Mitigating the Impact of Reference Quality on Evaluation of Summarization Systems with Reference-Free Metrics

Théo Gigant[*‡], **Camille Guinaudeau**[†], **Marc Decombas**[‡], **Frederic Dufaux**[*]
[*] Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes
{theo.gigant, frederic.dufaux}@l2s.centralesupelec.fr
[†] Université Paris-Saclay, Japanese French Laboratory for Informatics, CNRS
guinaudeau@limsi.fr
[‡] JustAI
marc@justai.co

Abstract

Automatic metrics are used as proxies to evaluate abstractive summarization systems when human annotations are too expensive. To be useful, these metrics should be fine-grained, show a high correlation with human annotations, and ideally be independent of reference quality; however, most standard evaluation metrics for summarization are reference-based, and existing reference-free metrics correlate poorly with relevance, especially on summaries of longer documents. In this paper, we introduce a reference-free metric that correlates well with human evaluated relevance, while being very cheap to compute. We show that this metric can also be used alongside reference-based metrics to improve their robustness in low quality reference settings.

1 Introduction

Given an input source, an abstractive summarization system should output a summary that is short, relevant, readable and consistent with the source. To reflect this, fine-grained human evaluations are split into different scores (Fabbri et al., 2021), such as fluency, faithfulness (sometimes called factual consistency), coherence and relevance. Fluency measures the linguistic quality of individual sentences, *eg* if they contain no grammatical errors. Coherence gauges if sentences in a summary are well-organized and well-structured. Faithfulness, or factual consistency, considers factual alignment between a summary and the source. Relevance is the measure of whether a summary contains the main ideas from the source.

Automatic summarization metrics are intended to capture one or multiple of these qualities (Zhu and Bhat, 2020; Vasilyev and Bohannon, 2021a), and used as a proxy to evaluate summarization systems when human annotations are too expensive.

These metrics can be compared on their different attributes such as the reliance on one or multiple references, the cost of inference (Wu et al., 2024), the dataset-agnosticism (Faysse et al., 2023) and their correlations with human judgment at system-level (Deutsch et al., 2022) or summary-level.

In this work, we introduce a new reference-free metric [1] that intends to capture the relevance of machine summaries using n-gram importance weighting. We rate n-grams of the source documents relative to how much semantic meaning they express, as measured by *tf-idf* (Sparck Jones, 1972), and score summaries according to their weighted lexical overlap with these n-grams.

We show that this metric is complementary to other metrics and can be mixed with reference-based metrics to alleviate their sensitivity to noisy and low quality references.

2 Related Work

2.1 Extractive summarization using word-importance estimation

A substantial amount of existing work investigated automatic extractive summarization using word-importance scores, based for instance on word statistics (Luhn, 1958), topic signatures (Lin and Hovy, 2000) or pretrained models (Hong and Nenkova, 2014). Our approach follows a similar line of thought by utilizing a word-importance score to identify and weigh the n-grams that should be included in an abstractive summary with high relevance.

2.2 Reference-based evaluation

Lexical overlap based metrics such as ROUGE (Lin, 2004), BLEU (Papineni et al., 2002) and chrF (Popović, 2015), or pretrained language model based metrics such as BERTScore (Zhang et al., 2019) and BARTScore (Yuan et al., 2021), are the standard way of evaluating abstractive summarization systems. However, these metrics rely on gold

[1] The code is available on github

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 19355–19368

standard reference summaries that can be costly, noisy, or missing altogether. We discuss some of the limits of these methods in section 3.

2.3 LLM-as-a-Judge evaluation

Large Language Models (LLMs) can perform many tasks effectively, even in few-shot or zero-shot settings. Recently, LLMs have also been used to evaluate natural language generation tasks, in replacement of human evaluation. LLM-as-a-Judge shows useful properties as an evaluation metric, for instance Faysse et al. (2023) illustrated using GPT-4 that it can be highly correlated with human judgement, format and task agnostic and comparable across tasks. Zheng et al. (2023) describe limitations of LLM-as-a-Judge, including position, verbosity and self-enhancement biases as well as poor performance at grading math or reasoning tasks. Other limitations are expressed by Kim et al. (2023) targeting proprietary LLMs such as GPT-4 for their closed source nature, uncontrolled versioning, and their high costs. Prometheus 2 (Kim et al., 2024) is designed for evaluating language models and shows high correlations with proprietary LLMs and human evaluations. Besides, its open-source nature mitigates some of the aforementioned issues. Liu et al. (2023) suggest that LLMs aligned from human feedback overfit to reference-less human evaluation of summaries, which they observed to be biased towards longer summaries and to suffer from low inter-annotator agreement.

2.4 Reference-free evaluation

Metrics designed to evaluate summaries without reference are useful when no gold reference are available, or when the property they intend to capture does not need a reference to be conveniently estimated.

GRUEN (Zhu and Bhat, 2020) aims at estimating the linguistic quality of a given summary by taking into account the grammaticality, non-redundancy, focus, structure and coherence of a summary. ESTIME (Vasilyev and Bohannon, 2021a) is evaluating the inconsistencies between the summary and the source by counting the mismatched embeddings out of the hidden layer of a pretrained language model. Info Diff (Egan et al., 2022) uses a pretrained model to compute the difference of Shannon information content between the source document and the source document given the summary. FEQA (Durmus et al., 2020) and SummaQA (Scialom et al., 2019) both compare how a model answers to questions about the document given the source document or a proposed summary.

Liu et al. (2023) observed that reference-free human evaluations have a very low correlation with reference-based human evaluations, and tend to be biased towards different types of systems.

2.5 Evaluating Summarization of Long Documents

Trained metrics usually generalize poorly to out-of-distribution tasks (Koh et al., 2022), and often cannot handle long contexts. In the long document summarization setting, Koh et al. (2022) showed that most automatic metrics correlate poorly with human judged relevance and factual consistency scores. Wu et al. (2024) use an extract-then-evaluate method to reduce the size of the long source document used as a reference for evaluation of factual consistency and relevance with LLM-as-a-Judge. They find that it both lowers the cost of evaluation, and improve the correlation with human judgement.

3 Limits of reference-based evaluation

Lexical overlap scores such as BLEU or ROUGE work under the implicit assumption that reference summaries are mostly extractive and contain no errors. This assumption is challenged by a study conducted by Maynez et al. (2020) on hallucinated content in abstractive summaries. In human written summaries from the XSum dataset, 76.9% of the gold references were found to have at least one hallucinated word.

Summarization methods can trade abstractiveness for faithfulness, creating a faithfulness-abstractiveness tradeoff curve that was illustrated and studied by Ladhak et al. (2022). They show that some metrics are more sensitive to the summary abstractiveness than others.

In the context of translations, *translationese* refers to source language artifacts found in both human and machine translations. This phenomenon is similar to extractive segments in summaries, as it is an artifact of the source document that can be mitigated through paraphrasing. Freitag et al. (2020) demonstrated that reference translations in machine translation datasets tend to exhibit this *translationese* language. They addressed this by creating new references through paraphrasing the existing ones. When tested, systems produced

much lower BLEU scores with the paraphrased references compared to the *translationese* ones, but the correlation with human judgment was higher. They observed that with *translationese* references, the n-grams with the highest match rates resulted from translations adhering to the source sentence structure. In contrast, using the paraphrased references, the most-matched n-grams were related to the semantic meaning of the sentence.

Following a *translationese* - extractiveness analogy, we assume that with highly extractive references, the most matched n-grams between proposed and reference summaries are artifacts of the extractiveness of the summaries. More abstractive references will yield much lower ROUGE scores, but might correlate better with human judgement.

We propose to use n-gram importance weighting methods, such as *tf-idf* (Sparck Jones, 1972) or *bm-25* (Robertson and Jones, 1976), to extract the n-grams expressing most of the semantic meaning of the source document. We believe that these n-grams should appear in relevant summaries, and are not artifacts of extractiveness.

4 Proposed Metric

Let $W_{t,d,D}$ be the importance of a n-gram t in a document d from a corpus D, defined as

$$W_{t,d,D} = \begin{cases} \tanh(\frac{w_{t,d,D}}{r_{t,d,D}}), & \text{if } t \in d \\ 0, & \text{otherwise,} \end{cases}$$

$w_{t,d,D}$ is an importance score obtained through word importance scoring methods (such as *tf-idf* and *bm-25*). The associated importance rank of the n-gram in the document is referred as $r_{t,d,D}$.

Given a proposed summary $\hat{s}$ of a document $d \in D$, we compute the metric:

$$m(\hat{s}, d, D) = \frac{\alpha_{\hat{s},d,D}}{N_{d,D}} \Sigma_{t \in \hat{s}} W_{t,d,D}$$

With $N_{d,D}$ the upper ceiling of the sum of weights, used to normalize the score: $N_{d,D} = \Sigma_{t \in d} W_{t,d,D}$.

By design this score will be maximized for a summary consisting of the full document. To alleviate this issue, we penalize longer summaries by multiplying with a term accounting for the length of the summary $|\hat{s}|$ relative to the length of the document $|d|$: $\alpha_{\hat{s},d} = f(|\hat{s}|, |d|)$[2]. We observe that this length penalty not only resolves the issue related to the scoring of entire documents but also shows a stronger correlation with human judgment at the system level.

It is relatively straightforward to devise a trivial heuristic that achieves a high score by employing the same n-gram importance weighting method to generate an extractive summary, with access to the full corpus. We do not consider this point to be a substantial issue, as such heuristic will result in a low score on metrics that measure other aspects of an abstractive summary, such as fluency.

5 Experiments

For our experiments, we work with different datasets of human evaluation of summarization systems. SummEval (Fabbri et al., 2021) contains human evaluations for 23 systems, each with 100 summaries of news article from the CNN/DailyMail dataset. Coherence, consistency, fluency and relevance are evaluated by experts and crowd-source workers. ArXiv and GovReport (Koh et al., 2022) contain annotations for 12 summarization systems, evaluated on 18 long documents for each dataset. Human evaluators rated the factual consistency and the relevance of the machine summaries. RoSE (Liu et al., 2023) is a benchmark consisting of 12 summarization systems evaluated on 100 news article from CNN/DailyMail. Each summary is annotated with different protocols, we are using the reference-based and reference-free human evaluations.

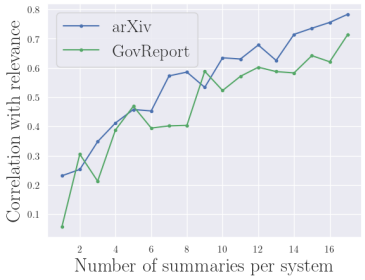
We describe the choice of settings for our metric in Appendix A, which takes into account system-level correlations on the four datasets, as well as the range of values taken by the metric.

5.1 System-level correlation scaling with number of summaries

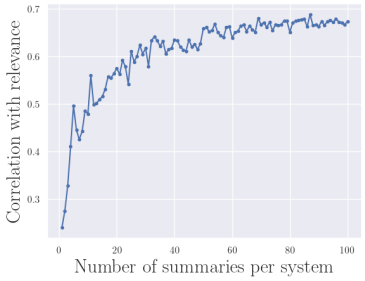
According to Deutsch et al. (2022), system-level correlations are usually inconsistent with the practical use of automatic evaluation metrics. To evaluate systems, usually only the subset of summaries judged by humans is used. However automatic metrics can be computed on summaries outside of this subset to give better estimates. Deutsch et al. (2022) also illustrates that testing with more examples will narrow down the confidence intervals of the evaluated scores, making it more convenient to compare systems. With a reference-free metric like ours, systems can be evaluated on more documents without the need for reference summaries. Figure 1

[2]The choice for f is illustrated in Appendix A, Figure 4

illustrates the increase of system-level correlation with human evaluated relevance when using more examples for each system.



(a) ArXiv and GovReport



(b) SummEval

Figure 1: System-level correlations with human judgement for our metric, depending on the number of summaries used for evaluation

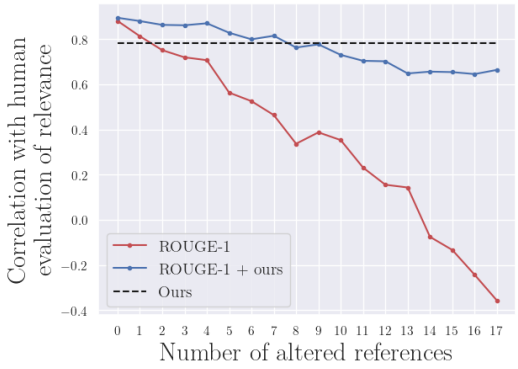


Figure 2: System-level correlation with human evaluation of relevance, depending on the number of altered references (RAND-3 alteration).

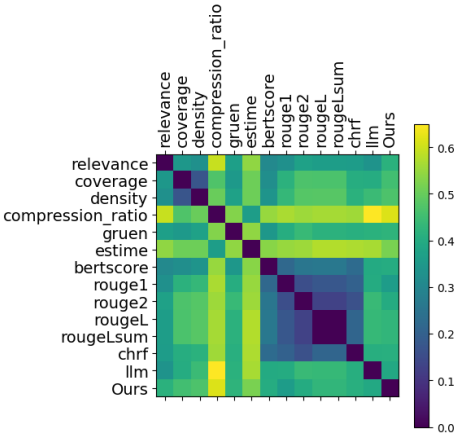


Figure 3: Complementarity between metrics on SummEval

5.2 Robustness to noisy references

Reference-based metrics such as ROUGE-1 are sensitive to the quality of the references. To evaluate the robustness of ROUGE-1 to noisy references, we gradually replace random reference summaries with altered references and compute the resulting system-level correlations. The references are altered by replacing them with three random sentences (RAND-3) from the source document. Results with the ArXiv dataset, averaged over 20 random draws, are reported in Figure 2. Results with different alteration methods and different datasets are reported in Figures 6, 7 and 8 in Appendix A. Our metric is not sensitive to altered references by design, contrary to ROUGE-1. When mixed with it, it improves the robustness of ROUGE-1 to low quality references. This aspect is beneficial in settings where the quality of the reference summaries is unknown or variable, for instance with web-crawled datasets.

5.3 Complementarity with other automatic metrics

We report the pairwise complementarity between each pair of metric[3] on SummEval in Figure 3, following Colombo et al. (2023). We observe that our metric has a high complementarity with most other metrics, noticeably with ROUGE and chrF scores, which are also based on lexical overlap, meaning that they capture different features of the evaluated summaries.

In Table 1 we report the system-level Spearman correlations using our metric, other metrics, and simple combinations of metrics. In the LLM-as-a-judge method, we are using the `gemini-1.5-flash` model (Gemini Team, 2024) following the prompt proposed by Wu et al. (2024) to evaluate the relevance of summaries.

[3]we use the `evaluate` implementation of ROUGE, chrF and BERTScore and official implementations of GRUEN and ESTIME

Table 1: System-level correlations of mixes of metrics

Metric	SummEval	ArXiv	GovReport
ROUGE-1	0.59	**0.88**	**0.92**
ROUGE-2	0.61	0.52	**0.91**
ROUGE-L	0.47	0.72	0.90
chrF	**0.75**	0.83	0.87
BERTScore	0.40	0.32	**0.91**
ROUGE-1 + chrF	**0.75**	**0.89**	0.90
ESTIME	-0.45	0.18	-0.69
GRUEN	0.59	0.32	-0.37
LLM-as-a-judge	**0.88**	**0.76**	0.63
Ours	0.67	**0.78**	**0.71**
Ours + ROUGE-1	**0.80**	**0.90**	**0.85**
Ours + chrF	0.74	0.83	0.82
Ours + BERTScore	0.74	0.77	0.76
LLM + ROUGE-1	**0.91**	**0.90**	**0.85**
LLM + chrF	0.89	0.87	**0.85**
LLM + BERTScore	**0.91**	0.81	0.70
Ours - ESTIME	0.71	-0.01	**0.77**
Ours + GRUEN	**0.83**	**0.71**	-0.12

Our simple metric achieves comparable results to LLM-as-a-Judge methods in term of correlations with human evaluations of summary relevance across various settings, at a significantly lower cost.

6 Conclusion and future works

In this work, we introduce a new reference-free metric based on importance-weighted n-gram overlap between the summary and the source. We demonstrated that it has high correlations with human judgement and can be used alongside other metrics to improve them and mitigate their sensitivity to low-quality references.

The prospects for future research include further exploration of the behaviour of reference-based, reference-free and hybrid metrics with references of varying quality, as well as potential extensions to multimodal settings such as the evaluation of vision-language systems.

7 Limitations

Like other lexical overlap metrics, ours works with the assumption that there is a vocabulary overlap between the source document and the summary, *ie* that the summary has a non-zero coverage. In order to evaluate the sensitivity of our metric to various levels of extractiveness of summaries, we would have wanted to compute the score on systems with varying values on the faithfulness-abstractiveness tradeoff curve presented in Ladhak et al. (2022); but their data was not made available yet.

Vasilyev and Bohannon (2021b) noticed that higher correlation with human scores can be achieved with "false" improvements, mimicking human behaviour. Using a referenceless evaluation metric, they limited the comparisons with the source text by selecting sentences to maximize their score, and observed a higher correlation with human judgement as a result. Wu et al. (2024) observe a similar consequence by first extracting sentences that maximize the ROUGE score with the original document and using the resulting extracted sentences along with the predicted summary as the input to be evaluated by a LLM-as-a-judge. Their interpretation however is different as they do not view this higher correlation with human scores as a "false" improvement, but as a way to mitigate the *Lost-in-the-Middle* problem of LLMs.

We believe that the relevant interpretation depends on the method that is used to extract sentences from the source document. Using comparisons with the summary to extract "oracle" spans of the original document, or selecting key sentences that span over the main information of the document are not motivated by the same reasons. Mimicking the human behaviour of referring only to the bits of the document that are relevant to the proposed summary *at first glance* to score marginally higher correlations is a different thing than filtering the most important bits of a document relative to a measure of word importance.

Our metric filters out the n-grams with little semantic significance in the document. This can mimick the human bias of comparing the summary to salient sentences only, but it will also lower the influence of the artifacts of extractiveness discussed in section 3.

Our metric is also specific to the task of summarization and might correlate differently with human judgement on summarization tasks with different compression ratio, extractiveness, or style. Table 2 in the Appendix A illustrates this.

LLM-as-a-Judge methods can solve the issues of sensitivity to extractiveness and task settings, while providing more interpretable results, but are not exempt from biases and come with a noticeably higher cost.

References

Pierre Colombo, Maxime Peyrard, Nathan Noiry, Robert West, and Pablo Piantanida. 2023. The Glass Ceiling of Automatic Evaluation in Natural Language Generation. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 178–183, Nusa Dua, Bali. Association for Computational Linguistics.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. Re-Examining System-Level Correlations of Automatic Summarization Evaluation Metrics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 6038–6052, Seattle, United States. Association for Computational Linguistics.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Esin Durmus, Faisal Ladhak, and Tatsunori Hashimoto. 2022. Spurious Correlations in Reference-Free Evaluation of Text Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1443–1454, Dublin, Ireland. Association for Computational Linguistics.

Nicholas Egan, Oleg Vasilyev, and John Bohannon. 2022. Play the Shannon Game with Language Models: A Human-Free Approach to Summary Evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10599–10607. Number: 10.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409. Place: Cambridge, MA Publisher: MIT Press.

Manuel Faysse, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2023. Revisiting Instruction Fine-tuned Model Evaluation to Guide Industrial Applications. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9033–9048, Singapore. Association for Computational Linguistics.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be Guilty but References are not Innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. Publication Title: arXiv e-prints ADS Bibcode: 2024arXiv240305530G.

Kai Hong and Ani Nenkova. 2014. Improving the Estimation of Word Importance for News Multi-Document Summarization. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 712–721, Gothenburg, Sweden. Association for Computational Linguistics.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2023. Prometheus: Inducing Fine-Grained Evaluation Capability in Language Models.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An Open Source Language Model Specialized in Evaluating Other Language Models. *arXiv preprint*. ArXiv:2405.01535 [cs].

Huan Yee Koh, Jiaxin Ju, He Zhang, Ming Liu, and Shirui Pan. 2022. How Far are We from Robust Long Abstractive Summarization? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2682–2698, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Faisal Ladhak, Esin Durmus, He He, Claire Cardie, and Kathleen McKeown. 2022. Faithful or Extractive? On Mitigating the Faithfulness-Abstractiveness Trade-off in Abstractive Summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1410–1421, Dublin, Ireland. 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.

Chin-Yew Lin and Eduard Hovy. 2000. The Automated Acquisition of Topic Signatures for Text Summarization. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*.

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023. Revisiting the Gold Standard: Grounding Summarization Evaluation with Robust Human Evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.

H. P. Luhn. 1958. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, 2(2):159–165. Conference Name: IBM Journal of Research and Development.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

S. E. Robertson and K. Sparck Jones. 1976. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–146. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.4630270302.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers Unite! Unsupervised Metrics for Reinforced Summarization Models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3246–3256, Hong Kong, China. Association for Computational Linguistics.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21. Publisher: MCB UP Ltd.

Oleg Vasilyev and John Bohannon. 2021a. ESTIME: Estimation of Summary-to-Text Inconsistency by Mismatched Embeddings. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 94–103, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Oleg Vasilyev and John Bohannon. 2021b. Is Human Scoring the Best Criteria for Summary Evaluation? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2184–2191, Online. Association for Computational Linguistics.

Yunshu Wu, Hayate Iso, Pouya Pezeshkpour, Nikita Bhutani, and Estevam Hruschka. 2024. Less is More for Long Document Summary Evaluation by LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 330–343, St. Julian's, Malta. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating Generated Text as Text Generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena.

Wanzheng Zhu and Suma Bhat. 2020. GRUEN for Evaluating Linguistic Quality of Generated Text. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 94–108, Online. Association for Computational Linguistics.

A Appendix

A.1 Spurious correlations

Durmus et al. (2022) observed that model-based reference-free evaluation often has higher correlations with spurious correlates such as perplexity, length, coverage or density, than with human scores. We report the correlations between metrics and spurious correlates in Table 2.

A.2 Correlations with human judgement on different settings

Figure 5 illustrate the distributions of system-level correlations of our metric with different settings.

For tokenization, we tested tokenizing texts as separated by space, using character tokenization, a pretrained GPT-2 tokenizer, or a custom tokenizer, trained on each corpus with a vocabulary of 100 tokens.

We included different sizes of n-grams in our tests, with bigrams, trigrams and 4-grams.

The two methods we considered for importance weigthing are *tf-idf* and *bm-25*.

The importance score is the weight used to score the overlapped n-grams, we included the following scores:

- importance: $t, d, D \mapsto w_{t,d,D}$
- exp-rank: $t, d, D \mapsto \exp(-r_{t,d,D})$
- inv-rank: $t, d, D \mapsto \frac{1}{r_{t,d,D}}$
- constant: $t, d, D \mapsto 1$
- tanh: $t, d, D \mapsto \tanh(\frac{w_{t,d,D}}{r_{t,d,D}})$

The options for the length penalty $\alpha_{\hat{s},\hat{d}}$ are no penalty or $\alpha_{\hat{s},d} = f(|\hat{s}|, |d|)$, with

$$f : |\hat{s}|, |d| \mapsto \frac{1}{1 + \exp(20 * \frac{|\hat{s}|}{|d|} - 10)}$$

f is illustrated in Figure 4.

We chose to use the corpus tokenizer, with trigrams, *tf-idf* and the tanh importance scoring with length penalty. These settings proved to be consistant in the tested conditions, and provided good ranges of values on different inputs. All the other experiments with our metric in this paper are using these settings.

Figures 6, 7 and 8 show the system-level correlation of our metric, ROUGE-1 and their combination

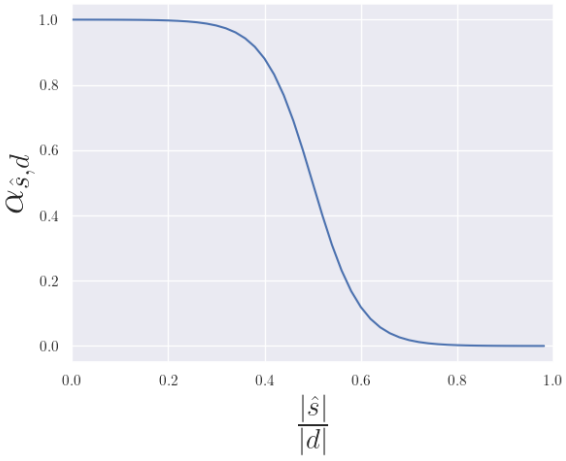


Figure 4: Length penalty $\alpha_{\hat{s},d} = f(|\hat{s}|, |d|)$ with

$$f : |\hat{s}|, |\hat{d}| \mapsto \frac{1}{1 + \exp(20 * \frac{|\hat{s}|}{|d|} - 10)}$$

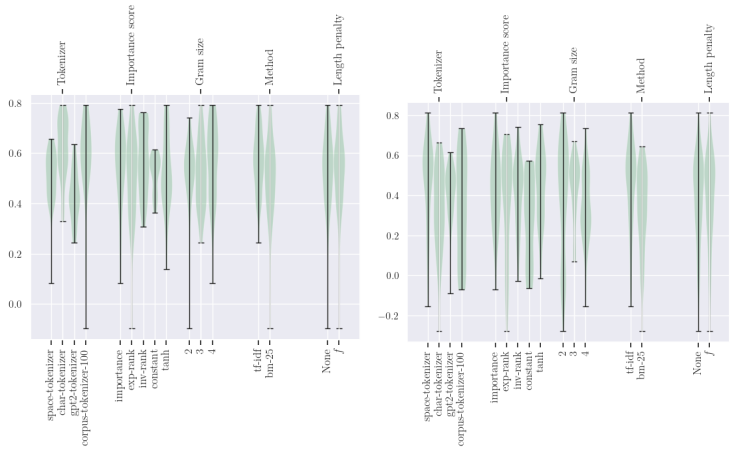
as we gradually replace the reference summaries with respectively three random sentences (RAND-3), the first three (LEAD-3) or last three (TAIL-3) sentences of the source document.

A.3 Range of values

We report the range of values taken by our metric, and ROUGE-1, for different inputs and on different datasets in Figures 9 and 10.

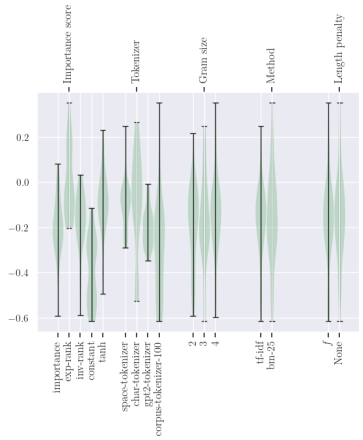
Table 2: Summary-level correlations between our metric, human evaluation metrics and spurious correlates. Values are bolded when the correlation with spurious correlate is higher than with human evaluation.

	SummEval		arXiv		GovReport		RoSE	
Metric	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
Relevance	0.24	0.23	0.42	0.45	0.15	0.18		
Coherence	0.20	0.23						
Consistency	0.03	0.04	-0.02	-0.11	-0.30	-0.32		
Fluency	0.01	-0.01						
Reference-based							0.17	0.14
Reference-free							0.18	0.15
Coverage	**0.26**	**0.28**	0.07	0.43	**-0.20**	**-0.19**	-0.08	0.05
Density	0.18	0.22	-0.04	-0.02	-0.03	0.17	0.04	0.02
Compression Ratio	-0.03	-0.06	0.01	0.02	**-0.54**	**-0.64**	**-0.28**	**-0.18**
Summary Length	**0.32**	**0.28**	0.40	0.28	0.02	-0.08	**0.30**	**0.26**

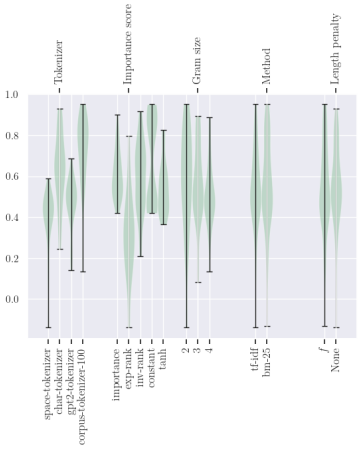


(a) ArXiv

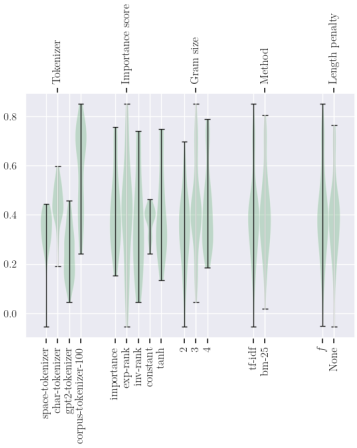
(b) GovReport



(c) RoSE (reference-based)

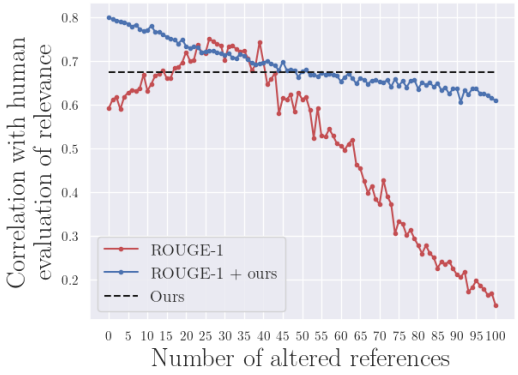


(d) RoSE (reference-free)

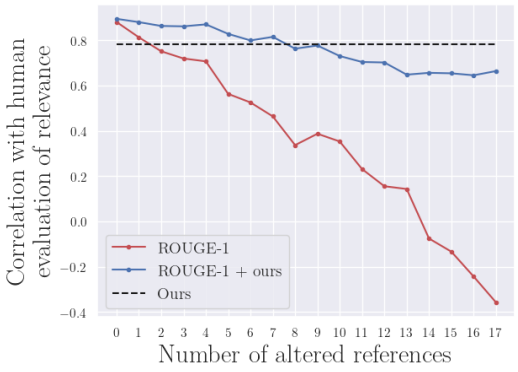


(e) SummEval

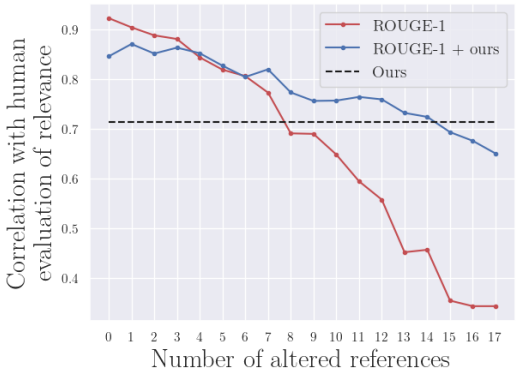
Figure 5: Distribution of system-level correlations of our metric in different settings



(a) SummEval

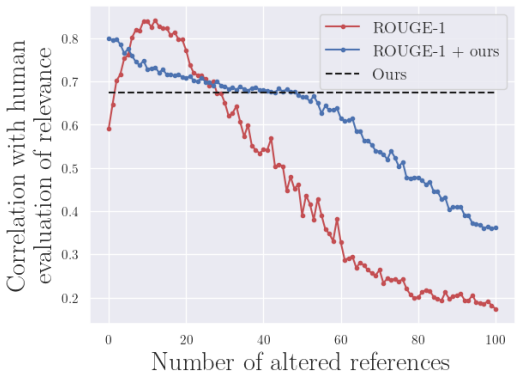


(b) arXiv

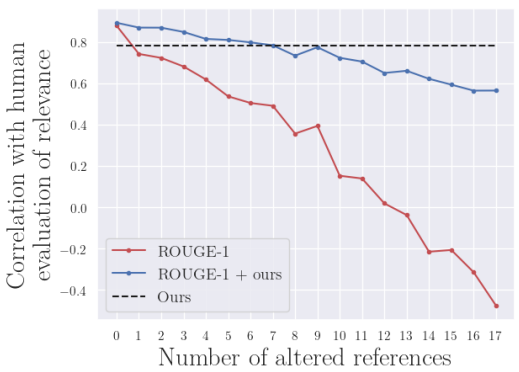


(c) GovReport

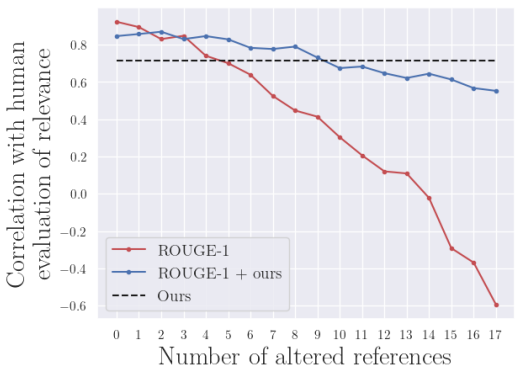
Figure 6: System-level correlation with human evaluation of relevance, depending on the number of altered references (RAND-3 alteration).



(a) SummEval

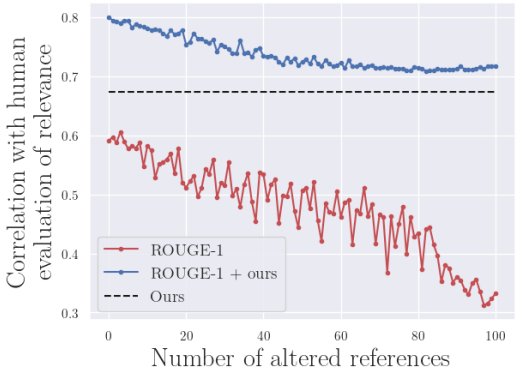


(b) arXiv

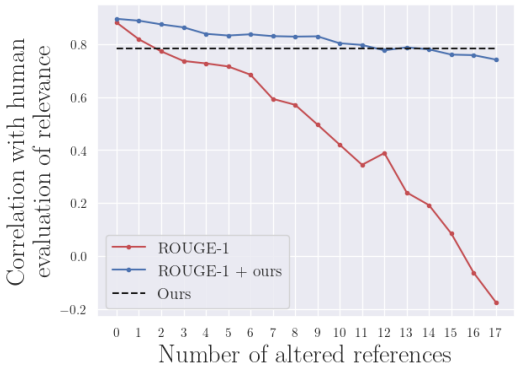


(c) GovReport

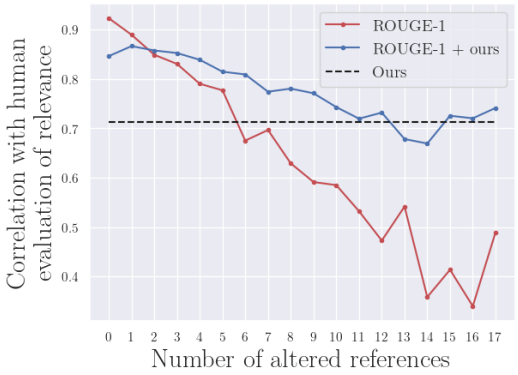
Figure 7: System-level correlation with human evaluation of relevance, depending on the number of altered references (LEAD-3 alteration).



(a) SummEval

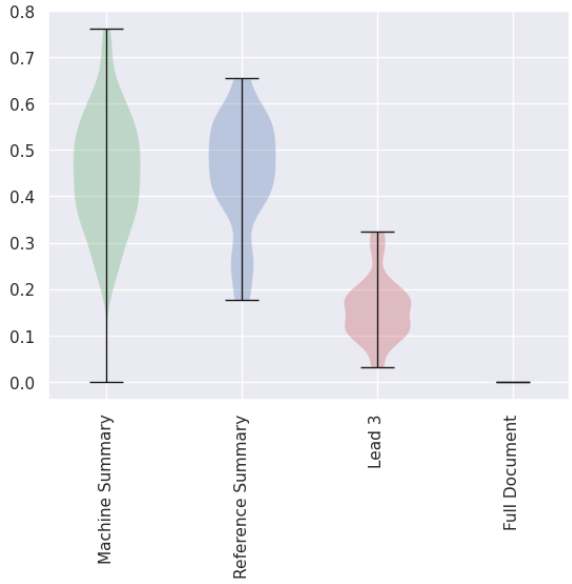


(b) arXiv

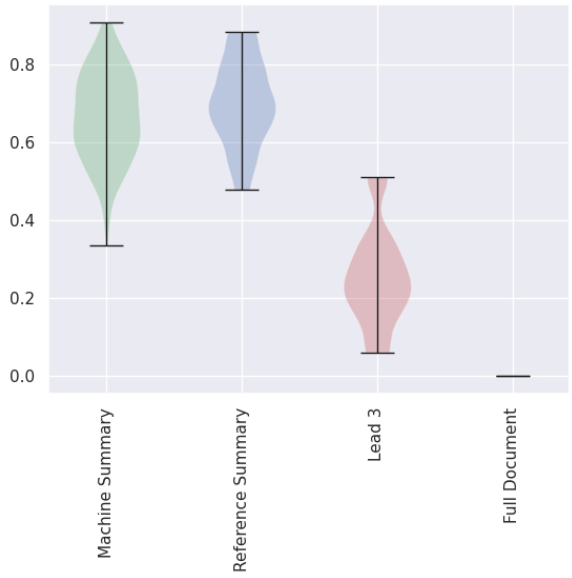


(c) GovReport

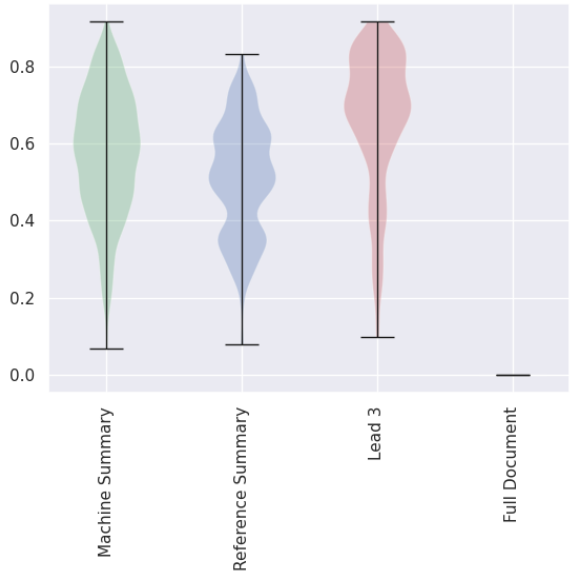
Figure 8: System-level correlation with human evaluation of relevance, depending on the number of altered references (TAIL-3 alteration).



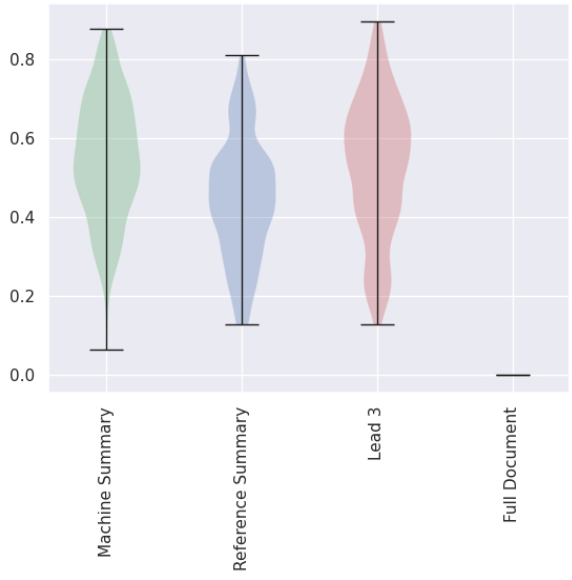
(a) ArXiv



(b) GovReport

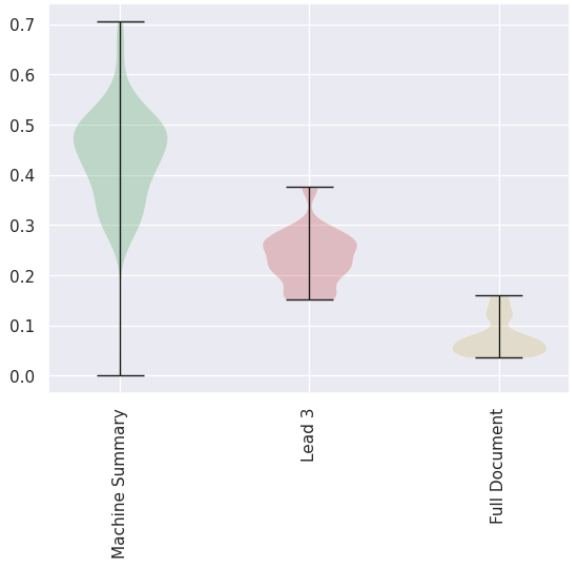


(c) SummEval

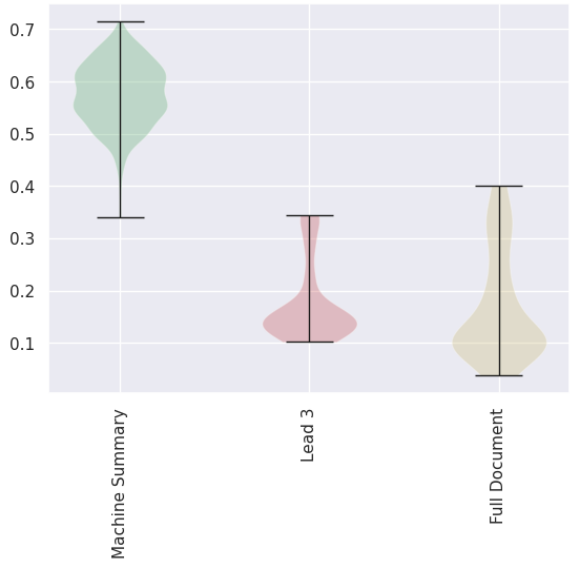


(d) RoSE

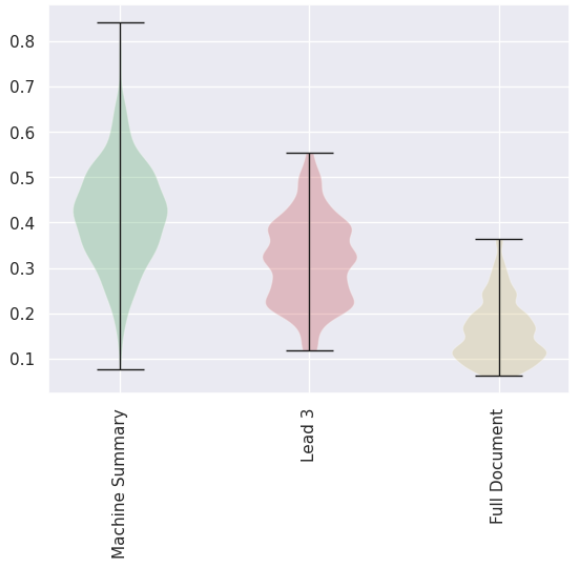
Figure 9: Range of values taken by our metric for different summaries



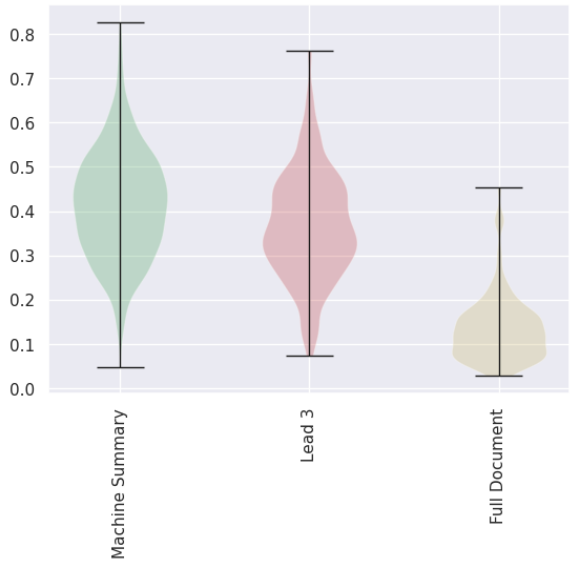
(a) ArXiv



(b) GovReport



(c) SummEval



(d) RoSE

Figure 10: Range of values taken by ROUGE-1 for different summaries