Journal of Autonomous Intelligence (2024) Volume 7 Issue 4
doi: 10.32629/jai.v7i4.1393

ORIGINAL RESEARCH ARTICLE

A comparative analysis of lexical-based automatic evaluation metrics for different Indic language pairs

Kiranjeet Kaur[1,2,*], Shweta Chauhan[1,2,3]

[1] *Department of Computer Science and Engineering, Chandigarh University, Mohali 140413, India*
[2] *University Centre for Research & Development, Chandigarh University, Mohali 140413, India*
[3] *Apex Institute of Technology, Chandigarh University, Mohali 140413, India*
*** Corresponding author:** Kiranjeet Kaur, kiranjeet.cse@gmail.com

ABSTRACT

With the rise of machine translation systems, it has become essential to evaluate the quality of translations produced by these systems. However, the existing evaluation metrics designed for English and other European languages may not always be suitable or apply to other Indic languages due to their complex morphology and syntax. Machine translation evaluation (MTE) is a process of assessing the quality and accuracy of the machine-translated text. MTE involves comparing the machine-translated output with the reference translation to calculate the level of similarity and correctness. Therefore, this study evaluates different metrics, namely, BLEU, METEOR, and TER to identify the most suitable evaluation metric for Indic languages. The study uses datasets for Indic languages and evaluates the metrics on various translation systems. The study contributes to the field of MT by providing insights into suitable evaluation metrics for Indic languages. This research paper aims to study and compare several lexical automatic machine translation evaluation metrics for Indic languages. For this research analysis, we have selected five language pairs of parallel corpora from the low-resource domain, such as English–Hindi, English-Punjabi, English-Gujarati, English-Marathi, and English-Bengali. All these languages belong to the Indo-Aryan language family and are resource-poor. A comparison of the state of art MT is presented and shows which translator works better on these language pairs. For this research work, the natural language toolkit tokenizers are used to assess the analysis of the experimental results. These results have been performed by taking two different datasets for all these language pairs using fully automatic MT evaluation metrics. The research study explores the effectiveness of these metrics in assessing the quality of machine translations between various Indic languages. Additionally, this dataset and analysis will make it easier to do future research in Indian MT evaluation.

Keywords: automatic machine evaluation; evaluation metrics; Indic languages; machine translation; natural language processing

ARTICLE INFO

Received: 27 October 2023
Accepted: 1 December 2023
Available online: 2 February 2024

1. Introduction

Machine Translation (MT) has become an essential and useful tool for communication in the increasingly globalized world. Communication among people plays a major role in the prosperity and advancement of the community, and MT is a very useful and efficient tool to break language barriers. However, it is a very difficult task to evaluate the quality of the MT. A machine translation system (MTS)[1] automatically translates the text from one language (source language) to another (target language) without the need for any human involvement. Traditionally, human experts have been used to assess the quality of translations. However, this approach is very time-consuming, expensive, and often subjective.

MT is an automated translation process that translates text from one language into another with the same meaning and similar construction. It can alternatively be described as an automated system that, preferably without any human involvement, analyzes text data from a source language (SL), performs various computational activities on that input, and then produces the original text data in the required target language (TL)[2].

To check the performance of MT; there are two main approaches, human evaluation, and automatic evaluation technique. Human evaluations are the gold standard but very time-consuming[3–5]. Evaluation of any MTS is the main step to improve the accuracy of these systems. This study evaluates the translation quality of the MTS. In MT research work, evaluation requires human intervention which is very time-consuming and expensive. This research analysis presents different automatic evaluations for the MT system[6]. Here, some of the key differences between these two are shown in **Table 1**:

Table 1. Key differences between human evaluation and automatic evaluation.

	Human evaluation	**Automatic evaluation**
Subjectivity	Involves subjective assessment by human evaluators	Relies on objective metrics
Linguistic nuances	Efficient in capturing linguistic nuances	May overlook subtle language and cultural references
Context sensitivity	Can disambiguate based on contextual understanding	Focuses more on lexical and syntactic similarities
Cost and time	Expensive and time-consuming	Efficient and cost-effective
Scalability	Limited	Highly scalable
Semantic accuracy	Better at assessing semantic accuracy	Focused on lexical and grammatical correctness
Continuous improvement	Provides feedback for system refinement	Allows for iterative improvement over time

Table 1 summarizes the key differences between human evaluation and automatic evaluation for Indian languages. So automatic evaluation metrics have been widely used like lexical, semantic, syntactic, character-level metrics, etc.

In simple words, human evaluation offers a more subjective and nuanced assessment, considering the linguistic and cultural intricacies of Indian languages. However, it is resource-intensive and less scalable. Automatic evaluation provides quick and objective evaluation at scale but may not fully capture the subtleties of Indian languages. A combination of both approaches can yield a comprehensive evaluation and help guide the development of effective MT systems for Indian languages.

MT is a challenging problem, especially for Indic languages, which are morphologically rich and have a low availability of parallel corpora. Therefore, it is important to have reliable and robust methods for evaluating the quality of MT systems for Indic languages.

To address this issue, automatic evaluation metrics have been developed to assess the quality of machine translations. These metrics are based on various criteria such as fluency, adequacy, and accuracy. However, different metrics may provide different results, and it is important to understand their strengths and limitations to select the most appropriate metric for a particular application.

One of the most widely used methods for Machine Translation Evaluation (MTE) is for comparing the output of a system with one or more reference translations using automatic metrics. But each metric has its limitations, such as relying on exact word matching; avoiding semantic similarity, and being sensitive to word-order variations. Moreover, these metrics may not capture the linguistic diversity and complexity of Indic languages. Automatic evaluation metrics are expected to be easy to compute and should mimic human evaluation[7].

There are several approaches to MT, including rule-based machine translation, statistical machine translation, and neural machine translation.

1) Rule-based machine translation (RBMT) uses a pre-defined set of grammatical and syntactic rules, for translating the text from a SL to a TL. This approach mainly uses some linguistic rules and dictionaries for generating translations based on established grammatical and syntactic rules and structures of different languages[8]. It is a very time-consuming and challenging task.
2) Statistical machine translation (SMT) uses statistical models to determine the most likely translation of a given text based on the vast amounts of bilingual data. It works well for identifying the patterns and probabilities for accurate translation with more training data and also for handling diverse language pairs[9–11].
3) Neural machine translation (NMT) is a most recent approach that makes use of artificial neural networks (ANNs) to discover the relationships between words and phrases in various languages, allowing for more accurate and natural translations[9–12]. NMT utilizes deep learning models, specifically sequence-to-sequence or transformer models that are used to learn translation patterns from training data.

MT systems are becoming increasingly important for Indian languages because they help to break down language barriers and enhance communication between various linguistic communities.

In this research paper, the main aim is to study and compare the different Lexical Automatic Machine Translation (LAMT) evaluation metrics for different language pairs. This research study explores the characteristics and performance of various metrics such as BLEU[13,14], METEOR[15,16], and TER[17]. In this study, we will also discuss their advantages and limitations and provide insights into their suitability for different translation tasks. This research paper aims to contribute to the advancement of MTE research by providing a comprehensive analysis of different lexical automatic evaluation metrics and their performance on various translation tasks.

1.1. Indic languages

MT for Indic languages has been gaining popularity in recent years due to the increasing demand for localization of digital content. In India, different states have different regional languages. In India, there are approximately 122 major languages that are spoken by people of different regions and communities daily. In India, only 22 languages are scheduled as official and many other languages are unofficial in use. Because of the different cultures and different regions in India, there is a great need for Inter-language translation for the transfer of ideas and sharing of any meaningful information. One of these languages is English, which is used as an associate official language along with Hindi. Despite the large number of native speakers in the majority of Indian languages, there are still not enough resources for language processing. The language of the vast majority of the resources is English[17].

These languages can be translated using some automated translation systems. to overcome the communication barriers. Indian languages are categorized into five different language families: Indo-Aryan (Hindi, Punjabi, Gujarati, Marathi, and Bengali), Dravidian (Tamil, Telugu, Malayalam, and Kannada), Austro-Asiatic (members include Khasi and Munda), and Sino-Tibetan (members include Manipuri and Bodo)[18] as shown in **Figure 1**.

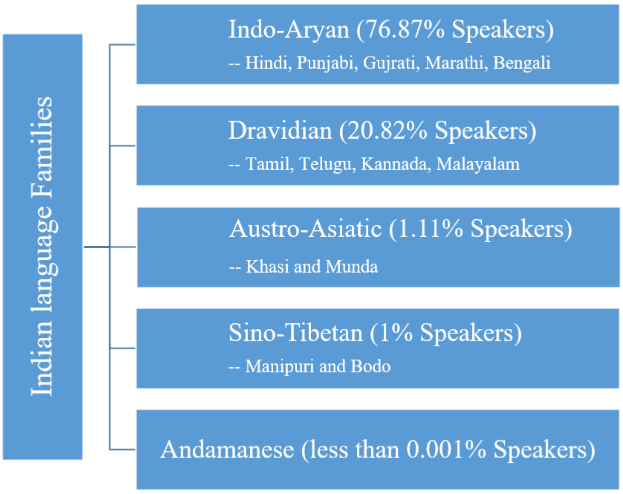
Indian language Families
Indo-Aryan (76.87% Speakers)
-- Hindi, Punjabi, Gujrati, Marathi, Bengali
Dravidian (20.82% Speakers)
-- Tamil, Telugu, Kannada, Malayalam
Austro-Asiatic (1.11% Speakers)
-- Khasi and Munda
Sino-Tibetan (1% Speakers)
-- Manipuri and Bodo
Andamanese (less than 0.001% Speakers)

Figure 1. Indian language families.

Most Indian languages have low resources and become a difficult task to develop MT systems for Indian languages[18]. There are numerous challenges that are associated with the MT of Indic languages, including variations in grammar and syntax, complex inflectional morphology, and a lack of parallel corpora, which can make it difficult to create accurate reference translations. Additionally, there may be differences in the level of fluency among human evaluators, which can lead to inconsistencies in evaluation scores. Indian languages have always suffered from the lack of annotated corpora in the past few years for different language processing tasks including MT which uses statistical models or different learning techniques. So this research study is very useful for investigating how to effectively use the available data that could be monolingual, noisy, and partially aligned[19].

Despite these challenges, there have been several advancements in MT for Indic languages. NMT has shown promising results for Indic languages, as it can capture complex language structures and improve translation quality[12]. Additionally, the development of parallel corpora and machine learning (ML) algorithms has facilitated the creation of more accurate and efficient MT systems. The Indian government also grants the status of classical language to six languages that have a long and rich literary tradition. These are Sanskrit, Tamil, Telugu, Kannada, Malayalam, and Odia.

Indic languages have a diverse and complex history and culture. They have developed various writing systems, such as Devanagari, Bengali-Assamese, Gurmukhi, Gujarati, Oriya, Sinhala script, Tamil, Telugu, Kannada, and Malayalam script. They have also produced many literary works of poetry, drama, epics, philosophy, and religion. Hindi, a mid-resource language with a huge amount of parallel resources in India, and Bengali, the second most spoken language can be classified as low-resource languages because they contain even fewer parallel resources.

For this research study, 5 Indian languages: Hindi, Punjabi, Gujarati, Marathi, and Bengali are used. Using these languages, we aim to provide a comprehensive idea of how different evaluation metrics will perform when used to evaluate Indic language as shown in **Figure 2**.

The following steps are used in MT for Indic languages:

1) Text pre-processing

The first step is pre-processing which is used to remove any irrelevant information, for instance, formatting or meta-data. And converted into a suitable format for input into the MT system. This step includes tokenization, sentence splitting into words, and normalization of these data.

2) Build translation model

The translation model is built using parallel corpora of both languages (source and target). It involves training the MT algorithm on a huge dataset of aligned parallel sentences that are very useful for learning the patterns and relationships among these languages.

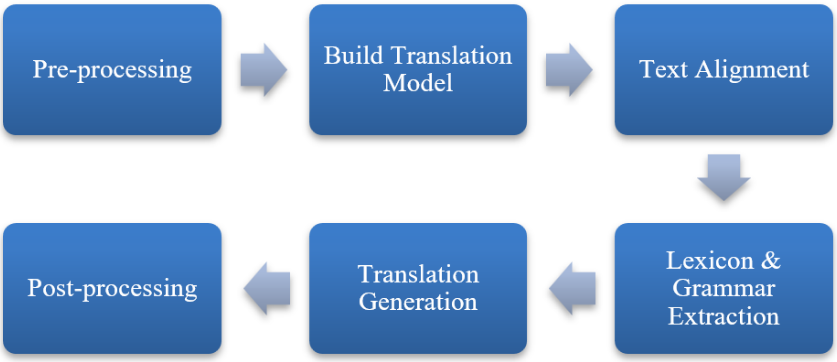

Figure 2. Machine translation for Indic languages.

3) Text alignment

In this step, the source and target texts are aligned to identify corresponding sentences, words, and phrases. It is very necessary for accurate translation, as it enables the MT system to predict the correct translations for each word or phrase.

4) Lexicon and grammar extraction

The MT system extracts a lexicon and grammar rules for each language pair. It involves identifying the grammatical structures and other linguistic features of each language.

5) Translation generation

The MT system generates translations for the source text using the translation model, lexicon, and grammar rules. The output is typically an initial translation that may require some post-processing.

6) Post-processing

The translated text is post-processed to ensure that it is fluent and reads naturally in the target language.

As the demand for localization of Indic language content continues to grow, the development of high-quality MT systems for Indic languages will become increasingly important.

1.2. Machine Translation Evaluation (MTE) Methods

The main role of MTE methods in the MT model development cycle is to assess and improve the quality of the machine-generated translations. MTE methods provide feedback to MT developers on how well their MT systems are performing for their intended use case. This information can be used to refine and improve the MT model, leading to better translation quality.

MTE methods are typically used at different stages of the MT model development cycle. In the initial stages, MTE methods are used to compare different MT systems and evaluate their performance against a reference translation. This helps developers choose the best MT system for their use case and identify areas that need improvement.

Once an MT system has been selected, MTE methods can be used in later stages of the development cycle to evaluate the quality of the translations that can be generated by the system. This helps developers identify issues such as errors in grammar, vocabulary, or syntax, as well as cultural and linguistic nuances that may require to be addressed in the MT model.

MTE methods play a crucial role in benchmarking the performance of MT systems against human translators. By comparing the output of the MT system with that of a human translator, developers can identify areas where the MT system needs improvement, and develop new techniques and models to improve the quality. Techniques for evaluating MT output are essential for determining its level of quality. The result of MT is evaluated to assess translation quality and pinpoint areas for improvement. Several MT evaluation techniques are available, including automatic, and human evaluation. Human evaluation entails having translators rate the output of the MT using multiple criteria, including fluency, sufficiency, and correctness. On the other hand, the automatic evaluation uses metrics like BLEU, METEOR, and TER to gauge how well the output of MT is done. These metrics evaluate the MT's output with the original text and give a score based on some factors, including grammatical correctness, sentence structure, and word overlap.

The most popular metrics used for evaluating MTS performance are BLEU, TER, and METEOR. BLEU measures the similarity of the machine-translated output to one or more reference translations[13]. TER measures the number of edits needed to change a machine-translated sentence to a reference sentence. METEOR combines precision, recall, and alignment-based metrics[14,15].

Human evaluation involves having human evaluators compare machine-generated translations to human-generated translations and provide feedback on the quality of the translation.

The quality of MT output is manually assessed using a combination of human and automatic evaluation methods. This method combines the advantages of human and MTE techniques, resulting in more precise and trustworthy outcomes as well as an evaluation of the caliber of translations produced by MT systems. The role of MT evaluation methods in MT evaluation is to provide feedback and guidance for MT developers, users, and researchers. MT evaluation methods can help to identify the advantages and limitations of different MT systems, compare and rank MT systems according to various criteria, monitor and improve the quality of MT outputs over time, and explore the impact of MT on various domains and applications. MT evaluation methods can also help to advance the scientific understanding of MT by providing empirical evidence and insights into the linguistic, cognitive, and social aspects of MT.

However, MTE methods also face several challenges and limitations. For example, human evaluation is costly, time-consuming, subjective, and inconsistent. Automatic evaluation is fast, cheap, objective, and consistent, but it may not capture the nuances and complexities of natural language and human communication. Moreover, different MTE methods may have different assumptions, objectives, and perspectives, which may lead to conflicting or incomparable results. Therefore, it is important to select appropriate MTE methods for different purposes and contexts and to combine multiple MTE methods to obtain a comprehensive and reliable assessment of MT quality and performance.

1.3. Problem statement

Machine translation (MT) is a critical component of several NLP applications, ranging from web translation services to voice-based applications. However, the evaluation of MT systems is an extremely difficult task. In India, different regions have different languages. Different languages have different grammatical structures and vocabulary of words, and a system that works well for one language pair might not work as well for other language pairs. So for this research analysis, we have used five Indic language pairs, all of which belong to the Indo-Aryan language family, and analyze the performance of different lexical-based machine translation evaluation metrics, such as BLEU, METEOR, and TER. Furthermore, the quality of MT can vary greatly depending on the specific system used, the architecture of the system, and the metrics used for evaluation. The need for a comparative analysis of lexical-based automatic evaluation metrics for different Indic language pairs is also evident, given the unique linguistic characteristics of these languages.

In this context, the problem is to conduct a comparative study of lexical-based automatic evaluation metrics for the different language pairs using different translation systems. The major goal is to comprehend the performance of these metrics for evaluating the quality of translations between each language pair. This will assist in identifying the evaluation metrics that work well for every language pair and may help guide the development of more precise and trustworthy machine translation systems.

In this paper, section 2 provides an overview of the different types of automatic evaluation metrics and their main features. We will analyze the strengths and weaknesses of each metric, including their sensitivity to different types of errors, their ability to capture various aspects of translation quality, and their robustness across different languages and domains. In section 3, the overview of the different renowned translators namely, Google, BING, Yandex, and ImTranslator are presented and shows which translator works better on these language pairs.

Next, we will conduct experiments to evaluate the performance of the selected metrics on a set of translation tasks by using the natural language toolkit (NLTK) tokenizers. We will use different evaluation datasets and compare the results obtained using each metric. We will also investigate the correlations between the metrics and machine-generated translation systems to assess their reliability and validity. These results are analyzed using fully automatic MT evaluation metrics for all language pairs.

Finally, we will conclude and provide recommendations for selecting the most appropriate metric for a given translation task based on our findings. Among all lexical-based evaluation metrics, widely used like BLEU, TER, and METEOR in which BLEU is the most widely used because its language is independent. We depict that apart from the major disadvantage of BLEU it is still widely used because of its language-independent nature and ease of implementation. The evaluation results depict an improved performance in the case of the proposed score. We will also discuss future research directions and potential improvements for automatic MTE metrics.

2. Lexical based evaluation metrics

Lexical-based evaluation metrics are commonly used to evaluate the quality of MT outputs based on lexical similarities between the candidate and reference translations. Here are a few widely used lexical-based evaluation metrics:

2.1. BLEU (Bilingual Evaluation Understudy)

BLEU is the most widely used metric that calculates the precision of n-gram overlap between the machine-generated and the reference translations and then gives a score based on how effectively the MT replicates the reference translations. This evaluation metric is used for measuring the quality of MT output[13]. It calculates the similarity between a candidate translation and one or more reference translations based on n-gram matches.

BLEU compares the contiguous sequences of words (i.e., n-grams) in the candidate translation to those in the reference translations. It takes into account both precision (how many n-grams in the candidate translation matched with the reference translation) and brevity penalty (to avoid favoring overly short translations).

The BLEU score ranges from 0 to 1, where 1 indicates a perfect match between the candidate and reference translations. However, in practice, it is difficult to achieve a BLEU score of 1, and a score of 0.5 or higher is generally considered to be a good indicator of translation quality. So, it’s important to note that BLEU is just one of many metrics used to evaluate MT, but it has its own limitations. While it provides a helpful quantitative measure, it does not capture aspects such as fluency, coherence, or overall comprehension.

The BLEU metric can be applied to evaluate MT output for Indic languages, just like any other language. However, there are a few considerations specific to Indic languages that the BLEU metric can be used for evaluating MT performance in Indic languages as well.

Indic languages may have longer n-grams compared to languages like English, which may affect the choice of n-gram order for calculating BLEU scores. Higher-order n-grams (e.g., 5-gram or 6-gram) might be more suitable to capture the linguistic nuances of Indic languages.

While BLEU can be used as a starting point for evaluating MT quality in Indic languages, it is important to consider domain-specific characteristics and cultural nuances that may impact the evaluation. Additionally, alternative metrics or language-specific evaluation methods may also be appropriate for evaluating Indic language translations. However, it is necessary to note that the availability and quality of linguistic resources (such as parallel corpora) for Indian languages may vary, which can impact the performance of these metrics.

However, the effectiveness of the BLEU metric for Indic languages can vary depending on the specific language and domain of translation. For example, some studies have suggested that BLEU may not be the most effective metric for evaluating the quality of translations in languages with complex morphology, due to the difficulty in accurately capturing the nuances of the language. Quality is considered to be the correspondence between a machine's output and that of a human: "The closer an MT is to a professional human translation, the better it is"[13]. BLEU was one of the first metrics to claim a high correlation with human judgments of quality[16] and remains one of the most popular automated and inexpensive metrics.

The BLEU score computes the degree of overlap between the machine translations and the reference translations at the n-gram level. The higher the BLEU score, the closer the machine-generated translation is to the reference translations in terms of the n-gram overlap.

While the BLEU score is a widely used metric, it is not without limitations. For example, it does not take into account the semantic or syntactic accuracy of the translations and can be biased toward shorter sentence lengths. It is important to use BLEU scores in conjunction with other evaluation metrics and human evaluation to get a more accurate assessment of the quality of machine-generated translations.

BLEU works by contrasting the candidate translation's n-grams (sequences of n words) with those of the reference translations. BLEU calculates a modified precision score for each n-gram size, which is typically between 1 and 4, i.e., the ratio of matching n-grams to the total number of n-grams in the candidate translation. However, this precision score can be biased by repeating words or phrases in the candidate translation that are not in the reference translations. To avoid this, BLEU uses a clipping function that limits the number of times an n-gram can be counted based on its maximum frequency in any reference translation.

Computing the same modified precision metric using n-grams is the major issue[14] in BLEU. Another problem with BLEU scores is that they frequently favor short translations, which may produce very high precision even when modified precision is used. The modified precision scores for different n-gram sizes are then combined using a weighted geometric mean, which gives more weight to longer n-grams. A brevity penalty (BP), which penalizes the candidate translations that are shorter than the reference translations, is also included in the final BLEU score. The ratio of the candidate translation length to the actual reference translation length is used to calculate the BP, which is usually the closest length to the candidate translation among all reference translations. However, in some versions of BLEU, such as NIST, the shortest reference translation length is used instead.

We compute the brevity penalty (BP),

$$\mathrm{BP} = \left\{ \begin{matrix} 1 & \quad \text{if } c > r \\ e^{(1-\frac{r}{c})} & \quad \text{if } c \leq r \end{matrix} \right\} \tag{1}$$

Then,

$$\mathrm{BLEU} = \mathrm{BP}.\exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{2}$$

The BLEU score ranges from 0 to 1, with higher scores indicating more similar translations. However, it is not necessary to achieve a score of 1, as it implies that the candidate translation is identical to one of the reference translations, which may not be possible or desirable. Moreover, adding more reference translations can increase the BLEU score, as there are more opportunities for matching n-grams.

BLEU has some limitations and challenges as a metric for evaluating MT quality. For instance, it does not account for grammatical correctness, semantic adequacy, or stylistic variation. It also assumes that there is a single best translation for each source sentence, which may not be true in practice. Furthermore, it relies on exact word matching, which can miss synonyms, paraphrases, or other linguistic variations that convey the same meaning. Additionally, it may not correlate well with human judgments at the sentence level, as humans may consider other factors besides lexical similarity.

Despite these drawbacks, BLEU is most widely used as a fast and easy method to compare different MT systems or approaches. It can also provide feedback for improving MT models or identifying errors. However, it should not be used as the sole criterion for assessing translation quality, and it should be complemented by other metrics and human evaluations. By averaging out individual sentence judgment errors throughout a test corpus rather than attempting to determine the exact human judgment for each sentence, the BLEU metric achieves excellent correlation with human judgments: quantity leads to quality[6].

Overall, while BLEU can be an effective evaluation metric for Indic languages, it is necessary to consider the specific characteristics of the language and the domain of translation and to use modified versions of BLEU or other evaluation metrics if necessary to accurately assess the quality of machine-generated translations.

Figure 3 shows the issue of the BLEU evaluation score in terms of divergences. The inductive translation has a different word order.

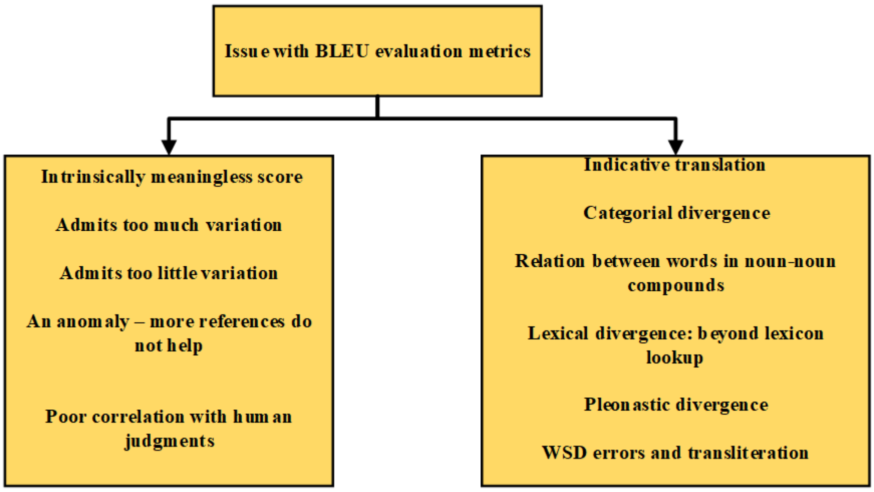
Issue with BLEU evaluation metrics
Intrinsically meaningless score
Admits too much variation
Admits too little variation
An anomaly – more references do not help
Poor correlation with human judgments
Indicative translation
Categorial divergence
Relation between words in noun-noun compounds
Lexical divergence: beyond lexicon lookup
Pleonastic divergence
WSD errors and transliteration

Figure 3. Issues with BLEU evaluation metrics.

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

This metric uses a combination of n-gram overlap, synonymy, and paraphrasing to evaluate machine translations[15,16]. METEOR computes a harmonic mean of unigram precision and recall along with various matching features like stemming synonymy, and word order. It considers more linguistic features than BLEU. To compute the METEOR score, the machine-generated translations, and reference translations are tokenized and stemmed to eliminate inflections and variations. The accuracy and recall of the alignments are then determined based on the alignment between the outputs of the machine translations and the reference translations. So the final score can be determined by taking the harmonic mean of accuracy and recall and applying an F-mean penalty to account for length disparities between the outputs of machine-generated translations and the reference translations. The METEOR is an MT evaluation metric that aims to measure the quality of MT results in a way that aligns with human judgments of translation quality[16]. This is accomplished by comparing the machine translation's output to one or more reference translations and assessing the output's quality using a mix of precision, recall, and alignment errors.

METEOR takes into account synonyms, paraphrases, and word order in addition to exact word matches, unlike other MT assessment metrics that emphasize word matching. To find semantic distinctions between words and to take into account variations in word order, this is accomplished by using multiple linguistic resources such as WordNet and synonym sets.

METEOR Score is calculated by:

$$\mathrm{Precision} = \frac{\mathrm{No.\,of\ matching\ unigrams}}{\mathrm{Total\ no.\,of\ unigrams\ in\ Hypothesis}} \tag{3}$$

$$\mathrm{Recall} = \frac{\mathrm{No.\,of\ matching\ unigrams}}{\mathrm{Total\ no.\,of\ unigrams\ in\ reference}} \tag{4}$$

$$\mathrm{F-Score} = \frac{10 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Recall} + 9 \times \mathrm{Precision}} \tag{5}$$

$$\mathrm{Penalty} = 0.5 \times \left[\left(\frac{\mathrm{No.\,of\ Chunks}}{\mathrm{No.\,of\ matched\ unigrams}}\right)^3\right] \tag{6}$$

$$\mathrm{METEOR\ Score} = \mathrm{F-Score} \times (1 - \mathrm{Penalty}) \tag{7}$$

To compute the METEOR score, the machine translation's output and reference translations are tokenized and stemmed to eliminate inflections and variations. The accuracy and recall of the alignments are then determined based on the alignment between the machine translation's output and the reference translations. The final score is determined by taking the harmonic mean of accuracy and recall and applying an F-mean penalty to account for length disparities between the machine translation's output and the reference translations.

2.3. TER (Translation Error Rate)

TER is another metric that measures the number of edits required to transform the candidate translation into the reference translation. It considers deletions, insertions, and substitutions at the word or phrase level[17]. TER as an alternative to the widely used BLEU metric. This is a metric commonly used in NLP to evaluate the quality of MT output. TER is a distance-based metric, which means it calculates the edit distance between the machine-generated translation and the reference translation. The edit distance is calculated by counting the number of operations that are essential to transform the machine-generated translation into the reference translation. These operations can be insertion, deletion, substitution, or reordering of words. So when the TER score is lower, the better the machine translations.

TER score is calculated by,

$$\mathrm{TER} = \frac{\mathrm{Minimum\ No.\,of\ edits}}{\mathrm{Average\ No.\,of\ reference\ words}} \tag{8}$$

TER was first introduced by Snover et al. in 2006 who proposed TER as an alternative to the widely used BLEU metric. Snover et al. demonstrated that TER had a stronger correlation with human judgments of translation quality than BLEU[17].

Since then, TER has been used by many researchers for evaluating the quality of MT output. These evaluation metrics will take the Machine-translated test sentences and the sentences translated by the Human language expert to give output. These outputs will be in the range of 0–1, so 0 presents the worst translations and 1 is the best possible translation output.

These metrics provide quantitative measures to evaluate the performance of machine translations. However, it needs to be noted that these metrics have some limitations and it's not fully capture the nuances and fluency of the translated text. Therefore, it is often recommended to complement these metrics with human evaluation and consider other factors such as grammar, syntax, and overall coherence to obtain a comprehensive evaluation of MT quality.

3. Methods

3.1. Machine translation evaluation process

There are the most important steps that we will follow to calculate and compare evaluation metrics for Indian languages as shown in **Figure 4**. Firstly, we will identify the lexical-based evaluation metrics that are commonly used in MT or NLP tasks. These evaluation metrics include BLEU, METEOR, and TER and then collect an appropriate dataset for evaluation. For this research work, we will use two test datasets dataset 1 and dataset 2 consisting of 100 and 150 sentences respectively, and translate these datasets by using 4 different MT systems for 5 Indian languages. The next step is Pre-processing these test datasets and then preparing them for evaluation. It may involve aligning the reference translations with the machine-generated translations and cleaning the data to remove any inconsistencies or errors. Calculating the different evaluation metrics for each MT system and all language pairs using the prepared dataset. We can use the existing libraries or implement the metrics. For this research analysis, we will use the nltk.translate module in Python to calculate scores for evaluation metrics. After that compare the results of different evaluation metrics for the same language pair and MT system. And also look for the correlations between the automatic metrics and human scores, as well as differences in performance between metrics. Finally, analyze the strengths and limitations of the evaluation metrics in capturing the quality of MT in Indian languages.

A basic flowchart for the MT evaluation process:

1) To select a MTS to evaluate.
2) To collect a set of test sentences.
3) To translate the test sentences using the MT system.
4) To collect reference translations for the same test sentences.
5) To evaluate the machine-generated translations using one or more of the following evaluation methods:
    - **Automatic evaluation:** Use algorithms and metrics such as BLEU, METEOR, and TER to measure how closely the machine-generated translations relate to the reference translations.
    - **Human evaluation:** Have human evaluators compare the machine-generated translations to the reference translations and provide feedback on the quality of the translation.
6) To analyze the evaluation results and identify areas where the MT system could be improved.
7) To refine the MT system based on the evaluation results.
8) To repeat the evaluation process to assess the effectiveness of the improvements made to the system.

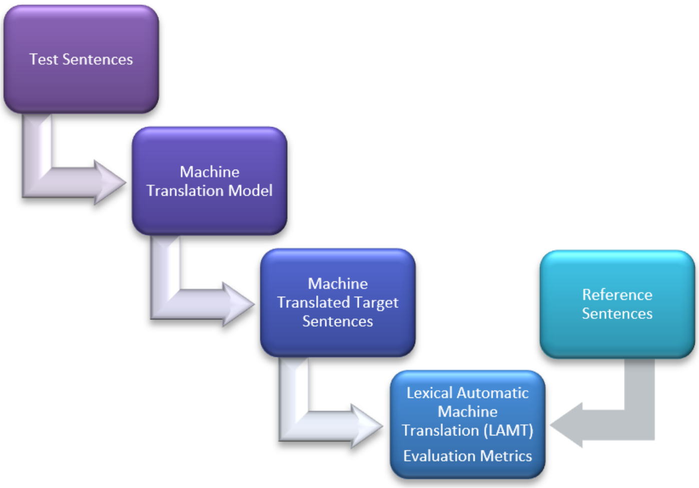
Test Sentences
Machine Translation Model
Machine Translated Target Sentences
Reference Sentences
Lexical Automatic Machine Translation (LAMT) Evaluation Metrics

Figure 4. MT evaluation process.

The above-given **Figure 4** shows the different lexical automatic machine translation (LAMT) evaluation metrics in the MT model. For evaluation, this **Figure 4** also has reference sentences and applied the test sentences and the machine-translated sentences. It shows different evaluation metrics namely, BLEU, METEOR, and TER. These evaluation metrics are fully automatic and evaluate the Machine-translated sentences by multiple reference sentences. Overall, the main task of MT evaluation is to ensure that machine-generated translations are accurate, fluent, and culturally appropriate.

3.2. Machine Translators

For this research analysis, we have used various famous translators that support Indic Language Translations and are easily available on the Internet. These translators are:

3.2.1. Google Translate

Google Translate is a powerful multi-lingual translation service developed by Google. It was launched in April 2006 as a statistical machine translation (SMT) service. Over time, it transitioned to a NMT[20,21] engine i.e., Google Neural Machine Translation (GNMT) that can translate whole sentences at a time for improved accuracy. It allows users to translate the text, handwritten text, images, and speech in over 100 languages.

Google Translate[22] supports several Indian languages (such as Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Odia, Punjabi, Tamil, Telugu, and Urdu), allowing users to translate the text to and from these languages. Google Translate is continuously improving and adding support for more languages. Recently, Google Translate has added Assamese and Bhojpuri, which are mostly spoken by millions of people in Northeast India and Northern India, respectively. It is a widely used translation tool that can be very helpful for various language-related tasks, especially when communicating with people who speak different languages. Google Translate provides a convenient way to translate text and conversation, but the main point to keep in mind is that the translations may not always be perfect and may require additional proofreading, especially for complex or context-dependent content. It's always a good practice to review the translations for accuracy.

To use Google Translate for Indian languages, we can visit the Google Translate website or the mobile app to translate text, speech, and even websites. It also offers features like camera translation and offline translation.

3.2.2. Bing Microsoft Translator

The history of Bing Translator dates back to the late 1990s when Microsoft started developing its MT system. Bing Translator[23] also known as Microsoft Translator, was launched in 2007 and has since evolved to provide multilingual MT services provided by Microsoft. The first version of Microsoft's MTS was developed between 1999 and 2000 within Microsoft Research.

It supports a wide range of Indian languages for translation. To use Bing Microsoft Translator for Indian languages, we can visit the Bing Translator website or download the Microsoft Translator app on a mobile device.

3.2.3. Yandex Translate

Yandex Translate[24] is a web service that has been provided by Yandex, a Russian multinational technology company founded in 1997. Yandex originally began as a search engine, it has since expanded its services to offer various products and services, including Yandex Translate. It is a translation service that supports different Indian languages and offers translation capabilities for text, web pages, and even text from images. The service utilizes a self-learning SMT system that has been developed by Yandex.

Yandex Translate offers translation services, but the quality and accuracy of translations can vary depending on the language pair and the complexity of the text data. It is always recommended to review and verify translations for accuracy when using any MT service.

3.2.4. ImTranslator

ImTranslator supports a variety of languages, including several Indian languages. It offers support for Hindi, Gujarati, Marathi, Kannada, Malayalam, Punjabi, Bengali, Tamil, Telugu, and Urdu. With the inclusion of these Indian languages, ImTranslator allows approximately 90% of the Indians to access information and work in their native languages.

ImTranslator[25] is a translation service that provides instant translation of texts, words, sentences, and webpages between more than 100 languages. It offers a Chrome extension, Firefox extension, and Opera extension, each with its own set of features. ImTranslator supports a wide range of languages, including Indian languages such as Bengali, Hindi, Kannada, Malayalam, Marathi, Punjabi, Tamil, Telugu, Urdu, and more. This will allow you to translate text, sentences, words, and webpages to and from Indian languages using the ImTranslator interface. ImTranslator uses multiple translation providers, including Google Translate, Bing Microsoft Translator, and Yandex Translate. These providers offer different translation algorithms and databases, which may result in variations in translation quality and accuracy. ImTranslator has been continuously updated and improved over the years, with different versions released to enhance its functionality and performance. And apply MTE metrics on the translations produced by these translators and then compare all these translators based on their output results.

3.3. Analysis by References Translation

Based on the provided reference and the hypothesis machine-translated data, the MTS will give scores on whether the hypothesis sentences align with the reference sentences in the range of 0 to 1. Here, the main criteria of the scores will be how close the hypothesis sentences match with the reference sentences. A score higher than 0.5 will indicate a high level of similarity with 1 indicating an absolute perfect translation and vice versa.

3.4. Normalization

The Normalization technique used here is Min-Max Normalization which makes use of Minimum and Maximum values from a given set of values to scale down the value to a specified range, usually between 0 and 1. With the help of scaling, we were able to improve the evaluation metrics which are somewhat sensitive to certain input features present in the dataset.

$$\mathrm{X}_{\mathrm{scaled}} = \frac{\mathrm{X} - \mathrm{X}_{\mathrm{min}}}{\mathrm{X}_{\mathrm{max}-X_{min}}} \qquad (9)$$

3.5. Pearson Correlation

Pearson Correlation Coefficient is a statistical measure of the linear relationship between two quantitative variables. It ranges between −1 to +1 with −1 indicating that there is a negative correlation, 0 indicating that there is no correlation, and +1 indicating that there is a positive correlation between the two variables.

So, this Coefficient is used to compare the results of the different Automatic MTE Metric Scores and the scores provided by the human language expert.

4. Experimental results and analysis

The experimental results are compared with Google, Microsoft BING, Yandex, and Im translators on two test datasets which are made up of 100 daily-use sentences and 150 wiki sentences. Google Translator is basically an SMT-type translator that has to be trained on a huge corpus of data for better efficiency and robustness. Bing Translator is fundamentally based on both SMT and RBMT approaches for translation. Similarly, Yandex uses SMT means that it relies on statistical data to perform the translation. Therefore, statistical learning is crucial for the most precise MT.

4.1. Dataset

Bharat Parallel Corpus Collection (BPCC) corpus of data was developed by AI4Bharat, a non-profit organization devoted to the promotion of artificial intelligence (AI) technologies for Indian languages, and is the source of the dataset that we used. The BPCC dataset by AI4Bharat is a publicly available parallel corpus that contains a combination of human-labeled datasets and automatically mined datasets. It consists of approximately 230 million bi-text pairs. BPCC includes both (existing and new data) for all 22 scheduled Indic languages. This dataset is divided into two parts: BPCC-Mined and BPCC-Human. The BPCC dataset is released under the Creative Commons CC0 license, which means it is free to use and has no rights reserved. To access and use the BPCC dataset, we can visit the AI4 Bharat website. They provide several other datasets, models, and applications for Indian languages as well[26].

Overall, the BPCC dataset is a valuable resource for NLP tasks in Indian languages, providing a large collection of parallel bi-text pairs for various languages. It can be used for MT, language generation, language understanding, and other language-related tasks in Indian languages.

The Corpus consists of a significant monolingual sentence-level corpus of Indian languages from Indo-Aryan language families, including English. However, we used only five languages that belong to the Indo-Aryan language family. We have compiled two different test datasets of 100 and 150 test sentences for respective languages. For reference translation evaluation we have taken four different machine-generated translation systems that evaluate the translation ranging from 0 to 1. 0 zero means a very poor score and 1 means a very good score. So we compare the scores given by all lexical-based evaluation metrics for evaluating the performance of MT evaluation.

4.1.1. Test Dataset 1

For Test Dataset 1, we have taken 100 test sentences that are most widely used for daily purposes. And then translated them on different machine-generated translation systems. We collect these datasets for 5 Indian languages, i.e., Hindi (hi), Punjabi (pbi), Gujarati (guj), Marathi (ma), and Bengali (ben). We have taken the test sentences from BPCC of the AI4Bharat dataset and obtained the translation outputs from 4 machine-generated translation systems for each of the 5 Indian languages.

1) **Pearson Correlation of different evaluation metrics for different language pairs with Google translate**

Pearson correlation of these metrics is represented for test dataset1, which consists of 100 sentences. In **Table 2**, BLEU has 0.142 whereas METEOR has 0.092 and TER has a 0.032 correlation for the eng-hi language pair. We can observe that flexible BLEU has maximum value rather than all other metrics for test dataset 1, which implies that the Google translation is good overall for that language pair as shown in **Figure 5**.

Table 2. Comparison of different language pairs using Pearson correlation of different evaluation metrics for Test dataset 1 with Google Translate.

Language pair	BLEU	METEOR	TER
eng-hi	0.142	0.092	0.032
eng-pbi	0.146	0.018	0.057
eng-guj	0.131	0.014	0.126
eng-ma	0.175	0.122	0.099
eng-ben	0.134	0.032	0.074

And here is a bar graph that can be generated from the above-given table data.

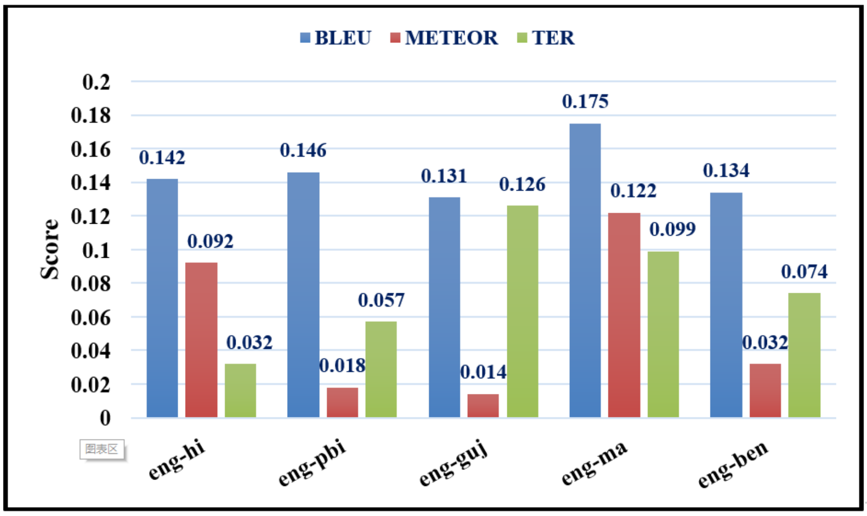

Figure 5. A bar graph comparison of different language pairs using different evaluation metrics for Test data1 with Google Translate.

2) Pearson Correlation of different evaluation metrics for different language pairs with Bing translate

The Pearson correlation for the BLEU score of eng-hi test dataset 1 for reference translation evaluation comes out to be 0.108, with METEOR 0.107, TER 0.046 for the eng-hi language pair, which implies that the correlation for BLEU, that is, 0.108 is the highest for that language pair that are shown in **Table 3**.

Table 3. Comparison of different language pairs using Pearson correlation of different evaluation metrics for Test data 1 with Bing Translate.

Language pair	BLEU	METEOR	TER
eng-hi	0.108	0.107	0.046
eng-pbi	0.111	0.078	0.007
eng-guj	0.120	0.086	0.057
eng-ma	0.145	0.111	0.106
eng-ben	0.115	0.147	0.066

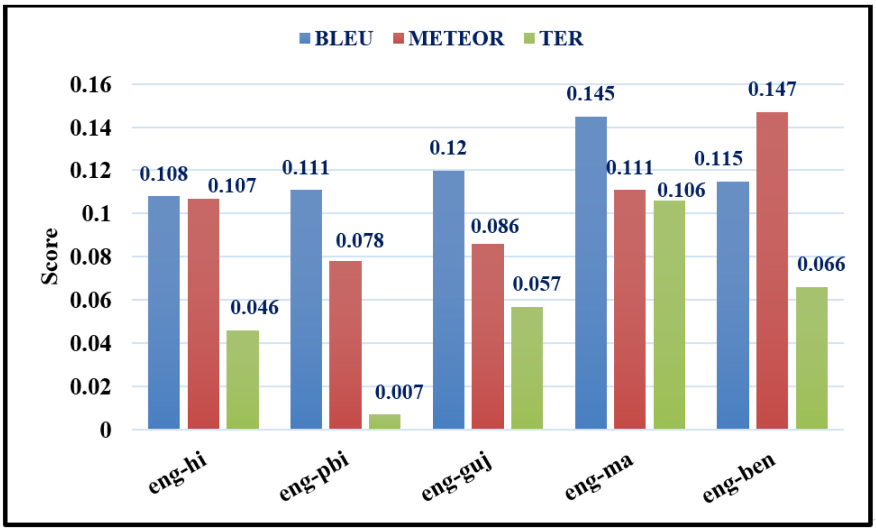

Figure 6. Comparison of different language pair using different evaluation metrics for Test data 1 with Bing Translate.

The Pearson correlation for the BLEU score of all these language pairs are higher than other evaluation metrics for test dataset 1 with Bing translator. Only eng-ben language pair, METEOR score has highest score which is 0.147 and BLEU score is 0.115 that are shown in **Figure 6**.

3) Pearson Correlation of different evaluation metrics for different language pairs with Yandex translate

Table 4. Comparison of different language pair using Pearson correlation of different evaluation metrics for Test data1 with Yandex Translate.

Language pair	BLEU	METEOR	TER
eng-hi	0.067	0.267	0.146
eng-pbi	0.094	0.155	0.14
eng-guj	0.02	0.253	0.2
eng-ma	0.128	0.182	0.06
eng-ben	0.041	0.16	0.014

Table 4 presents the comparisons of the correlation of lexical-based metrics for five Indian language pairs for Test dataset 1 with Yandex translate. For test dataset 1 the respective scores of METEOR are higher than all other metrics as shown in **Figure 7**. But **Figure 8** shows that the TER gives the better score among language pairs eng-pbi and eng-ma as compared to other language pairs, so it performs better machine translations. It has been observed that METEOR and TER have been providing better results as compared to others with Yandex and ImTranslators.

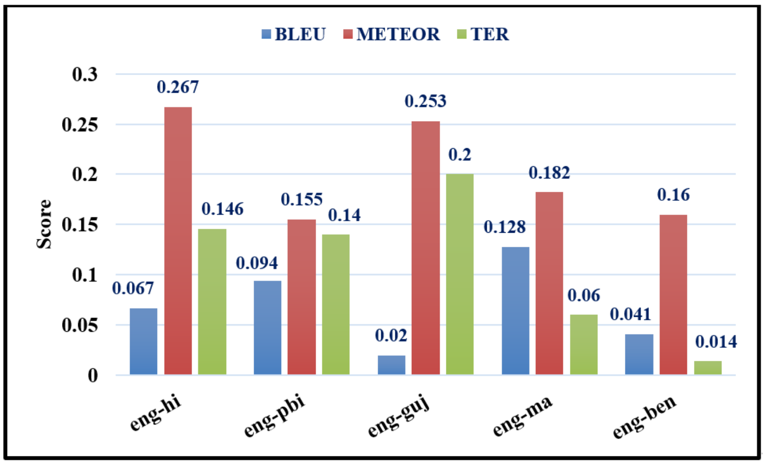

Figure 7. Comparison of different language pair using Pearson correlation of different evaluation metrics for Test data 1 with Yandex Translate.

4) Pearson Correlation of different evaluation metrics for different language pairs with ImTranslate

Table 5. Comparison of different language pair using Pearson correlation of different evaluation metrics for Test dataset 1 with ImTranslator.

Language pair	BLEU	METEOR	TER
eng-hi	0.142	0.041	0.105
eng-pbi	0.005	0.011	0.097
eng-guj	0.009	0.04	0.111
eng-ma	0.095	0.021	0.003
eng-ben	0.096	0.03	0.106

The comparisons of the correlation of lexical-based evaluation metrics for different language pairs for Test dataset 1 with ImTranslator are shown in **Table 5**.

To check the effectiveness of these translations we translated the test dataset 1 on ImTranslator, calculated Pearson correlation scores for all language pairs, and compared it with other scores which show that TER has the better score than the other metrics scores as shown in **Figure 8**.

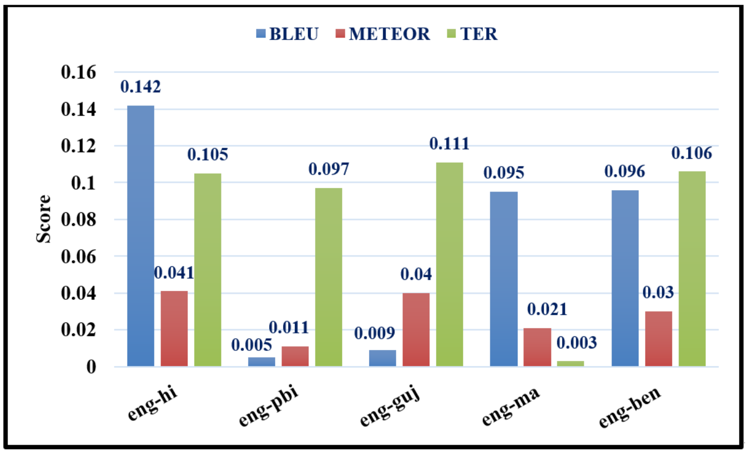

Figure 8. Comparison of different language pair using Pearson correlation of different evaluation metrics for Test data 1 with ImTranslator.

4.1.2. Test Dataset 2

For Test Dataset 2, we have taken 150 wiki sentences. And then translated them on different machine-generated translation systems. We collect these datasets for 5 Indian languages, i.e., Hindi (hi), Punjabi (pbi), Gujarati (guj), Marathi (ma), and Bengali (ben). We have also taken these test sentences from BPCC of the AI4Bharat dataset and obtained the translation outputs from 4 machine-generated translation systems for each of the 5 Indian languages.

5) Pearson Correlation of different evaluation metrics for test dataset 2 for different language pairs with Google translate

To determine the effectiveness of our translations we translated our test dataset 2 on Google Translate and calculated BLEU scores for those and compared it with other scores which are shown in **Table 6**.

Table 6. Comparison of different language pairs using Pearson correlation of different evaluation metrics for Test dataset 2 with Google Translate.

Language pair	BLEU	METEOR	TER
eng-hi	0.181	0.048	0.041
eng-pbi	0.054	0.005	0.063
eng-guj	0.129	0.019	0.075
eng-ma	0.112	0.002	0.104
eng-ben	0.062	0.041	0.057

Figure 9 shows that BLEU has a higher score than all other metrics, which implies that the Google translation is good overall for all these language pairs. For Instance, BLEU has 0.129 whereas METEOR has 0.019 and TER has 0.075 correlations for the eng-guj language pair.

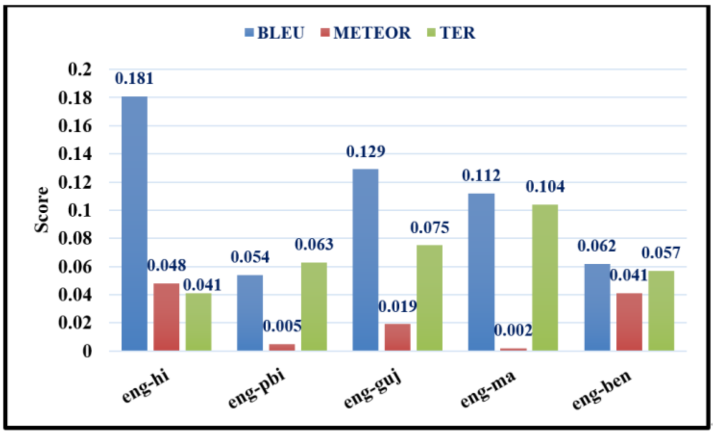

Figure 9. A bar graph comparison of different language pair using different evaluation metrics for Test Dataset 2 with Google Translate.

6) Pearson Correlation of different evaluation metrics for different language pairs with Bing translate

Table 7. Comparison of different language pair using Pearson correlation of different evaluation metrics for Test data 2 with Bing Translate.

Language pair	BLEU	METEOR	TER
eng-hi	0.112	0.104	0.042
eng-pbi	0.107	0.076	0.091
eng-guj	0.067	0.086	0.028
eng-ma	0.108	0.011	0.054
eng-ben	0.076	0.047	0.014

We translated our test dataset 2 on Bing Translate and calculated BLEU scores for all these language pairs and compared it with other metric scores which are shown in **Table 7**.

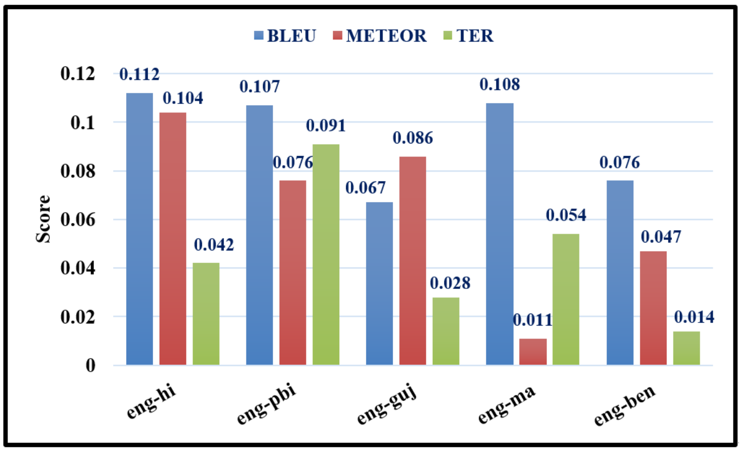

Figure 10. Comparison of different language pair using different evaluation metrics for Test data 2 with Bing Translate.

Figure 10 shows the BLEU score of 0.112 for eng-hi, 0.107 for eng-pbi, 0.067 for eng-guj, 0.108 for eng-ma, and 0.076 for eng-ben. So, the Pearson correlation for the BLEU score of all these language pairs is higher than other evaluation metrics for test dataset 2 with Bing translator, which are shown in **Figure 10**.

7) Pearson Correlation of different evaluation metrics for different language pairs with Yandex translate

Table 8. Comparison of different language pair using Pearson correlation of different evaluation metrics for Test data2 with Yandex Translate.

Language pair	BLEU	METEOR	TER
eng-hi	0.068	0.046	0.042
eng-pbi	0.094	0.012	0.056
eng-guj	0.111	0.022	0.034
eng-ma	0.118	0.082	0.062
eng-ben	0.121	0.146	0.071

Table 8 shows the effectiveness of our translations we translated our test dataset 2 on Yandex Translate, calculated BLEU scores for those, and compared it with other scores. The BLEU score is higher for all language pairs than other metric scores. But for only eng-ben language pair, the METEOR score has the highest score of 0.146, and the BLEU score is 0.121 as shown in **Figure 11**.

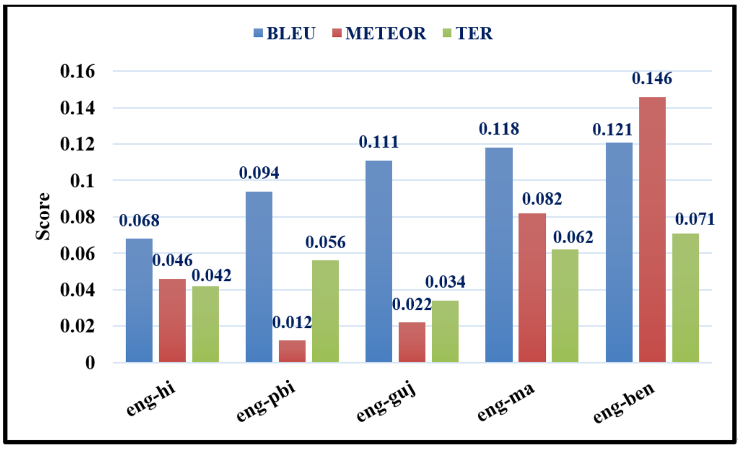

Figure 11. Comparison of different language pair using Pearson correlation of different evaluation metrics for Test Dataset 2 with Yandex Translate.

8) Pearson Correlation of different evaluation metrics for different language pairs with Im translate

Table 9. Comparison of different language pair using Pearson correlation of different evaluation metrics for Test Dataset 2 with ImTranslate.

Language pair	BLEU	METEOR	TER
eng-hi	0.142	0.097	0.105
eng-pbi	0.152	0.111	0.041
eng-guj	0.092	0.003	0.011
eng-ma	0.067	0.106	0.04
eng-ben	0.146	0.103	0.061

Table 9 shows that the BLEU has the highest score for all language pairs as compared to other metric scores. But for the eng-ma language pair, the METEOR score is higher, which is 0.106, and the BLEU score is 0.067.

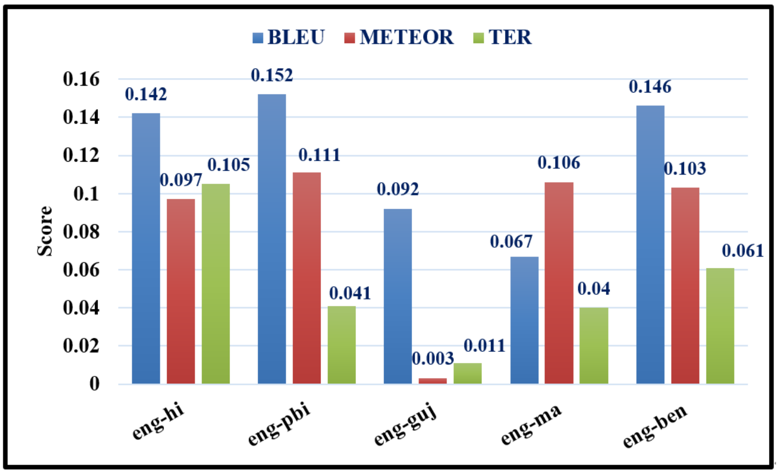

Figure 12. Comparison of different language pair using Pearson correlation of different evaluation metrics for Test Dataset 2 with ImTranslate.

The Pearson correlation for the BLEU score of eng-hi test dataset 2 with ImTranslate is 0.142, METEOR has 0.097, and TER has 0.105 score as shown in **Figure 12**.

So, we have analyzed that still BLEU has some issues in a few cases, but it still provides better correlation because it is language-independent. But in some cases, the respective scores of METEOR and TER also have a higher score as compared to other metrics which implies that METEOR and TER have been provided better results as compared to others with other MT systems.

5. Conclusion and future scope

In conclusion, the study aimed to explore various LAMT evaluation metrics for Indic languages. Several famous metrics such as BLEU, METEOR, and TER were compared and evaluated on their effectiveness in assessing the quality of the machine-translated text. The evaluation was carried out on multiple datasets, and the results were analyzed to determine which metric performed better. For this research work, we have applied MT techniques on several Indian language pairs and evaluated them on different automatic evaluation metrics. The findings revealed that BLEU performed relatively well on most datasets and was the most widely used metric for evaluating MT systems. However, the study also highlighted the limitations of BLEU and the need to use multiple metrics for a more comprehensive evaluation of MT quality.

The study recommends using a combination of BLEU, METEOR, TER, and NIST metrics to evaluate

MT systems for Indic languages. This approach provides a more comprehensive evaluation and a better understanding of the quality of the machine-translated text. Additionally, the study suggests that future research should focus on developing new evaluation metrics specifically for Indic languages to improve the accuracy and effectiveness of MT evaluation. From this analysis, we can conclude that, based on the BLEU scores, the MTS performs best for the English-Hindi and English-Punjabi language pairs with Google translation, followed by the other mentioned language pairs. But METEOR has the highest score and the MTS performs best for these language pairs with Yandex translation, and TER has the highest score and the MTS performs best for these language pairs with ImTranslator. In contrast, a higher BLEU or METEOR score represents better translation quality. We depict that apart from the major disadvantage of BLEU it is still widely used because of its language-independent nature and ease of implementation. Finally, we compare machine translations against reference translations by using evaluation metrics like BLEU, TER, or METEOR. These metrics measure the similarity between the machine-translated output and the reference translation.

In this paper, we have presented a comparative analysis of different LAMT evaluation metrics for the Indic language. The performance of these metrics is evaluated on five different datasets of English-Hindi, English–Punjabi, English-Gujarati, English-Marathi, and English-Bengali language pairs. This research study concludes that the research that has been presented will assist researchers studying machine translation in quickly determining the automatic machine translation evaluation metrics that are most effective for the improvement of the machine translation systems. Additionally, the study offers a general overview of the development of automatic machine translation evaluation research. The study also emphasizes the necessity for further research in this field to enhance the effectiveness of automatic evaluation metrics for different Indic language pairs.

As future work, we have to extend this research study to other Indic languages and domains. We also aim to incorporate syntactic and pragmatic features to capture the structural and contextual aspects of translation quality. Furthermore, we intend to explore the correlation of the metrics with other reference translations and conduct a user study to validate its usefulness and reliability. We hope that this research analysis will contribute to the advancement of MT research and evaluation for the Indic languages.

Author contributions

Conceptualization, KK; methodology, SC; software, SC; validation, KK, and SC; formal analysis, KK; investigation, KK; resources, SC; data curation, KK, and SC; writing—original draft preparation, KK; writing—review and editing, KK; visualization, KK; supervision, SC; project administration, SC; funding acquisition, KK. All authors have read and agreed to the published version of the manuscript.

Conflict of interest

The authors declare no conflict of interest.

References

1. Andrabi SAB, Wahid A. Machine Translation System Using Deep Learning for English to Urdu. Computational Intelligence and Neuroscience. 2022, 2022: 1-11. doi: 10.1155/2022/7873012
2. Khan NJ, Anwar W, Durrani N. Machine translation approaches and survey for Indian languages. arXiv. 2017, arXiv:1701.04290.
3. Hendy A, Abdelrehim M, Sharaf A, et al. How good are GPT models at machine translation? A comprehensive evaluation. arXiv. 2023, arXiv:2302.09210.
4. Rivera-Trigueros I. Machine translation systems and quality assessment: a systematic review. Language Resources and Evaluation. 2021, 56(2): 593-619. doi: 10.1007/s10579-021-09537-5
5. Rei R, Guerreiro NM, Treviso M, et al. The Inside Story: Towards Better Understanding of Machine Translation Neural Evaluation Metrics. Proceedings of the 61st Annual Meeting of the Association for Computational

Linguistics (Volume 2: Short Papers). Published online 2023. doi: 10.18653/v1/2023.acl-short.94

6. Sahaya V, Singh P. Evaluation of Performance Metric of Automatic Machine Translation. International Journal of Computer Science and Software Engineering. 2015, 1: 49-57.
7. Mrinalini K, P V, Thangavelu N. SBSim: A Sentence-BERT Similarity-Based Evaluation Metric for Indian Language Neural Machine Translation Systems. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2022, 30: 1396-1406. doi: 10.1109/taslp.2022.3161160
8. Garje GV, Bansode A, Gandhi S, et al. Marathi to English Sentence Translator for Simple Assertive and Interrogative Sentences. International Journal of Computer Applications. 2016, 138(5): 42-45. doi: 10.5120/ijca2016908837
9. Ramesh A, Parthasarathy VB, Haque R, et al. Comparing Statistical and Neural Machine Translation Performance on Hindi-To-Tamil and English-To-Tamil. Digital. 2021, 1(2): 86-102. doi: 10.3390/digital1020007
10. Hasler E, de Gispert A, Stahlberg F, et al. Source sentence simplification for statistical machine translation. Computer Speech & Language. 2017, 45: 221-235. doi: 10.1016/j.csl.2016.12.001
11. Xia Y. Research on statistical machine translation model based on deep neural network. Computing. 2019, 102(3): 643-661. doi: 10.1007/s00607-019-00752-1
12. Choudhary H, Rao S, Rohilla R. Neural Machine Translation for Low-Resourced Indian Languages. arXiv. 2020, arXiv:2004.13819.
13. Papineni K, Roukos S, Ward T, et al. BLEU. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics—ACL ’02. Published online 2001. doi: 10.3115/1073083.1073135
14. Ananthakrishnan R, Bhattacharyya P, Sasikumar M, Shah RM. Some issues in automatic evaluation of English-hindi MT: *More blues* for *BLEU*. 2007.
15. Denkowski M, Lavie A. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. Proceedings of the Ninth Workshop on Statistical Machine Translation. Published online 2014. doi: 10.3115/v1/w14-3348
16. Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Goldstein J, Lavie A, Lin CY, Voss C. Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Association for Computational Linguistics; 2005. pp. 65-72.
17. Snover MG, Madnani N, Dorr B, et al. TER-Plus: paraphrase, semantic, and alignment enhancements to Translation Edit Rate. Machine Translation. 2009, 23(2-3): 117-127. doi: 10.1007/s10590-009-9062-9
18. Kandimalla A, Lohar P, Maji SK, et al. Improving English-to-Indian Language Neural Machine Translation Systems. Information. 2022, 13(5): 245. doi: 10.3390/info13050245
19. Dewangan S, Alva S, Joshi N, et al. Experience of neural machine translation between Indian languages. Machine Translation. 2021, 35(1): 71-99. doi: 10.1007/s10590-021-09263-3
20. Philip J, Namboodiri VP, Jawahar CV. A baseline neural machine translation system for Indian languages. arXiv. 2019, arXiv:1907.12437.
21. Sai BA, Dixit T, Nagarajan V, et al. IndicMT Eval: A Dataset to Meta-Evaluate Machine Translation Metrics for Indian Languages. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Published online 2023. doi: 10.18653/v1/2023.acl-long.795
22. Google Translate. Available online: https://translate.google.com/ (accessed on 28 July 2023).
23. Bing Microsoft Translator. Available online: https://www.bing.com/translator (accessed on 29 July 2023).
24. Yandex Translator. Available online: https://translate.yandex.com/ (accessed on 31 July 2023).
25. ImTranslator. Available online: http://imtranslator.com/ (accessed on 31 July 2023).
26. AI4Bharat Open-Source Dataset. Available online. https://ai4bharat.iitm.ac.in/datasets/ (accessed on 26 July 2023).