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"[**?**], a paper 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 and analyze the results.

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

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.

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.

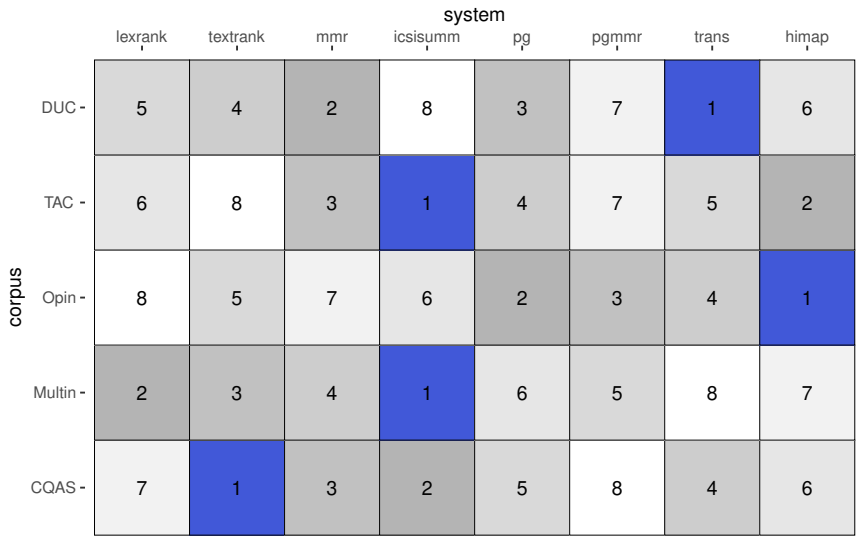

Fig. 1. Overview based on the results from the initial paper, where for every MDS system performance over each dataset is ranked and the best performing system for every dataset is highlighted. Lower number means higher ROUGE-1 score.

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

3 Alternative criteria

4 Current status

5 Summary

References

1. 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