International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 03 | March - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Enhancing En-X Translation: A Chrome Extension-Based Approach to Indic Language Models

Prof. Kirti Rane[1], Tanaya Bagwe[2], Shruti Chaudhari[3], Ankita Kale[4], Gayatri Deore[5].

[1] *Professor, Department of Information Technology Engineering, Datta Meghe College of Engineering, Airoli, Navi Mumbai, Maharashtra, India.*

[2345] *Students, Department of Information Technology Engineering, Datta Meghe College of Engineering, Airoli, Navi Mumbai, Maharashtra, India.*

ABSTRACT

Language translation is the lifeblood of any communication that crosses linguistic boundaries. Recent trends in the domain of neural machine translation (NMT) are already superior to the old traditions. In such circumstances, the works done by Prahwini et al. (2024) and Vandan Mujadia et al. (2024) highlight the application of NMT for resource-constrained Indian languages. In view of many challenges like parallel corpus scarcity, we present a real-time adaptable translation model that works on the Fairseq framework. It provides high-accuracy translations for Assamese, Gujarati, Kannada, Bengali, Telugu, Tamil, Malayalam, Marathi, Hindi, Oriya, and Punjabi, thereby aiding in accessibility as well as communication.

KEYWORDS: Machine Translation (MT), Neural Machine Translation (NMT), Indian Languages, Fairseq, Translation, Models, Parallel Corpora, Real-time Translation, Indic Languages, BLEU Scores, API (Application Programming Interface), Cross-Lingual Translation, Language Processing, Grammatical Structures, Morphology, Language Accessibility, Translation Evaluation, Under-resourced Languages, Chrome Extension, Speech-to-Text (STT), Text-to-Speech (TTS).

I. INTRODUCTION

Translating languages has always been an important part of Natural Language Processing(NLP) to help people from different linguistic background to communicate with each other. The area of machine translation(MT) has undergone a significant transformation over past few years with the adoption of deep learning techniques, particularly owing to the advancement made in Neural Machine Translation(NMT). These techniques are more result oriented and versatile as compared to older approaches that included rule-based machine translation(RBMT) and statistical machine translation (SMT). For instance, the work of Prahwini et al.(2024) and Vandan Mujadia et al (2024) have shown the potential that neural models possess while translating under resourced languages in the Indian scenario.

Nonetheless, many of the modern paradigms are either designed for particular language pairs or are limited in functionality due to dependence on hardware resources and scarcity of parallel corpora. On the other hand, we propose to extend this research by creating a real time adaptable translation model with the Fairseq framework that can be used with multiple Indian languages as well as English.

Although significant strides have been made in machine translation (MT), translating between Indian languages and English continues to present a challenging task. This is primarily due to the complex grammatical structures, rich morphology and the scarcity of parallel corpora for many languages. To tackle these challenges, our approach utilizes the Fairseq toolkit and Machine Translation, which supports state-of-the-art neural translation models, to develop a unified translation API. This system is designed to accommodate a wide array of Indian languages, ensuring that the unique linguistic features of each language are accurately captured. and efficient translation API capable of translating text between English and 11 major Indian languages, including Assamese (as), Gujarati (gu), Kannada (kn), Bengali (bn), Telugu (te), Tamil (ta), Malayalam (ml), Marathi (mr), Hindi (hi), Oriya (or), and Punjabi (pa).

II. RELATED WORK:

The study underlines the fact that there are no relevant translation tools for Tulu language thus points out the urgent need for research developments in the field.[1] One of the important outcomes of this research is the invention of an English-to-Tulu translation system, which functions with an accuracy rate of 89% for simple sentences.[1] Still, the paper acknowledges the necessity for further improvement, namely the one to the accuracy of more complex sentence structures.[1]

In the following attempts, translators will modify the translation system to not only handle the structural component of the language but also cover other more difficult ones in the communication.[1]

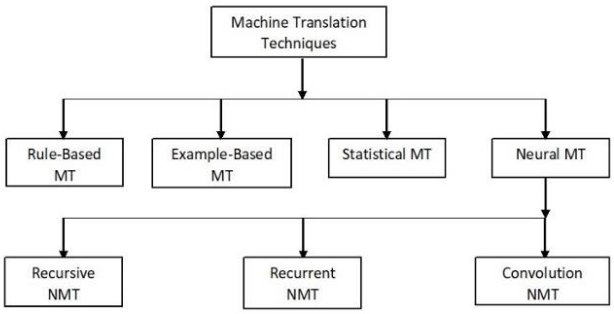

Fig 1 Machine Translation Techniques.

The field of translation has always involved complicated technology systems used together with both English and also Indian languages, a fact which has required significant research, one of the main studies is neural machine translation (NMT) and the involvement of large language models (LLMs).[1] The traditional as well as the new models of translation, along with the state-of-the-art neural methods, have been the object of investigation in order to make the translation process in different languages more accurate and efficient.[1]

For the past years, Large Language Models (LLMs)came out as the most natural and faithful way of translating without having to read any corpora first.[2] The study by Mujadia et al. (2024) emphasized the application of LLMs for reference-free translation evaluation, primarily for English and Indian languages.[2]They found that benchmark evaluators based on LLMs, such as LLaMA-2-13B, worked much better than the existing systems. For instance, BERT-Scorer and COMET were the traditional ones.[2] Their system, which took into account novel features such as zero-shot learning and fine-tuning, resulted in high accuracy rates (89% for words and 81% for sentences).[2] This study promotes the probable developments of LLMs in the translation evaluation domain by outlining methods to include other Indian languages to the framework of this evaluation.

Furthering this, Mujadia et al. (2023) covered the detailing LLMs translation capacities, first of all in terms of the Indian languages relation.[3] The research paper also unveiled a new set of prototypes that are being proposed for use in zero-shot as well as in-context translation experiments.[3]

They then exposed which capabilities of enabling LLMs to perform word alignment are using zero-shot corpora by transferring knowledge of the languages which have plenty of data to languages that do not, parallel one-to-many bilingual dictionaries, though they also admitted that LLLMs still need to be refined to showcase superiority over some traditional models like BLEU and chrF in performance.[3]

They also suggested future work including Indian-to-Indian language translation to better represent the diverse linguistic environment in the region.[3]

On the counter side, Samanantar is the fruit of one of the biggest parallel corpora collections in the Indic language group, which consists of 49.7 million sentence pairs between English and 11 Indic languages.[4] Besides, using of track, the corpus was augmented by 37.4 million sentences that were collected from the web, it, in turn, provided a significant contribution to the data source. The corpus was a subject to the thorough human scrutiny for excellence and trustworthiness before it was finalized.[4] The new Multilingual Neural Machine Translation (MT) models by Samanantar are used to train them, and they clearly outperform the existing models thus providing the desired areas for the improvement of machine translation systems in Indic languages.[4]

In summary, although incredible strides in machine translation are the product of NMT and LLMs, challenges persist, especially for languages that are under-resourced. The advancement of systems like Samanantar has seen good progress, but more work is needed to enable more complicated sentence structures and a larger percentage of supported languages. The goal of this project is to design a stable and flexible multilingual translation system for Indian regional languages by using NMT and NLP improvements. The metadata preprocessing, model development, system integration, evaluation, and scalability are among the objectives we aspire to accomplish to make online content easily accessible and inclusive.

III. METHODOLOGY

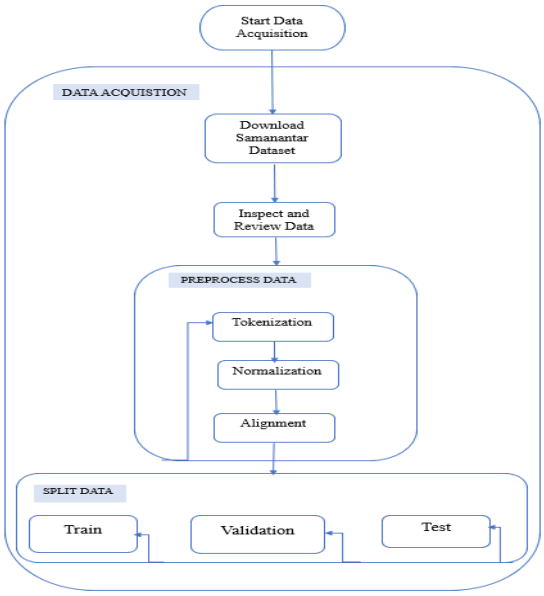

Fig 2 Data Acquisition.

Figure 2 depicts a multilingual translation system as an eminent approach to communicating in several languages, decoding the distinct phases comprising the data processing needed to train the translation model. The procedure is classified as follows:

A. Research Data Collection

1)Get the Latest Samanantar Dataset:

Samanantar is a collection of texts used for various Indian languages. Be sure to download this dataset from a reliable source that offers translations between Indian languages like Hindi and English or Marathi and English such as Samanantar's official site.

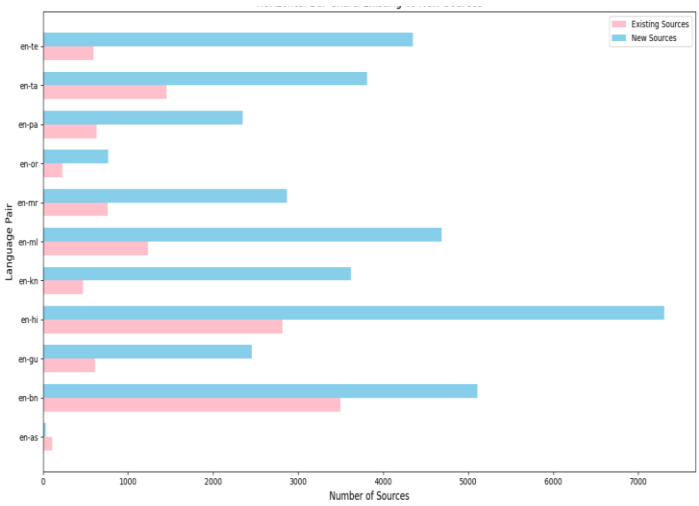

Fig 3. New vs Existing sources for each languages in Samanantar dataset

2) Clean Up the Data:

Once you have the dataset, look for any missing elements and other issues like wrong matches between the source and target sentences. You might also need to standardize some special characters. This step makes sure the data is good to use for training.

B. Data Preprocessing

1)Tokenization:

Tokenization as we know is the process of tokenizing the raw text into tokens or sub-words. It is a crucial part of converting input text into a form that the machine learning model can understand. It would be beneficial to tokenize text to ensure that the units are smaller and therefore the models can handle them efficiently.

```
[ ]  text = 'ઉઝઝસો'
     list(text)

['ઉ', 'ઝ', 'ઝ', 'સ', 'ો']
```

Fig 4 Tokenization example

2)Normalization:

This belongs to the step of transforming the text into a standard format, e.g., lowercasing, removing spaces, punctuation, and special characters handling. This helps the model to stay on the grounds of the most important data and ignore non-trivial sources of error.

3)Alignment:

Finally, the human of the law ends up being correctly translated into a piece of corresponding judicial language by the computer. Alignment between the source text and the target text must be established. Otherwise, the procedure of an error-free translation is likely to go askew.

4)Validation and Test Split:

Disaggregate your data into 3 divisions. Divide it into training, validation, and testing. Respectively, 80% is for training, 10% is for validation, and the remaining 10% is for testing. In this way, the model will be properly trained and it will also be tested on data even as it is not seen.

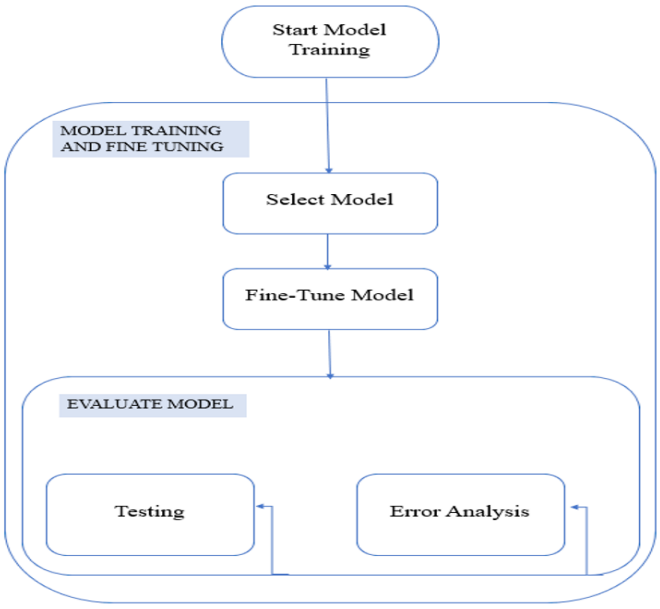

Fig 5. Model Training and Fine-tuning.

C. Model Training

1)Select Model:

In the task of machine translation, choosing a transformer-based model that is best adapted to your specific language pairs is the key. An example of such models is mBART (Multilingual BART), mT5 (Multilingual T5), and MarianMT, which are widely used for the translation of the languages in India. Accordingly, if you are working on each language separately, then Fairseq software is capable of supplying you with the Transformer models that are specifically geared for those languages.

2)Fine-Tune Model:

Fine-tuning is the process of taking a pre-trained model and adding more training data from your dataset to adapt it to the target task. By doing this, the weights of the model are optimized based on the data from your language pair and thus the accuracy is improved, making the model the most applicable for the pair at hand.

3)Evaluate Model:

After the training, you can also test the model using metrics like BLEU score like SacreBLEU, which is used to determine translation quality by comparing the machine's output to the correct human translation. On your side, each language pair will also undergo its own evaluation to verify the accuracy of the translation.

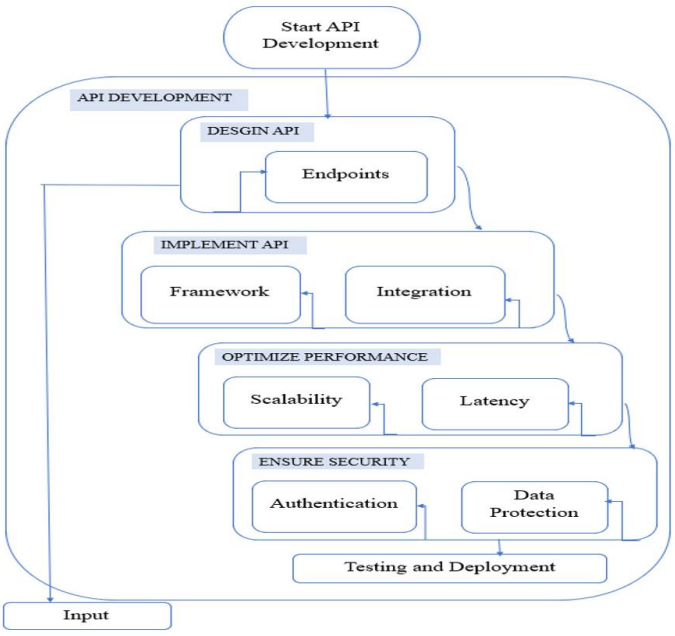

Fig 6. API Creation.

D. API Creation:

Above Fig 6 is the Chrome extension for translating the webpage, which includes the Flask-based API code to make real-time content translation possible. The work kicks off when the individual activates the Chrome extension, sending a request to the server-hosted Flask API. The extension extracts the text from the current web page in Java Script and then brings it to the Flask API. The service completes the request applying a translation service by using Google Translate or personal translation models to transform the text into the required language. When the option to translate the content is received, the Flask API returns the original text to the extension. The extension then changes the webpage on the fly and replaces the source text with the translation, eliminating the need for the page to be reloaded. This setup ensures easy and quick user interaction. The structure is in a client-server manner with the Chrome extension as the client, and Flask as the server, with both sides communicating over HTTP by sending requests and responses. Furthermore, the design takes into account the optimization of calls that are minimized and of the processing time as well as it guarantees the efficient performance of the extension regardless of the size and complexity of the web pages.

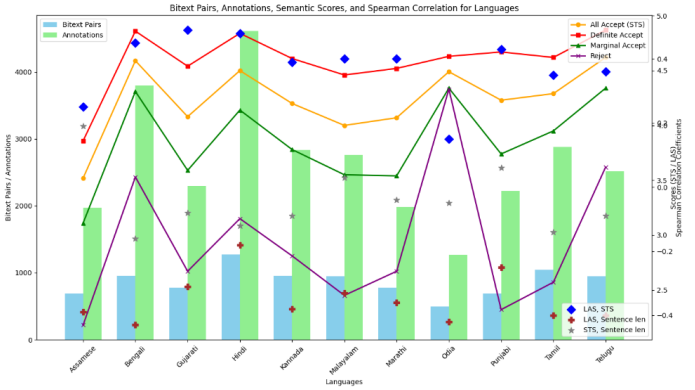

Fig 7. Graphical Representation of the Semantic Similarity Annotation Task Results Across 11 Languages (from Ramesh et al., 2021)

Using the data from Table 1 of [Ramesh et al., 2021] we have generated a graph which gives a comprehensive comparison of 11 linguistic indices. The table summarizes some key metrics such as bitext pairs and annotations, and it is shown as bar charts in the graph. Bitext pairs are depicted in the color of the sky, while annotations in the color of light green.

Moreover, the chart is used to depict both Semantic Textual Similarity (STS) scores and LAS (Language-Agnostic Sentence) scores with the line plots floating on the right y-axis. These lines represent the four types of semantic acceptability namely, All Accept, Definite Accept, Marginal Accept and Reject in the colors of orange, red, green, and purple, respectively. Through that, analysts can understand the semantic acceptability of sentences in different languages.

Moreover, the scatter plot indicates the Spearman correlation coefficients for LAS, STS, LAS with sentence length, and STS with sentence length. The coefficients are represented with different coloured markers (e.g., blue, brown, and grey). It is a strong show of interdependence between these various linguistic measures.

To summarize, the graph, which was the result of the research, carries the message of [Ramesh et al., 2024] table in a visual form depicting the relationships between dataset size, annotation quality, and semantic similarity evaluations, the correlations between the different linguistic variables across the languages studied.

Language	Annotation data		Semantic Textual Similarity score				Spearman correlation cofficient		
	Bittext pairs	Annotations	All accept	Definite accept	Marginal accept	Reject	LAS, STS	LAS, Sentence len	STS, Sentence len
Assamese	689	1972	3.52	3.86	3.11	2.18	0.25	-0.39	0.19
Bengali	957	3797	4.59	4.86	4.31	3.53	0.45	-0.43	-0.16
Gujarati	779	2298	4.08	4.54	3.59	2.67	0.49	-0.31	-0.08
Hindi	1276	4616	4.5	4.84	4.14	3.15	0.48	-0.18	-0.12
Kannada	957	2838	4.2	4.61	3.78	2.81	0.39	-0.38	-0.09
Malyalam	948	2760	4	4.46	3.55	2.45	0.4	-0.33	0.03
Marathi	779	1984	4.07	4.52	3.54	2.67	0.4	-0.36	-0.04
Odia	500	1264	4.49	4.63	4.34	4.33	0.15	-0.42	-0.05
Punjabi	688	2222	4.23	4.67	3.74	2.32	0.43	-0.25	0.06
Tamil	1044	2882	4.29	4.62	3.95	2.57	0.35	-0.4	-0.14
Telgu	949	2516	4.62	4.87	4.34	3.62	0.36	-0.4	-0.09
Overall	**9566**	**29149**	**4.27**	**4.63**	**3.89**	**2.94**	**0.37**	**-0.35**	**-0.04**

Table 1: Outcomes of the Semantic Similarity Annotation Task for 11 Languages (from Ramesh et al., 2021).

Having the annotation tasks and visualizations in place, we now move on to the analysis and subsequent results.

IV. RESULTS AND ANALYSIS

Model	En-X						
	GOOG	MSFT	CVIT	OPUS	Mbart	IT*	IT
as		13.6				7	6.9
bn	28.1	22.9	7.9		1.4	18.2	20.3
gu	25.6	27.7	14.1		0.7	19.4	22.6
hi	38.7	31.8	25.7	13.7	22.2	32.2	34.5
kn	32.6	22				9.9	18.9
ml	27.4	21.1	6.6	4.4	3	10.9	16.3
mr	19.8	18.3	8.5	0.1	1.2	12.7	16.1
or	24.4	20.9	7.9			11	13.9
pa	27	28.5				21.3	26.9
ta	28	20	7.9		8.7	10.2	16.3
te	30.6	30.5	8.2		4.5	17.7	22

Table 2: BleuScore for En-X translation based on FLORES devtest Benchmark, IT* is the research model developed.

In this section, we evaluate the capability of our En-X translation model with the results of the study conducted by Ramesh et al. (2024); the distance between BLEU scores with respect to languages such as Hindi and Tamil was similar. Given the consistent BLEU scores achieved in this language, it has corroborated the reported outcome, which indicates that core translation models-including ours-perform similarly across different studies. Our study results indicate that IT*, which we have developed as part of our research, generates a strong BLEU score, particularly in languages such as Hindi (34.5), Bengali (20.3), and Telugu (22)-close to the performance achieved by contemporary models such as GOOG, MSFT, CVIT, and mBART.

A key contrast that sets apart our research from that of the referenced paper is the translation platform. While Ramesh et al. (2024) had a webpage-based translation scheme, our study presents a Chrome extension for translating direct text on any webpage within the place of browsing. Obviously due to the different UIs, this model produced a consistent performance in both platforms. This gives strength to the translation models in quality results despite differing environments of users.

The Chrome extension we have developed features many real-life benefits over systems based on a webpage. The biggest merit is that it provides translation in real time as the users surf the Internet rather than asking the users to paste it into some other interface for translation. This does significantly improve accessibility of the service to the user, made all that much more valuable in real-time web scenarios, making real translation accessible and smooth for the everyday Web user.

Our competitive IT* model indeed gets such results across a number of Indic languages. For instance, Hindi scores BLEU 34.5, aligning our model closely with GOOG (38.7) and MSFT (31.8). In Tamil (ta), we achieve 16.3 on par with GOOG (28) and mBART (10.2). All of these results indicate very good performance of our model with high complexity language pairs, even with existing data for training and deploying it as a Chrome extension.

The performance of IT* is also competitive in other languages such as Bengali (bn), Gujarati (gu), and Telugu (te). For example, in Bengali, this model achieved a BLEU score of 20.3 whereas GOOG scored 28.1 and MSFT has 22.9. This indicates that the model scores well in language-pair performance. Likewise in Gujarati, the performance of the model is again quite high with a score of 22.6, which is closely followed by MSFT (27.7) and GOOG (25.6). Telugu also demonstrates a similar trend giving our model a BLEU score of 22, thus indicating the general robustness of IT* across Indic languages.

In contrast, however, performance in languages such as Malayalam (ml) and Marathi (mr) is, on average, lower for our model.

V. FUTURE SCOPE

Many suggestions for improving the translation system will contribute to its enhancement and wider practical applicability. One such change could be transformative handwriting transfer, whereby the integration of Optical Character Recognition (OCR) technology can be done. This would increase overall value for by letting the software understand, and convert, hieroglyphics into human-readable text, thus widening its outreach to potential users who probably find hand-writing as their main channel of communication. Additionally, Text-to-Speech (TTS) and Speech-to-Text (STT) could be excellent features for enhancing system usability. With TTS, the user can hear the translated text in the target language, allowing learners, or even those who have reading problems, the opportunity to listen to studies. Establishing STT, on the other hand, will allow users to speak aloud their input rather than having to write, making the system more accessible and user-friendly, particularly for those with communication challenges.

The system is highly scalable and could undertake further expansion to additional Indian languages and variations, increasing its user base to cater for a wider multilinguistic community. On the other hand, adopting these additional mechanisms like a feedback loop can enhance translation accuracy and user satisfaction. This feedback loop bases itself on incorporating dynamic learning (adaptive learning) mechanisms which will continue informing and updating the machine, thus aiding in the improvement of the initial machine-learned translation quality. Furthermore, another step that would certainly expand the potential of the system is developing mobile applications and companion software for offline use, thus substantially increasing accessibility. This is especially important for areas without internet accessibility, which at best give access to the required translation needs to its users.

Another area of improvement will involve collaborations with educational institutions to develop specialized translation tools specifically designed for academic users, fostering access to multilingual resources for students and researchers. Further advancement of these areas could considerably enhance the overall impact of multilingual communication across a variety of communities.

VI. CONCLUSION

This project successfully achieved the creation of a multilingual translation system that capitalizes on the Fairseq framework, allowing efficacious and accurate translations from English into various Indian regional languages, such as Punjabi, Hindi, and Bengali. The NMT system effectively tackles the language barrier challenge prevalent in a globalized setup using state-of-the-art Neural Machine Translation techniques.

While the intricacies of data preprocessing, model training, and implementation into a user-friendly translation API embedded in Flask speak for how well the translation system is able to preserve the intent of the original message while bearing in mind certain cultural aspects, the qualitative measures indicate that the system can compete with those that are already existent, thus representing a valuable aid to people and communities seeking to communicate across languages.

The design for the system includes a great emphasis on scalability and adaptability,

allowing for enhancements in future releases and the inclusion of more languages, thus extending its reach and applicability. This project promotes inclusiveness and facilitates communications in vital sectors like healthcare, education, and public services by improving access to vital information for speakers of less commonly represented languages.

Added to the fact that the multilingual translation system confirms the viability of modern NMT techniques is the fact that it thereby advances matters in the domain relating to the promotion of mutual understanding and support among multimodal linguistic groups.

VII. REFERENCES

[1] Prathwini, Anisha P. Rodrigues, P. Vijaya, Androshan Fernandes, **"Tulu Language Text Recognition and Translation"**, IEEE supported by NITTE (Deemed to be University), 2024.

[2] Vandan Mujadia, Pruthwik Mishra, Arafat Ahsan, Dipti Misra Sharma, **"Towards Large Language Model Driven Reference-less Translation Evaluation for English and Indian Languages"**, IEEE LTRC, IIIT Hyderabad India, 2024.

[3] Vandan Mujadia, Ashok Urlana, Yash Bhaskar, Penumalla Aditya Pavani, Kukkapalli Shravya, Parameswari Krishnamurthy, Dipti Misra Sharma, **"Assessing Translation Capabilities of Large Language Models Involving English and Indian Languages",** IEEE LTRC, IIIT Hyderabad, India, 2023.

[4] Sahinur Rahman Laskar, Abinash Dutta, Partha Pakray,Sivaji Bandyopadhyay, **"Neural Machine Translation: English to Hindi",** 2019 IEEE Conference on Information and Communication Technology (CICT).

[5] Ramesh, G., Doddapaneni, S., Bheemaraj, A., Jobanputra, M., AK, R., Sharma, A., Sahoo, S., Diddee, H., J, M., Kakwani, D., Kumar, N., Pradeep, A., Nagaraj, S., Deepak, K., Raghavan, V., Kunchukuttan, A., Kumar, P., & Shantadevi, M. (2024). *Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages*. *Transactions of the Association for Computational Linguistics (TACL)*, *12*, 1-16. https://doi.org/10.48550/arXiv.2104.05596

[6] K.M. Kavitha, **"Hybrid Approaches for Augmentation of Translation Tables for Indian Languages"**, IEEE International Conference on Machine Learning and Applications (ICMLA), 2020.

[7] Sahinur Rahman Laskar, Abinash Dutta, Partha Pakray, **"Neural Machine Translation: English to Hindi",** 2019 IEEE Conference on Information and Communication Technology (CICT).

[8] Sandeep Saini, Vineet Sahula, **"Neural Machine Translation: English to Hindi"**,2018 IEEE Conference on Information and Communication Technology (CICT).

[9] AI4Bharat, "Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages",https://www.kaggle.com/datasets/mathurinache/samanantar

[10] WAT 2021, "The 8th Workshop on Asian Translation (WAT 2021)," Kyoto University, 2021. Avaiable:https://lotus.kuee.kyotou.ac.jp/WAT/WAT2021/index.html

[11] J. Tiedemann, "OPUS: A collection of parallel corpora," 2025. [Online]. Available: https://opus.nlpl.eu/