Comparative Analysis of Google Translator and AI4Bharat Translator

CH. Srinivas Sai - 21BDS012, Abhiram Koppuravuri - 21BDS029, R. Vinay Kumar - 21BDS056
Comparitive Analysis of Google Translator and AI4Bharat Translator - GitHub
Data Science and Artificial Intelligence , Indian Institute of Information Technology Dharwad
Email:- 21bds012@iiitdwd.ac.in, 21bds029@iiitdwd.ac.in, 21bds056@iiitdwd.ac.in

***Abstract*—In today's modern era, access to precise and efficient information is crucial. Language translation tools have become increasingly popular in aiding individuals in understanding various content. Google Translate [1] is a widely used tool that has gained popularity among many users. Our project takes a different approach by introducing AI4 Bharat, specifically the IndicTrans2 model[2]. The primary focus of this project is to demonstrate the superior effectiveness of the IndicTrans2 model in translating languages, especially Indian languages, when compared to commonly used tools such as Google Translate.**

I. Introduction

Translation is essential in our interconnected world for facilitating communication, fostering cultural understanding, and sharing knowledge. Google Translate has revolutionized the field of translation and communication, although it faces challenges such as accuracy issues, static translations, and limited contextual understanding. Despite these obstacles, Google Translate continues to strive for improvement. However, it remains committed to enhancing its capabilities and overcoming its limitations. As technology advances, the role of translation in breaking down barriers and creating a more interconnected world will only continue to grow in importance.

In this highly competitive world, individuals are constantly innovating and revolutionizing various fields to propel the current generation into the future. One such remarkable initiative is the establishment of "AI4 Bharat" by the brilliant minds at IIT Madras. This open-source community of engineers is dedicated to advancing language technology, healthcare, education, and agriculture. As part of their efforts, they have successfully trained models on over 20 Indian languages. Notably, their website features a groundbreaking model called "IndicTrans2" [2] which is the first-ever transformer-based multilingual Neural Machine Translation model. This exceptional creation enables high-quality translations across all 22 scheduled Indic languages.

The discovery of this information online sparked our interest in comparing the performance of Google Translator and the AI4 Bharat model. We decided to conduct a study using a dataset of 10 lakh records of Hindi text from IIT Bombay to evaluate the accuracy of both translation tools.Our analysis revealed that the AI4 Bharat model outperformed Google Translator in terms of accuracy when translating[3] Hindi text. However, it is important to note that this comparison was limited to the dataset provided by IIT Bombay.

In the upcoming coming sections , Dataset explanation we look into the data how it is and what we did to the data for next step after that we will move on to the methodology we will discuss about the models we used in detail and the model architecture and at results and discussion section we will discuss about the comparison results and at last section we will conclude the report and give some references .

II. Dataset Explanation

The model utilized data from the IIT Bombay website, containing 10 lakh records of Hindi and English sentences across various topics. From this dataset, 995,109 records were extracted and divided equally among three individuals, each handling 331,703 records to begin the next phase of the process.

III. Methodology

A. Translation Workflow Implementation

Translation processing begins now. Initially, we utilized Google Translator to convert Hindi text into English text. To achieve this, we have invoked the 'translator' function from the 'googletrans' library. We specify the source language as Hindi and the destination language as English, instructing the model to perform the translation. Throughout this process, by incorporating the TQDM library, which introduces a progress bar to track the translation's advancement. Subsequently, the resulting translated text is stored in a new column labeled 'engtrans.'

B. Translation with ai4bharat Translator Integration

Moving on to performing translation using the ai4bharat translator, the first step is to import the 'indictrans' package, which serves as the primary framework for machine translation. Following this, it is essential to import all the necessary libraries for natural language processing in Indian languages. To facilitate efficient text processing, the Subword

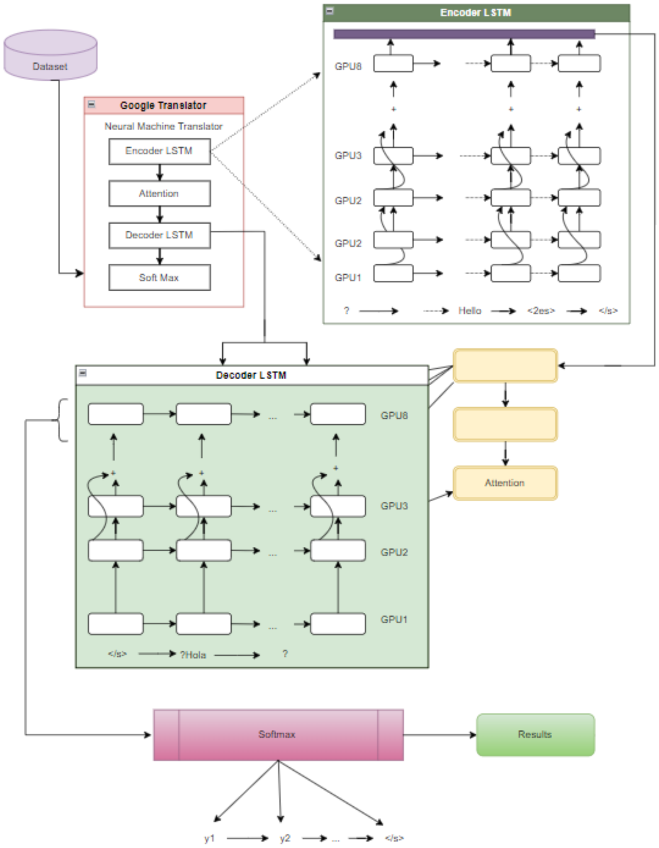

Figure 1. Architecture of Google Translator

NMT[4] model needs to be imported.

C. Integration of Essential Packages and fairseq Setup

Subsequently, certain packages such as ’sacremoses,’ ’mock,’ ’sacrebleu,’ ’tensorboardX,’ and ’fairseq,’ among others, must be installed. Notably, ’fairseq’ is a popular framework used for training and evaluating neural machine translation models. After installing the fairseq package, it’s essential to add it to the system path. Subsequently, import modules from fairseq, such as checkpoint utils, distributed utils, options, tasks, and utils. Proceed to download pre-trained IndicTrans models designed for machine translation between Indic languages and English. After that, navigate to the ’indicTrans’ directory for further processing, ensuring access to pre-trained IndicTrans models for various translation tasks involving Indian languages.

D. Model Initialization and Setup for Translation Tasks

Import the necessary class from the specific module within the ’indicTrans’ package. Initialize the constructor for the model object, specifying the directory containing the pre-trained Indic-English model files. This process ensures that the research project has the required pre-trained models and is ready for further development in translation tasks involving Indic languages.

E. Text Translation Process with Indic-English Model

The model is now prepared, and all that is required is to input the text into the function we’ve created. This function utilizes the ’translate paragraph’ from the ’indic2en model’ object to convert the input text from Hindi to English. Subsequently, it saves the translated text in a dedicated column within the DataFrame. The tqdm progress bar offers a visual representation of the translation process.

F. Semantic Textual Similarity (STS) & Bilingual Evaluation Understudy (BLEU) scores Analysis of Translations

We have both Google-translated text and AI4Bharat-translated text[5], and our goal is to compare them to the original English text in the dataset. We aim to determine which translation is most similar to the original Hindi text. To achieve this, we will calculate the STS score, which stands for Semantic Textual Similarity and BLEU score[6] which stands for Bilingual Evaluation Understudy . In analyzing the semantic similarity & Bilingual Evaluation between two sets of sentences within a DataFrame, one set being the original English sentences and the other being their translations by Google Translator and AI4Bharat Translator, we utilized various

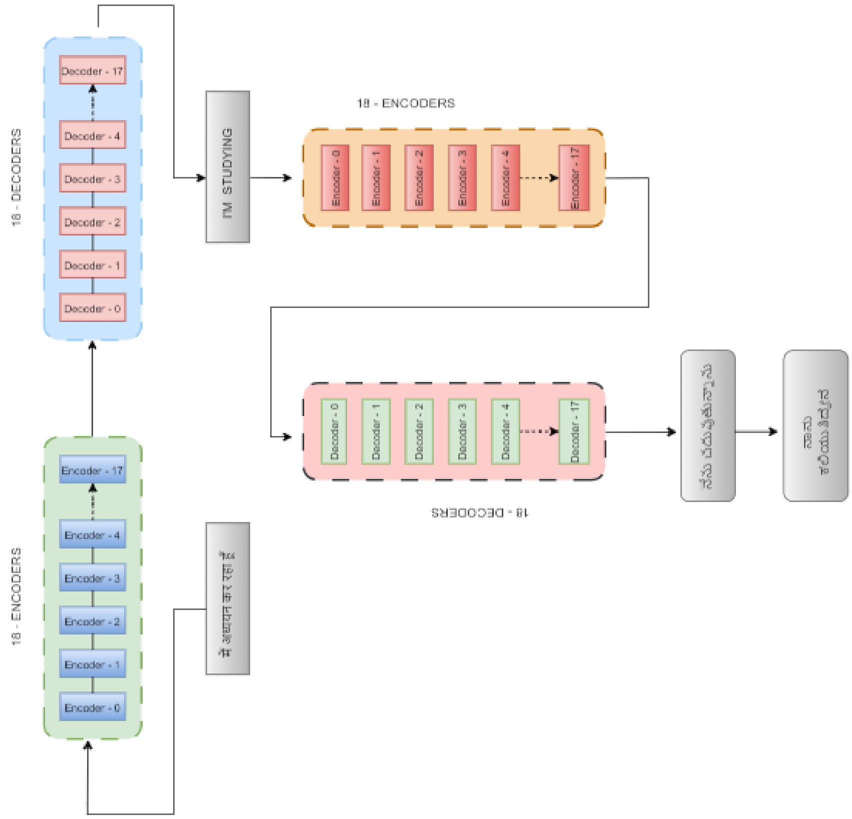

Figure 2. Architecture of AI4 Bharat Translator

natural language processing techniques. These techniques include tokenization, stop-word removal, and lemmatization to clean and standardize the text. Subsequently, we applied TF-IDF vectorization and cosine similarity metrics.

G. Evaluation and Analysis of Translation Performance

This process helps quantify the semantic similarity between each original sentence and its translated counterpart. The STS and BLEU scores that are calculated indicates how well the translation preserves the original meaning. Finally, we store the translated sentences along with their corresponding STS and BLEU scores in a new Excel file for further analysis and evaluation of the performance of both Google Translator and AI4Bharat.

IV. Results and Discussion

The STS and BLEU scores provide a valuable means of comparing the translations of Google and AI4Bharat with the original English dataset. Through the utilization of advanced natural language processing techniques such as tokenization, stop-word removal, lemmatization, TF-IDF vectorization, and cosine similarity metrics, the STS and BLEU analysis enables a comprehensive assessment of translation quality.

Based on preliminary results, it has been observed that both translation approaches yield positive outcomes. However, the IndicTrans model developed by AI4Bharat consistently surpasses Google Translator in maintaining the semantic subtleties of the original Hindi text. The STS and BLEU scores further support this, indicating a greater resemblance between the translations generated by AI4Bharat and the source sentences. This suggests that AI4Bharat's translations are more precise and contextually faithful representations.

The specialized machine translation frameworks show promise, particularly when customized for the unique linguistic characteristics of Indic languages. The research project's detailed methodology, involving progress monitoring with TQDM during translation and utilizing established NLP techniques, improves the accuracy of the outcomes. These discoveries add to the conversation about refining machine translation models for diverse linguistic settings and have implications for wider use in cross-language communication and content adaptation.

Table I
STS (Semantic Textual Similarity) Score based Comparison of Translator **Hindi to English**

Translators	**Total records**	**Equally Translated**	**Best Translated**	**Final**	**Percentage**
Google Translator	9,94,109	4,66,819	1,17,731	0.2232	22.32
AI4Bharat Translator	9,94,109	4,66,819	4,09,559	0.7767	77.67

Table II
BLEU (BiLingual Evaluation Understudy) Score based Comparison of Translator **Hindi to English**

Translators	**Total records**	**Equally Translated**	**Best Translated**	**Final**	**Percentage**
Google Translator	9,95,109	2,95,329	13,327	0.0190	1.90
AI4Bharat Translator	9,95,109	2,95,329	6,86,363	0.9809	98.0

V. Conclusion

Our research findings indicate that the IndicTrans model developed by AI4Bharat consistently outperforms Google Translator[7] in Hindi-to-English translation. This is evident from the higher scores in Semantic Textual Similarity (STS) and Bilingual Evaluation Understudy (BLEU). These results underscore the significance of tailored machine translation models in preserving the intricacies of language. The comprehensive methodology employed in this study, which includes progress tracking and NLP techniques, enhances the reliability of the outcomes[8]. The AI4Bharat IndicTrans model excels in maintaining semantic richness and contextual nuances, highlighting the necessity for language-specific approaches. By incorporating a progress bar, pre-processing steps, and leveraging frameworks like fairseq, our methodology demonstrates its robustness. Future research could focus on refining evaluation metrics to further enhance the accuracy of translation assessments. This study contributes to the continuous advancement of machine translation, emphasizing the effectiveness of language-specific frameworks in facilitating cross-language communication.

References

[1] J. of teaching English for specific and academic purposes 4, “Google translate in teaching english,” *Medvedev, Gennady*, vol. 151, pp. 181–193, 2016.

[2] P. A. C. R. A. S. D. V. G. A. K. J. N. e. a. Gala, Jay, “Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages,” vol. 151, pp. 181–193, 2023.

[3] A. Hegde and S. Lakshmaiah, “Mucs@ mixmt: indictrans-based machine translation for hinglish text,” *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 1131–1135, 2022.

[4] D. Britz, A. Goldie, M.-T. Luong, and Q. Le, “Massive exploration of neural machine translation architectures,” *arXiv preprint arXiv:1703.03906*, 2017.

[5] M. N. Jensen and K. K. Zethsen., “Translation of patient information leaflets: Trained translators and pharmacists-cum-translators–a comparison.,” *Medvedev, Gennady*, vol. 151, pp. 181–193, 2012.

[6] M. Post, “A call for clarity in reporting bleu scores.,” *Annals of Internal Medicine*, vol. 151, pp. 264–269, 2018.

[7] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, and the PRISMA Group, “To use or not to use google translate,” *Maulidiyah, Fitrotul*, vol. 151, pp. 1–6, 2018.

[8] G. L. A. D. d. I. Aranberri, Nora and K. Sarasola, “Comparison of post-editing productivity between professional translators and lay users.,” *arXiv preprint arXiv:1703.03906*, 2014.