A Benchmark of metrics for text summarization

Antoine Gorceix
CentraleSupelec
antoinegorceix@gmail.com

Gillian Keusch
CentraleSupelec
gillian.keusch@student-cs.fr

Abstract

In this article, we perform a Benchmark of the different automatic metrics for the evaluation of the text summarization task. This had already been done in the SummEval article [15], we propose an analysis that completes it. We reuse the same dataset and consider correlations between automatic metrics and human metrics not only at the system-level but also at the level of the type of summary studied (SumUp-level correlation). Moreover, we will also compute the Spearman and Pearson correlations and we will include in our analysis 3 new metrics: the Depthscore, the BaryScore and the InfoLM. The idea is to better understand the relationship between automatic evaluation metrics and human metrics for the text summarization task and thus direct research into creating metrics that are more correlated with human metrics.

1 Introduction

Generative AI, particularly natural language generation (NLG), has emerged as a critical technology for applications such as chatbots, customer service, fairness [9; 22; 1; 3], and content generation [28; 10; 14; 7; 32; 4; 29; 19] . However, evaluating the quality of generated text remains a challenging task, as it requires extensive human effort and expertise [30; 8; 31; 6; 26; 16; 2; 27; 11; 17] . To address this challenge, researchers have developed various automatic evaluation metrics to measure the quality of NLG models [**?**]. These metrics aim to provide a quantitative assessment of the generated text's quality and can help streamline the development process of NLG models.

The importance of automatic metrics for NLG models cannot be overstated, as they offer an efficient and scalable solution to evaluate the quality of generated text. Moreover, automatic metrics can provide valuable insights into the strengths and weaknesses of NLG models, helping researchers and developers identify areas for improvement. While automatic metrics have their limitations and may not always reflect human judgement accurately, they remain a crucial tool for evaluating the quality of NLG models and improving their performance.

Overall, the development of effective and reliable automatic metrics is crucial for the continued progress and adoption of generative AI, particularly in the field of natural language generation. As such, researchers continue to explore and refine automatic metrics to improve their accuracy and applicability in evaluating the quality of generated text.

2 Problem Statement

The focus of our study is on the assessment of the text summarization task. Unlike some tasks such as classification, there is no obvious way to measure the performance. Automatic evaluation is used as a substitute for human evaluation because it is easy to implement, reproducible, fast and cheap. An automatic evaluation metric is considered good when it has a significant correlation with human scores. There are several automatic methods to evaluate the performance : Edit Based, N-gram Based and Embedding-Based. The idea of this paper is to propose a benchmark of different automatic evaluation metrics.
There are several strategies to evaluate the relevance of an automatic evaluation metric :
we used a strategy similar to the one implemented by in [11]. Indeed, the relevance of the automatic evaluation metrics is evaluated at two levels: at the system-level and at the Sum-Up level.

Notations: Let's consider s_i^j the sum-up generated by the model $M_i \in M_1, ..., M_{23}$ for the original text $j \in 1, ..., N$. $m(s_i^j)$ is the score associated by a metric m to the sum-up s_i^j.

- **Sum-up level correlation** : The correlation between m1 and m2 is evaluated as a loss or reward for a model, by measuring how well-suited m1 is with respect to m2. This is done for each sum-up across all system outputs, and then the mean is calculated.

- **System-level correlation** : The suitability of m1 with respect to m2 is measured. This is done by applying correlation to the mean values of both metrics across all sum-up for all systems.

For each of these strategies we will calculate 3 different correlations:

- **Kendall's correlation** : non-parametric measure of the strength and direction of the relationship between two variables.
- **Spearman's correlation** : measure of the degree of association between two variables, based on the ranks of the values rather than the actual values themselves.
- **Pearson's correlation** : statistical measure of the strength and direction of the linear relationship between two continuous variables.

3 Experiments Protocol

We used the dataset proposed by the article [15]. To create the dataset, summaries were generated on the CNN/DailyMail dataset by 23 recent summary models. All models were trained on the CNN/DailyMail news corpus, and the summaries were generated without any restrictions on the length using the test split of the dataset. The detailed description of the 23 models can be found here .

At first we worked on pre-process data thanks to a script that we ran on the dataset provided by [15]. This allowed us to obtain the correlation at a system-level. In a second step, we refined the analysis on a smaller set of metrics by also calculating the Sum-Up level correlation. We also calculated the Spearman and Pearson correlations in addition to the Kendall correlation. For this second part, the rouge-1, rouge-2, meteor, bertscore and blue metrics are kept. We introduce 3 new ones: DepthScore [26], BaryScore [8] and InfoLM [11]. Here is a more detailed description of the metrics for which we have done further analysis:

- **Meteor** : computes an alignment between candidate and reference sentences by mapping unigrams in the generated summary to 0 or 1 unigrams in the reference, based on stemming, synonyms, and paraphrastic matches. Precision and recall are computed and reported as a harmonic mean.
- **Rouge-1** : refers to the overlap of unigrams (each word) between the system and reference summaries.
- **Rouge-2** : refers to the overlap of bigrams between the system and reference summaries.
- **Bleu** : is a corpus-level precision-focused metric which calculates n-gram overlap between a candidate and reference utterance and includes a brevity penalty. It is the primary evaluation metric for machine translation.
- **BertScore** : computes similarity scores by aligning generated and reference summaries on a token-level. Token alignments are computed greedily to maximize the cosine similarity between contextualized token embeddings from BERT.
- **BaryScore** : is a multi-layers metric based on pretrained contextualized representations. Similar to MoverScore, it aggregates the layers of Bert before computing a similarity score.
- **Depthscore** : is a single layer metric based on pretrained contextualized representations. Similar to BertScore, it embeds both the candidate and the reference using a single layer of Bert to obtain discrete probability measures. Then, a similarity score is computed using a specific pseudo metric.
- **InfoLM** : InfoLM is a metric based on a pretrained language model (PLM). Given an input sentence S mask at position i, the PLM outputs a discret probability distribution over the vocabulary. The second key ingredient of InfoLM is a measure of information that computes a measure of similarity between the aggregated distributions.

Models were evaluated and reviewed by both crowd-sourced and expert judges, resulting in a collection of human annotations. These annotations were obtained by scoring 100 randomly selected articles from the CNN/DailyMail test set, with each summary being evaluated by 5 crowd-sourced and 3 expert workers to ensure the accuracy and quality of the annotations. The judges were then asked to rate each summary on a scale of 1 to 5 (with higher scores indicating better quality) based on four different dimensions. Below is a more detailed description of each of these dimensions:

- **Coherence** : refers to the ability of the sentences to flow logically and build upon each other to create a cohesive summary.
- **Consistency** : the factual alignment between the summary and the summarized source. It penalizes summaries that contained fabricated facts that were not supported by the source material.
- **Fluency** : the quality of individual sentences. Can be assessed based on factors such as readability, clarity, and grammatical correctness.
- **Relevance** : The summary has to avoid redundancies and excess details that could detract from the summary's effectiveness. Evaluators penalized summaries that contained unnecessary or redundant information.

In our case, in order to extract a human metric from each summary, we average the scores given by the experts for this human metric. To simplify the problem we do not take into account the turker annotations. Also in this idea of ease of implementation, for each summary and each automatic evaluation metric, we take into account a single reference to calculate the score. The InfoML metric is not taken into account in the following experiments because it took a long time to implement. Nevertheless it is included in the provided code, so it is possible to implement it thanks to our repository.

4 Results

In this section, we assess the suitability of current automated metrics for Summarization evaluation. We will not comment on the results obtained with the data already processed, as this would mean reproducing the analysis carried out by [15]. Nevertheless these results are presented in the code associated with the project. As we said before, our analysis is done from two points of view: **system-level** and **sum-up level**. Correlation results show several trends.

4.1 System-level correlation with human judgements

metrics	coherence	consistency	fluency	relevance
meteor	**0.35**	0.42	0.57	**0.58**
rouge-1	0.2	**0.57**	**0.58**	**0.58**
rouge-2	**0.35**	0.55	**0.58**	**0.58**
bleu	0.1	-0.13	0.18	0.2
bert	0.28	-0.28	-0.008	0.016
barry	0.03	0.20	-0.04	-0.10
depth	0.28	-0.28	-0.008	0.016

Table 1: Kendall's tau correlation of human metrics with automatic metrics on a System-level

The results obtained for the **system-level** viewpoint are presented in the table above. For each human metric, the most important correlation with the automatic evaluation metrics is highlighted in **bold**. The majority of metrics show a correlation with human criteria that is either weak (below 25%) or moderate (between 25% and 50%). The effectiveness of metrics is constant across all criteria. Metrics with weaker capabilities will exhibit low correlation across the board, whereas metrics with greater strength will demonstrate uniformly superior performance. 3 automatic evaluation metrics stand out from the others and show significant correlations: meteor, rouge-1 and rouge-2.

4.2 Sum-Up level correlation with human jugements

metrics	coherence	consistency	fluency	relevance
meteor	0.10	0.14	**0.10**	0.20
rouge-1	0.1	**0.15**	0.07	**0.22**
rouge-2	0.08	**0.15**	0.08	0.18
bleu	0.05	0.04	0.03	0.11
bert	**0.18**	0.005	0.07	0.12
barry	-0.04	-0.09	-0.08	-0.14
depth	**0.18**	0.005	0.08	0.13

Table 2: Kendall's tau correlation of human metrics with automatic metrics on a SumUp-level

The results obtained for the **Sum-Up level** viewpoint are presented in the table above. The correlations at the Sum-Up level are lower than at the system level, ranging from 0% to 22% for most measures. Therefore, while the metrics are poor estimators of human criteria at the summary level, they can be relevant and useful for comparing systems. At the Sum-Up level, no one metric seems to particularly stand out. Indeed it depends on the human metric considered: for coherence the DepthScore and the BertScore perform well, for consistency it is rouge-1 and rouge-2, for fluency meteor and for relevance red-1.

4.3 Pairwise correlation for all automatic metrics

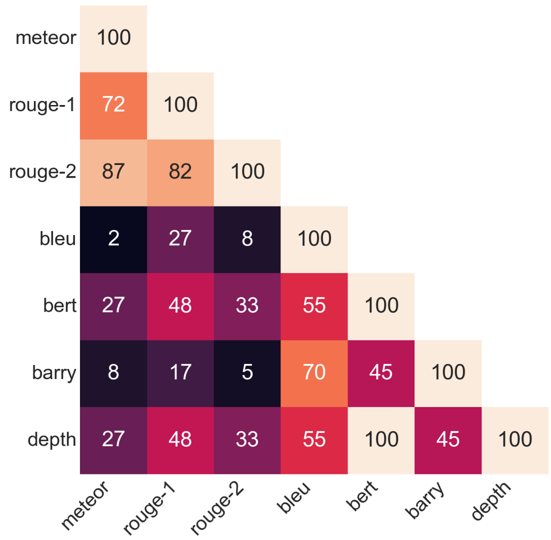

Figure 1: Pairwise Kendall's Tau correlations for all automatic evaluation metrics system level

System-level analysis : We notice that the Meteor metric is strongly correlated to the rouge-1 and rouge-2 metrics. The rouge-1 metric is highly correlated with the rouge-2 metric which seems coherent because their calculation is very similar. Although blue is n-gram based like rouge-1 and rouge-2, it is not highly correlated with them. On the other hand it has a high correlation with the BaryScore. The BertScore also appears to be highly correlated with the DepthScore.

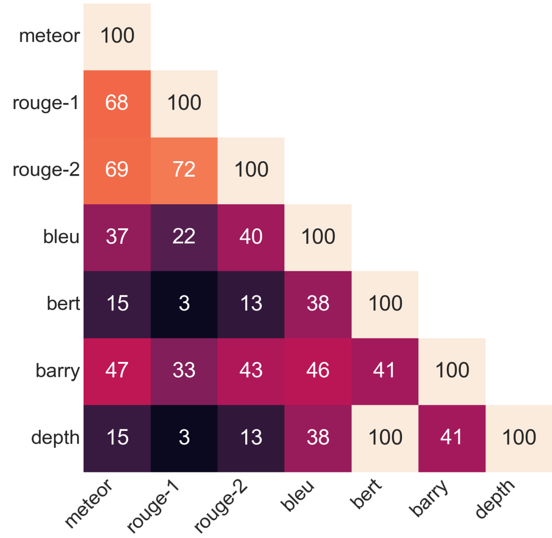

Figure 2: Pairwise Kendall's Tau correlations for all automatic evaluation metrics Sum-Up level

Sum-up level analysis : The system level analyses are still valid except for the correlation between blue and BaryScore which seems less important.

4.4 Influence of the correlation used on the final ranking

We place ourselves in the system-level to study the influence of the type of correlation. For each type of correlation, when we display the ranking of the 5 best performing metrics according to each human criterion, the rankings are identical for **Kendall** and **Spearman** for coherence, consistency and relevance and almost **identical** for fluency. On the other hand, there are **significant differences** between these two correlations and the **Pearson** correlation. Indeed, for example, for coherence, the Pearson correlation ranks the first three metrics as follows: bert, depth and blue while the Spearman and Kendall correlations lead to the following rankings: meteor, red-2, bert. The only common metric in this top 3 is the bert metric. We can try to explain this difference by the fact that the Pearson correlation measures the linear correlation between two continuous variables. It assumes a normal distribution and that the relationship between the two variables is linear. On the other hand, the Spearman and Kendall correlations measure the correlation between two variables without making any assumption about the distribution or the shape of the relationship between the two. These measures are therefore less sensitive to outliers and non-linear relationships. To overcome this decision problem it is possible to rely on the **Kemeny consensus** [20] which allows to combine several rankings i to form a common order of preference, which minimizes the sum of the deviations between the individual rankings and the common order.

We have implemented this algorithm for the human metrics coherence and relevance and we obtain the following top 3, **for coherence**: BertScore, Meteor and rouge-2 **for relevance**: Meteor, rouge-2 and rouge-1.

5 Discussion/Conclusion

To conclude, we proposed a study of various automatic evaluation metrics for the summarization task. We attempted to supplement the existing metric by adopting two viewpoints: a **Sum-Up level** viewpoint and a **system-level** viewpoint, by adding three new metrics and calculating three different correlations which we combine to obtain a final ranking.

There are certain aspects that could be improved to increase the robustness of the results obtained [12; 25; 21; 24; 5; 18; 13; 23] . Specifically, we only use expert ratings for human metrics, and we calculate scores for each summary based on a single reference only. It may be interesting to investigate the influence of these parameters. The project code is available here.

References

[1] P. Cheng, W. Hao, S. Dai, J. Liu, Z. Gan, and L. Carin. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, pages 1779–1788. PMLR, 2020.

[2] C. Chhun, P. Colombo, F. M. Suchanek, and C. Clavel. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5794–5836, Gyeongju, Republic of Korea, Oct. 2022. International Committee on Computational Linguistics.

[3] P. Colombo. *Learning to represent and generate text using information measures*. PhD thesis, Ph. D. thesis, Institut polytechnique de Paris, 2021.

[4] P. Colombo, C. Clavel, C. Yack, and G. Varni. Beam search with bidirectional strategies for neural response generation. In *Proceedings of the 4th International Conference on Natural Language and Speech Processing (ICNLSP 2021)*, pages 139–146, Trento, Italy, 12–13 Nov. 2021. Association for Computational Linguistics.

[5] P. Colombo, E. Dadalto, G. Staerman, N. Noiry, and P. Piantanida. Beyond mahalanobis distance for textual ood detection. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 17744–17759. Curran Associates, Inc., 2022.

[6] P. Colombo, N. Noiry, E. Irurozki, and S. Clémençon. What are the best systems? new perspectives on nlp benchmarking. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 26915–26932. Curran Associates, Inc., 2022.

[7] P. Colombo, P. Piantanida, and C. Clavel. A novel estimator of mutual information for learning to disentangle textual representations. 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)*, pages 6539–6550, Online, Aug. 2021. Association for Computational Linguistics.

[8] P. Colombo, G. Staerman, C. Clavel, and P. Piantanida. Automatic text evaluation through the lens of Wasserstein barycenters. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10450–10466, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics.

[9] P. Colombo, G. Staerman, N. Noiry, and P. Piantanida. Learning disentangled textual representations via statistical measures of similarity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2614–2630, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[10] P. Colombo, W. Witon, A. Modi, J. Kennedy, and M. Kapadia. Affect-driven dialog generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,*

Volume 1 (Long and Short Papers), pages 3734–3743, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[11] P. J. A. Colombo, C. Clavel, and P. Piantanida. Infolm: A new metric to evaluate summarization amp; data2text generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10554–10562, Jun. 2022.

[12] M. Darrin, P. Piantanida, and P. Colombo. Rainproof: An umbrella to shield text generators from out-of-distribution data. *arXiv preprint arXiv:2212.09171*, 2022.

[13] M. Darrin, G. Staerman, E. D. C. Gomes, J. C. Cheung, P. Piantanida, and P. Colombo. Unsupervised layer-wise score aggregation for textual ood detection. *arXiv preprint arXiv:2302.09852*, 2023.

[14] O. Dušek and F. Jurčíček. Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 45–51, Berlin, Germany, Aug. 2016. Association for Computational Linguistics.

[15] A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 2021.

[16] J. Fu, S.-K. Ng, Z. Jiang, and P. Liu. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*, 2023.

[17] S. Ghazarian, R. Weischedel, A. Galstyan, and N. Peng. Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7789–7796, 2020.

[18] D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[19] H. Jalalzai, P. Colombo, C. Clavel, E. Gaussier, G. Varni, E. Vignon, and A. Sabourin. Heavy-tailed representations, text polarity classification & data augmentation. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4295–4307. Curran Associates, Inc., 2020.

[20] J. G. Kemeny. Mathematics without numbers. *Daedalus*, 88(4):577–591, 1959.

[21] W. Liu, X. Wang, J. Owens, and Y. Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

[22] G. Pichler, P. J. A. Colombo, M. Boudiaf, G. Koliander, and P. Piantanida. A differential entropy estimator for training neural networks. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 17691–17715. PMLR, 17–23 Jul 2022.

[23] M. Picot, F. Granese, G. Staerman, M. Romanelli, F. Messina, P. Piantanida, and P. Colombo. A halfspace-mass depth-based method for adversarial attack detection. *Transactions on Machine Learning Research.*

[24] M. Picot, N. Noiry, P. Piantanida, and P. Colombo. Adversarial attack detection under realistic constraints.

[25] M. Picot, G. Staerman, F. Granese, N. Noiry, F. Messina, P. Piantanida, and P. Colombo. A simple unsupervised data depth-based method to detect adversarial images.

[26] G. Staerman, P. Mozharovskyi, P. Colombo, S. Clémençon, and F. d'Alché Buc. A pseudo-metric between probability distributions based on depth-trimmed regions. *arXiv preprint arXiv:2103.12711*, 2021.

[27] J. Wang, Y. Liang, F. Meng, H. Shi, Z. Li, J. Xu, J. Qu, and J. Zhou. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*, 2023.

[28] K. Wang, J. Tian, R. Wang, X. Quan, and J. Yu. Multi-domain dialogue acts and response co-generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7125–7134, Online, July 2020. Association for Computational Linguistics.

[29] W. Wei, J. Liu, X. Mao, G. Guo, F. Zhu, P. Zhou, and Y. Hu. Emotion-aware chat machine: Automatic emotional response generation for human-like emotional interaction. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1401–1410, 2019.

[30] Y.-T. Yeh, M. Eskenazi, and S. Mehri. A comprehensive assessment of dialog evaluation metrics. *arXiv preprint arXiv:2106.03706*, 2021.

[31] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[32] X. Zhou and W. Y. Wang. MojiTalk: Generating emotional responses at scale. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1128–1137, Melbourne, Australia, July 2018. Association for Computational Linguistics.