A Survey on the Machine Translation Methods for Indian Languages: Challenges, Availability, and Production of Parallel Corpora, Government Policies and Research Directions

Sudeshna Sani[1], Samudra Vijaya[2] and Suryakanth V Gangashetty[3]

[1,2,3]*Department of CSE, Koneru Lakshmaiah Education Foundation, Guntur, Andhra Pradesh, India*

Abstract: Since 1991, machine translation has been a prominent research area in India, with IIT Kanpur pioneering the original work which has since been expanded to several universities. Only 10 percent of India's 1.3 billion inhabitants can read, write and speak, English with varying degrees of competence, which makes machine translation crucial in overcoming the linguistic barrier. The Indian market for commercial products and events is greatly influenced by local languages, making the development and translation of region-based content an essential research topic nowadays. Several government-sponsored projects are being undertaken in this regard. However, there are limited sentence-aligned parallel bi-text resources available for the majority of Indian language pairs. This paper presents a detailed survey of the current trends of research on machine translation between Indian languages, along with their challenges over time. It also presents a timeline of recent research conducted and the key findings of past surveys conducted over a decade. Under a single canopy, this paper provides sources of data, the progress made in developing datasets for low-resource Indian languages and finally, new research directions.

Keywords: Machine Translation, RBMT, SMT, NMT, Low-Resource Indian languages, BLEU, METEOR, AI4Bharat, Bhashini

1. INTRODUCTION

Machine Translation (MT) is the method of automatically translation of one written human language to another, while maintaining the significance of the source text and generating fluent and proper text in the target language. MT has been developed as a subfield of Artificial Intelligence (AI) and is a part of computational linguistics and language engineering. MT techniques are further improved by utilizing concepts and methods from various fields such as statistics, computer science, AI, translation theory, and linguistics [1]. Figure 1 shows the basic structure of an MT system.

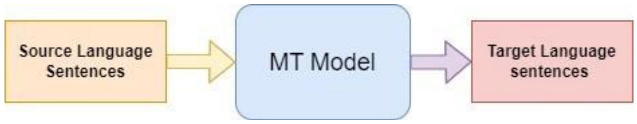


Figure 1. Diagram of a Basic Machine Translation System

Machine Translation (MT) research in Indian languages is relatively less developed as compared to other international languages such as English, Chinese, and Spanish. This is primarily due to the complexity and diversity of Indian languages, which makes MT a challenging task. Additionally, Indian languages have low resource availability, lack of parallel corpora, and limited research funding. However, in recent years, a growing MT research interest for Indian languages is observed, with several initiatives and collaborations between academia, industry, and government. Various research projects are underway to advance MT systems for Indian languages, and efforts are being made to improve the availability and quality of parallel corpora for Indian languages. Despite the challenges, MT research in Indian languages has great potential in the current global market scenario. India is a distinct country with more than 1.3 billion residents, and a growing economy with a huge demand for localization of content in regional languages. Indian languages are typically classified into five major language families [2] [3]:

- Indo-European: This family includes languages such as Bengali, Hindi, Marathi, Gujarati, Punjabi and Urdu.
- Dravidian: This family includes languages such as Tamil, Telugu, Kannada and Malayalam.

- Austroasiatic: This family includes languages such as Santali, Khasi and Mundari.
- Sino-Tibetan: Exemplar languages of this family are Manipuri, Lepcha and Bhutia.
- Andamanese: This family includes the languages spoken by the indigenous tribes of the Andaman and Nicobar Islands.

Each of these language families is further divided into numerous subgroups and dialects, reflecting the linguistic diversity of India.

India boasts a large diverse linguistic area with more than 22 official languages and over 1,600 mother-tongues [2]. However, only a small percentage of the Indian inhabitants can read, write, and speak English fluently. In the current global market scenario, where businesses and consumers operate on a global scale, language barriers can become a major obstacle for companies trying to reach out to new markets. Machine Translation (MT) technology can help bridge this gap by enabling communication in multiple languages. With the increasing importance of localization in the Indian market, there is a growing need for MT systems that can translate content from English to Indian languages and vice versa. Further, the availability of MT systems can make cross-border communication easier, faster, and more efficient, helping businesses to reach out to a wider audience and improve customer engagement. MT can also benefit government agencies, researchers, and individuals who need to communicate with people from different linguistic backgrounds. Therefore, the need for machine translation in India in the current global market scenario cannot be overstated, and efforts must be made to develop and improve MT systems to support Indian languages.

One of the significant institutions in India that have been working on Machine Translation research and development is the “Centre for Development of Advanced Computing” (CDAC) and its various centers, including the one in Pune, have been actively involved in developing MT systems for Indian languages. The CIS Department at the UoH and the IIIT in Hyderabad are also known for their research in MT for Indian languages. Additionally, the “Ministry of Communications and Information Technology” of the Government of India, via its TDIL Project, has supported the advancement of MT technologies for Indian languages. The Central Institute of Indian Languages in Mysore, the Amrita Vishwa Vidyapeetham in Coimbatore and AUKBC in Chennai are other notable institutions that have contributed to MT research in India. The efforts of these institutions are crucial for addressing the challenges and opportunities of MT for Indian languages, and for promoting the use of local languages in various domains [4] [5].

The objective of our paper is to perform a survey on the existing methods of Machine Translation for the Indian languages along with their challenges. In addition to that the key-findings from different surveys conducted on this topic are also highlighted along with current data-sources. In particular, the motivation is to answer a set of entire research questions regarding translating texts from one Indian language to another Indian language.

This paper’s contribution is divided into nine subsequent sections. Section-II describes different MT approaches suitable for Indian languages. Section-III and Section-IV contain details discussions about MT-challenges and evaluation metrics for MT-Models respectively. Section-V highlights the timeline of important surveys conducted on Machine Translation in Indian languages for last 10 years. Section-VI helps to find datasets from different sources. On the unavailability of proper data-source some methods of constructing new data-sets are discussed in section-VII. Recent encouragement from the Indian government, as well as valuable contributions from renowned Institutions, are discussed in Section-VIII which draws the direction for future research. Section-IX summarizes our work in the conclusion.

2. APPROACHES TO MT FOR INDIAN LANGUAGES

The field of MT comprises a range of techniques that are typically classified into different categories. Figure 2 displays several of these techniques and provides a timeline of their use over time.

A. Rule-based Machine Translation (RBMT)

RBMT relies on a set of human-created rules that specifies how a word or phrase in the source language should be translated into the target language. The rule set is determined by linguistic information such as morphology, vocabulary, syntax, phrase structure etc. RBMT works by matching the organization of the input sentence to that of the desired output sentence while preserving the original meaning of the input. After parsing the sentence in the source language, an transitional representation, like a parse tree or abstract representation, is generated. Figure 3 shows a general architecture of a RBMT system [6]. RBMT systems are further classified into Direct Translation, Transfer-Based Translation, and Interlingua categories based on the type of intermediate representation they use.

1) Direct Translation :

This simple method involves translating words directly from one language to another by using a bilingual dictionary, without considering the meaning or context of the source or target languages [7]. This approach can only handle one language pair at a time and is frequently unidirectional. From the late 1940s until the middle of the 1960s, the initial wave of machine translation was completely dependent on electronic or computer-readable dictionaries [8]. While this method works well for translating phrases, it is less successful when translating entire sentences.

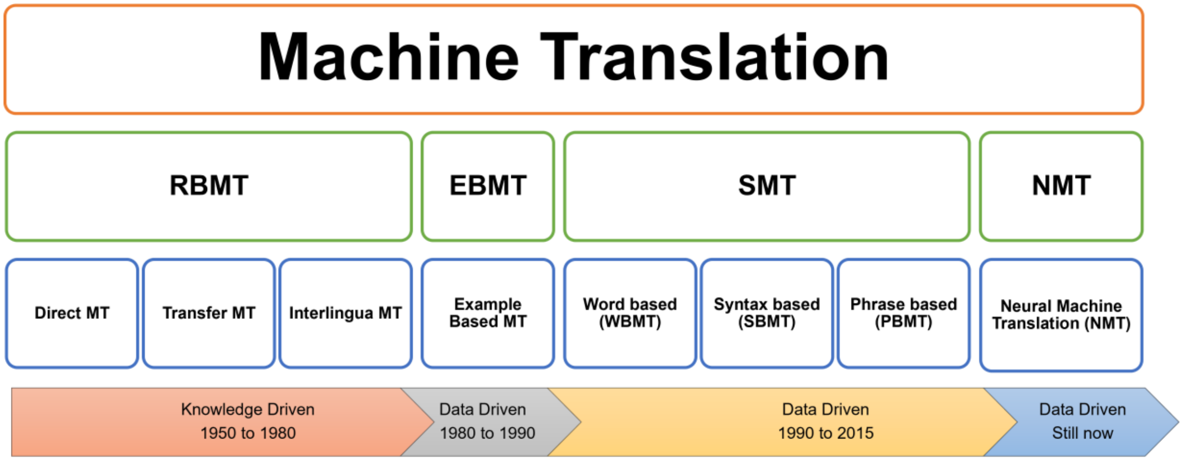


Figure 2. Approaches to Machine Translation with a Timeline

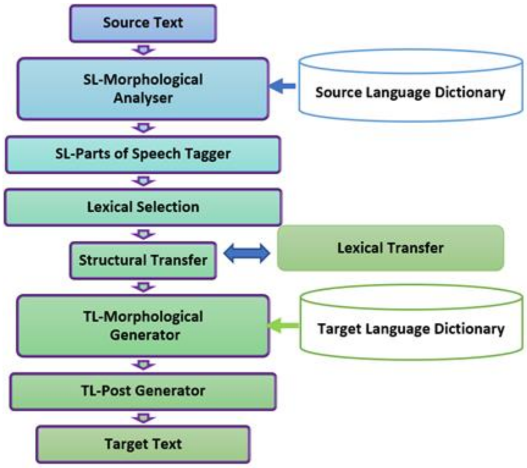

Source Text
SL-Morphological Analyser
Source Language Dictionary
SL-Parts of Speech Tagger
Lexical Selection
Structural Transfer
Lexical Transfer
TL-Morphological Generator
Target Language Dictionary
TL-Post Generator
Target Text

Figure 3. Architecture of RBMT approach

2) *Transfer Based Translation*

Transfer-based machine translation is referred to as the second generation of MT's core (mid-1960s to 1980s). Transfer-based machine translation implies translating a sentence from the input language to a pivot language, and then from that pivot language to the output language. This approach allows for the use of more advanced translation techniques and takes into account the differences between the source and destination languages. However, it has the potential to introduce errors or lose meaning in the process of translating through a pivot language [8].

3) *Interlingua Based*

The Interlingua approach to MT prioritizes semantics and pragmatics above syntax. This method achieves the translation into two phases, the first of which involves converting the Source Language (SL) into an Interlingua (IL) form. The primary benefit of the Interlingua technique is that the SL analyzer and parser is not dependent on the Target Language (TL) generation and vice versa [9].

B. *Example Based Machine Translation (EBMT)*

The EBMT system produces novel translations by extracting pertinent examples from its existing translation repository. This process encompasses three stages: matching, alignment, and recombination. In the matching phase, the system seeks comparable instances within the example base. Subsequently, in the alignment process, the pertinent segment of the example is identified and aligned with other relevant examples. Ultimately, the reusable components generated during the alignment phase are employed to formulate the translation in the target language [10].

C. *Statistical Machine translation (SMT)*

SMT method uses statistical models to learn patterns in a parallel corpus. A parallel corpus is a set of texts in two or more languages that are translations of each other. SMT system analyzes big amounts of bilingual parallel texts and forms the probabilistic model of how words, phrases, and sentences in one source language are related to the another target language. The statistical approach gained popularity recently due to the availability of large parallel corpora and the development of powerful statistical models and algorithms. The main benefit of SMT is that it can produce high-quality translations without the need for explicit linguistic knowledge or rules. Figure 4 shows the architecture of a typical SMT model. An SMT system aims to find the target sentence (comprising m words) y: y1, y2,...,ym, given a source sentence (comprising n words) x: x1, x2,...,xn, such that the conditional probability $p(y|x)$ is maximized. To achieve this, the Bayes rule is used.

$$\hat{y} = argmax_y P(y|x) = argmax_y P(x|y)P(y) \ldots\ldots\ldots\ldots\ (1)$$

P(y): a language model
$P(x|y)$: a translation model
$argmax_y$ = a decoder

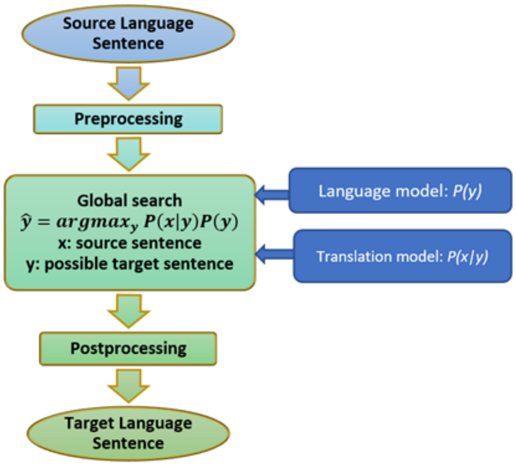


Figure 4. Architecture of a typical statistical machine translation system

The language model gets trained on monolingual data of the target language sentences to maintain the fluency. Meanwhile, the translation model gets trained on parallel corpus of the source language and target language to identify lexical correspondences between them and their probabilities. A decoder is then used to combine the information from the language and translation models, and search for the best possible translation among all possible translations [11].

D. Neural Machine Translation (NMT)

NMT is the newest form of MT modeling that has succeeded in producing more accurate translations by exploiting huge amount of parallel text corpora. It relies on neural networks and deep learning techniques to create models based on existing reference translations. NMT requires a single sequence model, which leads to increased productivity. Using conditional probability modeling, NMT models the source phrase to the target sentence, producing a context vector c.

Source phrase : $x_1, x_2, x_3, ..., x_m$
The target sentence : $y_1, y_2, y_3, ..., y_n$

$$logP(y|x) = \sum_{m=1}^{n} logP(y_k | y_{k-1}, \ldots y_1, x, c) \qquad (2)$$

$P(y|x)$ represents the likelihood of obtaining the target sentence words y given a source language word x, where c denotes the context of that specific word. The essence of NMT consists of two key elements: the ”encoder” and the ”decoder.” The encoder transforms the input texts into a context vector (c), and subsequently, the decoder processes this vector to produce single word at a time for the output sentence with a length of n. Unlike other machine translation approaches, NMT requires minimal domain expertise [12]. The encode-decoder model for NMT can be represented in a block diagram with figure 5.

3. CHALLENGES OF MT FOR INDIAN LANGUAGES:

Indian languages present a diversity of linguistic phenomena in terms of tense, gender, numbers, and other concepts. Due to structural and morphological complexity machine translation from English to Indian languages and vice versa is a challenging task. There are some challenges and problems faced during translation between ILs.

A. Syntactic Divergence

A fundamental structural distinction between English and Indian languages lies in the order of words in sentence. English follows the ’subject-verb-object’ order, but the majority of Indian languages follow the ’subject-object-verb’ order. Certain Indian languages have a trait called free word order. Sense of prepositions in Indian languages are founded on specific symbolic conjunctive words however in English phrases, prepositions plays that role [13]. In English, prepositions come before the noun or pronoun they modify, whereas in the majority of Indian languages, they come after the noun or pronouns, which are also referred to as postpositions. Table-1 shows the divergence in word-order and use of prepositions in English and some Indian languages along with transliteration and word meaning [14].

B. Morphological Divergence

The field of morphology investigates the inner composition of words and their ability to take on unique shapes within different types of texts. The recognition, analysis, and description of morphemes as well as other linguistic constructions like words, affixes, and parts of speech are collectively referred to as ”morphology” in the study of language. The term ”morpheme” alludes to the lowest semantically significant item in a language. Words in the Indian language vary in terms of lemma, person, number, gender, case, tense, aspect, and modality. Languages with poor morphology typically use word order and syntax to convey various meanings. As a result, these languages have a smaller lexicon than languages with a rich morphological structure. Richer languages have more nuanced words that accurately communicate various meanings, which increases the language’s complexity. Hebrew, Turkish, Dravidian languages, and other languages are thought to be morphologically rich, whereas English, Mandarin, and other languages are thought to be morphologically poor. Due to a bigger vocabulary, sparser data, and increased complexity, morphologically rich languages are more difficult for neural networks to model than poor ones. The Stochastic Morph Analyzer (SMA) is a Morph Analyzer that forecasts the morph information using machine learning [15] [16]. In India, Dravidian languages such as Telugu and Tamil exhibit greater morphological complexity compared to Indo-Aryan languages like Hindi, Punjabi, and Gujarati. Translating text into Dravidian languages like Telugu, Tamil, and Malayalam often yields lower BLEU scores, whereas translations into Indo-Aryan languages like Hindi, Gujarati, Punjabi,

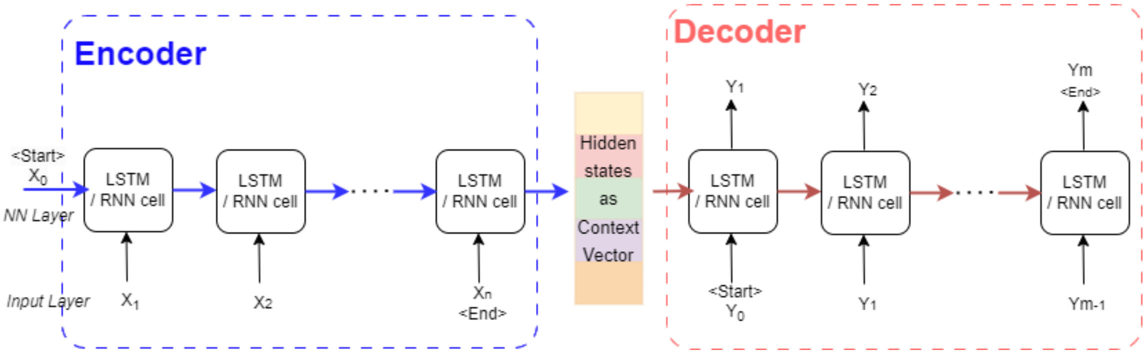
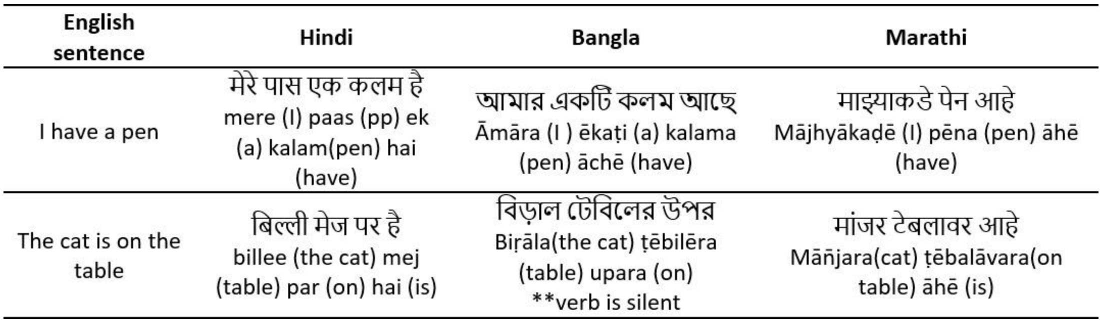


Figure 5. A General Encoder-Decoder Model



English sentence	Hindi	Bangla	Marathi
I have a pen	मेरे पास एक कलम है mere (I) paas (pp) ek (a) kalam(pen) hai (have)	আমার একটি কলম আছে Āmāra (I) ēkaṭi (a) kalama (pen) āchē (have)	माझ्याकडे पेन आहे Mājhyākaḍē (I) pēna (pen) āhē (have)
The cat is on the table	बिल्ली मेज पर है billee (the cat) mej (table) par (on) hai (is)	বিড়াল টেবিলের উপর Biṛāla(the cat) ṭēbilēra (table) upara (on) **verb is silent	मांजर टेबलावर आहे Māñjara(cat) ṭēbalāvara(on table) āhē (is)

Figure 6. Word-order divergence in between Indian languages

and Bengali tend to achieve relatively higher BLEU scores. A larger number of distinct words can be found in the richer languages within a multilingual parallel corpus. Morphological complexity can be measured by Type-Token ratio. Here is the increasing order of morphological complexity for different languages:- Hindi<Punjabi<Gujarati<Tamil<Telugu [17].

C. Data scarcity

Building of Corpus can be expensive for users with limited resources. When the word order is significantly diverse between two languages, statistical machine translation struggles. NMT does not come up to the mark for morphologically diverse languages.

D. Interpreting the intentions of speakers is challenging

Depending on the speaker's aim (such as sarcasm, sentiment, metaphor, etc.), phrases or words might have many interpretations.

E. Code-mixed language

Processing code-mixed language is difficult because users often utilize numerous languages in a single statement or utterance. E.g.: User tweet : "Hi friends, keyse ho? Ayo chill kare."

F. Idioms

Sometimes idioms may not be interpreted idiomatically. Indian regional languages are rich with idioms.

4. EVALUATION METRICS OF MT-ALGORITHMS

To measure the goodness of a MT-model several metrics such as BLEU, METEOR, ROUGE, TER, NIST etc. are available for automatic evaluation. Evaluation metrics can be categorized into 2 types, Intrinsic Evaluation and Extrinsic Evaluation.Both intrinsic and extrinsic evaluation metrics are focused on the performance of the final objective, which is the performance of the NLP component on the entire application, whereas intrinsic evaluation metrics are more concerned with intermediate objectives, such as how well an NLP component performs on a specified subtask. We discussed some common intrinsic evaluation metrics used for MT systems.

A. Bilingual Evaluation Understudy (BLEU)

The BLEU metric calculates the score by comparing n-grams of the candidate translation of text to one or more n-grams reference translations. The BLEU metric ranges from 0 to 1. A score of 1.0 denotes a perfect match, whereas a score of 0.0 denotes a perfect mismatch. Sometimes BLEU score is expressed as a percentage rather than a decimal between 0 and 1. The following interpretation of BLEU scores (expressed as percentages rather than decimals) is followed in general [18].

The provided color gradient can serve as a broad representation of the BLEU score on a scale.

TABLE I. Interpretation of BLEU scores in percentage

BLEU Score	Interpreted as
Less than 10	Not useful
10 to 19	It’s hard to obtain the meaning
20 to 29	The sense is clear, but it has large grammatical errors
30 to 40	Translations quality is good
40 to 50	Translations quality is high
50 to 60	Very high-quality, acceptable, and smooth translations
Greater than 60	Quality is quite acceptable than human-efforts

Figure 7. BLEU Score Table

It is the most widely accepted, inexpensive and easily understandable metric.

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

METEOR is based on the unigram matching and calculated by the harmonic mean of precision and recall. The recall is higher weighted than precision. It overcomes some of the drawbacks of the BLEU score, as because it can perform stemming- and synonymity matching, as well as standard exact word-matching [19]. This is a perfect metric for Machine translation. Once the final alignment is computed, the score of Unigram precision P and Unigram Recall R is calculated as:

$$P = \frac{m}{w_t} \quad R = \frac{m}{w_r} \tag{3}$$

where m = no. of unigrams in the observed translation that are also available in the reference translation, w_t = no of unigrams in the observed translations, w_r = no of unigrams in the reference translations. The harmonic mean (F) is calculated as :

$$F_{mean} = \frac{10PR}{(R + 9P)} \tag{4}$$

C. *Recall-Oriented Understudy for Gisting Evaluation (ROUGE)*

There is a set of metrics and software available for analysing automatic summarization and machine translation software in NLP. It basically measures the “recall”. It is used in machine translation projects to assess the quality of the text that is produced [20].

D. *Translation Error Rate (TER)*

TER quantifies the number of editing operations needed to align a translated segment with a reference translation. TER score ranges from 0% to 100%. The quality of the translation improves with decreasing TER scores. A higher BLEU or METEOR score, on the other hand, indicates better translation quality. A better MT system achieves higher BLEU scores with lower CDER, TER and PER scores [20] [21].

1) *National Institute for Standards and Technology (NIST) from US*

It is based on BLEU metric with some features. The n-gram precision calculation is differently taken. In contrast to BLEU, which assigns equal weight to all n-grams, NIST takes into account the relevance of each n-gram. It assigns higher weight to n-grams that are considered less likely to occur [22]. Metrics for automatic evaluation are quick, tuneable, affordable, and require less human labour. But these automatic evaluation metrics are not adequate for evaluating MT systems in Indian languages. Due to the many intricacies involved with Indian languages, they will not generate reliable results, but same measures produce excellent evaluation results for Non-Indic western languages. For evaluating the quality of translated phrases, human evaluation metrics are preferred for particularly morphologically rich languages, despite being time-consuming and costly. Human evaluation entails bilingual expertise in both the source and target languages, offering a level of consistency often deemed superior to automatic translation assessments [20].

5. RECENT MT RESEARCH FOR INDIAN LANGUAGES

In this section we highlight important research work done for Indian languages with a focus on low-resource languages.

Jindal et al. 2018 used SMT based MT model for translation between English and Punjabi using three sets of parallel-sentence corpus achieving 0.8767 BLUE score [23].

Mahata et al. 2018 implemented RNN encoder-decoder architecture to improve the quality of translation done by traditional SMT. English-Hindi parallel corpus from MTIL2017 was used as dataset to analyse the scores of phrase-pairs by a comparative experiment between two models. It was found that SMT performs fine for long sentences and NMT performs well for short sentences [24].

Pathak et al. 2019 exploited OpenNMT system architecture for English to Punjabi, English to Tamil, and English to Hindi translations. They observed the betterment of performance of NMT model with the growth in the training data and length of test sentences [25]

Dewangan et al. 2021 worked for Indian Language NMT using one of the popular subword methods i.e., BPE based NMT model. They used ILCI dataset to derive BLEU scores for different pairs of languages . The authors proposed a data augmentation technique which combined NMT and SMT [26].

Laskar et al. 2021 participated in Workshop on Asian Translation 2021 multimodal translation task of English to Hindi. An investigation was done for phrase pairs through data augmentation approach in both multimodal and text-only NMT. The results were evaluated by BLUE, Rank-based Intuitive Bilingual Evaluation (RIBES), and Adequacy Fluency Metrics (AMFM) which scored better than the previous works [27].

A Chowdhury et al. 2022 used Transfer Learning approach for translation between a low-resource Indian language called Lambani and other Indian languages. The BLEU score was improved when the TL was used and the authors have observed that freezing the initial layers of the TL model improved the BLUE score further [28].

Some important points from past recent surveys on Machine translation in Indian languages have been summarized in the following Table II

6. AVAILABILITY OF DATASET

This section discusses some open-source datasets for the automatic translation between Indian languages. A parallel text corpus is comprised of pairs of sentences, one in source language and another in target language and the meaning of the both sentences are same.

A. The EMILLE Corpus

The EMILLE (Enabling Minority Language Engineering) Corpus was created by collaboration among the CIIL, Mysore, India, Lancaster University, UK. The corpus is made up of three parts: parallel, monolingual, and annotated corpora. The fourteen monolingual corpora for fourteen south Asian languages are Bengali, Assamese, Hindi, Gujarati, Malayalam, Telegu, Kannada, Tamil, Kashmiri, Punjabi, Marathi, Oriya, Sinhala, and Urdu. They contain written and (for some languages) spoken data. The EMILLE/CIIL Corpus (ELRA-W0037) is provided without charge for use in exclusively non-commercial research [35].

B. IJCNLP-2008 data set

This dataset was developed for the Named Entity Recognition (NER) challenge in a workshop about NER for South and South East Asian languages which was hosted by IIIT, Hyderabad. It included Hindi, Bengali, Oriya, Telugu, and Urdu databases [36].

C. Tatoeba

The Tab-delimited Bilingual Sentence Pairs datasets are created by Tatoeba project by compiling statements from many languages. They paid particular attention to the creation of numerous linguistic datasets that included translations of sentences in various low-resource languages. Many low-resource language to English translation can be done using this dataset. The tab key serves as a line between the original and translated sentences. Each dataset contains at least 100 sentences and their translations [37]. Table III highlights a few sample snapshots of the accessible data sources.

D. Anuvaad

It is an open-source platform for translating court papers at scale in the judicial sector. Supreme Courts of India (SUVAS) and Bangladesh (Supreme Court) have separate Anuvaad instances deployed (Amar Vasha). Now Anuvaad have high quality NMT models for nine Indian languages [38] [39].

E. AI4Bharat

AI4Bharat is the recent initiative of IIT Madras. It aims on building a rich open-source language AI system for Indian languages, including datasets, models, and applications. Samanantar is an extensive parallel corpus collection for Indic languages that is accessible to the public [40] [41].

F. Mann ki Baat

"Mann Ki Baat" – is a monthly program of All India Radio in which the Prime Minister of India speaks and addresses the citizens in Hindi language. Later the speech is converted to different other Indian languages. The Textual Data or Parallel corpus for Indian languages can be mined from multilingual articles called ”CVIT Mann Ki Baat” [42] [43] [44].

7. INITIATIVE OF CONSTRUCTING PARALLEL CORPORA

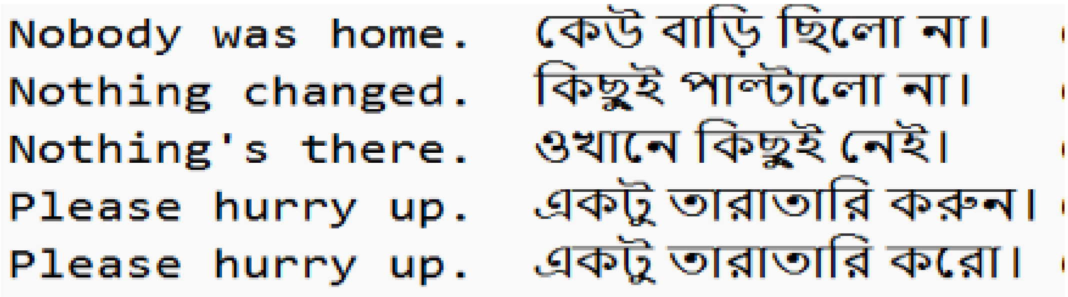
Indic languages often have an abundance of monolingual corpora but a scarcity of parallel corpora, making it challenging to apply machine-engineered techniques for dataset creation. The following are some of the reasons that make the creating parallel data a difficult task:

1) Many data are not in digital format. Some of them are either in PDF files or in image format;
2) Texts are not in Unicode. they use proprietary font formats;
3) Many datasets are not in format that can be directly used for MT. The incomplete sentence, invalid

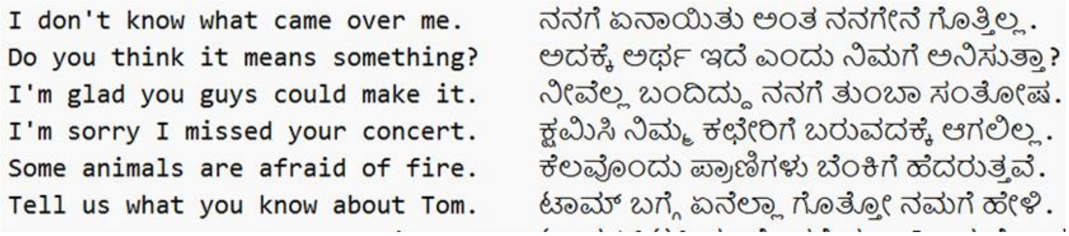
TABLE II. Key points of past few surveys on ML in Indian Languages

Year	Key observations and limitations	Ref No.
2015	Transferred-Based approach is more flexible. Most of work has been done in Aryan languages. Dravidian languages are yet to be explored.	[29]
2018	Automatic Performance metrics for MT algorithms are not adequate. Human Evaluation metrics are suitable for Indian Languages. Existing systems performance is not satisfactory.	[30]
2018	Machine Translations are carried out between English and Indian languages, with the exception of Google Translator.	[31]
2019	The USA leads the world in MT research followed by Japan, China. India is still now in the infancy stage of MT due to it’s language-diversity. MT-research can be improved by govt. policies for the benefit of society.	[32]
2019	Low-resource languages should be focused more for future studies in terms of the availability of data-sources, translation methods, and challenges for translation.	[33]
2020	Hybrid and NMT methods show better performance as compared to other techniques.	[34]
2021	SMT performs well for translation among Indo-Aryan family, but is poor for Dravidian family.	[26]

This data is from tatoeba project
Link : ”http://tatoeba.org/files/downloads/sentences$_d$*etailed.csv*”
Date of this file: 2022-09-06



Nobody was home.	কেউ বাড়ি ছিলো না।
Nothing changed.	কিছুই পাল্টালো না।
Nothing's there.	ওখানে কিছুই নেই।
Please hurry up.	একটু তারাতারি করুন।
Please hurry up.	একটু তারাতারি করো।



I don't know what came over me.	ನನಗೆ ಏನಾಯಿತು ಅಂತ ನನಗೇನೆ ಗೊತ್ತಿಲ್ಲ.
Do you think it means something?	ಅದಕ್ಕೆ ಅರ್ಥ ಇದೆ ಎಂದು ನಿಮಗೆ ಅನಿಸುತ್ತಾ?
I'm glad you guys could make it.	ನೀವೆಲ್ಲ ಬಂದಿದ್ದು ನನಗೆ ತುಂಬಾ ಸಂತೋಷ.
I'm sorry I missed your concert.	ಕ್ಷಮಿಸಿ ನಿಮ್ಮ ಕಛೇರಿಗೆ ಬರುವದಕ್ಕೆ ಆಗಲಿಲ್ಲ.
Some animals are afraid of fire.	ಕೆಲವೊಂದು ಪ್ರಾಣಿಗಳು ಬೆಂಕಿಗೆ ಹೆದರುತ್ತವೆ.
Tell us what you know about Tom.	ಟಾಮ್ ಬಗ್ಗೆ ಏನೆಲ್ಲಾ ಗೊತ್ತೋ ನಮಗೆ ಹೇಳಿ.

TABLE III. Example dataset snap of sentence pairs from the Tatoeba Project

character sequence, spell errors, mixed with other language etc. create immature dataset for machine translation.

Thus, in order to construct machine translation systems for Indic languages, it is imperative to either create synthetic parallel corpora or use language models in the system's training.

Steps to create Bilingual Parallel corpora:

1) Selection of the Source and the Target Language
2) Collection of source and target texts from books, newspapers, websites and other documents.
3) Preprocessing: cleaning errors, formatting, and extraneous characters.
4) Alignment of source and their corresponding target texts by different automated tools (Bluealign, Giza++, Ugarit) [45]
5) Annotation: After alignment, the parallel corpus needs to be annotated with metadata such as a sentence or phrase-level information, part-of-speech tags, named entities, and other linguistic features.
6) Quality control: Finally, the parallel corpus needs to be checked for quality control to ensure accuracy and consistency in translations.

Under the project MTIL-2017 Shared Task an initiative was taken by M. Anand Kumar et. al to develop parallel corpora between English and Indian languages in September 2017 by conducting a shared task among 29 teams of people. The team worked with Hindi, Tamil, Malayalam, and Punjabi languages and employed Neural Network based system. The output evaluation was done by human beings [46].

Philip et al. [47] built a standard NMT system, a retrieval module, and an alignment module make up the iterative alignment pipeline. This pipeline is used to interact with publicly accessible websites, such as government news releases. As more articles are published to PIB and additional tools are put in place to gather more sentences, the corpus will undoubtedly grow in size.

8. INDIAN GOVT. ENCOURAGEMENT AND FUTURE SCOPE OF MT

The following 22 languages are listed in the Constitution's Eighth Schedule: (1) Assamese, (2) Bengali, (3) Gujarati, (4) Hindi, (5) Kannada, (6) Kashmiri, (7) Konkani, (8) Malayalam, (9) Manipuri, (10) Marathi, (11) Nepali, (12) Oriya, (13) Punjabi, and (14) Sanskrit are among the other languages. (15) Sindhi, (16) Tamil, (17) Telugu, and(18) Urdu, (19) Bodo, (20)Santhali, (21) Maithili, and (22) Dogri are among the 19th and 22nd groups. The Central Institute of Indian Languages (CIIL), the Ministry of Human Resource Development's (MHRD) nodal organisation, is responsible for the promotion and preservation of Indian languages. In Mysore, Karnataka, the CIIL was established to oversee the development of Indian languages [48]. Some newer projects of the CIIL are:

- New Language Survey of India (NLSI).
- LDC-IL.
- National Translation Service.
- Development and promotion of minor Indian languages.
- Development of Pali.
- National Testing Mission.

To lower the barriers to communication, various organisations in India are supporting the adoption and integration of MT technologies and programmes. India is positioned to experience tremendous growth in the international IT sector with the launch of the government's ”Digital India” plan. Initiatives like Digital India promise to provide plenty of chances for national and international businesses to broaden and deepen their penetration into Indian markets.

A. ILCI

The Indian Languages Corpora Initiative (ILCI), a massive effort started by the Indian government, aims to compile parallel annotated corpora in each of the 17 languages listed in the Indian Constitution. ILCI project aims to provide a common language platform by developing parallel annotated corpora in the tourism and health sectors in 11 Indian languages, with Hindi serving as the source language. The project's primary goal is to create an annotated parallel corpus from source Hindi to Indian languages with English [26].

B. C-DAC

C-DAC is a research and development organization that operates under the MeitY of the Government of India. Its mission is to develop tools for multilingual translation and methods to bridge the gap between Indian languages due to the country's multilingual nature. C-DAC provides users with access to these resources for their research projects. Additionally, it offers dictionaries and corpora for Indian languages, among other resources [49].

C. TDIL

The Government of India's Meity initiated the Technology Development for Indian Languages (TDIL) Program. The primary objective is to facilitate the creation and accessibility of multilingual knowledge resources. The program also strives to develop tools and techniques for information processing, fostering human-machine interaction devoid of language barriers. An additional goal involves the integration of these advancements to craft innovative user products and services. The program also actively participates in national and international standardization bodies such as UNICODE, ISO, the World Wide Web Consortium (W3C), and BIS (Bureau of Indian Standards) to promote language technology standardization and ensure appropriate description of Indian languages in current and future standards [4].

Though research in MT for Indian languages has grown tremendously since past one decade, still some facts are there which is still unexplored, such as Code-mixed IL processing, Opinion mining, sarcasm translation, idioms extraction for Indian languages.

9. CONCLUSION

In this paper, we projected some light on the previous works related to Machine translation for Indian languages by keeping in mind the rising demand for research in the multilingual translation process of India. We presented a systematic as well as comprehensive review of the different methods of MT for Indic languages and the challenges faced by other researchers in this regard. To establish a rigorous evaluation process, this review engages in an in-depth exploration of various evaluation metrics employed in the domain of machine translation. We have also enriched this paper with the most recent references of a detailed source of available datasets, The importance of parallel corpora is crucial for MT research in India. Yet, it has been noted that there are still no suitable techniques for producing parallel corpora datasets. We also provided some insight into earlier attempts made in this area. Finally, there are many opportunities for machine translation research in India because to Indian government's strong encouragement and assistance.

References

[1] R. A. Sinhal, Dept. of CSE Shri Ramdeobaba College of Engineering and Management Nagpur, and K. O. Gupta, "Machine translation approaches and design aspects," *IOSR J. Comput. Eng.*, vol. 16, no. 1, pp. 22–25, 2014.

[2] Department of Higher Education, "Language Education," https://www.education.gov.in/sites/upload files/mhrd/files/upload _ document/languagebr.pdf, Data Collected on 29-11-2023.

[3] Ministry of Home Affairs, Government of India, "Report of the Committee of Parliament on Official Language," https://rajbhasha.gov.in/sites/default/files/cpol7threporteng.pdf, 2005.

[4] Ministry of Electronics Information Technology, Govt. of India, "Final draft standard on machine translation acceptance. Version 4.0." https://tdil-dc.in/index.php?option=com rff article&task=view-article&article id=11&lang=en, Data Collected on 29-11-23.

[5] G. N. Jha, "The tdil program and the indian language corpora initiative (ilci)," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2010.

[6] N. Bhadwal, P. Agrawal, and V. Madaan, "A machine translation system from hindi to sanskrit language using rule based approach," *Scalable Computing: Practice and Experience*, vol. 21, pp. 543–554, 2020.

[7] R. Sankaravelayuthan and G. Vasuki, "English to tamil machine translation system using parallel corpus," *http://www.languageinindia.com/*, 2013, unpublished Manuscript.

[8] V. S.-C. Yang, "Electronic dictionaries in machine translation," in *Encyclopedia of Library and Information Science*, A. Kent and et al., Eds. New York: Marcel Dekker, 1991, vol. 48, no. Suppl. 11, pp. 74–92.

[9] A. Chatterjee, *Elements of Information Organization and Dissemination*, 2017, doi:10.1016/B978-0-08-102025-8.00025-9.

[10] K. M. Anwarus Salam, M. Khan, and T. Nishino, "Example based english-bengali machine translation using wordnet," in *Conference Proceedings - Centre for Research on Bangla Language Processing*. BRAC University, 2010.

[11] M. Mumin, M. Hanif, M. Iqbal, and M. J. Islam, "shu-torjoma: An englishbangla statistical machine translation system," *Journal of Computer Science*, vol. 15, pp. 1022–1039, 08 2019.

[12] S. Sharma and M. Diwakar, "Machine translation for indian languages utilizing recurrent neural networks and attention," in *Distributed Computing and Optimization Techniques*, ser. Lecture Notes in Electrical Engineering, S. Majhi, R. P. d. Prado, and C. Dasanapura Nanjundaiah, Eds., vol. 903. Springer, Singapore, 2022.

[13] M. D. Okpor, "Machine translation approaches: Issues and challenges," *International Journal of Computer Science Issues (IJCSI)*, vol. 11, no. 5, p. 159, 2014.

[14] R. N. Patel, P. B. Pimpale, and M. Sasikumar, "Machine translation in indian languages: Challenges and resolution," *Journal of Intelligent Systems*, vol. 28, no. 3, pp. 437–445, 2019.

[15] S. Srirampur, R. Chandibhamar, and R. Mamidi, "Statistical morph analyzer (sma++) for indian languages," 2014, pp. 103–109.

[16] A. Bharati, R. Sangal, S. Bendre, P. Kumar, and Aishwarya, "Unsupervised improvement of morphological analyzer for inflectionally rich languages." 01 2001, pp. 685–692.

[17] A. Tanwar and P. Majumder, "Translating morphologically rich indian languages under zero-resource conditions," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 19, no. 6, p. Article 85, November 2020.

[18] G. Cloud. (current year) Automl api documentation. Accessed: dd Month yyyy. [Online]. Available: https://cloud.google.com/translate/automl/docs/evaluate

[19] S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005, pp. 65–72.

[20] A. Kandimalla, P. Lohar, K. Maji, and A. Way, "Improving english-to-indian language neural machine translation systems," *Information*, vol. 13, p. 245, 2022.

[21] P. Madaan and F. Sadat, "Multilingual neural machine translation involving indian languages," in *Proceedings of the WILDRE5–5th Workshop on Indian Language Data: Resources and Evaluation*. Marseille, France: European Language Resources Association (ELRA), 2020, pp. 29–32.

[22] G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," in *Proceedings of the Second International Conference on Human Language Technology Research*. San Diego, California: Morgan Kaufmann Publishers Inc., 2002, pp. 138–145.

[23] S. Jindal, V. Goyal, and J. S. Bhullar, “English to punjabi statistical machine translation using moses (corpus based),” *Journal of Statistics and Management Systems*, vol. 21, no. 4, pp. 553–560, 2018.

[24] S. K. Mahata, D. Das, and S. Bandyopadhyay, “MTIL2017: Machine translation using recurrent neural network on statistical machine translation,” *Journal of Intelligent Systems*, vol. 28, no. 3, pp. 447–453, 2019.

[25] A. Pathak and P. Pakray, “Neural machine translation for indian languages,” *Journal of Intelligent Systems*, vol. 28, no. 3, pp. 465–477, 2019.

[26] S. Dewangan, S. Alva, and e. a. Joshi, Nisarg, “Experience of neural machine translation between indian languages,” *Machine Translation*, vol. 35, pp. 71–99, 2021.

[27] S. R. Laskar, A. F. U. R. Khilji, D. Kaushik, P. Pakray, and S. Bandyopadhyay, “Improved English to Hindi multimodal neural machine translation,” in *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 155–160.

[28] A. Chowdhury, D. K. T., S. V. K, and S. R. Mahadeva Prasanna, “Machine translation for a very low-resource language - layer freezing approach on transfer learning,” in *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*. Gyeongju, Republic of Korea: Association for Computational Linguistics, 2022, pp. 48–55.

[29] S. Saini and V. Sahula, “A survey of machine translation techniques and systems for indian languages,” in *2015 IEEE International Conference on Computational Intelligence* & *Communication Technology*, 2015.

[30] D. Chopra, N. Joshi, and I. Mathur, “A review on machine translation in indian languages,” *Eng. Technol. Appl. Sci. Res.*, vol. 8, no. 5, pp. 3475–3478, October 2018.

[31] N. Kharate, “Survey of machine translation for indian languages to english and its approaches,” *3*, pp. 613–623, 2018.

[32] B. M. Gupta and S. Dhawan, “Machine translation research: A scientometric assessment of global publications output during 2007-16,” *DESIDOC Journal of Library* & *Information Technology*, vol. 39, pp. 31–38, 2019.

[33] B. S. Harish and R. K. Rangan, “A comprehensive survey on indian regional language processing,” *SN Applied Sciences*, vol. 2, 2020.

[34] 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*, 2020.

[35] “Emille dataset,” https://www.lancaster.ac.uk/fass/projects/corpus/emille/, 2003.

[36] “Ijcnlp dataset,” http://ltrc.iiit.ac.in/ner-ssea-08/, 2008.

[37] “Manythings.org dataset,” https://www.manythings.org/anki/, 2015.

[38] “Anuvad dataset,” https://www.anuvaad.org/.

[39] “Anuvad parallel corpus,” https://github.com/project-anuvaad/anuvaad-parallel-corpus.

[40] “Samanantar parallel corpus,” https://ai4bharat.org/samanantar.

[41] G. Ramesh, S. Doddapaneni, A. Bheemaraj, M. Jobanputra, A. Raghavan, A. Sharma, S. Sahoo, H. Diddee, M. J, D. Kakwani, N. Kumar, A. Pradeep, S. Nagaraj, K. Deepak, V. Raghavan, A. Kunchukuttan, P. Kumar, and M. S. Khapra, “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.

[42] G. of India. (Accesed on : 30/12/2023) Mann ki baat. [Online]. Available: https://www.pmindia.gov.in/en/mann-ki-baat/

[43] A. Project, “Indicnlp catalog,” https://github.com/AI4Bharat/indicnlp catalog.

[44] T. Tiwari. (Year of access) Pm india mann ki baat. [Online]. Available: https://www.kaggle.com/datasets/taruntiwarihp/pm-india-mann-ki-baat

[45] A. Sati, “Word alignment using giza++ and cygwin on windows,” *International Journal of Engineering Research* & *Technology (IJERT)*, vol. 02, no. 05, May 2013.

[46] M. A. Kumar, B. Premjith, S. Singh, S. Rajendran, and K. P. Soman, “An overview of the shared task on machine translation in indian languages (mtil) – 2017,” *Journal of Intelligent Systems*, vol. 28, no. 3, pp. 455–464, 2019.

[47] J. Philip, S. Siripragada, V. P. Namboodiri, and C. V. Jawahar, “Revisiting low resource status of indian languages in machine translation,” in *Proceedings of the 3rd ACM India Joint International Conference on Data Science* & *Management of Data (8th ACM IKDD CODS* & *26th COMAD)*, January 2021.

[48] “Central institute of indian languages (ciil),” https://www.ciil.org/default.aspx.

[49] “Cdac dataset,” https://www.cdac.in/index.aspx?id=products _ services, 2003.

Sudeshna Sani was born in 1984, received her Diploma in Computer Science and Technology from WBSCTE, AMIE (in CSE) degree from The Institution of Engineers (India), Kolkata and thereafter M.Tech. in CSE from MAKAUT, West Bengal, India in the year of 2005, 2016 and 2019 respectively. Her project work was based on the field of Sentiment Analysis through Deep learning. Presently, she is working as an Assistant Professor in the Dept of Computer Science and Engineering at Koneru Lakshmaiah Education Foundation, Andhra Pradesh. Before that she worked as Lecturer in the Department of Computer Science and Technology at Technique Polytechnic Institute, Hooghly under WBSCTE, India. She also carries 8 years’ experience as Technical Staff in different Engineering Colleges. She is a lifetime member of the Indian Society for Technical Education, New Delhi. Her research interest is focused on NLP, speech and text processing through machine learning and deep learning-based systems.

Author 2 Name short biography ….

Dr. Suryakanth V Gangashetty is a Professor in the Department of Computer Science and Engineering at KLEF (KL University) Vaddeswaram, Guntur District, Andhra Pradesh, India. He completed his PhD (in Neural Network Models for Recognition of Consonant-Vowel Units of Speech in Multiple Languages) from IIT Madras in 2005. Before joining KLEF Vaddeswaram, he worked as a member of faculty at IIIT Hyderabad, Telangana, from 2006 to 2020. Previously he has worked as a Senior Project Officer at Speech and Vision Laboratory, IIT Madras. He has worked as a member of faculty at BIET Davangere Karnataka, from 1991 to 1999. He has also worked as a visiting research scholar at OGI Portland (USA) for three months during the summer of 2001. He has done his post-doctoral studies (PDF) at Carnegie Mellon University (CMU) Pittsburgh (PA, USA) during April 2007 to July 2008. He is an author of about 180 papers published in national as well as international journals, conferences, and edited volumes. He is a life member of the CSI, IE, IUPRAI, ASI, IETE, ORSI, and ISTE. He has reviewed papers for reputed journals and conferences.

Dr. Suryakanth V Gangashetty has about 24 years of experience working with speech processing technologies. He has participated in Speaker Identification, Speaker Verification, Anti Spoofing challenge and Text to speech synthesis benchmarks. He has worked on Speech Signal Processing for a Large Vocabulary Continuous Speech Recognition, while at CMU Pittsburgh (USA). He has been part of ASR and TTS projects funded by MeitY. He has undertaken many projects sponsored by the Indian Government (such as MHRD, TDIL) and the IT Industry (such as Samsung). He has also executed International Collaborative projects (UKIERI) in the area of speech processing. He has guided PhD scholars in the areas of Speech recognition, Speech synthesis, Voice conversion, Language identification. Speaker recognition, Speech enhancement, Audio scene classification. Code mixed speech processing, Dialect identification. Emotion recognition and prosody processing. He has experience in the development of speech to speech translation system for the languages English, Hindi, and Telugu. His research interests include Speech Processing, Neural Networks, Machine Learning, Natural Language Processing, Artificial Intelligence. He was local Organizing Chair for the INTERSPEECH-2018 conference which happened in India in September 2018 held at Hyderabad.