International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE) ISSN: 0976-1353 Volume 32 Issue 03 – MAY 2026.

Emerging Research Trends for Machine Translation and Different Evaluation Metrics for Indian Languages

Kiranjeet Kaur[#1] and Sandeep Kaur[*2]

[#] *Department of Computer Science, Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India*

[*] *Department of Computer Science, Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India*

***Abstract*--Machine translation systems translate text automatically across languages while maintaining the original context by employing a range of Natural Language Processing techniques. Due to their growing use, it is now more important than ever to assess the translation quality of machine translation systems. However, due to their complex morphology and syntax, other low-resource languages could not always fit or apply to the current evaluation measures for English and other languages. This approach to evaluating machine translation is known as the machine translation evaluation (MTE) method. Similarity and accuracy levels can be ascertained by comparing the machine translation output with the MTE reference translation. The research assesses the metrics on many translation systems using datasets for low-resource languages. The study contributes to machine translation (MT) research by providing useful assessment standards for low-resource languages. Data from 1993 to 2024 were collected and analyzed for this study using the keywords "machine translation" AND "evaluation metrics" from the Scopus database. This article computes many application areas, yearly published documents, and source-wise analysis using Scopus databases. The study and evidence presented in this article demonstrate the importance of evaluation metrics for various low-resource languages.**

Keywords*-- *Automatic Machine Evaluation, Evaluation Metrics, Indic Languages, Machine Translation.

I. INTRODUCTION

Natural language processing advances and the development of extensive language models have driven the field of machine translation (MT) into an amazing upsurge in recent years. With a major emphasis on high-resource languages like English, this growth has not been evenly divided across all languages. Over a billion people speak Indian languages globally, hence there is a shortage in research and assessment measures on these languages as a result of this imbalance. When it comes to MT systems, Indian languages have different barriers than European languages because of their distinctive linguistic traits, such as variable word order and rich morphology [1].To address this gap, researchers have begun working on developing and assessing MT systems for Indian languages, with a focus on defining metrics that faithfully represent translation quality.

By developing a comprehensive dataset for assessing MT measures specifically for Indian languages, the IndicMT Eval project, for example, seeks to close this research gap. This dataset called the Multidimensional Quality Metric (MQM) dataset, includes seven MT systems and 7000 fine-grained annotations in five Indian languages. Pre-trained models like COMET are among the metrics that the project assesses. It finds that these metrics have strong relationships with annotator scores, suggesting that they are useful in determining the quality of MT [2].

MT is a difficult endeavor, particularly for low-resource languages with rich morphologies and few parallel corpora. Therefore, it's critical to have trustworthy methods for evaluating MTS efficacy for low-resource languages. To address this issue, automated assessment tools for MT quality have been developed. These metrics are based on accuracy, adequacy, and fluency, among other things [3]. Understanding the advantages and disadvantages of the various metrics available is necessary to choose the best one for a certain application.Due to the quick development of MT systems, thorough meta-evaluations of assessment measures are now required to guarantee that the metrics most closely represent the quality of MT are chosen. However, an enormous amount of the most recent literature on assessment metrics overlooks the demands of low-resource languages in favor of high-resource language focus. In the context of Indian languages, which differ linguistically from English and have distinctive traits like agglutination in Dravidian languages and comparatively free word order, this disparity is especially prominent [4].

Furthermore, these evaluation criteria may not accurately capture the linguistic richness and complexity of low-resource languages. Assessing someone automatically is easy and should be similar

to assessing someone manually [5]. Moreover, MT quality may be significantly impacted by the particular system being utilized, its architecture, and the assessment criteria that are applied.

The major contributions of this work are as follows:

• We have examined the bibliometric analysis indicating study and development in the field of machine translation and evaluation metrics gradually increased from 1993 to 2024.
• A graphical depiction of country-specific publications and citations has been analyzed and addressed.
• Analysis has been done on the keyword clustering of "Machine Translation" and "Evaluation Metrics".
• A graphical representation of reputable journals and SCOPUS publications with keywords machine translation for assessment measures has been discussed.

II. RELATED WORK

Over the past thirty years, machine translation and assessment metrics have drawn a lot of attention from researchers. As part of this research endeavor, papers and conferences that were released between 1993 and 2024 are analyzed using the VOS viewer application. By searching for "machine translation" and "evaluation metrics" together, 5871 papers in the (.CSV) file type were found in the Scopus database. This study work aims to further MT assessment research by carefully analyzing several lexical automatic evaluation measures and their performance on various translation tasks.

A. Co-occurrence Keywords Network Visualization

This analysis considers keywords that were present in at least five of the collected documents. Out of all the keywords, 6759, only 338 were able to meet the requirements displayed in the co-occurrence network diagram to form the vital areas for evaluation metrics and machine translation. As demonstrated in Fig. 1, the co-occurrence keyword network consists of distinct keyword clusters that can be built with the VOS viewer software. In this network representation, every cluster has a different color.

B. Country-wise Research Trends

Many studies throughout the world use automatic machine translation evaluation metrics. Publications from different countries have been released within the past 30 years. Table I lists the top 25countries together with the number of MT-related research papers published there. In this table, the Serial number is denoted as Sr. No.

Table I.Publications and citations (Country-wise)

Sr. No.	Country	Documents	Citations	Total Link Strength
1	China	1535	26769	609
2	United States	1450	103282	772
3	India	495	4276	144
4	United Kingdom	495	17398	490
5	Germany	316	10842	315
6	Japan	232	2695	120
7	Canada	231	19976	171
8	France	194	4363	214
9	Australia	191	7152	198
10	Spain	180	2720	170
11	Hong Kong	147	3505	123
12	Netherlands	147	6939	202
13	Singapore	134	5672	142
14	Ireland	128	1684	124
15	Italy	125	3324	145
16	South Korea	114	1770	42
17	Switzerland	93	1880	136
18	Czech Republic	66	1052	100
19	Saudi Arabia	66	1014	86
20	Taiwan	58	1280	43
21	Turkey	52	972	47
22	Israel	50	1131	50
23	Portugal	43	370	52
24	United Arab Emirates	42	430	49
25	Brazil	40	286	33

Most of the Researchers do their research work globally in implementing various applications of machine translation and evaluation metrics. Based on the Scopus database, Fig. 2 depicts the network visualization for country-specific research publications and citations. As we can see, compared to other countries, the top five countries—China, the United States, India, the United Kingdom, and Germany—have strong relationships. China has 1535 documents and 26769 citations which arethe highest, United States has 1450 documents and 495 documents for both India and the United Kingdom respectively. Thus, compared to the other countries, their clustering is the highest [6].

In conclusion, we will discuss our findings and offer suggestions for choosing the best metric for a given translation work.

III. LEXICAL-BASED EVALUATION METRICS

Large language models and multilingual models have led to a rise in the development and use of MT systems. Nonetheless, assessing these systems continues to be a major difficulty, especially for low-resource languages like those spoken in the Indian subcontinent. The development of evaluation metrics that are uniquely suited to Indian languages is required due to their distinct linguistic characteristics, which include morphological richness and various degrees of word order [7]. This is especially important because of the quick development of MT systems, which necessitates the use of thorough and precise evaluation techniques to guarantee the relevance and quality of translations. A few popular lexical-based assessment metrics are as follows:

A. BLEU (Bilingual Evaluation Understudy)

An MT output measure called BLEU counts the amount of n-grams that the generated translation overlaps with one or more reference translations. Next, based on how closely the machine translation matches the reference translations, it offers a score. This evaluation metric is used to evaluate the quality of MT output [8]. A significant issue emerges when BLEU computes the same adjusted precision measure using n-grams [9]. Another problem is that the BLEU scores also tend to favor short translations, which result in high precision ratings, even when adjusted precision is applied. The weight assigned to longer n-grams increases with the number of n-gram sizes when the adjusted precision scores are combined.The final BLEU score additionally takes into account brevity penalties, which punish candidate translations that are shorter than reference translations.

A translation with a high morpheme usage rate may yet have a low BLEU score even when it accurately captures the meaning of the original language. Bengali sentences are also superior to Hindi sentences in terms of structure [10]. Greater scores indicate more comparable translations using fewer resources (BLEU value, 0 to 1). It is not necessarily necessary for the candidate translation to precisely match one of the reference translations to receive a score of 1, as this may not be possible or intended.

B. METEOR (Metric for Evaluation of Translation with Explicit Ordering)

By taking synonyms, paraphrasing, and n-gram overlap into account, the METEOR metric—which expands on BLEU—has been used to assess machine translations [11–12]. Recall is given more weight by computing the harmonic mean of unigram recall and precision, along with an additional penalty for word order errors.For machine-generated translations, length differences with reference translations can be taken into consideration by adding an F-mean penalty to the final score. The accuracy and recall of harmonic means can be used to achieve this.

C. *NIST(NationalInstituteofStandardsandTechnology)*

NIST is a collection of guidelines and software tools for evaluating MT systems. Apart from the METEOR score, it offers further resources to assess translation quality.

NIST gives fewer common n-grams a higher weight to improve correlation with human assessment [13]. The NIST assessment measure assigns a score to the machine-generated phrases depending on how similar they are to the reference sentences. This enables systems with constrained resources to assess the performance of the MT system. It is important to remember that additional measures are available, and researchers typically combine different metrics to assess the MT system more comprehensively. The NIST assessment meter is one of them. The evaluation metric to be used will depend on the specific criteria and assessment objectives.

D. TER (TranslationErrorRate)

TER is a metric that calculates the minimum number of modifications (such as additions, substitutions, etc.) essential to translate a machine-generated translation into a human-produced reference translation [14].

E. ROUGE(Recall-Oriented Understudy for Gisting Evaluation)

ROUGE evaluates the longest common sequences and word n-gram recall by comparing machine-generated and reference text. This metric is mainly used for text summarization.

This evaluation metric gives weight to the higher proportion of N-grams. This evaluation metric has many types like ROUGE S, L, W, and U [15-16].

Table II presents a comparison of various lexical-based automation evaluation criteria utilized in MT and NLP activities. These metrics are used to evaluate the efficacy of MT systems, enabling researchers and practitioners to compare various strategies and determine which is most appropriate for their particular requirements. It demonstrates how to effectively work with Indian languages and highlights the salient characteristics of certain assessment criteria.

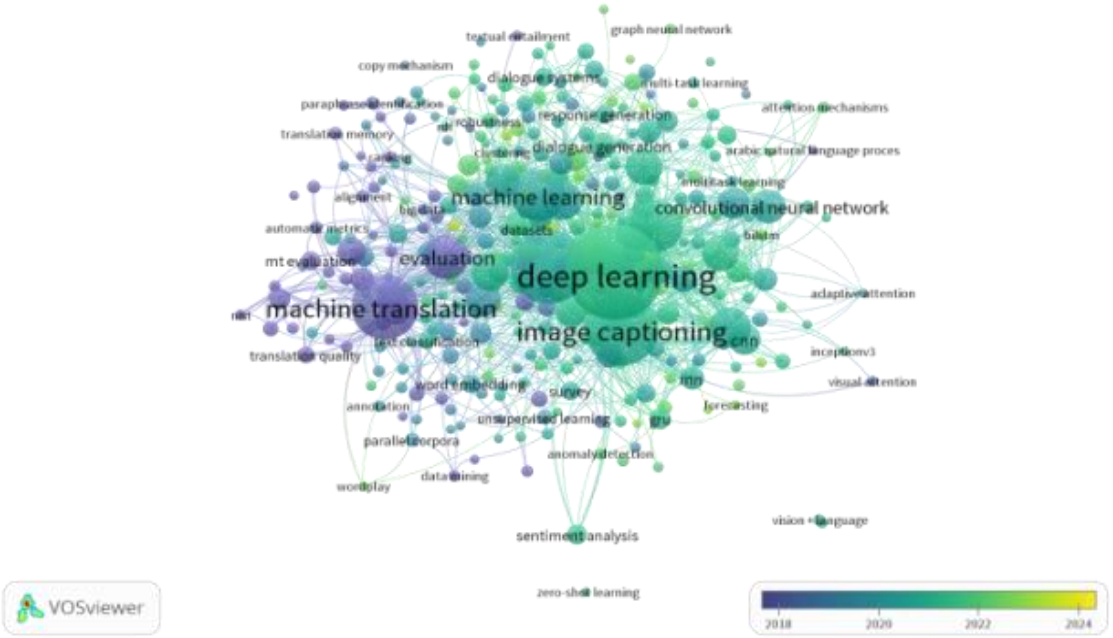

Fig 1. Co-occurrence keyword Network on Machine Translation and Evaluation Metrics

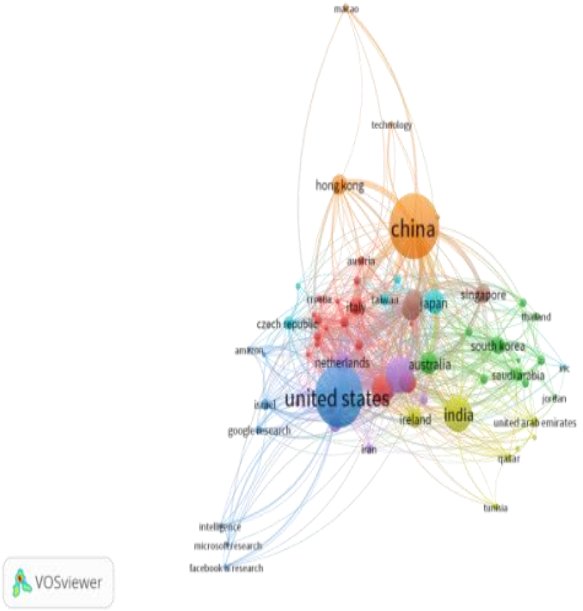

Fig 2. Country-wise Publications and citations

Table II. Comparison of key features of BLEU, TER, METEOR, and NIST evaluation metrics.

Metric	Description	Key Features
BLEU (Bilingual Evaluation Understudy)	This metric calculates the similarity between the machine-translated and the reference sentences using some factors such as fluency, integrity, and adequacy.	1) Uses n-gram co-occurrence statistics 2) Calculates unigram matches 3) Applies a brevity penalty, and aggregates precision on various n-grams.
TER (Translation Edit Rate)	This metric addresses BLEU's limitations by counting the number of modifications required to transform the translation system's output into a reference.	1) Uses a word-based system 2) Matches are counted as edits that remove a word from the system output.
METEOR (Metric for Evaluation of Translation with Explicit ORdering)	This metric is designed to improve the segment-level correlation of MT quality with human involvement.	1) Uses word-to-word matches, 2) Exact matches and flexible matches based on stemming and WordNet.
NIST (National Institute of Standards and Technology)	This metric is used to determine the quality of translated sentences using machine translation. The National Institute of Standards and Technology in the US is the source of this metric name.	1) Based on the BLEU metric, but with certain modifications. In contrast to BLEU, which only determines n-grams assigning each one an identical weight, NIST determines the level of information contained in a given n-gram. For instance, when a correct n-gram is determined, the greater its rarity, the greater the weight assigned to it. 2) In terms of how the brevity penalty is calculated, NIST differs from the BLEU in that the overall score is less affected by little variations in translation length.
ROUGE (Recall-Oriented Understudy for GistingEvaluation)	This metric is mainly used for text summarization. ROUGE evaluates the longest common sequences and word n-gram recall by comparing machine-generated and reference text.	1) A strongcorrelation with human assessments of text summarization quality. 2) Effectively manages numerous variations and references.

IV. CONCLUSION

The research on emerging trends for MT and different evaluation metrics for Indian languages has shown significant advancements and challenges. The significance of automated translation tools in a country as linguistically diverse as India has been brought to light by research on MT systems for Indian languages, particularly low-resource languages. The study highlights the necessity for enhanced and efficient translation systems that accommodate the wide range of languages used throughout the country while promoting enhanced communication and preserving cultural legacy.

A significant role of MT evaluation is assessing sentences that have been machine-translated. Many nations and academics have completed their studies in the field of MT and various evaluation measures with success in recent years. The terms "machine translation" AND "evaluation metrics" were used to gather data for this study between 1993 and 2024. This article computes many application areas, yearly

published documents, and country-wise analysis using Scopus databases. This study supports and validates the significance of evaluation tools for a range of low-resource languages. Furthermore, the study and dataset will facilitate future work in low-resource MT evaluation. Future research must extend this study to additional languages, such as low-resource languages and domains. We also wish to include pragmatic and syntactic criteria to capture the contextual and structural aspects of translation quality. We also intend to do user research and look into the metrics' correlation with other reference translations to confirm their usefulness and reliability.

REFERENCES

1. Rivera-Trigueros, I. (2022). Machine translation systems and quality assessment: a systematic review. *Language Resources and Evaluation*, *56*(2), 593-619.
2. Sai, A. B., Nagarajan, V., Dixit, T., Dabre, R., Kunchukuttan, A., Kumar, P., & Khapra, M. M. (2022). IndicMT Eval: A Dataset to Meta-Evaluate Machine Translation metrics for Indian Languages. *arXiv preprint arXiv:2212.10180*.
3. Kaur, K., & Chauhan, S. (2024). Original Research Article A comparative analysis of lexical-based automatic evaluation metrics for different Indic language pairs. Journal of Autonomous Intelligence, 7(4).
4. Khan, N. J., Anwar, W., & Durrani, N. (2017). Machine translation approaches and survey for Indian languages. *arXiv preprint arXiv:1701.04290*.
5. Mrinalini, K., Vijayalakshmi, P., & Nagarajan, T. (2022). SBSim: A Sentence-BERT Similarity-Based Evaluation Metric for Indian Language Neural Machine Translation Systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 1396-1406.
6. Wu, Y., & Li, W. (2018). Automatic audio chord recognition with MIDI-trained deep feature and BLSTM-CRF sequence decoding model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(2), 355-366.
7. Singh, S. M., & Singh, T. D. (2022). Low resource machine translation of english–manipuri: A semi-supervised approach. *Expert Systems with Applications*, *209*, 118187.
8. Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318).
9. Ananthakrishnan, R., Bhattacharyya, P., Sasikumar, M., & Shah, R. M. (2007). Some issues in automatic evaluation of english-hindi mt: more blues for bleu. *Icon*, *64*.
10. Mahapatra, S., Datta, D., Soni, S., Goswami, A., & Ghosh, S. (2023). Improving Access to Justice for the Indian Population: A Benchmark for Evaluating Translation of Legal Text to Indian Languages. *arXiv preprint arXiv:2310.09765*.
11. Denkowski, M., & Lavie, A. (2014, June). Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation* (pp. 376-380).
12. Banerjee, S., & Lavie, A. (2005, June). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization* (pp. 65-72).
13. Doddington, G. (2002, March). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research* (pp. 138-145).
14. Snover, M. G., Madnani, N., Dorr, B., & Schwartz, R. (2009). Ter-plus: paraphrase, semantic, and alignment enhancements to translation edit rate. *Machine Translation*, *23*, 117-127.
15. Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74-81).
16. Lin, C. Y., & Och, F. J. (2004, June). Looking for a few good metrics: ROUGE and its evaluation. In *Ntcir workshop*.