artificial intelligence, NLP, Language model, GPT-3, ChatGPT

ACM Reference Format:
Maryam Lotfigolian, Christos Papanikolaou, Samaneh Taghizadeh, and Frode Eika Sandnes. 2023. Human Experts' Perceptions of Auto-Generated Summarization Quality. In *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '23), July 05–07, 2023, Corfu, Greece.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3594806.3594828

1 INTRODUCTION

A summary is a short representation of a larger text that summarizes readers about main ideas in the source text. Reading a summary typically involves active reading of a text noting down key points, and then later synthesizing the text purely based on the notes. The ability to write summaries is a skill that requires practice. Moreover, it is time-consuming. Readers may sometimes only be interested in making rapid decisions without having to read an entire text, but rather by getting the gist of the text through a summary. Consequently, there has been much interest in algorithms for automatically summarizing texts. Recent developments in artificial intelligence have resulted in impressive demonstrations of the technology.

Researchers have made many attempts at various ways of automatically summarizing texts for several decades. Typically, such methods are evaluated using deterministic metrics. Although such metrics are convenient and pragmatic, they also do not give insight into how they will be perceived by readers.

In this study we wanted to go beyond the typical deterministic metrics and positive impressions one may get from ad-hoc testing of such technology through toy demos. We wanted to explore how domain experts would perceive automatically generated summaries of text with which they were familiar, and to what degree they would be willing to rely on such automatically generated summaries.

2 RELATED WORK

Many studies are published on automatic summarization of text [11]. Some works focus on improving the summarization algorithms [9] and others are exploring domain specific application areas, for example, summarization of micro blogs [16], summarization of lectures and meetings [3], multi-document summarization [18].

Advances in deep learning and very large language models have led to impressive improvements in automatic text summarization and related text processing tasks, such as grammar checking [14]. In particular, ChatGPT [1] has received much attention recently. It has been applied to a large range of tasks, including writing [2, 4], translation [13], mathematics [10], and education [22].

Automatic summarization techniques are usually evaluated using metrics such as ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [17] where the machine generated summary is automatically compared to a summary written manually by a human [11, 15]. The human generated summaries serve as the ground truth. Metrics allows large amounts of text to be objectively and consistently compared without the cost, time, and effort involved with manual assessment. Assessments can easily be run at each

PETRA '23, July 05–07, 2023, Corfu, Greece

ACM ISBN 979-8-4007-0069-9/23/07.
https://doi.org/10.1145/3594806.3594828

step of tweaking a summarization algorithm. Clearly, it is hard to develop summarization algorithms if human assessors are used at each stage. The trade-off between the convenience and cost of human assessors versus automatic evaluation with such text-based technologies is discussed in several studies [19]. It is argued that automatic evaluations have been viewed with some mistrust [20]. It has also been pointed out that human assessment is not without problems [12] as there are few established practices for such assessments. Key weaknesses identified include lacking information about demographics, task design, experimental protocol, and reliability assessments [12]. Clearly, human perceptions are inconsistent and variable, yet it is the human perception of the technology that will determine to what degree users will trust, accept, and use a technology [5, 7]. This study attempts to avoid some of the pitfalls raised in [12] as the participants are recruited from a relatively homogenous cohort of participants with in-depth insight about the summarized texts.

3 METHOD

3.1 Experimental design

This study involved two stages. First, we had to select the most suitable summarization engine from a set of available options. This was done using a commonly used measure from the literature. Next, the results produced by the selected summarization engine were assessed using a panel of human readers with in-depth familiarity with the source texts.

3.2 Selecting a summarization engine

First, we decided to select one of the many summarization engines available to reduce the burden on the participants. Decided to independently evaluate four publicly available summarization engines intended for academic texts, namely Paper Digest (https://www.paperdigest.com/), Scholarcy (https://www.scholarcy.com/), Bundle IQ (https://app.bundleiq.com/), and Quillbot AI (https://quillbot.com/) using their respective web interfaces.

These are all modern automatic summarization applications for which it is claimed they provide human-like extractive summaries of scientific papers. It is claimed that these locate crucial information and sum up articles and papers into the most valuable points while maintaining the original context. They have slightly varying interfaces.

Scholarcy provides a summary, context, and highlights key sections. Bundle IQ identifies the key points in a document and generates a summary for either the entire document or for specific pages. QuillBot provides summaries at sentence level or per paragraph. A slider allows the user to interactively adjust the length of the summary. Paper Digest presents the user with key bullet points.

We tested the four engines using 30 academic papers randomly drawn from the reading lists of the specialization topics of the master programme in applied computer science, thereby covering a wide range of computer science topics covering mathematical modeling, data science, artificial intelligence, human computer interaction, etc. First, the paper abstracts were removed. Each academic paper (without abstract) was run through the four engines. The results were compared with the actual abstracts (representing the ground truth authored manually by the authors). The comparisons were made using a python implementation of ROUGE-L [18].

The result of the evaluation is shown in Figure 1. The results reveal that overall, Paper Digest yielded the highest F-scores of the four methods (highest mean, max, min, and second quartile). Bundle IQ had the highest fourth quartile point. Based on these results we decided to use Paper Digest in the subsequent user study. A repeated measures ANOVA omnibus test flags a significant difference, but with a relatively moderate effect size ($F(3, 87) = 3.702$, $p = .015$, $\eta^2 = 0.113$). A Holm post-hoc test reveals that the significant difference occurred between Paper Digest and Quilbot AI ($p = .012$).

3.3 Participants

A total of 11 students enrolled in a course in Intelligent User Interface were recruited for this study, from a class of 18 students (61% participation rate). This is a research-oriented course where students actively present the material. We classify these participants as experts due to their familiarity and exposure to the material used in this study and specific training in interpreting scientific literature [8]. The participants' unique insight thus provided a rare opportunity to manually assess summaries. We decided to omit any detailed demographic information related to gender and age to preserve the privacy of this relatively small cohort.

3.4 Material

The reading list from the course was used as the source material. The reading list comprises one paper pre-assigned to each student by the teacher, and one self-selected paper. All the papers were peer reviewed academic texts from the past proceedings of the Intelligent User Interface conferences and CHI conferences. All the papers were on the topic of intelligent user interfaces.

It was assumed that each participant would be especially familiar with the two assigned texts as the student had studied the texts, presented these in plenary to the other students in the class and led the subsequent in-class discussion. This activity was compulsory for all the students. A total of 22 papers from the 36-paper reading list were used to generate 22 summaries for the 11 participants (two summaries for each student).

3.5 Procedure

This study adheres to the authors' institutional privacy and ethics regulations. The participants were informed about the purpose and content of the study and provided their oral consent to participate. They were also informed of their rights to withdraw at any time. The study was conducted in a single session and no linking data or personal information were collected [21]. The results are thus anonymous. The sessions were conducted remotely.

Each participant was presented with the two summaries matching their two individual papers. After reading and assessing the two summaries they were asked four closed (5-item Likert-style) questions related to the quality of each summary, namely comprehensiveness, conciseness, coherence, and the likelihood the student would use Paper Digest in the future. We therefore solicited N = 22 (2 x 11) responses.

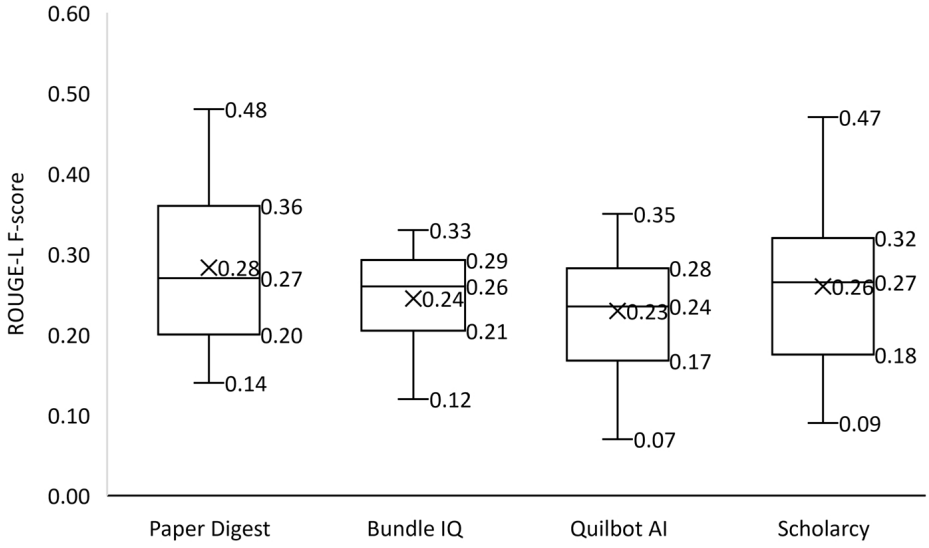

Figure 1: Box and Whisker plot of the summarization engine evaluation results (ROUGE-L F-scores). N = 30.

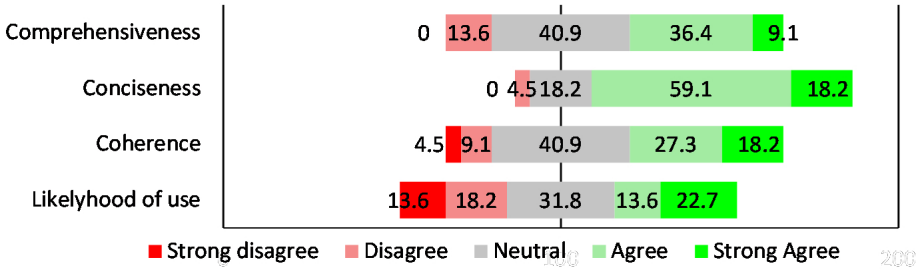

Figure 2: Diverging stacked bar graph summarizing the results of the human assessment of automatic summaries ($N = 22$).

3.6 Analysis

Our sample size was too small to attempt any inferential statistics. We therefore assessed the trends using visual inspection of the data plotted using a diverging stacked bar graph.

4 RESULTS

Figure 2 summarizes the responses from the participants. The results show that participants were most positive regarding conciseness of the automatically generated summaries with most participants indicating agree or strongly agree (close to 77.3%). A one-sample Wilcoxon signed rank test shows that the median is significantly different from the neutral response of 3 ($V = 146.0$, $p < .001$, $ES = 0.908$). Note that the rank-biserial correlation effect size is relatively high. Next, the results for comprehensiveness and coherence are quite similar with a slight positive skew (both with 45.5% positive responses). The median comprehensiveness responses were significantly different from neutral with a medium effect size ($V = 73.0$, $p = .042$, $ES = 0.604$), while the median coherence responses were not significantly different from neutral ($V = 71.0$, $p = .071$). The responses to likelihood of future use were more divided and less skewed with 36.3% positive and 31.8% negative responses. The median responses were not significantly different to neutral ($V = 69.5$, $p = .599$). The portion of neutral responses were high for all four questions with comprehensiveness of coherence both constituting 40.9% of the responses.

5 DISCUSSION

The results do not seem to suggest that automatic summarization technology is yet sufficiently mature to replace human generated summaries. This is especially evident in the participants' responses to how likely they are to use the technology in the future. Also, the observation that text conciseness was rated most favorably is what one could expect as the summarization engines indeed makes texts short. However, the attributes relating to the substantial contents of the summaries, namely comprehensiveness and coherence is not as positively rated.

We are unaware of the implementation details or exact technology used in the summarization engine. However, we assume it is

based on a contemporary trained language model due to its power. Such modern language models have limitations. For example, biases of the results could be related to the data on which the model was trained. Moreover, extractive summaries are typically formed by locating key sentences within the original text while information within discarded sentences are not included in the summary.

Another issue is that the choice of summarization engine was based on the ROUGE metric. Clearly, the ROUGE metric lacks semantic and factual attributes. Yet, we assume that the four engines were all based on very similar underlying artificial intelligence techniques, and the differences are thus likely to be minimal.

Another weakness of this experiment was the relatively small sample, although this is within the norm of typical human computer interaction [6]. However, we would argue that the quality of the assessments are of relatively high quality due to the participants' invested efforts with the texts, and one may argue that a small sample of high quality measurements are preferable over a larger number of measurements with lower quality.

In hindsight, we should also have recorded the time of the student's presentation as the participants probably can recall details from recently presented work more accurately than work that was presented less recently. The first presentations were given in late August, while the study was conducted in late November (range of 0 to 3 months). It could be relevant to correlate the recency of working with these papers to the responses. However, each student had to present two papers in two phases of the semester and one presentation was thus further in the past and the other presentation more recent. This has probably helped counterbalance any effects of recall decay with time.

6 CONCLUSIONS

Modern language models such as GPT-3 and similar technologies have undoubtedly contributed to changing people's perceptions of artificial intelligence. However, despite such technologies really impressing abilities to automatically summarize texts, our results suggest that this technology does not yet seem capable of fully replacing the process of manually reading papers. For that reason, they may serve a valuable role as a human-in-the-loop assistive tool to complement manual reading.

REFERENCES

[1] Mohammad Aljanabi, *et al.* 2023. ChatGpt: Open Possibilities. Iraqi Journal For Computer Science and Mathematics, 2023, 4.1: 62-64.

[2] Ömer Aydin and Enis Karaarslan. 2022. OpenAI ChatGPT generated literature review: Digital twin in healthcare. Available at SSRN 4308687, 2022.

[3] Chidansh Bhatt, Andrei Popescu-Belis, and Matthew Cooper. 2016. Audiovisual Summarization of Lectures and Meetings Using a Segment Similarity Graph. In Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval (ICMR '16). Association for Computing Machinery, New York, NY, USA, 261–264. https://doi.org/10.1145/2911996.2912047

[4] Som Biswas. 2023. ChatGPT and the Future of Medical Writing. Radiology, 2023, 223312.

[5] Josieli Aparecida Marques Boiani, *et al.* 2019. On the non-disabled perceptions of four common mobility devices in Norway: a comparative study based on semantic differentials. Technology and Disability, 2019, 31.1-2: 15-25.

[6] Kelly Caine. 2016. Local Standards for Sample Size at CHI. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16). Association for Computing Machinery, New York, NY, USA, 981–992. https://doi.org/10.1145/2858036.2858498

[7] Aline Darc Piculo dos Sandos, *et al.* 2022. Aesthetics and the perceived stigma of assistive technology for visual impairment. Disability and Rehabilitation: Assistive Technology, 2022, 17.2: 152-158.

[8] Evelyn Eika, and Frode Eika Sandnes, 2022. Starstruck by journal prestige and citation counts? On students' bias and perceptions of trustworthiness according to clues in publication references. Scientometrics, 2022, 127.11: 6363-6390.

[9] Thérèse Firmin and Inderjeet Mani. 1998. Automatic text summarization in TIPSTER. In Proceedings of a workshop on held at Baltimore, Maryland: October 13-15, 1998 (TIPSTER '98). Association for Computational Linguistics, USA, 179–180. https://doi.org/10.3115/1119089.1119119

[10] Simon Frieder, *et al.* 2023. Mathematical Capabilities of ChatGPT. arXiv preprint arXiv:2301.13867, 2023.

[11] Mahak Gambhir and Vishal Gupta. 2017. Recent automatic text summarization techniques: a survey. Artificial Intelligence Review, 2017, 47: 1-66.

[12] Neslihan Iskender, Tim Polzehl, and Sebastian Moller. 2021. Reliability of human evaluation for text summarization: Lessons learned and challenges ahead. In: Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval). 2021. p. 86-96.

[13] Wenxiang Jiao, *et al.* 2023. Is ChatGPT a good translator? A preliminary study. arXiv preprint arXiv:2301.08745, 2023.

[14] Hitesh Mohan Kaushik, Evelyn Eika, and Frode Eika Sandnes. 2020. Towards universal accessibility on the web: do grammar checking tools improve text readability?. In: Universal Access in Human-Computer Interaction. Design Approaches and Supporting Technologies: 14th International Conference, UAHCI 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part I 22. Springer International Publishing, 2020. p. 272-288.

[15] Farshad Kiyoumarsi. 2015. Evaluation of automatic text summarizations based on human summaries. Procedia-Social and Behavioral Sciences, 2015, 192: 83-91.

[16] Sanghoon Lee, Sunny Shakya, Raj Sunderraman, and Saeid Belkasim. 2013. Real Time Micro-blog Summarization Based on Hadoop/HBase. In Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT) - Volume 03 (WI-IAT '13). IEEE Computer Society, USA, 46–49. https://doi.org/10.1109/WI-IAT.2013.148

[17] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. 2004. p. 74-81.

[18] Peng Li, Yinglin Wang, Wei Gao, and Jing Jiang. 2011. Generating aspect-oriented multi-document summarization with event-aspect model. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11). Association for Computational Linguistics, USA, 1137–1146.

[19] Selina Meyer, David Elsweiler, Bernd Ludwig, Marcos Fernandez-Pichel, and David E. Losada. 2022. Do We Still Need Human Assessors? Prompt-Based GPT-3 User Simulation in Conversational AI. In Proceedings of the 4th Conference on Conversational User Interfaces (CUI '22). Association for Computing Machinery, New York, NY, USA, Article 8, 1–6. https://doi.org/10.1145/3543829.3544529

[20] Karolina Owczarzak, *et al.* 2012. An assessment of the accuracy of automatic evaluation in summarization. In: Proceedings of workshop on evaluation metrics and system comparison for automatic summarization. 2012. p. 1-9.

[21] Frode Eika Sandnes. 2021. HIDE: Short IDs for Robust and Anonymous Linking of Users Across Multiple Sessions in Small HCI Experiments. In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (CHI EA '21). Association for Computing Machinery, New York, NY, USA, Article 326, 1–6. https://doi.org/10.1145/3411763.3451794

[22] Teo Susnjak. 2022. ChatGPT: The End of Online Exam Integrity?. arXiv preprint arXiv:2212.09292, 2022.