Evaluation of Summarization Systems across Gender, Age, and Race

Anna Jørgensen[*]
Findest
anna.jorgensen@findest.eu

Anders Søgaard
Department of Computer Science
University of Copenhagen
soegaard@di.ku.dk

Abstract

Summarization systems are ultimately evaluated by human annotators and raters. Usually, annotators and raters do not reflect the demographics of end users, but are recruited through student populations or crowdsourcing platforms with skewed demographics. For two different evaluation scenarios – evaluation against gold summaries and system output ratings – we show that summary evaluation is sensitive to protected attributes. This can severely bias system development and evaluation, leading us to build models that cater for some groups rather than others.

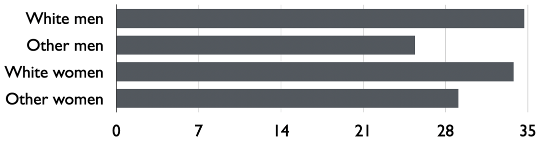

Figure 1: **Social bias in automatic summarization**: We take steps toward evaluating the impact of the gender, age, and race of the humans involved in the summarization system evaluation loop: the authors of the summaries and the human judges or raters. We observe significant group disparities, with lower performance when systems are evaluated on summaries produced by minority groups. See §3 and Table 1 for more details on the Rouge-L scores in the bar chart.

1 Introduction

Summarization – the task of automatically generating brief summaries of longer documents or collections of documents – has, so it seems, seen a lot of progress recently. Progress, of course, is relative to how performance is measured. Generally, summarization systems are evaluated in two ways: by comparing machine-generated summaries to human summaries by text similarity metrics (Lin, 2004; Nenkova and Passonneau, 2004) or by human rater studies, in which participants are asked to rank system outputs. While using similarity metrics is controversial (Liu and Liu, 2008; Graham, 2015; Schluter, 2017), the standard way to evaluate summarization systems is a combination of both.

Both comparison to human summaries and the use of human raters naturally involve human participants, and these participants are typically recruited in some way. In Liu and Liu (2008), for example, the human subjects are five undergraduate students in Computer Science. Undergraduate students in Computer Science are not necessarily representative of the population at large, however, or of the end users of the technologies we develop. In this work, we ask whether such sampling bias when recruiting participants to evaluate summarization systems, is a problem? In other words, do different demographics exhibit different preferences in rater studies of summarization systems? NLP models are only fair if they do not put certain demographics at a disadvantage (Larson, 2017), and it is therefore crucial our benchmarks reflect preferences and judgments across those demographics (Ethayarajh and Jurafsky, 2020).[1]

Contributions We present the, to the best of our knowledge, first in-detail evaluations of summarization systems across demographic groups, focusing on two very different extractive summarization systems – TextRank (Mihalcea and Tarau, 2004) and MatchSum (Zhong et al., 2020). The groups are defined by the three protected attributes: gender, age, and race. While the systems are reported to perform very differently, we show that the system rankings induced by performance scores or user preferences differ across these groups of human summary authors and summary raters. We analyze what drives

[*] The work was done while the author was at the University of Amsterdam.

[1] We thereby challenge the widely held position that lay people cannot be used for summary evaluation, because they exhibit divergent views on summary quality (Gillick and Liu, 2010). We, in contrast, believe such variance is a product of social differences and something we need to worry about in NLP.

Proceedings of the Third Workshop on New Frontiers in Summarization, pages 51–56
November 10, 2021. ©2021 Association for Computational Linguistics

these differences and provide recommendations for future evaluations of summarization systems.

2 Experiments

We present two evaluations in this short paper: an **automated scoring against human summaries** (Exp. A) and a **human rater study** (Exp. B). In both experiments, we use Amazon Mechanical Turk to recruit annotators from different demographic groups, and the first paragraphs of biographies from English Wikipedia as our input data, using the Wikidata API for extraction.[2] We create a dataset of biographies of women and men, obtain human summaries, and generate summaries of these biographies using two out-of-the-box extractive summarization systems. In Exp. A, we compare the system summaries directly to the human summaries (from different groups); in Exp. B, we let human raters compare and rate the two system summaries. To ensure differences between the two summarization systems, we use the 2004 graph-based TextRank (Mihalcea and Tarau, 2004) and the 2020 state-of-the-art, BERT-based MatchSum (Zhong et al., 2020).[3] We follow the MatchSum guidelines described in (Zhong et al., 2020) and limit the length of the input biographies to a maximum 5 sentences and force the output summaries to be between 2-3 sentences long. Our final dataset consists of the original 975 biographies (700 men and 275 women), along with two automatic summaries, as well as human 3 sentence summaries, and is made freely available.[6]

Our evaluations rely on annotations and ratings from Amazon Mechanical Turk. For quality control, we rely on a control question, as well as analyzing annotation time: If a task is completed faster than one standard deviation of the average time spent, the answers in that task are discarded. We collected one manual summary and two system rankings per biography, resulting in 3,135 annotations.

Gender	Race	Rouge-1	Rouge-L
♀		0.407	0.326
♂		0.417	0.326
♀	White	0.418	0.338
	Other	0.371	0.291
♂	White	**0.436**	**0.347**
	Other	0.347	0.254

Table 1: Automated scoring of MatchSum (Zhong et al., 2020) across self-reported protected attributes: **gender**, with values ♀, ♂, and other (all our annotators identified as either male or female), **race**, binarized here as white and other (in order to achieve rough size balance). The ROUGE scores of MatchSum are clearly higher when evaluated against reference summaries created by white men. We also considered **age** (binarized as ±30, to achieve size balance): Here we see slightly better performance when evaluated against summaries of older participants across all genders annotators identified with.

Human summaries In Exp. A, participants were asked to enter the three most important sentences in the document and in three blank text fields; for quality control, we check that these sentences occur in the input document. We collect a total of 1,185 summaries, 53% of which are written by women (0.5% identified neither as male or female). 74% of summaries are written by participants older than 30 years of age. 76% identified as white; 11% as Blacks; 5% as American Indians; 4% as Asians, and 4% as Hispanics.[7] We binarize race as white and other to achieve rough size balance across groups. Aggregating scores across multiple races is not ideal, but by doing so, we compensate for poor representation of some demographics.

Rater study In Exp. B, we present participants with two 2-3 sentence machine summaries and ask them to a) pick their preferred summary and b) rank the two summaries on 4-point forced Likert scales, for fluency, informativeness and usefulness. 40.2% of our raters identified as female. 37.5% were below 30 years of age. 70.8% of ratings identified as white, the rest as American Indians (2.3%), Asians (3.5%), Blacks (19.1%), Hispanics

[2] https://query.wikidata.org/

[3] We use the implementation of TextRank by Barrios et al. (2016)[4] and the original MatchSum implementation.[5] MatchSum obtains state-of-the-art performance across a range of benchmarks by learning to produce summaries whose document encoding is similar to that of the input document. TextRank is a much simpler extractive algorithm; it adopts PageRank to compute node centrality recursively based on a Markov chain model. While MatchSum obtains a Rouge-1 score of .44 on CNN/Daily Mail, TextRank obtains a Rouge-1 score of .33 (Zheng and Lapata, 2019). We use both systems with recommended parameters, as was done in Zheng and Lapata (2019). Note that TextRank, in contrast to MatchSum, is unsupervised. Our Rouge-1 scores below for Wikipedia biographies are generally comparable.

[6] https://github.com/ajoer/summary_preferences

[7] Our race taxonomy was standard, based on https://www.census.gov/prod/2001pubs/cenbr01-1.pdf, but all annotators identified as either American Indian, Asian, black, Hispanic, or white.

Gender	Age	TextRank	MatchSum	N/A
♀	≥30	0.379	0.565	0.056
	<30	**0.481**	**0.454**	0.065
♂	≥30	0.397	0.511	0.092
	<30	0.396	0.531	0.073

Table 2: System ratings across participant gender and age. We highlight the outlier: Younger women significantly preferred TextRank over MatchSum ($p < 0.01$).

Age	Race	TextRank	MatchSum	N/A
<30	Asian	34.1	39.0	26.8
	black	**49.0**	**43.1**	7.8
	Hispanic	40.7	59.3	0.0
	white	43.6	53.5	2.9
≥30	Amer. Ind.	40.0	51.3	8.7
	white	43.6	53.5	2.9

Table 3: System ratings across participant race and age. We highlight the outlier: Young blacks significantly preferred TextRank over MatchSum ($p < 0.01$).

		Informative		Useful		Fluent	
	Age	T	M	T	M	T	M
All	≥30	0.94	0.96	0.94	0.96	0.9	0.95
	<30	0.77	0.81	0.72	0.79	0.81	0.83
♂	≥30	0.88	0.92	0.86	0.91	0.84	0.89
	<30	0.86	0.9	0.82	0.89	0.85	0.91
♀	≥30	0.89	0.91	0.88	0.92	0.88	0.91
	<30	0.83	0.84	0.8	0.83	0.86	0.83

Table 4: Rater study results with respect to age, on all biographies, as well as on biographies of men (♂) and women (♀) only.

	Informative		Useful		Fluent	
Race	T	M	T	M	T	M
Amer. Indian	0.5	0.6	0.7	0.7	**1.2**	1.0
Asian	0.7	1.0	0.8	0.9	1.0	0.8
black	0.7	0.8	1.0	0.8	0.9	0.8
Hispanic	**1.4**	0.9	**1.5**	**1.2**	0.9	1.0
white	0.8	0.7	0.8	0.8	0.9	0.8

Table 5: Rater study results on All for race

(2.0%), or as others (2.2%).

We ask all participants to voluntarily submit their race and gender information, and require that they be US-based. We asked the participants in the rater study to also include age information.

Results In Table 1, we present the results of Exp. A: Rouge-1 and Rouge-L results are significantly better when evaluated on summaries produced by white men than when evaluated on summaries produced by any other group. MatchSum summaries also align better with those written by white women compared to those written by non-white women. Generally, MatchSum aligns better with men than with women.

Exp. 2 includes three demographic variables (gender, age, and race). Table 2 presents ratings across gender and age. Most participants prefer the reportedly superior system (with a Rouge-1 advantage of 0.11 on a standard benchmark; see §2), but younger women significantly preferred TextRank over MatchSum ($p < 0.01$). Table 3 presents the ratings across age and race. Here, we again find a single outlier group: Younger blacks significantly prefer TextRank over MatchSum ($p < 0.01$). Our results imply that our standard evaluation methodologies do not align with the subjective evaluations of younger women and younger blacks.

We try to explain these two observations in §5.

We checked for significant group rating differences using bootstrap tests (Efron and Tibshirani, 1994; Dror et al., 2018). Across 1000 rounds, with Bonferroni correction, we find significant ($p < 0.05$) differences in preferences for these groups: ≥30, American Indian, white ♂, American Indian ♀, ≥30 ♂, Asian< 30, Asian< 30♂, white≥30♂, and American Indian ≥ 30♀. All these subdemographics exhibit significantly different ranking behavior from their peers. So, for example, our results show a significant difference between young and old raters.

We also bin our results by gender of the subjects of the biographies. We rely on Wikidata gender information to make this classification. There are 1409 preferences and ratings of men's biographies (Men), and 585 of biographies of women (Women). This of course means we see fewer significant differences in ratings of female biographies. For Men, we find significant differences across a wide range of groups, and with stronger effects for some demographics, suggesting that the gender of the subject of the biography *does* impact ratings differently across subdemographics. We find significant results for Women only for the subdemographic white ($p = 0.004$). This result is interesting, though, since it shows that on female biographies, white and non-white annotators prefer different systems.

Finally, we also asked our annotators to rank the two systems based on fluency, informativeness and usefulness. We used a 4-point forced Likert scale.

One observation is that even across fine-grained dimensions, younger annotators rate summaries lower; see Table 4. Interestingly, however, this difference is only observed with female biographies (rows 3–6). See Table 5 for the results on ALL across race. While ratings are generally low, we see clear differences, with Hispanics finding TextRank significantly more informative and useful, and American Indians finding TextRank significantly more fluent. Interestingly, Hispanics exhibit significant differences across WOMEN and MEN, finding TextRank summaries of female biographies significantly more informative and useful than TextRank summaries of male biographies.

3 Analysis

In order to analyze the differences between the rating behavior of subdemographics, we learn which features are significant for each demographic by training a simple logistic regression text classifier trained on the summaries ranked by each of the subdemographics with significantly different ranking behavior. As task representation, we represent each ranking instance as a vector of 2*149 features, one 149-sized subspace for each summary. Each subspace is made up of a one-hot vector of 145 frequent words (from the English stop words list in NLTK[8]), as well as four task specific features: the summary's average word length, whether the first sentence of the biography is included in the summary, the type/token ratio, and the text complexity of the summaries. We concatenate the 149 features from each system and scale them. We extract the top 20 most salient features for each demographic group and analyze them manually:

The **average word length** of the MatchSum system correlates positively to annotators preferring MatchSum across several demographics, e.g., OVER 30 and MALE WHITE, but this effect is absent with female annotators. Since the inductive bias of TextRank does not explicitly prohibit redundancy (Mihalcea and Tarau, 2004), this finding indicates that MatchSum is preferred among older men, especially whites, when it is informative, introduces main entities, etc. However, other subdemographics seem less sensitive to this variation. MatchSum is *not* generally rated more informative and useful across demographics (Table 5). In other subdemographics, e.g., AMERICAN INDIAN, MatchSum summaries with **pronouns** are rated higher, indicating it is better than TextRank at extracting sentences with pronouns without breaking coreference chains. Referential clarity, e.g., dangling pronouns, is a known source of error in summarization (Pitler et al., 2010; Durrett et al., 2016). TextRank summaries are often preferred by AMERICAN INDIAN and ASIAN, when they include **negation**. This is unsurprising, since negated sentences can often be very informative, and may seem more sophisticated in the context of machine-generated summaries. Negation is also a known source of error (Fiszman et al., 2006). In our data, however, this effect varies across subdemographics.

Our main observation is that female and black participants under 30 prefer TextRank over MatchSum. What drives this? The main predictors in our logistic regression analysis are a) TextRank extracting the **first sentence** of the biography (*twice* as frequently than MatchSum, in more than half of its summaries); and b) TextRank sentences containing **negation**. The former suggests a need for anchoring or framing of the summary, as initial sentences tend to provide this; the latter could suggest that young female or black participants are less prone to the common bias of evaluating negated sentences as less important (Kaup et al., 2013).

4 Conclusion

Our paper is, as far as we know, the first to evaluate summarization systems across different subdemographics. We did so in two different evaluation scenarios: automatic evaluation against gold summaries and system output ratings by human evaluators. We made the gold summaries and the ratings available for future research.

What did we learn from our experiments? Most importantly, of course, we learned that performance numbers differ when evaluated on summaries written by different subdemographics, and that the preferences of rathers from different subdemographics differ. In our experiments with automatic evaluation against gold summaries written by different subdemographics, we saw that summarization systems achieve higher performance scores when evaluated on summaries produced by white men, highlighting an unfortunate bias in these systems. In our rater studies, we also saw significant differences across subdemographics. Most surprisingly, perhaps, we saw that a summarization system from 2004 was rated better than a state-of-the-art system from 2020 by some subdemographics, and effect

[8] nltk.org

that was found to relate to the occurrence of first sentences (providing anchoring or framing of summaries) and negation (often evaluated as less important by majority groups). For now, we can only speculate what a summarization system optimized to perform well across *all* subdemographics would look like, e.g., a system minimizing the worst-case loss across subdemographics rather than the average loss. Our results show very clearly, however, the current state of the art in summarization is biased toward some demographics and therefore fundamentally unfair.

Acknowledgement

We thank the anonymous reviewers, the workshop organizers, and our colleagues at CoAStaL for their helpful comments on early drafts of this paper. The second author was funded by Innovation Fund Denmark.

Ethics Statement

We present two evaluations of summarization systems in which we bin participants by gender, age, and race. All demographic information was self-reported, and we payed annotators equally who chose *not* to report this information. Our work highlights the importance of recruiting balanced pools of participants in evaluations of summarization systems, an issue that has previously been ignored. A major limitation of this work is the under-representation of some groups, which led us to binarize all three social variables. We think of this study as a first attempt to highlight an important issue and hope that others will follow up with large-scale studies with better representation for more groups. Such studies could include many other social variables, e.g., income or level of education.

References

Federico Barrios, Federico López, Luis Argerich, and Rosa Wachenchauzer. 2016. Variations of the similarity function of textrank for automated summarization. *CoRR*, abs/1602.03606.

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.

Greg Durrett, Taylor Berg-Kirkpatrick, and Dan Klein. 2016. Learning-based single-document summarization with compression and anaphoricity constraints. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1998–2008, Berlin, Germany. Association for Computational Linguistics.

Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap*. CRC Press.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Online. Association for Computational Linguistics.

Marcelo Fiszman, Thomas C Rindflesch, and Halil Kilicoglu. 2006. Summarizing drug information in medline citations. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, page 254—258.

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.

Barbara Kaup, Rolf Zwaan, and Jana Lüdtke. 2013. The experiential view of language comprehension: How is negation represented?

Brian Larson. 2017. Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain. 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.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. 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.

Emily Pitler, Annie Louis, and Ani Nenkova. 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, Uppsala, Sweden. 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.

Hao Zheng and Mirella Lapata. 2019. Sentence centrality revisited for unsupervised summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6236–6247, Florence, Italy. Association for Computational Linguistics.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, Barcelona, Spain. Association for Computational Linguistics.