Corpora Evaluation and System Bias Detection in Multi-document Summarization

A critical approach

Yannick Lang
Study Program: B.Sc. Applied Computer Science

University of Bamberg, Bamberg, Germany
yannick-stephan.lang@stud.uni-bamberg.de

Abstract. This paper focuses on metrics for quantifying bias in multi-document summarization corpora and systems, compares different proposals and assesses their usage.

Keywords: Bias · Summarization · Corpora · Language Models

1 Introduction

With the recent rise in popularity of large language models such as GPT3 or DALL-E, and companies beginning to bridge the gap between research and commercial use, it is important to keep in mind that such models may not be free of biases.

This paper will focus on bias in both actual systems and underlying corpora of multi-document summarization (MDS) models. MDS has use cases such as review or news aggregation, and as such biases can have noticeable impact.[**?**] Starting from a 2020 paper[**?**] dealing with this topic, the metrics for quantifying biases proposed in this paper as well as its impact and possible alternative criteria will be discussed.

2 Original Paper

The starting point and primary source for this paper will be "Corpora Evaluation and System Bias Detection in Multi-document Summarization"[**?**], hereafter referenced as "original paper", which was published in late 2020. In it, the authors propose several metrics for quantifying biases in corpora for multi-document summarization models.

They then apply their metrics to several high profile corpora, analyze the results and request researchers consider those metrics when publishing new corpora in order to facilitate comparisons.

2.1 Metrics

The paper proposes a list of metrics which can be used to gauge the quality of a corpus. These metrics are:

– **Inter Document Similarity** shows the similarity between each documents
– **Pyramid Score** "defined as the ratio of a reference summary score and an optimal summary score"[**?**], i.e. how good the reference summaries are
– **Inverse Pyramid Score** measures the influence of documents on a given summary
– **Redundancy** describes the density of information in the documents

In addition, they also suggest a list of metrics that can be used to evaluate the performance of a given MDS systems:

– **ROUGE** one of the most basic metrics in text summarization, used to measure similarity between generated summaries and the references included in the dataset via recall [**?**]
– **F1 Score -** similar to ROUGE but considers both recall and precision
– **Inter Document Distribution -** similar to Inverse Pyramid Score, this measures the influence of each document on the generated summary
– **Redundancy** describes the summaries coverage of information from the documents

There are also metrics that can be applied to both corpus and system, where the reference summary is used for corpus evaluation and the generated summary for system evaluation.

– **Abstractness** quantifies the similarity between the generated or reference summary and the associated documents, where less similarity means higher abstractness
– **Layout Bias** quantifies the distribution of information within a document for corpora and the distribution of sections in the documents that provide the information in the generated summary for systems

The authors later group those metrics into subjective and objective metrics. They assign the highest importance to the objective metrics of Pyramid Score and Inverse Pyramid Score, and propose that scientists introducing new MDS corpora should at least report values for those, although ideally all metrics should be considered.

2.2 Results

The authors apply their metrics to a range of corpora and MDS systems. Their main finding is that corpus used has a high impact on the the performance of each system.

This can be seen in 1, which ranks system performance as measured by ROUGE-1 for every corpus under consideration. For every dataset, a different MDS system performed best, as denoted by value 1 and blue color.

The only outlier is ICSI_Summ, which managed to be the best-performing system on two separate corpora, but also achieved the worst ROUGE-1 rating on the DUC corpus and the third-worst on Opin.

The other collected metrics, such as ROUGE-2 or F1 Score, do not show the exact same pattern. When using ROUGE-2, for example, there are two systems that achieve the highest score on two separate Corpora instead of just one. The main observation, that no system strictly outperforms all others, still holds.

Therefore, to make reliable statements about the capabilities of a MDS system, it is not sufficient to just choose one dataset and report ROUGE scores for that. Instead, performance over multiple datasets should be considered and reported.

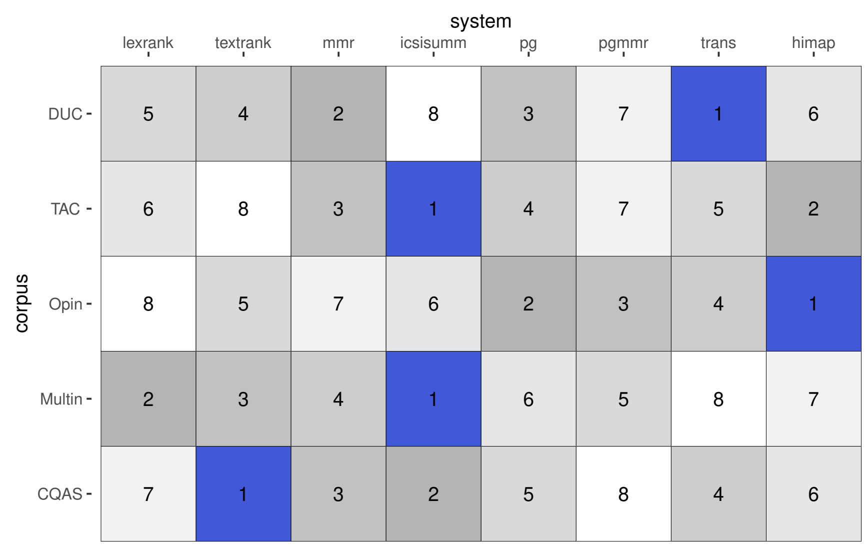

Fig. 1. Overview based on the results from the initial paper, where for every MDS system performance over each dataset is ranked and the best performing system for every dataset is highlighted. Lower number means higher ROUGE-1 score.

2.3 Key Takeaways

The authors state that a lack of clear definition for MDS tasks leads to a lack of a single standardized dataset. Instead, most scientists provide their own custom dataset with newly proposed MDS systems. This leads to a lack of comparability for the performance of systems, as this heavily depends on the actual dataset.[**?**]

2.4 Impact

But did this paper have any impact? Did the other scientists build upon these proposed metrics or apply them to their own system or corpus?

To answer this question, we will perform a forward search, i.e. look at other papers referencing "Corpora Evaluation and System Bias detection in Multi Document Summarization".

During the almost two years since its publication in October 2020, this paper was referenced 5 times. SemanticScholar.org classifies those citations into three categories: [?]

– Background - 3 citations
– Methods - 1 citation
– Results - 1 citation

It also identifies one citation, namely "ConvoSumm: Conversation Summarization Benchmark and Improved Abstractive Summarization with Argument Mining" (ConvoSumm)[?], as "highly influential"[?]. ConvoSumm propose a benchmark of summaries and performances for four "widely-used datasets". The authors use the Inter-document Similarity, Redundancy and Layout Bias metrics proposed in the original paper to evaluate their data. They also publish the results and their key finding from applying the metrics, which is a reduced layout bias in their data compared with the mostly news focused datasets that were used in the original paper.

The choice of metrics sounds reasonable, but deviates from the original papers authors assessment of the metrics importance. Dey et al. consider Pyramid Score and Inverse Pyramid score to be the most objective and most important metrics to report.[?]

Another paper, "Two-phase Multi-document Event Summarization on Core Event Graphs" [?], also uses a selection of proposed metrics to evaluate their datasets.

The other three papers that reference the original paper do so in a minor way. They use it to back up general statements or disclaimers about MDS corpora potentially containing bias the possibility of bias propagating to the generated content. [?,?,?]

An overview over the types of citations is given in 1.

In summary, it can be stated that the original paper has not yet had a lot of impact, which is not surprising given its somewhat recent publishing at the end of 2020. Two of the five studies that reference this paper actually use the proposed metrics, however it stands out that neither of those use the two metrics the original authors deemed most important.

uses metrics	IDS	Abstractness	Layout Bias	Redundancy
ConvoSumm [?]	x		x	x
2-Phase [?]	x	x		

generic citation	citation for:
Segmentation [?]	existence of bias
What is a Summary? [?]	existence of bias
Counseling [?]	past studies

Table 1. Breakdown of the context in which the five papers reference the original paper. In the left table, only metrics that were used are displayed, the x character indicates that this paper used this metric.

3 Alternative criteria

4 Current status

5 Summary

References

1. Adams, G., Alsentzer, E., Ketenci, M., Zucker, J.E., Elhadad, N.: What's in a summary? laying the groundwork for advances in hospital-course summarization. Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting **2021**, 4794–4811 (2021)
2. Chen, Z., Xu, J., Liao, M., Xue, T., He, K.: Two-phase multi-document event summarization on core event graphs. J. Artif. Intell. Res. **74**, 1037–1057 (2022)
3. Cohan, A., Ammar, W., van Zuylen, M., Cady, F.: Structural scaffolds for citation intent classification in scientific publications. In: NAACL (2019)
4. Dey, A., Chowdhury, T., Kumar, Y., Chakraborty, T.: Corpora evaluation and system bias detection in multi-document summarization. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 2830–2840. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.findings-emnlp.254, https://aclanthology.org/2020.findings-emnlp.254
5. Fabbri, A.R., Rahman, F., Rizvi, I., Wang, B., Li, H., Mehdad, Y., Radev, D.: Convosumm: Conversation summarization benchmark and improved abstractive summarization with argument mining. In: ACL (2021)
6. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), https://aclanthology.org/W04-1013
7. Moro, G., Ragazzi, L.: Semantic self-segmentation for abstractive summarization of long documents in low-resource regimes. In: AAAI (2022)
8. Srivastava, A., Suresh, T., Lord, S.P., Akhtar, M.S., Chakraborty, T.: Counseling summarization using mental health knowledge guided utterance filtering. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. p. 3920–3930. KDD '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3534678.3539187, https://doi.org/10.1145/3534678.3539187

9. Valenzuela, M., Ha, V.A., Etzioni, O.: Identifying meaningful citations. In: AAAI Workshop: Scholarly Big Data (2015)