Automatic Metrics in Natural Language Generation: A Survey of Current Evaluation Practices

Anonymous ACL submission

Abstract

Automatic metrics are extensively used to evaluate natural language processing systems. However, there is an increasing focus on how they are used and reported. This work presents a survey on the use of automatic metrics, focusing on natural language generation (NLG) tasks. We report the used metrics, the rationale for choosing them, and how their use is reported. Our findings reveal significant shortcomings, including inappropriate metric usage, lack of implementation details, and missing correlations with human judgments. We conclude with recommendations that we believe authors should follow to enable more rigor within the field.

1 Introduction

Evaluation practices in Natural Language Processing (NLP) are increasingly coming under scrutiny. Concerns have been raised about automatic metrics, such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), which often show poor correlation with human judgments (Reiter and Belz, 2009; Novikova et al., 2017; Reiter, 2018). Additionally, interpreting and reproducing these metrics is challenging due to variations in implementations and parameters (Post, 2018).

Given these well-documented shortcomings, this paper offers a snapshot of the current state of metric-based evaluations in natural language generation (NLG). We analyze published works to understand how metrics are used and reported, and identify gaps and potential improvements.

2 Survey Method

To provide an up-to-date overview of automatic evaluation practices in NLG, we analyzed papers from the 2023 International Conference on Natural Language Generation (INLG) and NLG track of the Annual Meeting of the Association for Computational Linguistics (ACL). Each paper was annotated to capture key evaluation features: which automatic or human evaluation methods were used, whether the metric was newly introduced, and the specific tasks evaluated using the metrics. Additional features annotated included whether the results from automatic metrics were correlated with human evaluations, whether implementation details were provided, whether metrics were only reported in the appendix and whether the authors explained the rationale for their metric choices.

3 Analysis and Results

Our analysis covers 102 papers that included the evaluation of an NLG system. 94% used automatic metrics, and 57% incorporated human evaluations, but only 51% used both.

We identified 634 uses of automatic metrics across these papers, grouped into 34 metric families (see Table 1). As shown in Figure 1, the most frequently used metrics were BLEU and ROUGE despite concerns about their validity (Reiter, 2018). Trainable metrics like BERTScore (Zhang et al., 2019) and BLEURT (Sellam et al., 2020) have not yet overtaken traditional metrics in popularity.

Furthermore, 64% of papers using BLEU and 63% of those using ROUGE did not provide specific implementation details, raising concerns about reproducibility, especially given challenges in reproducing original evaluation scores across different implementations (Post, 2018).

A significant issue is the lack of rationale for metric usage: 77% of the metrics were used without justification. This is concerning, as it suggests a tendency to follow established practices without critical evaluation of metric relevance.

Metric Family Name	INLG	ACL	Total
BLEU	26	69	95
ROUGE	27	65	92
N-gram diversity	6	49	55
Style Classifier	5	37	42
BERTScore	8	32	40
Perplexity	3	29	32
METEOR	6	21	27
Semantic Similarity	9	12	21
Overlap	6	21	27
Factuality	5	13	18
Accuracy	8	8	16
Quality Estimation	7	7	14
Combination	0	14	14
BARTScore	2	10	12
NLI	44	8	12
F1	4	7	11
BLEURT	5	5	10
CIDEr	2	6	8
N-gram repetition	2	6	8
SARI	2	6	8
Sequence Length	3	5	8
MAUVE	0	8	8
Unieval	0	8	8
Distribution Comparison	0	7	7
NIST	0	7	7
MoverScore	1	5	6
PARENT	1	5	6
Recall	2	44	6
Edit Distance	1	5	6
Flesch Readability	1	3	4
Inference Speed	0	4	4
Precision	1	2	3
loss/error	0	3	3
chrF++	1	1	2

Table 1: Total automatic metric usage counts of each of the metric families for both INLG and ACL conferences.

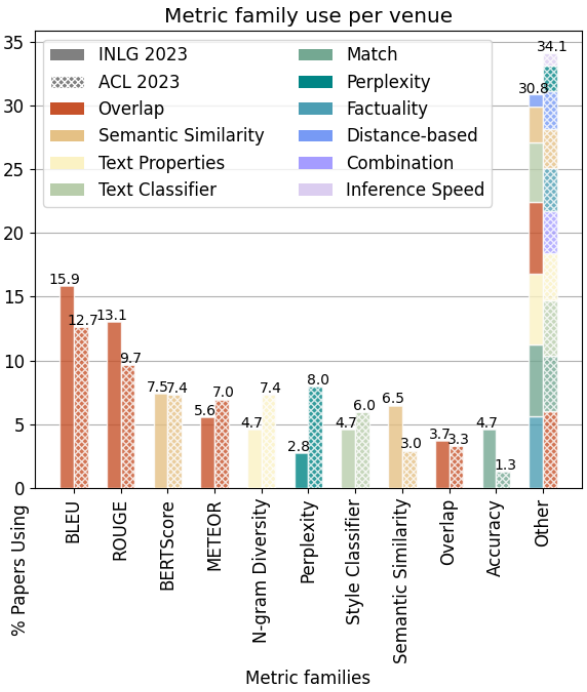

Figure 1: Usage percentages of top 10 metric families in INLG and ACL, with metric category color-coded.

We also investigated whether researchers correlated automatic and human evaluations. The majority of papers either did not conduct a human evaluation (42%) or did not comment on any correlation between human and automatic metrics (37%). Only a minority offered either quantitative or qualitative correlation analysis.

4 Discussion and Recommendations

Our analysis highlights key issues and offers actionable recommendations for improving evaluation practices in NLG research.

Metric Use, Rationale, and Combinations The use of automatic metrics often lacks a clear rationale, with only 13% of metrics justified beyond merely following prior work. Researchers should explicitly state the purpose and expected insights of each metric to avoid reliance on popular metrics without justification.

Additionally, metrics frequently have blind spots, so it is important to comment on how metrics are combined to ensure a comprehensive evaluation. Justifying the use of metrics—whether new or repurposed—based on empirical evidence or theoretical foundations is crucial for robust evaluations.

Automatic vs. Human Evaluation Discussing correlations between automatic and human evaluations is important. Papers should address any similarities or differences in the results to better understand the evaluation's effectiveness and limitations.

Reproducibility Authors should disclose metric implementation details, including library versions and parameters. Whenever possible, they should also release code. Sharing example outputs and human annotations, when available, is also recommended. Releasing datasets with evaluated outputs would allow future researchers to apply and develop new metrics.

5 Conclusion

Our analysis of metrics used at INLG and ACL 2023 reveals significant issues with the types of metrics used, the lack of comparison between automatic and human evaluations, and insufficient justifications for metric choices.

We offer recommendations to address these issues, but progress will depend on their adoption by the research community. Ultimately, enhancing transparency about metric usage and its rationale is crucial. This will clarify evaluation practices and improve the quality and reproducibility of evaluations in our field.

References

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.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. 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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Ehud Reiter and Anja Belz. 2009. An Investigation into the Validity of Some Metrics for Automatically Evaluating Natural Language Generation Systems. *Computational Linguistics*, 35(4):529–558.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.