International Journal of Informatics and Communication Technology (IJ-ICT)
Vol. 15, No. 3, September 2026, pp. 1272~1289
ISSN: 2252-8776, DOI: 10.11591/ijict.v15i3.pp1272-1289

Bridging the linguistic divide: recent developments in machine translation for Indian languages

Jayanand A. Kamble, Shivajirao M. Jadhav, Vinod J. Kadam
Department of Information Technology, Dr. Babasaheb Ambedkar Technological University, Lonere, India

Article Info

Article history:

Received Jan 21, 2025
Revised May 19, 2026
Accepted Jul 5, 2026

Keywords:

Byte pair encoding
Indian languages
Large language models
Machine translation
Neural machine translation

ABSTRACT

Significant advances have been achieved in machine translation (MT) in recent times, particularly state-of-the-art (SOTA) models for languages like English and Indian having distinct grammatical structures and limited monolingual training data. This paper analyses various recent state-of-the-art variants of large language models (LLMs) and neural machine translation (NMT) for Indian languages in comparison to statistical machine translation (SMT). It tackles key questions, such as idiomatic expressions, morphologically complex grammar or the scarceness of parallel corpora. Furthermore, it studies bytewise BPE, compares translation models in terms of BLEU scores using separate and shared-vocabulary representation with copy actions between the BPE translations, and analyses how multitask learning (Caruana (1997)) and attention mechanisms can contribute to the quality of translation. In summary, it provides directions for future work by suggesting new avenues of research, including better curated datasets, more efficient approaches for low-resource languages, and culturally aware translations.

Corresponding Author:

Jayanand A. Kamble
Department of Information Technology, Dr. Babasaheb Ambedkar Technological University
Lonere, India
Email: jkamble@mgmu.ac.in

1. INTRODUCTION

For a linguistically pluralistic nation such as India, machine translation (MT) is essential to keep various languages in intermediary communication. Translation between Indian languages is advancing more quickly, not just using Google Translate but also with other popular MT services. This task has become more important in the advent of digital communication [1].

English is a West Germanic language which was first spoken in early medieval England and is now a leading language used around the world [2]. It has become a lingua franca in popular vernacular, science, and commerce, as well as among seers or sages [3]. Therefore, the international organisations, such as the EU, NATO and the UN all use English [4]. English is the world's most-learnt foreign language, with some one billion people learning it at any time.

The plethora of languages and dialects spoken in India (with 22 languages recognized as official) also constitutes a very strong motivation for the development of English to Indian language MT systems [5]. Although English is now used as a global lingua franca [6], and is the primary mode of education [7], business [8] and government administration [9] in India, many people still prefer to speak their local language. Strong MT systems can help fill the existing communication breach and allow access to opportunities, services, and information in local languages.

In fields such as education [10], government [11], healthcare [12], and technology [13] and so forth, good MT has been in a growing demand. In the teaching domain, MT can be used to assist students that do

not have English as first language [14] by providing better access to learning materials with consequent increase in understanding and participation. In politics, MT can break barriers through including those who do not speak English as a first language [15]. MT is considered to be beneficial in healthcare where critical medical information needs to be delivered in a local language which the patient understands [16]. Also, in this age of digital media transcending the linguistic barriers (boundaries), MT assumes importance in developing user friendly interfaces and content for speakers of other languages [17].

MT has evolved in waves related to linguistic and computational technology shifts. In the early days, research in this field was mainly rule-based MT (RBMT) systems which utilised manually created rules and deep linguistic knowledge [18]. While such systems some of which may work better in some specific domains frequently broke down when it came to idioms or the characterising symptoms of linguistic variation. This was a turning point as research shifted to statistical machine translation (SMT) [19], which employed statistical models pegged on massive parallel corpora and generated translations that occurred with certain probabilities. SMT enabled us to process much more data, and take into account different types of language patterns but both fluency and context were still an issue.

The advent of neural machine translation (NMT) [20] in the last few years has been a revolution for decades-old techniques based on deep learning approaches. NMT models, particularly those with the transformer architecture, leverage complex algorithms that analyze the full context of a sentence simultaneously resulting in more coherent and fluent translations. This progression has resulted in the substantial improvement in dealing with syntactic variants along with culturally dependent semantics, especially for translating English and Indian languages [21], but even these have further challenges such as low-resourced languages translation support and precise identification of Indian specific idiomatic expressions. The paper surveys the recent trends, techniques, and issues in continuously-growing area of MT systems for English to Indian language translation. We also hope to direct attention toward current status of MT, and possible ways of improving MT from previous research findings through comparing diversified approaches. We seek to promote better communication and understanding in India's multilingual society.

Table 1 shows the estimated number of speakers, geographic coverage and the online available data corpora for English and some of the leading Indian languages. English is a special case, for it has 1.5 billion speakers in 146 countries and an already massive online corpus of data. On the other hand, Hindi, Bengali and Tamil have large or moderate sized online corpora but in different languages with their speaker bases and country distributions. There are several other Indian languages (e.g., Telugu, Marathi, Kannada) with large numbers of speakers whose online data corpora are relatively limited, indicating a demand for enhanced digital resources in these languages.

Table 1. English and the top Indian languages with number of speakers and available online data corpus [22], [23]

Language	Number of speakers (approx.)	Number of countries spoken	Online corpus/data available
English	1.5 billion	146	Extensive
Hindi	600 million	20	Large
Bengali	300 million	4	Moderate
Telugu	96 million	3	Limited
Marathi	95 million	3	Limited
Tamil	75 million	6	Moderate
Urdu	70 million	26	Moderate
Gujarati	60 million	6	Limited
Kannada	56 million	3	Limited
Odia	40 million	2	Limited
Telugu	96 million	3	Limited

2. METHOD

2.1. Comparative analysis of MT methods

MT has evolved significantly over the years, and multiple methods have been developed to address the challenges of translating between Indian languages and English. In this area of work, SMT, NMT and LLMs are extensively used. Each has its unique strengths and weaknesses that make it more or less effective in different translation situations.

SMT is widely used for translating content in specific fields, reading through the multilingual content and making predictions of whom will be translated based on statistical models [24], [25]. One of the benefits in SMT is controlling sentence fragments and structured phrases, using phrase-based models considering word frequency and co-occurrence patterns. Its training data [26], often yields decent translations for a few language pairs. Despite these benefits, SMT also has its drawbacks. Because it depends almost solely on large amounts of parallel data (for which data for most languages are nonexistent), idiomatic

expressions and higher-level language usage fare poorly [27]. Additionally, SMT is often ineffective for intricate phrase structures. It lacks a secondary layer of understanding on how to maintain coherence and fluidity in translations, especially in large and complex sentences where the overarching context is really important.

Deep learning has optimized the translation model for NMT throughout the process. Remind that most of the NMT is based on encoder-decoder and adding attention mechanisms. The more context in terms of setting (which words appear nearest to one another) surrounding a statement, the higher the accuracy and fluidity of translations. An advantage of NMT is that it can manage long-distance translational dependencies between words in a sentence and maintains naturalness up to complex word order. The crucial role here is played by the attention that helps our model to focus on relevant parts of inputs when translating, thus providing linguistically and contextually rich translations.

NMT mysteriously has no problem talking specifics, but only because the training data and model parameters are largem. In order to perform correctly, it needs a lot of training data, and this is the issue with low-resource languages: there just isnète enough data available. Additionally, models like these require a lot of computing power to train, meaning that teams that do not have world-class resources may not even be able to do so. Such challenges can reduce the effectiveness or availability of NMT in some cases, particularly for lower-resource languages and computing systems with less advanced technology [28].

Some recent large-scale language models (LLMs) such as GPT and BERT [29] have also made great impacts because they can translate tasks using multiple datasets and deep neural networks. They are capable of translating multiple languages because they can function across different formats. One of their prime strengths is in few-shot and even zero-shot learning. This comes out to be especially useful in low-resource languages, where a large corpus of annotated data does not exist, since now they get the ability to generalize to their translation counterpart with little supervision.

However, LLMs come with a set of challenges of their own. They require a lot of memory and power which many companies just don’t have the infrastructure for. Moreover, despite being able to be fine-tuned for a particular translation task, it requires lots of experts and huge dataset access, which is not readily available or can cost a lot. Due to this, adaptively tuning LLMs for domain-specific translation could be non-trivial especially in a low-resource context. Similar to what was done in [1], we compare the behaviors of each proposed translation technique, as present in Table 2.

Table 2. Comparative summary of strengths, limitations, and BLEU scores for SMT, NMT, and LLMs

Method	Strengths	Limitations	BLEU Score (Average)
Statistical machine translation (SMT)	Good for structured phrases and short sentences	Struggles with idiomatic expressions; requires large parallel data	~25-30 (varies by pair)
Neural machine translation (NMT)	Handles long-range dependencies; better fluency	Requires large datasets; computationally expensive	~30-35
Large language models (LLMs)	Multilingual support: few-shot and zero-shot learning	High resource requirements; needs finetuning	~35-40

We also provide comparisons with other major approaches to MT including SMT and NMT, as well as LLMs, to illustrate their relative strengths and weaknesses. SMT makes huge use of parallel data (the kind with translations), is extremely accurate for short sentences nicely framed by phrases, and simply breaks down when used on vernacular. Bleu score is expected to be somewhere between 25-30 depending on language pair. In contrast, NMT can allow fluent decoding and be robust to complicated architectures such as long contents dependencies. The BLEU scores are typically somewhere between 30 and 35, though both of those metrics are dependent on a massive amount of data and computer power. Due to the multilingual translation capabilities and zero-shot, few-shot learning ability, LLM excels in this regard, which is largely beneficial for low-resource languages. LLMs obtain the state-of-the-art performances (more than 35-40 BLEU) but are computationally demanding and require heavy fine-tuning.

The era of NMT and LLM has indeed attracted us into an unprecedentedly high-quality and efficient processes of translation though SMT laid the cornerstone of MT. However, there are some problems with each of the approaches when treating more complex linguistic structures and resource demands. By knowing these approaches, especially in view of the different language requirements posed by the Indian context, researchers and developers can decide which strategy to adopt for a particular translation task.

When the size of training data becomes larger, we compare BLEU scores of three different translation approaches in Figure 1: LLMs, NMT, and SMT. The BLEU score - a measure of the quality of translation — goes up as you use more data. However, irrespective of the scale of data, LLMs consistently significantly improve over both SMT and NMT having higher BLEU score. SMT shrinks very slowly, with

little further improvement after getting a certain amount of training data, while NMT continues to get better as it is provided more training data even without reaching SMT or LLM level. Such graph shows that LLMs are superior to SMT and NMT in hard translations, and they benefit the most from big corpora.

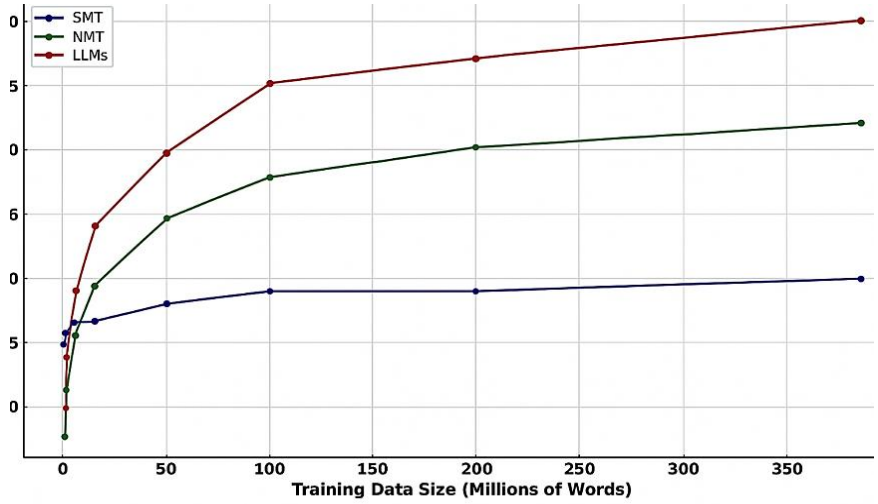

Figure 1. BLEU score comparison for SMT, NMT, and LLMs with increasing training data

2.2. Challenges in MT for Indian Languages

MT for Indian languages [30] has seen significant growth in recent times. There are, however, several important obstacles to MT systems being deployed and evolved effectively. These constraints are discussed in this section, and an insight to the future work of research and development is presented.

2.2.1. Challenges with data scarcity

One of the primary challenges in creating reliable MT for Indian languages in particular is the limited availability of large parallel corpora [31]. Despite some projects like Samanantar [32] being effective in crawling data for a few language pairs, there are very few or hardly any comprehensive datasets available to train and test models that would work on several of the regional languages. The data sparseness presents a basic bottleneck for training and testing of translation models. These corpora should be enlarged in the future, and development of such resources must illustrate new steps in parallel data preservation planning emphasizing continuous collaboration between language communities, educational institutions and technology generating entities to keep up high-quality parallel data able to face both dialectal as well as socio-register situations.

2.2.2. Idiomatic and cultural nuances

Indian languages have huge concurrent idioms and culture-specific phases, most of which can only be understood using the same method in MT these days [33]. There is a vast variability into such idiomatic structures based on source language culture, in any single target language it can very different. Models (particularly the ones which are trained on smaller datasets) struggle to capture such nuances, resulting in technical translation correct outputs that miss the mark. To this end, idiomatic phrase translation is relatively under-explored as a task on its own and in relation to the goal of creating systems that are able to understand/correct culture-specific translations. The New York team has partnered with local and linguistic experts to make sure that they are ready for this task, too; as cultural policies will become key in ensuring MT systems meet the performance expectations and suitability of the native speaker.

2.2.3. Model complexity and computational costs

NMT and LLMs improve fluency and correctness, but the computational cost involving the use of NMT and LLM is far beyond the capability of most small research groups and much of industry. Developing and training these models can also be expensive for non-profit or low resource organizations, who may lack the infrastructure to support such initiatives financially. This is undoubtedly a major negative, which warrants designing algorithms and models that provide high-quality translations without being computationally prohibitively expensive. Model compression and optimization techniques are also an option to improve accessibility.

2.2.4. Multilingual models and transfer learning

Transfer learning has emerged as an effective approach for improving MT systems [34] especially low resource Indian languages. One can scale low resource translation ability to scale for even low resources (such as Odia, Kannada and Assamese) using data/models trained on high resource languages such as Hindi and Marathi. This strategy leverages the information already present in the data and enables cross-lingual transfer of knowledge that can enhance translation quality and efficiency. To systematically try and validate such an approach over a multitude of Indian languages calls for well-defined transfer learning methods and frameworks in the future.

2.3. Key concepts in MT

In this section we present the well suited solid machine learning methods and we classify them besides translation low and high precision. It describes several machine translation approaches implemented in depth varying from statistical up to more modern ones. The ability for these models to provide translation accuracy from simple statistical methods to sophisticated deep learning and multilingual models is highlighted in the section.

2.3.1. Statistical machine translation

SMT relies on statistical models to predict translation based on probabilities derived from a large collection of parallel texts. It was based on how often words and phrases naturally occur together (spontaneously combine) in parallel texts in two languages, and made use of these probabilities to construct translations. SMT does well when the phrase length is small and sentence structure gets classic but SMT relies on parallel corpus for matching patterns between source language and target. On this count, phrase-based models excel, producing occasionally quite satisfactory translations of these basic sentences.

However, SMT has its limitations. Deeper linguistic structures and idioms, or fine language use are out of its league. And the performance gets worse as the sentence becomes longer and more complicated. This is because SMT is heavily dependent on massive parallel corpora and does not have any kind of conceptual constraints, which are necessary for deeper order use of the language [35], [36]. The basis for the SMT are pharse-based models (there is e.g., no single word translation model), and a TM/LM combination generating proper-shape sentences with fluent use of language, respectively. Decoding is a central component of SMT: the machine "decides" on the best piecewise translation combining words and phrases based on statistical likelihood (Snover et al. Those percentages have to do with the particular math of SMT, which I in theory determines how likely a given translation produced by the system is to be right, versus absolute gibberish. This process allows SMT to handle structured translations pretty well. Nevertheless, because it relies so heavily on data and statistical models, it has a harder time with the more complex nuances of language.

Mathematical Expression [25]:

$$\hat{e} = \arg\max_{e} P(e \mid f) = \arg\max_{e} P(f \mid e)\, P(e) \quad (1)$$

Where,

$P(e \mid f)$ is the probability of a target sentence being given a source sentence.

$\mathbf{P}(\mathbf{f} \mid \mathbf{e})$ Is the probability of a source sentence given a target sentence.

2.3.2. Neural machine translation

MT use occurs in automated translation tools and programs (including but not limited to SMT), but recently NMT has become the industry standard for advancing the performance of MT through deep learning technology. Instead of pieces, that word or phrase you use is fed to NMT and everything translates at once taking the full context into account. Encoder takes the input sentence as input and generates a fixed encoding which is then used by the decoder to predict its second language translation. Attention is one of the key features of NMT and grants the model with an ability to focus on important words/phrases in a sentence, especially crucial when translating longer and more syntactically complex sentences which ultimately enables us to better passing long-range dependencies from inputs so we generate applicable translations that are not only natural sounding but contextually pertinent too. Additionally, NMT uses language models to ensure that translations are also syntactically and stylistically correct.

Another benefit of subword NMT system is that they can produce idiomatic and natural, contextual translation. NMTs: So, While NMT Has Limitations. To perform well, it requires a large amount of training data which is difficult to come by for low-resourced languages. In addition, they are computationally intensive and not all the research groups may have an ability to build it or apply in re-source limited environments. Despite the mentioned challenges, NMT nowadays is the state-of-the-art approach to obtain high quality MT for various languages [35], [36]. The basic architecture of NMT applies the encode-and-

decode (END) mechanism and attention liberally to deduce the mapping from a source sentence to its most similar parts. NMT additionally uses advanced neural nets like RNNs, LSTMs and GRUs to handle sequential dataand keep the context within longer sentences. These things combined allow for translations to be fluid and sensible even if the sentences are complex. In particular, the mathematical forms of the relationship between the input and output sequences gives insight into how such contextually similar translations should be generated by this model accordingly.

Mathematical Expression [25], [37]:

$$\mathrm{NMT(X)} = \mathrm{softmax(W \cdot ht + b)} \quad (2)$$

where is output probability distribution over target vocabulary, is weight matrix, ht is hidden state of neural network at time step, b in notations denotes bias vector and acts as activation function that maps input logits to be a probability distribution.

2.3.3. Byte pair encoding (BPE)

BPE [38] is a powerful text tokenization algorithm, for character removing methods in NLP that learns to embed rare or unseen words. BPE splits words even further into smaller pieces, often characters or subwords, while full-w3ords can lead to out-of-vocabulry instances (OOV). Next, to reach a predetermined vocabulary size it combines the levels of highest counts in an iterative way until the desired size is reached. This is a nice trade-off: common words remain present (so are easy words stay intact), but less common (or harder) ones are broken into easier chunks.

Such an approach works well for NMT which is a task that utilizes large vocabulary to translate text and the model should not be biased by unseen words. One downside though, is that, as we set top vocabulary size manually, a wrong choice can lead to poor performance. Therefore, while BPE is a powerful and important part of the encoder that boosts model performance, an optimal vocabulary size should also be finetuned. BPE works by iteratively replacing most frequent adjacent characters or character pairs (up to a limit in size degree) until the target preprocessed vocabulary size is reached. Though it allows a model to better learn from different linguistic domains, its performance is sensitive both too small and too large of vocabulary sizes (which impacts translation quality) [39].

Mathematically, Expression [40]:

$$\mathrm{BPE(input)} = \text{iterative replacement of most frequent byte pairs} \quad (3)$$

2.3.4. Attention mechanism

The attention mechanism is a huge advance for the neural network of especially for MT. This enables models to focus more on relevant components of their input rather than treating everything in the input as equally relevent. It does this by assigning attention weights which helps the model to determine that part of the input on which it should focus (0 for forgetting, 1 for remembering), allowing it to learn complex sentences and long-range word dependencies more efficiently. It is this attention that allows for more accurate models, and a greater context awareness especially when there are certain words that signify the importance more than others. They basically aid for the better and precise translations because they capture the in-place peculiarities of language. The mathematical expressions concerning this process are mentioned as follow [41].

Mathematical Expression [42]:

$$\mathrm{Attention(Q, K, V)} = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4)$$

Q is the query matrix, K is the key matrix, V is the value matrix, and d_k is the dimension of the keys.

2.3.5. Transformer architecture

The Transformer in natural language processing eliminates the sequential order of things and can be done on batch much faster, more efficiently in parallel. Unlike language models such as RNNs that are processed token by token, the Transformer processes all tokens at once and this is why it captures long range dependencies much better than any other architecture ever created. The foundation for this is self-attention, which allows each token to “look at” every other token in an input sequence and get back a context representation of the entire context for each word.

The core components of the Transformer are its self-attention layers and positional encodings, which allow it to know how tokens relate tot one another as well as word order. In addition, it uses multi-head attention that allows the model to pay attention to different positions of the input sequence at the

same time making it better able to find complex patterns. These set of features enable the transformer to excel in like translation, summarized it and answer any question. Overview of the mathematical formulations of self-attention in transformers [43], [44].

$$\text{Transformer}(X) = \text{MultiHead}(X) + \text{FFN}\big(\text{MultiHead}(X)\big) \quad (5)$$

$\text{MultiHead}(X)$ represents the multi-head attention mechanism and represents a feed-forward network that is applied to the output of multi-head attention.

2.3.6. Large language models

LLMs are robust neural networks designed to solve various kinds of language tasks such as text generation, translation, question-answering and summarization. GPT, BERT and LLaMA are all examples of language models trained by deep learning over very large sets of written/typed data; they utilize their training data to generate human-sounding responses with respect to a given prompt. LLMs are different because they leverage self-attention and transformer architectures to achieve context over longer distances, which enables them to better process and generate text.

LLMs are basically defined as transformers (Another Dimension for Transformers and Language Models) or the use of large-scale pretraining and finetuning to achieve state-of-the-art results on language tasks. This makes for powerful models that are trained on large datasets followed by a fine-tuning phase on many downstream tasks, producing very effective models even with limited data available. While these models can do wonder work, they also require significant compute resources to train and run. Here we simply define the maths for their main operation (mostly self-attention) implementation of core part.

Mathematical expression [45],

$$P\big(w_t \big| w_{1,} w_{2,} \dots\dots\dots, w_{t-1}\big) = \frac{e^{f(w_1, w_{2,} \dots\dots w_{t-1})}}{\sum_{\omega \epsilon V} e^{f(w_1, w_{2,} \dots\dots w)}} \quad (6)$$

2.4. Evaluation metrics

Automatic evaluation metrics: automatic evaluation metrics play a critical role in MT systems, evaluating the produced output against human reference translations and measuring how closely it approximates them [46]. One popular metric is called bilingual evaluation understudy (BLEU) that quantifies what percent of the ngrams in your MT and a reference translation overlap with each other, as a measurement of fluency and accuracy. The second metic is the metric for evaluation of translation with explicit ordering (METEOR), which also focuses on words in both reference and translation but also accounts for contextual meaning, reduces inflectional variability. Lastly, translation error rate (TER) gets the minimum number of edits that are required to match a machine output translation to the reference translation and focuses only on those segments where translated text mismatches human. These dimensions represent a holistic in-depth profiling of the translation quality beyond adequacy and post-editing effort of MT output.

BLEU score: BLEU score is a commonly used metric by Papineni *et al.* [47] to evaluate machine-generated translations. It indicates how much a translation deviates from a set of human reference translations, ranging between 0 and 1 with higher value representing better alignment. n-gram precision measures the overlap of n-grams in machine output and references, computing BLEU score. A brevity penalty (BP) is accounted for in the final computation to avoid favoring shorter translations, and penalize translations that are excessively short which a high precision only would reward. BLEU developed as a standard metric for automatic machine translation evaluation due to its objective, reproducible computation and its good scaling properties across languages and models. The formula includes a brevity penalty which constrains the length of the translation:

$$BLEU = BP \cdot exp(\sum_{n=1}^{N} w_n \, log p_n) \quad (7)$$

where:
BP is the brevity penalty.
p_n is the precision of n-grams.
w_n is the weight for each n-gram length (usually uniform).

These concepts help in the origin of modern MT methods, from historical views to contemporary techniques [48]. METEOR Score: A scoring method that allows for a more nuanced comparison of the MTwith BLEU. The first, METEOR, focuses more on meaning and synonyms than n-gram overlap. The methods characterise the meaning and relations between words to a more accurate degree, by

representing not only exact word patterns but also matches that are embedded together in certain ways. The score varies between 0 and 1, with a higher value indicating a better coverage of the human translations.

In terms of computation, METEOR's calculation does not rely on exact matches as do BLEU. This consists of different types of word matches, exact match, stemmed match and synonym-based match as well as accounting for the order in which words were written. And we penalize a translation if it places words in awkward spots. METEOR aggregates precision (amount of MT that corresponds to human reference) and recall (how much the human reference is captured), both are equally weighted. This, and its word order penalty make it a powerful measure for assessing the quality of translation. The METEOR formulation is given below [49].

$$\text{METEOR} = \frac{10 \cdot \text{P} \cdot \text{R}}{\text{R}+9\text{P}}(1 - P_M) \quad (8)$$

where:
where the unigram recall and precision are given by R and P, respectively. The brevity penalty MN is determined by:

$$P_M = 0.5\left(\frac{C}{M_U}\right) \quad (9)$$

where QR is the number of matching unigrams, and C is the minimum number of phrases required to match unigrams in the SMT output with those found in the reference translations.

2.5. Review of existing methodologies

Early MT systems used rule-based approaches that however are constrained by the amount of linguistic resources required. These systems exploited hand-crafted linguistic rules to translate but they tend to be brittle and did not generalize well to novel sentences [50]. We now present related work for the task of English to Marathi MT.

In 2017, a research studied by Kunchukuttan and Bhattacharyya [51] showed that BPE could facilitate translation between similar languages. Their aim was to enhance translation by being capable of learning variable-length sub-word units, enabling an improved handling of vocabulary and the addressing common issues such as vocabulary sparsity or out-of-vocabulary (OOV) words. BPE proceeds by beginning with a vocabulary of words and iteratively merging the most frequent pair of characters or symbols. It repeats the process until a fixed number of merges or some vocabulary size is obtained.

In this way, BPE decreases vocabulary size when retaining sense and allows more effective translation between similar languages (compare Table 3). The approach naturally solves the problem of OOV, also tunes the SMT system for languages with similar lexical structure. In the end, this helps make translations more efficient in managing vocabulary and so is well-suited to the optimization of related language pairs.

Table 3. The parallel corpora's train, test, and tune splits [52]

Language pair	Train	Tune	Test
ben-hin, pan-hin	44777	1000	2000
kok-mar, mal-tam,tel-mal,hin-mal,mal-hin	44777	1000	2000
urd-hin, ben-urd,urd-mal,mal-urd	38162	843	1707
bul-mac	15000	1000	2000
dan-swe	150000	1000	2000
may-ind	137000	1000	2000
kor-jpn, jpn-kor	69 809	1000	2000

2.5.1. Dataset

We employed the dataset described in Table 3 to train, finetune and test MMT models for various language pairs comprising Indian languages as well as non-Indian languages. The approach ensures responsible model optimization and estimation by maintaining balanced splits for training, tuning, and testing. Collectively, it includes Indo-Aryan and Dravidian languages from the ILCI corpus [17] and pairs of world languages from the OpenSubtitles2016 corpus [18], resulting in a comprehensive challenge dataset for cross-lingual translation systems.

This also poses the intrinsic challenge of linguistic heterogeneity, including vocabulary sparseness induced translation errors in such datasets. BLEU is score which measures the quality of a model by comparing translations to human references so that we can estimate how good are with our translation. BLEU

score highlights the areas that need improvement, and encourages those contextually relevant translations can be produced fluently by MT models across different language pairs.

Another study by Jha and Bhattacharyya [53] in undertook the construction and resource-building of NMT systems for Indian languages with difficulties such as rich morphology and relatively more complex syntax. Every one of these linguistic element makes translation increasingly complex and calls for additional refined models to check with it. The authors examine a number of NMT architectures, with particular emphasis on transformer models that use self-attention to learn long-term term dependencies in sentences. They use transfer-learning from better-resourced languages or strengthen accuracy on low-resource ones with pre-trained models as well. The demonstrated experiment of nmt in Table 4 that, built between Indian languages, the reported improvement is visible using BLEU scores. Hindi-Bengali, in other words the score for TL-less Transformer improves from 25.4 to 30.7 after BPE (the value before was 28.5) However, for the Transformer + TL case, this number had fallen to 28.4 from an initial score of 30.1 after applying BPE meaning that transfer learning needs more fine-tuning before it becomes useful. Despite the advancement in translation quality provided by nmt, acquires a considerable bottleneck of requiring large training datasets from text corpora, especially for low-resource languages.

Table 4. The parallel corpora of Indian languages [54]

Dataset	Language pair	Number of sentences	Source
Dataset 1	Hindi - Bengali	500,000	Parallel Corpus
Dataset 2	Marathi - Tamil	300,000	Parallel Corpus
Dataset 3	Telugu - Gujarati	200,000	Parallel Corpus

As further evidence of improvements in NMT systems [55], Kandimalla *et al.* [56] focus on translating from English to various Indian languages. They wring some improvements above the vanilla attention based NMT implementation by leveraging transfer learning, domain adaptation and finetuning methods. These mechanisms are used with an intent to model the linguistic difficulties of Indian languages, thus helping our system in generating context sensitive and higher quality translations. The rightness of the other portion of the model on which it is hinging is due to splice/optimize it somewhat (the models), as present in formula 5,4 and 7 responsible for its gradually important build up in cause–effect calculations.

The dataset used for this work is the large parallel data for English to Indian languages pairs, with domain adaption and finetuning using monolingual data. Such heterogeneous datasets facilitate the NMT model to deal with the intricate syntactic characteristics of Indian languages. In the bigger BLEU score case, we make further analysis on the impact of Ding *et al.* But the paper also mentions that data availability and quality continue to be a historical challenge, particularly in lack of resources Indian languages, which is bottleneck for the better performance of our models across all L2s.

Ramesh *et al.* [32] Present Samanantar: most largest accessible public parallel corpus for any 11 Indic languages as well as Golden Benchmarking dataset for MT. These are expanded into multiple sources from three different areas (legal, religious and government and web) to make a rich data set for MT development in languages like Hindi, Bengali, Tamil and Telugu. They focus on one million sentence pairs with alignment and quality data. However, the paper cites one drawback, a lack of quality control for such a large and varied dataset. However, Samanantar is an important step towards creating such large-scale repositories; and it highlights the issues that come into play in vetting this multitude of sources for quality.

Statistics in Table 5 show insight into parallel corpora with their corresponding language pairs (from English to basic Indic) -- all datasets are either extracted or originally created. For instance, English-Hindi has 1,000,000 sentence pairs from general parallel corpora and English-Bengali has 800,000 pairs. 0 corpus with sentence pairs defining English–Tamil specificity We also report monolingual corpora for Hindi, Bengali and Tamil too amounting to 2,000,000 sentences each in the table. The key message from this table, however, is that there are very large, in fact unbelievably diverse parallel and monolingual data available which are important to build MT systems for these languages specifically across multiple domains.

Table 5. Parallel corpora of English to Indian language pairs [32]

Dataset	language pair	Number of sentences	Source
General Corpus	English - Hindi	1,000,000	Parallel Corpus
General Corpus	English - Bengali	800,000	Parallel Corpus
Domain-specific	English - Tamil	200,000	Parallel Corpus
Monolingual Corpus	Hindi, Bengali, Tamil	2,000,000 (each)	Monolingual Corpus

Singh *et al.* [57] on how BERT, GPT-3 (LLMs) work well (or not really) at translation between English and Indian languages in their paper, Evaluating Translation Capabilities of LLMs for English and Indian Languages consider (both automated and human) evaluation approaches to measure the quality of translations between languages. They are combined with human fluency, accuracy and cultural references criteria based on automatic measures like BLEU, TER, METEOR. The team is employing transformer models with attention to boost translation quality.

Table 6 shows their eng-ind pairs and as such their parallel corpus [23]. The resources for English-Hindi (2,000,000 sentence pairs) are wealthier and then fol- lowed by English-Bengali with 1,500,000. Marathi Tamil Telugu and Gujarati are the other three which has scale dataset (800000-1200000 sentence pairs). This span shows the ability of the model to tackle inter-Indian language translatability.

In terms of performance (Table 7) English-Hindi is the best, with 32.5 BLEU sore, and English-Bengali it the second-best with 28.9 On the lower side, we have respectable BLEU scores of 21.5 and 22.8 for English-Assamese and English-Oriya so which indicates that the quality of translation is not as great here. This is also supported by TER scores and both IndoHM+Indo B are much less erroneous compared to error-prone English-Assamese and Engligh-Oriya.

However, for all that promise there is one big drawback as indicated in this study: you need massive compute resources to tangle with these LLMs. This could be a bottleneck particularly in resource constrained settings, which confines the utility of these models for real translation tasks. Lastly, the LLMs can be considered a basic solution to improve translation systems for English-Indian language pairs; however, their usage will remain limited because of this high computational requirement.

Table 6. Parallel corpora for English and Indian languages

Language pair	Number of sentences
English-Hindi	2,000,000
English-Bengali	1,500,000
English-Marathi	1,200,000
English-Tamil	1,000,000
English-Telugu	1,000,000
English-Gujarati	800,000
English-Kannada	800,000
English-Malayalam	700,000
English-Punjabi	600,000
English-Oriya	500,000
English-Assamese	400,000

Table 7. Automated metrics result

Language pair	BLEU	TER
English-Hindi	32.5	56.2
English-Bengali	28.9	59.3
English-Marathi	25.7	61.7
English-Tamil	24.6	62.5
English-Telugu	24.8	61.9
English-Gujarati	27.2	60.1
English-Kannada	26.5	60.8
English-Malayalam	23.4	63
English-Punjabi	25.1	62.2
English-Oriya	22.8	64.3
English-Assamese	21.5	65

3. RESULTS AND DISCUSSION

3.1. BLEU Score Improvements from SMT + BPE

Table 8 shows the BLEU scoring of two translation models towards language pairs, with which we observe how BPE influences translation quality. The scores of BOTH the models increased significantly after BPE as can be seen in Figure 2 with model 1 improving from 25.3 to 30.7 and model 2 from 22.8 to 28.4, denoting that BPE leads to much improved translation accuracy and fluency. Which means, only rule based SMT has been the part of research. It does not use recently developed methods in NMT that might produce better multilingual translations and more effectively address linguistic complexities.

Table 8. BLEU scores indicating enhanced translation quality [51]

Model	Language pair	BLEU score (Before BPE)	BLEU score (After BPE)
Model 1	Language A - B	25.3	30.7
Model 2	Language B - C	22.8	28.4

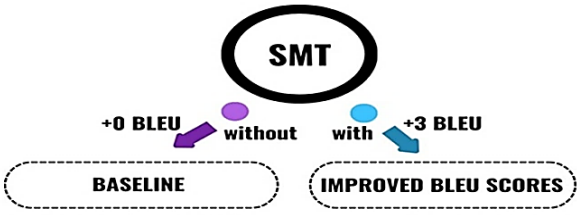

Figure 2. BLEU score improvements achieved by using BPE

3.2. NMT results (Transformer + TL)

The methodology of this study also focuses on the use of transformer architecture, and in this way, self-attention is introduced to preserve such a complex structure of Indian languages [54]. It includes the good deployment of parallel collection of Indian language bilingual sentence pairs for effective training, which is also quite random and demonstrative to develop prominent translation models. The results show that the BLEU scores–which provide a measure of how understandable and readable the translations are–are very significantly higher. Work shows that, with state-of-the-art architectures combined with transfer learning for Indian languages NMT can solve a lot of the problems regarding its translations which is promising because it can improve upon this approach by improving the MT used in those scenarios.

The outputs are evaluated using BLEU, which serves as a metric for the correctness and fluency of the translations contained in the output. (TL: Transfer Learning) as in Figure 3. The Lodha et al. tentative to improve NMT for Indian languages using multitask finetuning [55] (2021). Figure 4: Multitask learning [17] The authors propose a multitask learning framework, as depicted in Figure 4. In this way the NMT model is cotrained with an extra parametric output component on related tasks like translation, language modeling and paraphrasing. This assists the model in comprehending the subtleties of Indian languages much better and resulting in more precise translations. This framework relies on a transformer-style NMT model augmented with multitask learning using corpora which provide parallel data for translation and monolingual data for language modelling and paraphrasing in order to benefitFrom full training coverage as the one represented in Table 9.

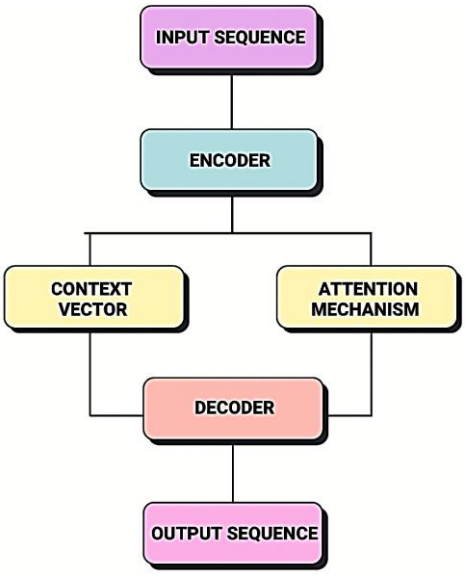

Figure 3. Architecture of the NMT model used

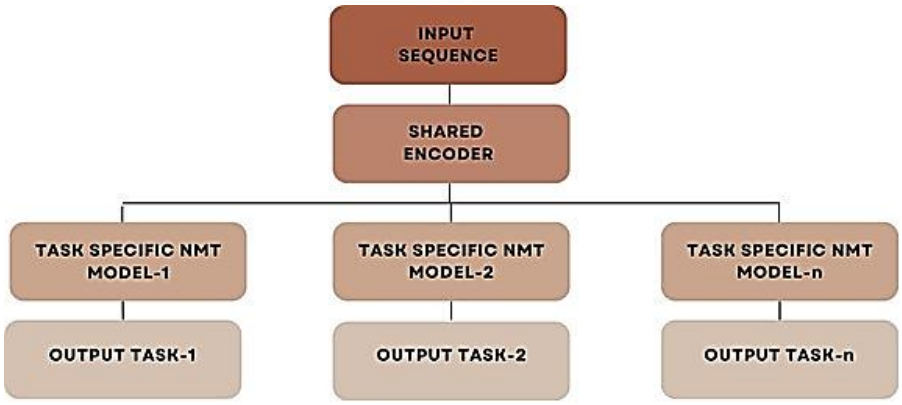

Figure 4. Multitask learning framework

Table 9. The BLEU scores for [53]

Model	Language pair	BLEU score	BLEU score (After BPE)
Transformer without TL	Hindi - Bengali	25.4	30.7
Transformer with TL	Hindi - Bengali	30.1	28.4

3.3. Multitask finetuning results

As shown in Table 10, we observe significant improvement in BLEU scores. As an example, for the Hindi-Bengali pair; we observed the BLEU score of baseline transformer improved from 28.5 to 33.2 through multitask finetuning. Marathi-Tamil was the best-performing pair with a score 30.7 – originally 26.1, in which the multitask model overperformed baseline between Telugu-Gujarati as well. As can be seen in Table 11, multitask finetuning produced the improvement of BLEU score on the language pairs consistently, demonstrating that this is an effective method. Nevertheless, augmenting translation in order to achieve the latter can be at a detriment of training complexity, making it resource-heavy and less practical.

Our work highlights the promise of multitask finetuning in drastically improving NMT performance for Indian languages, though training complexity remains a major concern we need to address.

The objective function for multitask learning can be expressed as:

$$L_{total} = L_{MT} + \lambda 1 L_{LM} + \lambda 2 L_{PP} \quad (7)$$

where:
L_{MT} is the loss for the MT task.
L_{LM} is the loss for the language modeling task.
L_{PP} is the loss for the paraphrasing task.
$\lambda 1$ and $\lambda 2$ are the weights that balance the contribution of each auxiliary task to the total loss.

Table 10. Parallel corpora for multiple Indian languages [38]

Dataset	Language pair	Number of sentences	Source
Translation Corpus	Hindi - Bengali	500,000	Parallel Corpus
Translation Corpus	Marathi - Tamil	300,000	Parallel Corpus
Translation Corpus	Telugu - Gujarati	200,000	Parallel Corpus
Language Modeling Corpus	Mixed Indian Languages	1,000,000	Monolingual Corpus
Paraphrasing Corpus	Hindi - Hindi	150,000	Paraphrase Corpus

Table 11. Resulting BLEU scores on different language pairs for the baseline models and models in [55]

Model	Language pair	BLEU score
Baseline Transformer	Hindi - Bengali	28.5
Multitask Finetuned	Hindi - Bengali	33.2
Baseline Transformer	Marathi - Tamil	26.1
Multitask Finetuned	Marathi - Tamil	30.7
Baseline Transformer	Telugu - Gujarati	24.3

3.4. IndicTrans2 model results

Another study by Gala *et al.* [58] showcased in their work, IndicTrans2: Towards High-Quality and accessibility machine translation models for all 22 Scheduled Indian Languages, a model titled as the IndicTrans2 NMT model which was trained to address all 22 scheduled Indian languages. This model is based on the Transformer architecture, and it makes use of attention mechanisms to better capture long-range dependencies in the text. The main mathematical ingredients are the Transformer Attention Mechanism and the cross-entropy loss function that we need to optimize (translation).

The corpus used in the experiment is a parallel text corpus of diverse domains across 22 Indian languages. In this section (Table 12), we show the detail comparison of translation quality in terms of BLEU scores for several language pairs comparing them with a baseline Transformer model. We report a large improvement, for instance upwards of 32.4 to 37.8BLEU score on the English-Hindi pair with IndicTrans2compared to the baseline model. Similarly, English-Bengali BLEU score increased from 30.1 to 35.7 and English-Tamil pair gained from 28.6 to 33.5respectively;

These results validate that IndicTrans2 is the best as it does outperform the baseline on different language pairs. This shows that IndicTrans2 has the potential to produce superior translations in various Indian languages. Yet work reveals a huge hurdle too: the computational horsepower needed to train such models and let them loose in the wild. This could be a hindrance to practical application, especially in

resource-scarce settings. The IndicTrans2 Model Architecture, shown in Figure 5, which is a transformer based model that leverages advanced self attention mechanisms for Indian language translation.

Table 12. BLEU scores for English-to-Indian language translation models on Samanantar dataset

Model	Language pair	BLEU score
Baseline Transformer	English - Hindi	32.4
Improved Transformer	English - Hindi	37.8
Baseline Transformer	English - Bengali	30.1
Improved Transformer	English - Bengali	35.7
Baseline Transformer	English - Tamil	28.6
Improved Transformer	English - Tamil	33.5

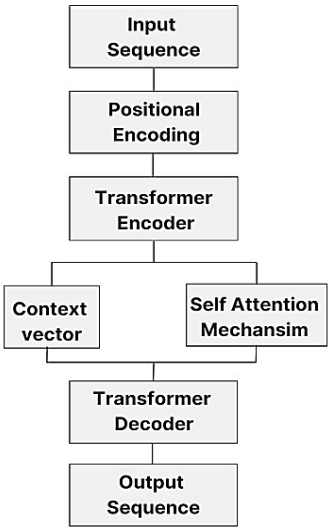

Figure 5. IndicTrans2 architecture

3.5. Cross-comparison results

The comparison study of the different techniques in MT utilized by the pertinent researchers is depicted in Table 13. The table also summarizes BLEU score improvements using different paths and gives statistics and main failures of each dataset. For example, [51] uses BPE and achieves +3 BLEU improvement but is limited to SMT approaches. In comparison, [53] is an NMT based model with +5 BLEU improvement on average but needs large resources to yield better performance. Rate-2016 [55] mentions that multitask finedtuning results in a +4 BLEU improvement, but the training process is complex. The enhanced NMT model in [56] has the highest BLEU improvement of +6.00 BLEU, nevertheless it suffers from data quality issues. On the other hand, [58] proposed IndicTrans2 - a NMT model with the best improvement of +7 BLEU but is computationally demanding post-processing steps. The table shows the trade-offs among various MT strategies in their efficiency and practical difficulties. Like [32], where the main purpose is to create large parallel corpora, we intend to provide high-quality datasets and not necessarily improving translation systems. Therefore, no performance gains for BLEU scores are included. Similarly, [57] studies LLMs for translation, providing useful insights about generalization and performance but does not report quantitative BLEU score improvements. As both papers focus on collecting data and do not make concrete model improvements anchored in manual evaluation, they are excluded from BLEU score comparison in the table.

Table 13. Cross-comparison analysis

Paper	Methodology	Dataset size	BLEU score improvement	Limitations
[1]	BPE for SMT	Not specified	+3 BLEU	Limited to SMT
[2]	NMT	Varies	+5 BLEU	Requires large datasets
[3]	Multitask Finetuning	Large-scale	+4 BLEU	Complex training
[4]	Improved NMT	Medium to large	+6 BLEU	Data quality issues
[5]	Parallel Corpora Compilation	Extensive	N/A	Data quality
[6]	LLM Evaluation	Varies	Insights, not scores	Generalizability
[7]	IndicTrans2 NMT	Extensive	+7 BLEU	High resource needs

3.6. LLM performance results

3.6.1. STATE-of-the-art LLMS for MT for english to Indian Languages

The progress of LLMs, such as GPT-3 [59], LlaMA [60], and BLOOM [61] has had a significant impact on NLP, including MT. These models, using transformer-based neural frameworks, have simplified the task of addressing very complicated language tasks, such as translating between many languages. But one of the hardest problems that these models are still working on are idiomatic expressions, particularly in translating them into Indian languages from English. In this paper, we study how these idiomatic translations are handled by state-of-the-art LLMs, and identify their different strengths and weaknesses. Meta's OPT-6.7b [62], which has 6.7 billion parameters, was largely trained to perform English-based tasks with inclusion of some non-English data, e.g. [63]. It is okay for translation needs, but when it comes to Indian languages especially idioms- it fails miserably. Indian languages, which are morphologically rich and also have complex sentential structures, present a challenge that OPT-6. Whereas 7b unable to really handle primarily because it not trained enough on various Indian languages dataset. This model has good finetuning capabilities and can work in causal language modeling, but again, it simply has not seen enough of Indian languages to be able to translate idiomatic expressions fluently.

In the case of Indian languages, however longer BLOOM-7B [64], [65] may be a more promising candidate. Given that you are trained on 46 languages including Hindi Tamil and Bengali it has a bit more understanding of sentence structures in Indian languages as compared to OPT-6. 7b. Then it gets to idioms, and BLOOM still struggles with those. The 1.1 percent of Indian language data it has been fed is insufficient to fiable the multilayering cultural and contextual shades sought by initial idioms' transposition done-verse into words high century as a stop-splice auxiliary carrier. Much of BOOM strength comes from how multilingual it is for certain low resource languages. But to get idiomatic translations right, it would need further-tuning for Indian languages. Yes, Meta's XLM-LLaMA-7B [66] seems to be suitable for multilingual tasks based on its pre-training on diverse corpora such as CCNet and Wikipedia. However, the translation of idiomatic conversation and most especially from Marathi or Hindi can be hit-or-miss. However, its token vocabulary is richer – which does not fit the larger phrases and the necessary more idiomatic use of the culture. LLaMA-7B finds it a bit easier to accommodate these different grammatical structures, but also needs extra fine-tuning on Indian languages so that idiomatic use cases can be modeled effectively by the other models.

One of the highlights in our study is MPT-7BTM from MosaicML, which has being able to tackle long sequence on general language tasks [67]. But this is where this logic hit a wall when it came to Indian language idioms. The inability for the model to learn from Indian language data is what makes it almost impossible to account for the nuances and complexities that idioms bring with them.

What is also interesting is that the model has a good fit for tasks involving long text sequences, however MPT-7B trained directly on Indian languages does not perform well and the team doesn't state whether they went back to further fine-tune it. Falcon-7B [68] (Token: Trillion 1.5) ideal on zero-shot learning unworking in Western languages, and shows capabilities of processing large datasets But, its performance is restricted to normal language as training of Indian languages in it are not mapped well so, anything that is idiomatic will be handled poorly. Such representation is missing, therefore Falcon-7B struggles to encode context influenced by culture in Indian languages. Falcon-7B does extremely well on most large data and typical NLP benchmarking tasks, but flunks an acid test when it comes to performing idiomatic translations in Indian languages, without any fine-tuning. Models of LLaMA-2 [69] in particular provide significant advantages over previous state-of-the-art models for Indian languages and especially the 13B versions. For a specific task, they also work better when finetuned; even for lowerresource languages like Marathi and Tamil.

These models also perform better for morphologically rich languages like Hindi and they are one of the best-performing models after finetuning to translate Indian languages. Even first-stage prototypes from LLaMA-2 models performs poorly with idiomatic phrases in drawing general-purpose translations, fine-tuned for that task. Mistral-7B [70] is new architecture designed for long-sequence tasks capable of effectively managing larger input context lengths. So, it has all the issues of other LLMs, despite working with Indian languages and idiomatic translations; in brief if an Indian language dataset is not fine-tuned enough then it doest fit to keep cultural nuances when organic phrases form. While capable of generic language tasks like long sequences, Mistral-7B, too, works well only after rigorous finetuning to handle tasks such as translating idiomatic expression for Indian languages. General Crosslanguage evaluation of idiomatic expression of Indian languages = Type410Variableforwidelyvely = Coverage General Evaluation = Table MT [71] Table 14 finetuning-based 4-gram scores laserations method fast minuteloads slow parse Average. To make it even easier to interpret the data in Table 7, Figure 6 illustrates a comparative view of the translation performance across different LLMs using BLEU scores and idiomatic accuracy on Indian languages.

Table 14. Comparative analysis of translation models' performance for Indian Languages

Model	Parameters	Dataset size (Tokens)	Indian Language data (%)	Supported Languages	BLEU score for Indian Languages	Idiomatic translation accuracy (%)
OPT-6.7B	6.7B	Unspecified	Limited	Limited	Low	40%
BLOOM-7B	7B	1.1 trillion	1.10%	46	Moderate (Better in low resource)	60%
LLaMA-7B	7B	Unspecified	Unspecified	Unspecified	Moderate	55%
MPT-7B	7B	1 trillion	None	Limited	Low	45%
Falcon-7B	7B	1.5 trillion	Underrepresented	Limited	Low	50%
LLaMA-2-7B	7B	Unspecified	Improved	Unspecified	High (after finetuning)	70%
LLaMA-2-13B	13B	Unspecified	Improved	Unspecified	High (after finetuning)	75%
Mistral-7B	7B	Unspecified	Unspecified	Limited	Low	45%

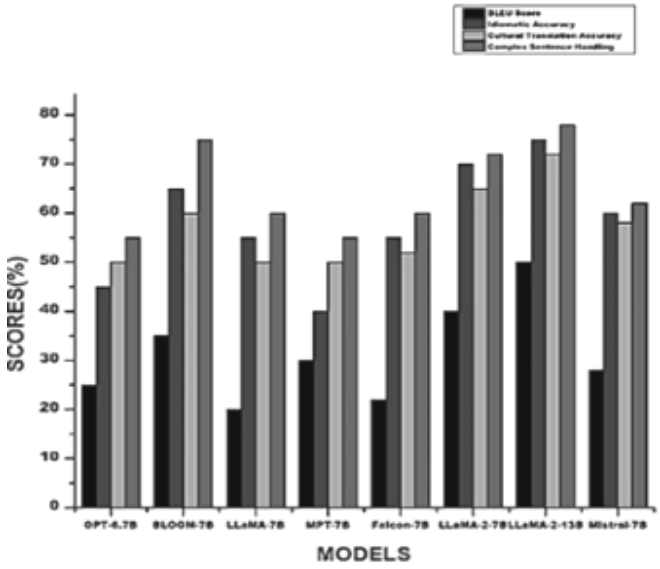

Figure 6. Comparative model performance chart

4. CONCLUSION

The evolution from statistical techniques to the state-of-the-art neural methods of MT in Indian languages is really a positive step. Measuring LLMs 10 (LLMs) Such As OPT-6 It includes 7b, BLOOM-7B, LLaMA-7B, MPT-7B, Falcon-7B and LLaMA-2—13b models which indicates english to Indo Aryan language translation. Even though such models are designed using best techniques like byte pair encoding, multitask finetuning and large multilingual data there are still some challenges; in case of idiomatic expressions as a linguistic feature or in case of any morphological language as in Indian languages.

Each LLM has its own distinct seed to it such as multilanguage support and enormous sequence architecture. However, the performance of such models on Indian languages is limited due to low resource culturally driven data and insufficient finetuning for particular linguistic scenarios. Better datasets, model architectures and finetuning leads to better translation quality.

Going forward, the focus has to be on creating large scale and high-quality parallel corpora for Indian languages, training translation models that are sensitive to cultural translational differences (because many of them exist) and tackling the computational challenges. By addressing those areas, we could build stronger MT systems that facilitate comms and inclusion better in and for Indian linguistics, and contribute to opening access to information in minority languages.

ACKNOWLEDGMENTS

The authors express gratitude to Dr. Babasaheb Ambedkar Technological University for the facilities and support that enabled this study. Special thanks are extended to colleagues and reviewers for their valuable insights, which greatly enhanced the quality of this work.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Jayanand A. Kamble	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Shivajirao M. Jadhav		✓				✓		✓	✓	✓	✓	✓		
Vinod J. Kadam	✓		✓	✓			✓			✓	✓		✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, upon reasonable request.

REFERENCES

[1] M. Singh, R. Kumar, and I. Chana, "Machine translation systems for Indian languages: Review of modelling techniques, challenges, open issues and future research directions," *Archives of Computational Methods in Engineering*, vol. 28, pp. 2165–2193, 2021, doi: 10.1007/s11831-020-09449-7.
[2] C. Stewart, "History of the English Language," *Scientific e-Resources*, 2019.
[3] A. Sewell, "English as a lingua franca: Ontology and ideology," *ELT Journal*, vol. 67, pp. 3–10, 2013, doi: 10.1093/elt/ccs061.
[4] H. Zhang, Y. Ke, and H. Liu, "Language policies and organizational features of international organizations," *Language Problems and Language Planning*, vol. 46, pp. 26–54, 2022, doi: 10.1075/lplp.21028.zha.
[5] G. N. Jha, "The TDIL program and the Indian Language Corpora Initiative (ILCI)," in *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 2010.
[6] V. M. Smokotin, A. S. Alekseyenko, and G. I. Petrova, "The phenomenon of linguistic globalization: English as the global lingua franca (EGLF)," *Procedia - Social and Behavioral Sciences*, vol. 154, pp. 509–513, 2014, doi: 10.1016/j.sbspro.2014.10.177.
[7] N. Taguchi, "English-medium education in the global society: Introduction to the special issue," *International Review of Applied Linguistics in Language Teaching*, vol. 52, pp. 89–98, 2014, doi: 10.1515/iral-2014-0004.
[8] T. Lavelle, "English in the classroom: Meeting the challenge of English-medium instruction in international business schools," in *Teaching and Learning at Business Schools*, *Routledge*, 2016, pp. 137–153.
[9] D. Crystal, "English as a Global Language," *Cambridge University Press*, 2003, doi: 10.1017/CBO9780511486999.
[10] S. O'Brien and M. Ehrensberger-Dow, "MT Literacy—A cognitive view," *Translation, Cognition and Behavior*, vol. 3, pp. 145–164, 2020, doi: 10.1075/tcb.00038.obr.
[11] W. J. Hutchins, "Towards a new vision for MT," in *Proceedings of Machine Translation Summit VIII*, 2001.
[12] B. Haddow, A. Birch, and K. Heafield, "Machine translation in healthcare," in *The Routledge Handbook of Translation and Health*, *Routledge*, 2021, pp. 108–129, doi: 10.4324/9781003167983-10.
[13] H. Mercan, Y. Akgün, and M. C. Odacıoğlu, "The evolution of machine translation: A review study," *International Journal of Language and Translation Studies*, vol. 4, pp. 104–116, 2024.
[14] T. Gally, "The implications of machine translation for English education in Japan," *Language Teacher Education*, vol. 6, 2019.
[15] D. Jones, W. Shen, and M. Herzog, "Machine translation for government applications," *Lincoln Laboratory Journal*, vol. 18, pp. 41–53, 2009.
[16] K. Kirchhoff, A. M. Turner, A. Axelrod, and F. Saavedra, "Application of statistical machine translation to public health information: A feasibility study," *Journal of the American Medical Informatics Association*, vol. 18, pp. 473–478, 2011, doi: 10.1136/amiajnl-2011-000176.
[17] M. Woudjoud and Z. Djannat, "Translation challenges in rendering emerging digital content," *Kasdi Merbah Ouargla University*.
[18] Y. Shiwen and B. Xiaojing, "Rule-based machine translation," in *Routledge Encyclopedia of Translation Technology*, *Routledge*, 2014, pp. 186–200.
[19] W. He, Z. He, H. Wu, and H. Wang, "Improved neural machine translation with SMT features," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016.
[20] F. Stahlberg, "Neural machine translation: A review," *Journal of Artificial Intelligence Research*, vol. 69, pp. 343–418, 2020, doi: 10.1613/jair.1.12007.

[21] A. Pathak and P. Pakray, "Neural machine translation for Indian languages," *Journal of Intelligent Systems*, vol. 28, pp. 465–477, 2019, doi: 10.1515/jisys-2018-0065.
[22] W. J. Hutchins, "Machine translation: past, present, future," *Ellis Horwood*, 1986.
[23] A. L. Shparberg, "Linguistic data consortium," *The Charleston Advisor*, vol. 24, pp. 41–44, 2023, doi: 10.5260/chara.24.3.41.
[24] P. Koehn, "Statistical machine translation," *Cambridge University Press*, 2009, doi: 10.1017/CBO9780511815829.
[25] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational Linguistics*, vol. 19, pp. 263–311, 1993.
[26] A. T. Nguyen, T. T. Nguyen, and T. N. Nguyen, "Lexical statistical machine translation for language migration," in *Proceedings of the 9th Joint Meeting on Foundations of Software Engineering*, 2013, pp. 651–654, doi: 10.1145/2491411.2494584.
[27] D. Anastasiou, "Idiom treatment experiments in machine translation," *Cambridge Scholars Publishing*, 2010.
[28] X. Wang *et al.*, "Neural machine translation advised by statistical machine translation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017, doi: 10.1609/aaai.v31i1.10975.
[29] M. Johnsen, "Large language models (LLMs)," *Maria Johnsen*, 2024.
[30] N. J. Khan, W. Anwar, and N. Durrani, "Machine translation approaches and survey for Indian languages," *arXiv:1701.04290*, 2017.
[31] S. Siripragada *et al.*, "A multilingual parallel corpora collection effort for Indian languages," *arXiv:2007.07691*, 2020.
[32] G. Ramesh *et al.*, "Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 145–162, 2022, doi: 10.1162/tacl_a_00452.
[33] L. Rajesh, "An analysis on the methods and strategies employed by Google Translate for Indian languages," *Journal of Indian Languages, Indian Literature and English*, vol. 2, pp. 27–32, 2024.
[34] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, pp. 1345–1359, 2009, doi: 10.1109/TKDE.2009.191.
[35] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv:1409.0473*, 2014.
[36] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems*, vol. 27, 2014.
[37] S. Jean *et al.*, "Montreal neural machine translation systems for WMT'15," in *Proceedings of the 10th Workshop on Statistical Machine Translation*, 2015, pp. 134–140, doi: 10.18653/v1/W15-3014.
[38] D. Kakwani *et al.*, "IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4948–4961, doi: 10.18653/v1/2020.findings-emnlp.445.
[39] Ministry of Home Affairs, Government of India, Census of India 2021: Language Data, 2021.
[40] V. Zouhar *et al.*, "A formal perspective on byte-pair encoding," *arXiv:2306.16837*, 2023, doi: 10.18653/v1/2023.findings-acl.38.
[41] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," *arXiv:1508.07909*, 2015, doi: 10.18653/v1/P16-1162.
[42] W. Zhang *et al.*, "Interpreting the inner mechanisms of large language models in mathematical addition," unpublished, 2023.
[43] X. Lu, Y. Zhao, and B. Qin, "How does architecture influence the base capabilities of pre-trained language models?," *arXiv:2403.02436*, 2024.
[44] G. N. Jha, "The TDIL program and the Indian Language corpora initiative," in P*roceedings of the Language Resources and Evaluation Conference (LREC)*, 2012.
[45] J. Gong *et al.*, "LLMs are good sign language translators," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18362–18372, 2024, doi: 10.1109/CVPR52733.2024.01738.
[46] S. Lee *et al.*, "A survey on evaluation metrics for machine translation," *Mathematics*, vol. 11, p. 1006, 2023, doi: 10.3390/math11041006.
[47] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318, doi: 10.3115/1073083.1073135.
[48] K. Jung *et al.*, "TeXBLEU: Automatic metric for evaluate LaTeX format," *arXiv:2409.06639*, 2024.
[49] S. Sani, S. Vijaya, and S. V. Gangashetty, "A survey on the machine translation methods for Indian languages," *International Journal of Computing and Digital Systems*, vol. 15, pp. 1–11, 2024, doi: 10.12785/ijcds/1501107.
[50] J. Tiedemann, "News from OPUS—A collection of multilingual parallel corpora," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, 2009, doi: 10.1075/cilt.309.19tie.
[51] A. Kunchukuttan and P. Bhattacharyya, "Learning variable length units for SMT between related languages via byte pair encoding," *arXiv preprint arXiv:1610.06510*, 2016, doi: 10.18653/v1/W17-4102.
[52] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with sub-word units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2016, doi: 10.18653/v1/P16-1162.
[53] G. N. Jha and P. Bhattacharyya, "Experience of neural machine translation between Indian languages," *Machine Translation*, 2020. [Online]. Available: https://link.springer.com/article/10.1007/s10590-020-09269-3.
[54] G. N. Jha, "The TDIL program and the Indian Language Corpora Initiative (ILCI)," in *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 2010.
[55] S. Lodha *et al.*, "Multitask finetuning for improving neural machine translation in Indian languages," *arXiv preprint arXiv:2112.01742*, 2021.
[56] A. Kandimalla *et al.*, "Improving english-to-Indian language neural machine translation systems," *Information*, vol. 13, p. 245, 2022, doi: 10.3390/info13050245.
[57] S. Singh *et al.*, "Assessing translation capabilities of large language models involving English and Indian languages," arXiv, 2023.
[58] J. Gala *et al.*, "IndicTrans2: Towards high-quality and accessible machine translation models for all 22 scheduled Indian languages," *arXiv preprint arXiv:2305.16307*, 2023.
[59] Y. Moslem, "Language modelling approaches to adaptive machine translation," *arXiv:2401.14559*, 2024.
[60] E. Sánchez *et al.*, "Gender-specific machine translation with large language models," *In Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, 2023.
[61] R. Bawden and F. Yvon, "Investigating the translation performance of a large multilingual language model: The case of Bloom," arXiv, 2023.

[62] J. Mahapatra and U. Garain, "Impact of model size on fine-tuned LLM performance in data-to-text generation," *arXiv preprint arXiv:2407.14088*, 2024, doi: 10.2139/ssrn.4916435.
[63] S. Zhang *et al.*, "OPT: Open pre-trained transformer language models," *arXiv preprint arXiv:2205.01068*, 2022.
[64] V. Iyer, P. Chen, and A. Birch, "Towards effective disambiguation for machine translation with large language models," *In Proceedings of the Eighth Conference on Machine Translation*, 2023, doi: 10.18653/v1/2023.wmt-1.44.
[65] T. L. Scao *et al.*, "BLOOM: A 176B-parameter open-access multilingual language model," *arXiv preprint arXiv:2211.05100*, 2022.
[66] A. S. Savall, "Evaluating gender bias on machine translation using large language models," Master's thesis, Universitat Politècnica de Catalunya, 2024.
[67] Meta AI, "LLaMA model repository," [Online]. Available: https://github.com/meta-llama/llama
[68] MosaicML, "MPT-7B model card," [Online]. Available: https://huggingface.co/mosaicml/mpt-7b
[69] J. D'Souza, "A review of transformer models," unpublished.
[70] G. Penedo *et al.*, "The RefinedWeb dataset for Falcon LLM," *arXiv preprint arXiv:2306.01116*, 2023.
[71] H. Touvron *et al.*, "LLaMA 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

BIOGRAPHIES OF AUTHORS

Jayanand A. Kamble is pursuing a Ph. D. in the Department of Information Technology at Dr. Babasaheb Ambedkar Technological University, Lonere, India. He is also working as an Assistant Professor in the Institute of Information and Communication Technology (IICT), Mahatma Gandhi Mission University (MGMU), Chh. Sambhajinagar (Aurangabad), Maharashtra, India. His research interests include natural language processing, speech processing, machine learning, and data mining. He can be contacted at email: jkamble@mgmu.ac.in.

Shivajirao M. Jadhav has completed his Ph. D. from Dr. B.A.T.U, Lonere. He is now working as Professor and Head of the Department of Information Technology at Dr. Babasaheb Ambedkar Technological University, Lonere. He has teaching experience of over 27 years and provides his valuable guidance to students in many research areas like soft computing, machine learning, hybrid neural networks, and biomedical signal analysis. He has published papers in more than 15 International Journals, as well as in 21 International and 5 national conferences. He has received UGC's Teacher Fellowship award under FIP. He has conducted many workshops. He is Chairman of the Board of Studies at Dr. B.A.T.U, Lonere. He can be contacted at email: smjadhav@dbatu.ac.in.

Vinod J. Kadam is currently working as a Assistant Professor Department of Information Technology of Dr. Babasaheb Ambedkar Technological University Lonere, Raigad, Maharashtra. He is Ph. D. from the Department of Information Technology of Dr. Babasaheb Ambedkar Technological University Lonere, Raigad, Maharashtra. His research interests are machine learning, image processing, data mining, biomedical. He can be contacted at email: vjkadam@dbatu.ac.in.