Vol. 14 (2026): Proceedings of the 2nd International Conference on Artificial Intelligence, Modern Engineering and Environmental Sustainability (IC-AIMEES 2026)
E-ISSN: 3105-0522 P-ISSN：3105-0514
DOI: https://doi.org/10.70267/ ic-aimees.20260266270

A Review of Automatic Text Summarization Quality Evaluation Based on Attribution Explainability

Yingying Qi*

Mathematics and Applied Mathematics, School of Mathematics and Computer Sciences, Nanchang University, Nanchang 330000, China

**Corresponding author: Yingying Qi.*

Abstract

As automatic text summarization systems become more widely used, there is growing interest in how to evaluate their outputs in an interpretable and explainable way. The paper reviews existing work on automatic text summarization, explainable evaluation, and attribution-based interpretability methods. Particular attention is given to the differences between *ROUGE* and *BERTScore* in evaluating summary quality. Based on these metrics, the study further explores how *SHAP* can be used to interpret evaluation results generated by *ROUGE* and *BERTScore*. Attribution analysis based on *ROUGE* mainly reflects lexical overlap between generated and reference summaries, and can help identify missing or overemphasized information. In comparison, attribution analysis based on *BERTScore* is more sensitive to semantic similarity and may provide additional insights into paraphrasing behavior, hallucinations, and missing key information. Future research may focus on improving the efficiency and reliability of attribution-based evaluation methods, as well as exploring how attribution results can be incorporated into summarization model training and refinement.

Keywords

natural language processing, automatic text summarization, explainability, attribution analysis, SHAP

1. Introduction

The rapid development of large language models has greatly changed current approaches to automatic text summarization [1]. However, summarization systems still frequently generate hallucinated content that is either unsupported by or only weakly related to the source document [2]. This problem is particularly concerning in domains where factual accuracy is essential, such as medical and intelligence-related applications. As a result, researchers have increasingly focused on how to explain the decisions made by summarization models.

Most previous studies have focused on improving summarization performance rather than explaining or evaluating the quality of generated summaries. Currently, *ROUGE* is one of the most widely adopted methods for evaluating the quality of automatic text summaries. Because *ROUGE* mainly measures lexical overlap, it often fails to capture semantically similar expressions written in different forms. In addition, *ROUGE* usually produces only an overall score, making it difficult to understand why a summary is rated highly or poorly [3].

Later metrics such as *BERTScore* introduced contextual semantic matching, allowing evaluation beyond simple word overlap. Compared with *ROUGE*, *BERTScore* is more effective for evaluating abstractive summaries because it captures semantic similarity at the contextual level [4]. In this paper, attribution analysis is combined with *ROUGE* and *BERTScore* to improve the interpretability of summary evaluation [5]. Using *SHAP*, the contribution of individual tokens or sentences to the evaluation result can be estimated more explicitly. This makes it easier to identify which parts of the source text most strongly influence the generated summary and helps reveal hallucinated or missing information [6].

This paper reviews explainability research in automatic text summarization evaluation. It first introduces common attribution methods and existing evaluation metrics such as *ROUGE* and *BERTScore*, then discusses how attribution analysis can improve the interpretability of summary evaluation. Finally, the paper outlines current challenges and possible future directions.

2. Automatic Text Summarization and Its Explainability

2.1 Definition of Automatic Text Summarization

The rapid growth of online information has increased the demand for efficient text summarization tools. Effectively distilling this vast amount of data has attracted growing public attention. A summary is a concise and accurate representation of a document that enables readers to grasp its main content in minimal time. Automatic text summarization refers to the use of computational algorithms to generate such summaries, helping users grasp the main content of a document more efficiently [1].

2.2 Explainability in Automatic Text Summarization

As large language models continue to improve, automatic summarization is being used in a wider range of applications. Key concerns for users include how models construct summaries and whether the generated content accurately reflects the source material. Intrinsic hallucinations can undermine trust and affect decision-making [2]. To help users understand the model's decision process and better assess the reliability of its outputs, post-hoc attribution methods therefore provide a useful way to explain how summarization models generate their outputs.

3. Attribution Analysis Methods

Traditional evaluation methods for automatic text summarization mainly focus on the similarity between generated summaries and reference texts, but they provide limited insight into how summarization models produce their outputs. As a result, the interpretability of summarization systems has attracted increasing attention in recent years. This paper adopts post-hoc explainability methods commonly used in deep learning research. Among these methods, attribution-based approaches are selected because they are relatively intuitive and supported by clear mathematical formulations. As Chen Chong et al. point out, attribution methods assign contribution scores to input features such as tokens or sentences, making it possible to estimate how strongly different parts of the source text influence the generated summary [5]. Existing attribution methods can generally be divided into gradient-based, backpropagation-based, perturbation-based, and *Shapley* value-based approaches. In this study, *Shapley* value-based methods are employed to analyze the explainability of automatic summarization outputs.

4. Attribution Analysis for Quality Assessment in Automatic Text Summarization

4.1 Rouge

ROUGE is currently one of the most widely used automatic evaluation metrics for text summarization. It essentially measures how much information from the source document is covered by the generated summary by calculating recall over matching units such as *n-gram*s or longest common subsequences [1, 4]. *ROUGE* has been adopted as a primary evaluation metric in the assessment of most automatic summarization models. This method calculates the *ROUGE-N* score as the ratio of the number of overlapping *n-gram*s between the generated summary and the reference summary to the total number of *n-gram*s in the reference summary.

Because *ROUGE* is fully automatic, it can evaluate large numbers of summaries quickly and at relatively low cost. However, *ROUGE* mainly relies on lexical overlap and therefore often fails to recognize semantically equivalent expressions such as synonyms or paraphrases. In addition, *ROUGE* does not adequately evaluate coherence, fluency, or factual consistency, making it difficult to identify summaries that contain misleading or unsupported information. More importantly, *ROUGE* usually produces only an overall score and provides little information about which parts of the source text contribute to the evaluation result or which important content is missing from the generated summary.

4.2 Bertscore

To overcome these limitations of *ROUGE*, especially its weak ability to capture semantic similarity, researchers proposed *BERTScore*. Compared with *ROUGE*, *BERTScore* evaluates summaries at the semantic level and is generally better at measuring whether the generated summary preserves the meaning of the source text [4]. Unlike *ROUGE*, *BERTScore* calculates similarity using contextual embeddings produced by pretrained language models. Specifically, each token is first converted into a contextual embedding vector through a pretrained *BERT* model. It then computes the cosine similarity between each reference token and all tokens in the generated summary to derive recall, and between each generated token and all tokens in the reference summary to derive precision.

Compared with *ROUGE*, *BERTScore* is better at capturing semantic relationships between words and sentences. It demonstrates greater robustness to synonym substitution and varying syntactic structures expressing the same meaning. Previous studies have also shown that *BERTScore* correlates more strongly with human judgments of summary quality, especially in fluency and semantic relevance. *BERTScore* can also provide some clues about common summarization problems. For instance, low precision scores often suggest that the summary contains irrelevant or hallucinated content, while low recall suggests the omission of critical information from the source document. However, *BERTScore* is computationally more expensive than *ROUGE* and its performance depends heavily on the quality of the pretrained language model and reference summaries.

4.3 Attribution-Based Quality Assessment for Automatic Text Summarization

Among existing attribution methods, this study adopts the *Shapley* value approach because of its clear theoretical basis in cooperative game theory. This method is widely used in explainable machine learning research. Originating from cooperative game theory, where it addresses the fair allocation of total payoff among coalition members, the *Shapley* value was adapted to machine learning explainability by Lundberg et al. It estimates how much each feature contributes to the final prediction result [6]. Although the *Shapley* value is theoretically sound, its exact computation requires enumerating all possible feature subsets, which becomes computationally expensive for high-dimensional text inputs. *SHAP* improves the practicality of *Shapley*-value analysis by approximating feature contributions through an additive attribution framework. In attribution-based evaluation, the evaluation metric is treated as the target function, while the attribution method is used to explain how the final score is produced — that is, how the summary score is determined by specific tokens and sentences from the source document. Importantly, different evaluation metrics, when used as target functions, emphasize different aspects of model behavior.

When *ROUGE* is adopted as the target function *f* for *SHAP* analysis, the analysis focuses on which parts of the source text contribute most to lexical overlap between the generated and reference summaries. *SHAP* values directly quantify the contribution of each source document unit to the final *ROUGE* score, making token- or phrase-level correspondences easier to observe. This approach performs particularly well for extractive summarization. However, due to *ROUGE*'s limited ability to recognize synonym substitutions, reasonable paraphrasing or syntactic restructuring performed by the model may be incorrectly assigned low or zero contribution, even when such operations represent high-quality abstractive behavior.

When *BERTScore* is adopted as the target function *f* for *SHAP* analysis, the analysis instead focuses on semantic similarity: which parts of the source document contribute most to the semantic similarity between the generated summary and the reference summary in the deep embedding space. In this case, *SHAP* values quantify each unit in the source document's contribution to the semantic similarity between the generated and reference summaries. Unlike surface-level approaches, this method does not require high-contribution source

segments to match the reference summary lexically; instead, they must exhibit strong semantic relatedness. This makes it easier to observe how the model handles paraphrasing and semantic abstraction. Attribution results can reveal whether the model successfully identifies and utilizes semantically equivalent but lexically distinct source content, which may reflect the model's ability to capture semantic relationships beyond surface wording. Moreover, low-precision cases may help identify source regions associated with hallucinated content, while attribution on low-recall cases clearly identifies semantically important content that was omitted. In some cases, these analyses provide more informative results than *ROUGE*-based attribution. However, each attribution computation requires multiple calls to *BERTScore*, resulting in substantially higher computational overhead. Furthermore, this attribution method inherently depends on the reliability of the underlying model. Any biases or limitations in the pretrained *BERT* model will correspondingly propagate to and undermine the quality of the attribution explanations.

In practice, the two methods can also be used together. Researchers may first apply *ROUGE*-based attribution for preliminary identification of potential problems, followed by *BERTScore*-based attribution for more detailed semantic analysis. Alternatively, both can be used jointly to validate model performance. Cases where a source segment receives low *ROUGE* attribution but high *BERTScore* attribution may indicate strong synonym replacement and semantic abstraction capabilities. In summary, using *ROUGE* and *BERTScore* as target functions for attribution analysis provides complementary explanations of the summarization decision process at both lexical and semantic levels. Future research directions include developing more efficient semantic evaluation models to reduce attribution costs and integrating additional attribution metrics for more comprehensive and rigorous model explanations.

5. Challenges and Future Directions

Although the attribution-based evaluation framework proposed in this paper offers a new perspective for explainable summarization assessment, the discussion remains largely conceptual and still requires empirical verification. Two issues deserve particular attention. First, further experiments are needed to examine whether attribution scores can faithfully reflect the decision behavior of evaluation metrics and whether the explanations remain stable when the source text undergoes slight modifications. Second, the computational cost of *BERTScore* itself is relatively high, and incorporating attribution analysis further increases the complexity of evaluation.

Future research may therefore focus on two directions. One is to design more systematic experimental settings for testing the reliability and robustness of attribution explanations. The other is to explore lighter semantic evaluation frameworks that can reduce computational overhead while preserving interpretability. In addition, attribution results could potentially be incorporated into model training as auxiliary feedback, allowing explainability analysis to move beyond post-hoc interpretation and participate more directly in summarization optimization.

6. Conclusion

This paper reviewed two widely used evaluation metrics for automatic text summarization, namely *ROUGE* and *BERTScore*, and discussed their respective strengths and limitations in summarization quality assessment. On this basis, attribution analysis based on *SHAP* values was introduced into the evaluation process to provide more interpretable explanations of model behavior. The analysis suggests that *ROUGE*-based attribution is more suitable for identifying surface-level lexical overlap and is therefore particularly useful in extractive summarization tasks. By contrast, *BERTScore*-based attribution focuses more on semantic similarity and can better capture paraphrasing, synonym substitution, and other abstractive behaviors commonly found in generative summarization models. At the same time, the study also notes several limitations of attribution-based approaches, including their dependence on underlying pretrained models and the relatively high computational cost involved in semantic-level attribution analysis. Overall, attribution analysis provides a possible direction for improving the interpretability of summarization evaluation. Future work may further examine its practical reliability in different summarization settings and explore more efficient methods for combining explainability with automatic evaluation frameworks.

References

[1] Li, J. P., Zhang, C., Chen, X. J., et al. (2021). A review of research on automatic text summarization. *Journal of Computer Research and Development*, 58(01), 1-21.

[2] He, J., Shen, Y., & Xie, R. F. (2025). Identification and optimization of hallucination phenomena in large language models. *Journal of Computer Applications*, 45(03), 709-714.

[3] Yue, Y. F. (2020). Research and application of automatic summarization algorithms based on multi-models [Master's thesis, China Academy of Electronics and Information Technology]. https://doi.org/10.27728/d.cnki.gdzkx.2020.000013.

[4] Wang, Y. R. (2022). Research on generative text summarization methods based on deep learning [Master's thesis, North China University of Technology]. https://doi.org/10.26926/d.cnki.gbfgu.2022.000619.

[5] Chen, C., Chen, J., Zhang, H., et al. (2023). A review of interpretability in deep learning. *Computer Science*, 50(05), 52-63.

[6] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 4768-4777). Curran Associates Inc.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Open Access