MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, 2026, Pages 6621-6624

EVALUATION CHALLENGES IN AI-DRIVEN ABSTRACTIVE TEXT SUMMARIZATION FOR LOW-RESOURCE LANGUAGES: A SURVEY

Shilpa Serasiya, Assistant Professor, Department of Information Technology, Sankalchand Patel University
Sudha Patel, Assistant Professor, Department of Computer Engineering, Monark University
Nimisha Raval, Assistant Professor, Department of Computer Engineering, Sankalchand Patel University

ABSTRACT
Abstractive text summarization has gained increasing attention within natural language processing as a means to automatically condense large volumes of text while preserving key information and meaning. The rapid adoption of deep learning models has led to noticeable improvements in summary generation; however, determining how well these systems perform remains a persistent challenge. This issue is particularly evident in low-resource language environments, where data scarcity and linguistic variability complicate both model training and evaluation. This survey paper aims to critically examine existing evaluation practices for AI-based abstractive text summarization, with an emphasis on their applicability and reliability in low-resource language settings. The goal is to identify strengths, limitations, and open challenges associated with current evaluation methodologies. A comprehensive review of the literature is conducted, covering widely used automatic evaluation metrics, such as n-gram overlap–based measures, as well as more recent semantic and learning-based evaluation approaches. The survey also analyzes studies that incorporate human evaluation criteria, including content adequacy, coherence, fluency, and readability. Comparative insights are drawn from reported experimental results and evaluation frameworks across diverse summarization systems. The analysis reveals that while automatic evaluation metrics offer efficiency and reproducibility, they often provide an incomplete representation of true summary quality, especially in low-resource scenarios. Human evaluation remains indispensable for assessing semantic fidelity and linguistic naturalness, despite its high cost and limited scalability. This survey highlights the need for balanced evaluation strategies that integrate automatic and human-centered approaches, and it outlines future research directions toward more reliable and context-aware evaluation frameworks for abstractive text summarization.

Keywords: Abstractive Text Summarization, Evaluation Metrics, Low-Resource Languages, Automatic Evaluation, Human Evaluation, Natural Language Processing, Artificial Intelligence

1. INTRODUCTION

Abstractive text summarization has emerged as a prominent and intellectually demanding area within the domains of Artificial Intelligence (AI) and Natural Language Processing (NLP). Unlike extractive summarization, which selects and concatenates portions of the original text, abstractive summarization aims to generate concise summaries that capture the underlying meaning of the source content while potentially introducing novel phrases and reformulations. This process requires a deeper level of semantic understanding, contextual interpretation, and natural language generation, making it inherently more complex and reliant on advanced AI capabilities [1], [2]. The task closely resembles human summarization behavior, where comprehension, reasoning, and paraphrasing play a central role.The rapid evolution of deep learning techniques has significantly advanced the performance of abstractive summarization systems. In particular, sequence-to-sequence (Seq2Seq) architectures enhanced with attention mechanisms, along with more recent Transformer-based models, have demonstrated remarkable success in high-resource languages such as English and Chinese [3], [4]. These models leverage large-scale annotated datasets, pre-trained language representations, and well-established evaluation benchmarks, enabling them to generate fluent, coherent, and contextually accurate summaries. However, such progress is heavily dependent on the availability of extensive linguistic resources, which are often absent in low-resource languages. The application of abstractive summarization to low-resource languages introduces a unique set of challenges. These include limited availability of high-quality parallel corpora, lack of standardized benchmarks, and insufficient pre-trained models tailored to these languages. Additionally, linguistic complexities such as rich morphology, free word order, dialectal variations, and frequent code-switching further complicate both model training and evaluation. As a result, the performance and reliability of summarization systems in such contexts remain significantly constrained.Evaluation plays a critical role in determining the effectiveness, usability, and trustworthiness of generated summaries. Automatic evaluation metrics such as ROUGE, BLEU, and METEOR are widely adopted due to their computational efficiency, reproducibility, and ease of implementation [5]. However, these metrics primarily rely on surface-level lexical overlap between generated and reference summaries, often failing to capture deeper aspects such as semantic adequacy, factual consistency, coherence, and linguistic fluency. These limitations are particularly pronounced in abstractive summarization, where paraphrasing and novel expression are common, and even more so in low-resource languages where reference summaries may be scarce or inconsistent [6]. Human evaluation, while generally considered more reliable, introduces its own set of challenges. It is time-consuming, costly, and inherently subjective, often leading to variability in judgments across evaluators and contexts. Furthermore, designing standardized evaluation protocols that can be consistently applied across different languages and domains remains a significant hurdle. In low-resource settings, these issues are further exacerbated by the scarcity of linguistic expertise and the diversity of language usage patterns [7].

Given these challenges, the evaluation of AI-generated abstractive summaries, particularly in the context of low-resource languages—remains an open and evolving research problem. There is a growing need for more robust, semantically aware, and language-agnostic evaluation frameworks that can better capture the true quality of generated summaries. Considering these observations, this survey paper aims to provide a comprehensive and critical examination of existing evaluation methodologies used in abstractive text summarization, with a specific focus on AI-driven approaches for low-resource languages. The study seeks to identify current limitations, highlight research gaps, and explore emerging directions, including hybrid evaluation techniques and semantic-based metrics, that can contribute to more reliable and meaningful assessment of summarization systems in linguistically diverse and resource-constrained environments.

1.1 Scope of This Survey

The scope of this survey is to provide a comprehensive overview of the evaluation challenges and methodologies associated with AI-driven abstractive text summarization in low-resource language scenarios. Rather than focusing on architectures summarization model alone, this survey emphasizes the assessment mechanisms used to measure summary quality and their suitability for linguistically diverse and data-scarce environments. Specifically, this survey covers the following aspects:

1) Automatic Evaluation Metrics: An analysis of widely used metrics such as ROUGE, BLEU, METEOR, and BERTScore, discussing their strengths, limitations, and applicability to low-resource languages. Particular attention is given to issues related to lexical dependency, reference scarcity, and semantic inadequacy.
2) Human Evaluation Practices: A review of human-centric evaluation criteria, including fluency, coherence, relevance, and factual correctness, along with challenges related to subjectivity, inter-annotator agreement, and scalability in low-resource settings.
3) AI-Based and Semantic Evaluation Approaches: Examination of recent AI-driven evaluation techniques that leverage pretrained language models, semantic similarity, and contextual embeddings to address the shortcomings of traditional metrics, especially in languages with limited resources [8].
4) Challenges Specific to Low-Resource Languages: Discussion of linguistic, cultural, and infrastructural challenges such as morphological complexity, lack of standard datasets, domain mismatch, and limited evaluation benchmarks.
5) Future Research Directions: Identification of open research problems and potential directions, including multilingual evaluation frameworks, cross-lingual transfer learning, and human-in-the-loop evaluation strategies.

By synthesizing findings from existing literature, this survey aims to serve as a reference for researchers and practitioners working on AI-based summarization systems for low-resource languages, and to encourage the development of more reliable, fair, and semantically informed evaluation methodologies.

2. SURVEY METHODOLOGY

This survey follows a systematic literature review methodology to analyze evaluation metrics and associated challenges in AI-driven abstractive text summarization, with a particular emphasis on low-resource languages. The methodology is designed to ensure coverage of foundational works, recent advances, and emerging trends in evaluation practices while maintaining relevance to multilingual and data-scarce settings.

2.1 Survey Plan

The survey was conducted in four phases. First, a comprehensive literature search was performed using academic digital libraries. Second, relevant studies were screened based on predefined inclusion and exclusion criteria. Third, selected papers were analyzed and categorized according to evaluation methods, language focus, and challenges addressed. Finally, insights were synthesized to identify research gaps and future directions in evaluation methodologies for low-resource languages.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, 2026, Pages 6621-6624

2.2 Research Questions

The survey is guided by the following research questions:

RQ1: What automatic and human evaluation metrics are commonly used for abstractive text summarization?
RQ2: How effective are existing evaluation metrics when applied to low-resource languages?
RQ3: What challenges arise in evaluating AI-driven summarization systems for low-resource languages?
RQ4: What emerging evaluation approaches show promise for addressing these challenges?

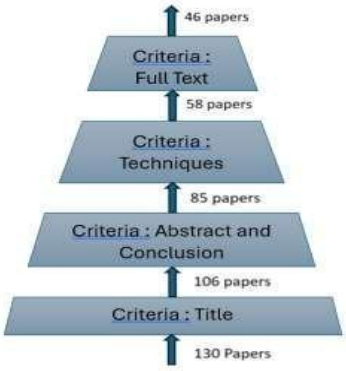
2.3 Source Selection

The literature for this review was collected from well-established and reliable academic databases, including IEEE Xplore, ACM Digital Library, Springer Link, Elsevier Science Direct, and the ACL Anthology. These platforms were chosen due to their strong reputation for publishing high-quality research in the fields of artificial intelligence and natural language processing. In addition to direct database searches, a citation-based approach was also adopted to ensure comprehensive coverage of relevant studies. This involved examining the reference lists of selected papers (backward citation) as well as identifying newer studies that cite these works (forward citation), thereby expanding the scope of the review.

2.4 Inclusion and Exclusion Criteria

The study focuses on peer-reviewed journal articles and conference papers published between 2015 and 2025 to capture recent developments in the field. As illustrated in Figure 1, the selection process was carried out in multiple stages to ensure relevance and quality. Initially, papers were screened based on their titles and abstracts.

FIGURE 1: Study Selection Process



This was followed by a more detailed assessment of their conclusions and full content. Only those studies that specifically address the evaluation of abstractive text summarization, particularly in the context of low-resource or multilingual languages—and discuss related challenges were included. Any papers that did not meet these criteria were systematically excluded from the final analysis.

3. NEED FOR LOW-RESOURCE LANGUAGE

Low-resource languages constitute a significant portion of global linguistic diversity, yet they continue to be inadequately represented in artificial intelligence and natural language processing research. Unlike high-resource languages, these languages typically lack large-scale annotated corpora, standardized evaluation benchmarks, and robust pretrained language models, all of which are critical for developing and assessing modern neural summarization systems [9]. As a result, both model training and performance evaluation are constrained by data scarcity and limited linguistic resources. Evaluation metrics commonly used in abstractive summarization research are predominantly designed and validated on high-resource languages, particularly English. These metrics often rely on surface-level textual overlap or pretrained semantic representations that assume relatively stable word order, limited morphological variation, and abundant training data. In contrast, many low-resource languages exhibit rich morphology, flexible syntax, and diverse semantic structures, which can significantly affect how meaning is expressed. Consequently, evaluation metrics developed for high-resource settings may fail to accurately capture summary quality in low-resource language contexts [10].The lack of reliable and language-sensitive evaluation frameworks makes it difficult to measure progress, compare models fairly, or reproduce results across studies. Inconsistent evaluation practices can also discourage research efforts and slow innovation in low-resource language summarization. Moreover, without appropriate evaluation mechanisms, improvements in fluency or semantic adequacy may go unnoticed, while models optimized for unsuitable metrics may produce misleadingly high scores. Addressing these challenges is essential for building inclusive and equitable AI systems. Developing evaluation approaches that account for linguistic diversity, semantic variation, and limited resources can enable more meaningful assessment and foster broader participation in NLP research. Such efforts are crucial not only for advancing abstractive summarization but also for ensuring that AI technologies benefit speakers of all languages, regardless of resource availability.

4. EVALUATION METRICS FOR ABSTRACTIVE SUMMARIZATION

Evaluating abstractive summarization systems is inherently complex due to the open-ended nature of generated summaries. Unlike extractive summarization, abstractive models often paraphrase, reorganize, or infer information, making direct comparison with reference summaries challenging. Existing evaluation approaches can be broadly categorized into lexical overlap–based metrics, semantic similarity–based metrics, and human evaluation metrics.

4.1 Lexical Overlap–Based Metrics:Lexical overlap based metrics are the most commonly used evaluation techniques in abstractive summarization research. Among them, ROUGE (Recall-Oriented Understudy for Gisting Evaluation) remains the de facto standard due to its simplicity, interpretability, and reproducibility. ROUGE variants such as ROUGE-1, ROUGE-2, and ROUGE-L measure n-gram overlap, bigram overlap, and longest common subsequence, respectively, between system-generated summaries and reference summaries [11]. Despite their widespread adoption, lexical metrics suffer from notable limitations. They assume that high-quality summaries share significant surface-level overlap with reference texts, which is often not the case in abstractive generation. These metrics fail to adequately capture paraphrasing, semantic equivalence, logical consistency, and factual correctness, leading to weak correlation with human judgment in many scenarios. As a result, models optimized solely for ROUGE may generate lexically similar but semantically inaccurate or unnatural summaries.

4.2 Semantic Similarity–Based Metrics :To address the shortcomings of lexical overlap metrics, recent research has proposed semantic similarity–based evaluation methods. Metrics such as BERTScore, MoverScore, and embedding-based cosine similarity leverage contextual representations from pretrained language models to assess semantic alignment between generated and reference summaries. These metrics are more robust to lexical variation and can better recognize paraphrases and semantically equivalent expressions. Empirical studies have shown that semantic metrics often exhibit higher correlation with human judgments compared to ROUGE. However, their effectiveness strongly depends on the availability and quality of pretrained language models. In low-resource language settings, pretrained models may be unavailable, undertrained, or biased, limiting the reliability and generalizability of semantic evaluation approaches [12], [13].

4.3 Human Evaluation Metrics:Human evaluation remains the most reliable method for assessing the quality of abstractive summaries. Human evaluators typically assess summaries along multiple qualitative dimensions, including fluency, coherence, relevance, adequacy, and faithfulness to the source text [14]. These evaluations provide nuanced insights into linguistic quality and semantic correctness that automatic metrics often fail to capture. However, human evaluation is time-consuming, cost-intensive, and inherently subjective. Inter-annotator agreement can vary significantly, and scaling human evaluation across large datasets or multilingual settings is often impractical. The scarcity of expert annotators further complicates human evaluation in low-resource and multilingual contexts.

Overall, while each category of evaluation metrics offers distinct advantages, none provides a complete solution. This has motivated recent research toward hybrid evaluation frameworks that combine automatic metrics with selective human assessment to achieve more reliable and scalable evaluation. Table 1 gives a comparison of evaluation metrics for abstractive summarization. Figure 2 brief the comparative analysis of evaluation methods.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, 2026, Pages 6621-6624

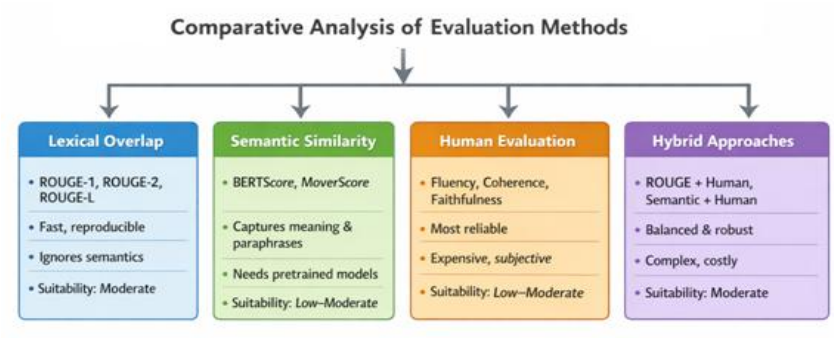


FIGURE 2: Comparative analysis of Evaluation method

TABLE 1: Comparison of Evaluation Method for Abstractive Summarization

Metric Category	Examples	Strengths	Limitations	Suitability for Low-Resource Languages
Lexical Overlap - Based	ROUGE-1, ROUGE-2, ROUGE-L	Simple, fast, reproducible, widely accepted	Ignore semantics, paraphrasing, and factual correctness	Moderate (language-independent but shallow)
Semantic Similarity - Based	BERTScore, MoverScore, Embedding Similarity	Captures semantic similarity, robust to paraphrasing	Requires high-quality pretrained models, computationally expensive	Low to Moderate (depends on model availability)
Human Evaluation	Fluency, Coherence, Relevance, Faithfulness	Most reliable, captures linguistic and semantic quality	Expensive, subjective, not scalable	Low (annotator scarcity and cost)
Hybrid Approaches	ROUGE + Human / Semantic + Human	Balanced evaluation, improved reliability	Increased complexity and cost	Moderate (promising research direction)

5. LITERATURE REVIEW

Previous work on the evaluation of abstractive text summarization focused mainly on the ROUGE metric due to its simplicity, interpretability, and wide usage in benchmarking various summarization models [15]. ROUGE assesses summaries by calculating the overlapping n-grams between the summaries produced by the model and one or more human-written references. For many years, this metric provided a convenient quantitative basis for comparing different summarization systems, especially in extractive and early neural approaches. However, with the emergence of neural sequence-to-sequence models and Transformer-based architectures, several studies began to highlight fundamental limitations of ROUGE for abstractive summarization [16], [17]. Neural models often generate paraphrased, restructured, or semantically equivalent summaries that differ lexically from reference summaries, resulting in low ROUGE scores despite high human-perceived quality. This discrepancy has raised concerns regarding the metric's ability to accurately reflect semantic adequacy, coherence, and factual consistency. To address these shortcomings, researchers have proposed a range of semantic and neural evaluation metrics. Metrics like BERTScore leverage contextualized embeddings generated using pretrained language models to quantify the semantic similarities between token pairs. They exhibit higher correlations with human assessments compared to conventional lexical-based metrics [18]. Likewise, MoverScore makes use of semantic embeddings and word movers' distances to identify meaning equivalencies between summaries [19]. The above examples illustrate the transition from surface matching to semantic assessments. Despite these advances, the application of semantic metrics to low-resource languages presents additional challenges. Most pretrained language models employed for metric computation are trained primarily using high-resource languages, rendering them ineffective in linguistic diversity and morphologically complex low-resource scenarios [20]. Studies have shown that evaluation reliability decreases significantly when reference summaries are scarce, domain-specific, or linguistically inconsistent [21]. Moreover, cross-lingual and multilingual evaluation methods, while promising, have demonstrated uneven performance across languages and domains [22]. These findings underscore the need for evaluation frameworks that are both language-agnostic and resource-efficient. Table 2 shows a comparative analysis of various metrics for abstractive text summarization.

TABLE 2: Summary of Evaluation Metrics for Abstractive Text Summarization

Metric	Category	Evaluation Focus	Advantages	Limitations
ROUGE-N / ROUGE-L	Lexical overlap	N-gram and sequence overlap	Simple, fast, widely accepted	Ignores semantic equivalence; biased toward lexical similarity
BLEU	Lexical precision	N-gram precision	Easy to compute	Poor correlation with human judgment in summarization
METEOR	Lexical + linguistic	Synonyms, stemming	Better than BLEU for paraphrasing	Still surface-level dependent
BERTScore	Semantic (neural)	Contextual similarity	High correlation with human judgment	Requires pretrained models; high computation
MoverScore	Semantic + distance	Meaning alignment	Captures paraphrasing effectively	Computationally expensive
Human Evaluation	Manual	Fluency, coherence, relevance	Most reliable	Subjective, costly, not scalable

Figure 3 gives comparative analysis on different Evaluation metrics based on limitations and key point of the evaluation.

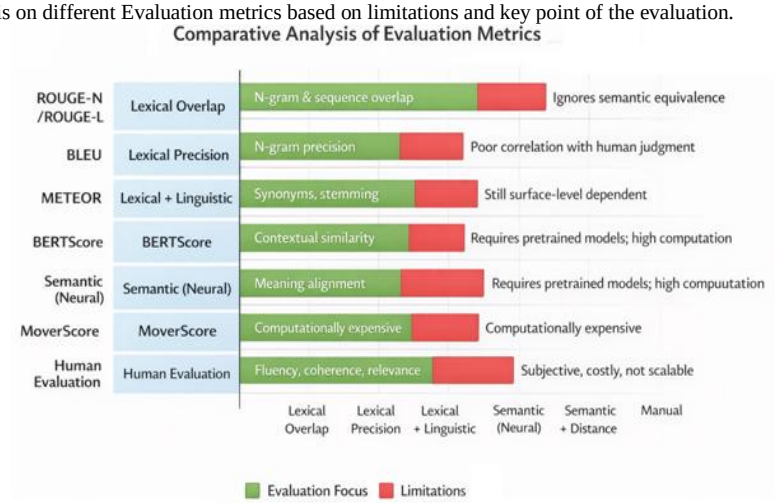


FIGURE 3: Comparative Analysis of Evaluation Metrics

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, 2026, Pages 6621-6624

The literature collectively indicates that evaluation metrics designed for high-resource languages do not generalize effectively to low-resource contexts. Lexical metrics underestimate semantic quality, while neural metrics suffer from inadequate pretrained representations. This gap presents a critical research opportunity to design evaluation methods that balance semantic fidelity, scalability, and linguistic inclusivity.

6. ISSUES AND CHALLENGES

Several challenges hinder effective evaluation of abstractive summarization in low-resource languages:

Reference Scarcity: Limited availability of high-quality human summaries
Metric Bias: Overdependence on lexical overlap metrics
Semantic Mismatch: Difficulty capturing meaning equivalence across languages
Morphological Complexity: High inflection and free word order reduce metric reliability
Human Evaluation Constraints: Cost, subjectivity, and annotator availability

These challenges highlight the inadequacy of current evaluation practices for low-resource scenarios.

7. APPLICATION DOMAINS AND RESEARCH OPPORTUNITIES

Abstractive summarization for low-resource languages is increasingly applied in domains of healthcare and clinical documentation, Legal and government records, educational content summarization, News and social media analysis, Digital inclusion and accessibility tools, Reliable evaluation is crucial in these domains, where inaccurate summaries may lead to misinformation or decision-making errors [22].

This survey identifies several promising research directions:

- Development of language-agnostic semantic evaluation metrics
- Human-in-the-loop evaluation frameworks
- Cross-lingual transfer of evaluation models
- Task-specific evaluation for critical domains
- Creation of standardized low-resource evaluation benchmarks

Advances in these areas can significantly improve fairness and reliability in summarization evaluation.

8. CONCLUSION

This survey reviewed existing evaluation metrics and key challenges associated with AI-driven abstractive text summarization, with particular attention to low-resource language scenarios. The analysis shows that widely used automatic metrics, originally developed for high-resource languages, often struggle to reflect semantic accuracy, factual consistency, and human-perceived quality when applied to low-resource contexts. While human evaluation provides richer and more reliable insights, its high cost and limited scalability restrict widespread adoption. These findings highlight the need for evaluation frameworks that are semantically aware, language-inclusive, and adaptable across resource settings. The insights presented in this survey are intended to inform and guide future research toward developing more robust, fair, and human-aligned evaluation methodologies for abstractive summarization.

REFERENCES

1. R. Nallapati, B. Zhou, C. dos Santos, Ç. Gülçehre, and B. Xiang, “Abstractive text summarization using sequence-to-sequence RNNs and beyond,” Proc. 20th SIGNLL Conf. Comput. Natural Lang. Learn. (CoNLL), 2016, pp. 280–290.
2. A. See, P. J. Liu, and C. D. Manning, “Get to the point: Summarization with pointer-generator networks,” in Proceedings of 55th Annual Meeting of the Association for Computational Linguistics (ACL), Vancouver, Canada, 2017, pp. 1073–1083.
3. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention Is All You Need,” in *Advances in Neural Information Processing Systems (NeurIPS 2017)*, vol. 30, 2017. Available: https://arxiv.org/abs/1706.03762.
4. Y. Liu and M. Lapata, “Text summarization with pretrained encoders,” in Proceedings of Conference on Empirical Methods in Natural Language Processing and the 9th Internation Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, 2019, pp. 3730–3740.
5. Dhaval Taunk1, Vasudeva Varma, “Summarizing Indian Languages using Multilingual Transformers based Models”, CEUR Workshop Proceedings, arXiv:2303.16657v1, 2023 DOI: 10.48550/arXiv.2303.16657
6. D. Novikova, O. Dušek, A. Curry, and V. Rieser, “Why we need new evaluation metrics for NLG,” in *Proc. 2017 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 2017, pp. 2241–2252.
7. S. Joshi, E. Santy, A. Budhiraja, K. Bali, and M. Choudhury, “The state and fate of linguistic diversity and inclusion in the NLP world,” in *Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020, pp. 6282–6293.
8. T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “BERTScore: Evaluating text generation with BERT,” in *Processing in international conference on Learning. Representations (ICLR)*, 2020.
9. A. Ghosh and A. Das, “A survey on text summarization for Indian languages,” ACM Transactions on Asian Low-Resource Language Information Processing., vol. 22, no. 3, pp. 1–34, 2023.
10. P. Kumar et al., “Challenges in NLP for low-resource languages,” IEEE Access, 2021.
11. M. P. Shrivastava and P. Bhattacharyya, “Hindi resources for natural language processing,” Proceedings International Conference on Language Resource and Evaluation (LREC), pp. 124–129, 2008.
12. W. Zhao, M. Peyrard, F. Liu, Y. Gao, C. Meyer, and S. Eger, “MoverScore: Text generation evaluating with contextualized embeddings,” in *Proc. 2019 Conf. on Empirical Methods in Natural Language Processing and 9th Int. Joint Conf. on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 563–578.
13. A. Kryscinski, “Evaluating the Factual Consistency of Abstractive Text Summarization,” in Proceedings at Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 9332–9346.
14. K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: a Method for Automatic Evaluation of Machine Translation,” Proceedings at 40th Annual Meeting of the ACL, Philadelphia, USA, 2002, pp. 311–318.
15. Dong Qiu Bing Yang “Text summarization based on multi-head self-attention mechanism and pointer network”, Complex & Intelligent Systems Volume 8, pages 555–567, 2022, DOI: 10.1007/s40747-021-00527-2
16. Shilpa G V, Shashi Kumar D R, “Abs-Sum-Kan: An Abstractive Text Summarization Technique for an Indian Regional Language by Induction of Tagging Rules”, International Journal of Recent Technology and Engineering Volume-8, Issue-2S3, July 2020, DOI : 10.35940/ijrte.B1193.0782S319
17. S. Novikova, P. Mantrach, F. Artzi, and M. Lemon, “Why We Need New Evaluation Metrics for NLG,” in Proceedings at Conference on Empirical Methods in Natural Language Processing (EMNLP), 2017, pp. 2241–2252.
18. W. Zhao, M. Peyrard, F. Liu, Y. Gao, C. M. Meyer, and S. Eger, “MoverScore: Text generation evaluating with contextualized embeddings and Earth Mover Distance,” in Proc. EMNLP, 2019.
19. Davoodijam, E., Alambardar Meybodi, M., “Evaluation metrics on text summarization: comprehensive survey”, Knowledge and Information Systems-Springer 66, 7717–7738 (2024). https://doi.org/10.1007/s10115-024-02217-0
20. M. Hasan, T. Joty, and E. Hoque, “Low-resource neural abstractive summarization,” in Proc. ACL, 2021.
21. X. Wang, W. Gao, M. Yang, Z. Zhang, and J. Li, “Multilingual and cross-lingual abstractive summarization,” *IEEE Access*, vol. 9, pp. 120123–120135, 2021.
22. D. Deutsch and D. Roth, “Understanding the extent to which summarization evaluation metrics measure information quality,” Trans. ACL, vol. 9, pp. 394–409, 2021.