Corpora Evaluation and System Bias Detection in Multi-document Summarization

Impact & Alternatives

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.

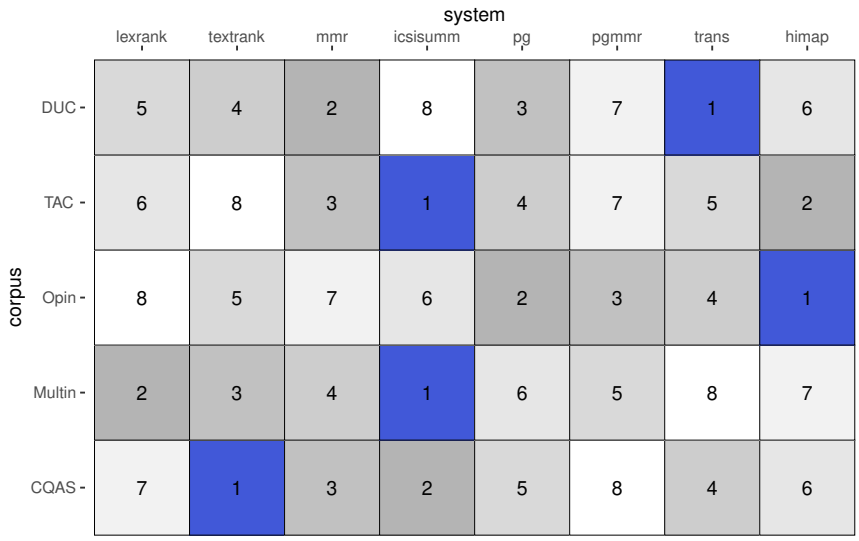

Fig. 1. Overview based on the results from the initial paper, where for every MDS system performance over each dataset is ranked and the best performing system for every dataset is highlighted. Lower number means higher ROUGE-1 score.

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.

For corpora, the authors highlight differences between datasets that were quantified by their metrics. They also discern trends over time and between crowd-sourced and non crowd-sourced datasets. Since three of the five considered corpora consist of news articles, the main bias mentioned throughout the document is layout bias.

2.3 Key Takeaways

Regarding biases, the paper focuses mostly on structural as opposed to contentual biases. The key bias highlighted in the paper is layout bias, i.e. the influence a tokens position in the document has on the likelihood of making it into the summary. The authors prove empirically, that layout bias present in a corpus carries over to trained models.

Layout Bias can be especially prominent int news articles, which make up a lot of (particularly early, i.e. early 2000) MDS corpora. This prominence can be explained by the availability of news articles as well as the combination and summarization and news articles being a valid real-world use case for MDS. Layout bias in news articles usually has a simple cause: reporters aim to provide readers with an overview over the topic within the first few sentences. This leads to a higher amount of summary relevant information in these sentences. A system trained on a news corpus may now perform significantly worse on a collection of documents that either have no layout bias at all or a different, not as front-loaded, kind of layout bias.

A further key observation from the paper is that abstractness in the training corpus is correlated with the abstractness of system generated content.

For the explicit causes of the corpus bias to system prediction bias pipeline and potential counter measures, the authors point to future research possibilities.

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 severely reduces the comparability between systems for two reasons.

For one, the authors have established the influence a specific dataset can have on common performance of MDS system (see 2.2). Therefore, the reported performance of a newly proposed system could be inflated when it is only evaluated on a custom corpus as compared with established corpora.

However, using an existing corpus may not always be viable, due to the unstandardized nature of MDS tasks. To increase comparability between datasets, the authors thus aim to establish metrics for comparing datasets. In order to be useful, these metrics should be reported by researchers creating a new corpus. For this, the authors divide the metrics into objective and subjective. Objective metrics include Pyramid and Inverse Pyramid Scores and are assigned a higher importance by the authors, who state those two "must be reported as they are strong indicators of generic corpus quality"[**?**].

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.

3 Alternative Criteria

In the last section we determined that, as of August 2022, no paper has built upon these metrics, either proposing changes to them, criticizing or extending them. But are there any other metrics or biases not without direct reference to the original paper?

At the beginning of their paper, Dey et al. state the problem of missing standard corpora due the non-standardized tasks of MDS. The same is currently true for metrics of bias. While the authors aimed to establish a framework for this, the previous section showed that they have not yet succeeded.

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.

As a result, some papers use different metrics, that may aim at similar attributes. For example, a 2016 paper proposing a new MDS corpus and uses textual heterogeneity as a metric. [?] While not exactly the same, it is introduced to determine the similarity of documents, just as IDS in the original paper.

Similarly, for the "Wikipedia Current Events Portal" (WCEP) dataset, which will be analyzed later in in this paper, the authors used "coverage" and "density" metrics to quantify abstractness, very similar to the Abstractness metric.[?] These coverage and density have been used for single document summarizations earlier.[?]

While Dey et al. focus on structural biases, especially Layout Bias, corpora generated from human content, i.e. most if not all corpora, also display biases in their content. These usually carry over through MDS systems into the generated summaries. [?]

Even if the training corpus does not contain biases, systems have been shown to introduce biases, such as hallucinations, i.e. statements in the generated summary that are not included in the candidate documents and do not follow from them. This is correlated with exposure bias, caused by differences in documents between training corpus and data used for actual work. [?]

4 Current Status

In this section, the status quo shall be outlined, i.e. how researchers, that propose MDS corpora handle bias and the overall quality of their dataset.

As previously stated, there is no widely accepted framework for this, leading to different approaches by different authors. I have selected recent (published in 2020 or later) papers introducing MDS corpora, and will now show their approach to bias and comparability.

5 Own Analysis

For this paper, I will calculate the metrics Dey et al. proposed for MDS corpora to a dataset myself. The corpus was selected on the following criteria:

\textbf{**MS2: Multi-Document Summarization of Medical Studies**}*cite*{MS2-https://doi.org/10.48550/arxiv
\textbf{**HowSumm: A Multi-Document Summarization Dataset Derived from WikiHow Articles**
}\cite{wikihow-https://doi.org/10.48550/arxiv.2110.03179
}
\textbf{**QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization**
}\cite{qmsum-https://doi.org/10.48550/arxiv.2104.05938}

- must be MDS corpus - the metrics were specifically designed for MDS corpora, and some, like Inter Document Similarity, only make sense in this context.
- must not already have been analyzed with these metrics, either in the original paper or by its creators
- should be somewhat recent
- dataset must be publicly available

This last criterion was actually the most constraining. I opted for the "Wikipedia Current Events Portal" (WCEP) dataset, proposed in 2020. [**?**] It meets all of the conditions mentioned above, as it was published in 2020, was designed explicitly for MDS, does not include the metrics from the original paper (which was published after WCEP) and the authors provide a public download link to the entire dataset on their GitHub.[**?**]

Although the original papers authors state that they "develop an interactive web portal for imminent corpora to be uploaded and evaluated based on [their] proposed metrics", I was unable to find any link or other kind of reference to this. The paper does, however, include a link to the source code for their analysis on GitHub, which serves as the starting point for analysis of WCEP.

For this paper I have forked the repository and updated the code for the corpus metrics to work with the WCEP dataset. I have also added documentation on how to use the code and which format the data is expected to have, which is missing in the original version.

I have not been able to measure the Layout Bias and the Pyramid Scores, as it is unclear to me what format the original authors code expect for the input to calculation.

The WCEP dataset is split into 3 parts, a vary large training set and the test and validation sets, which each make up about 10 percent of the total amount. For this evaluation, I treated them as separate datasets, so that I could determine if they are homogenous.

6 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. complementizer: complementizer/wcep-mds-dataset (Feb 2022), https://github.com/complementizer/wcep-mds-dataset
5. 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
6. 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)
7. Gholipour Ghalandari, D., Hokamp, C., Pham, N.T., Glover, J., Ifrim, G.: A large-scale multi-document summarization dataset from the Wikipedia current events portal. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 1302–1308. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.120, https://aclanthology.org/2020.acl-main.120
8. Grusky, M., Naaman, M., Artzi, Y.: Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 708–719. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/N18-1065, https://aclanthology.org/N18-1065
9. 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
10. Moro, G., Ragazzi, L.: Semantic self-segmentation for abstractive summarization of long documents in low-resource regimes. In: AAAI (2022)
11. Nadeem, M., Bethke, A., Reddy, S.: StereoSet: Measuring stereotypical bias in pretrained language models. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 5356–5371. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.acl-long.416, https://aclanthology.org/2021.acl-long.416
12. 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
13. Valenzuela, M., Ha, V.A., Etzioni, O.: Identifying meaningful citations. In: AAAI Workshop: Scholarly Big Data (2015)

14. Wang, C., Sennrich, R.: On exposure bias, hallucination and domain shift in neural machine translation. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.326, https://aclanthology.org/2020.acl-main.326/
15. Zopf, M., Peyrard, M., Eckle-Kohler, J.: The next step for multi-document summarization: A heterogeneous multi-genre corpus built with a novel construction approach. ACL Anthology p. 1535–1545 (Dec 2016), https://aclanthology.org/C16-1145/