Comparative Analysis of Google Translator and AI4Bharat Translator

A. Saras Chandrika - 21BDS003, CH. Srinivas Sai - 21BDS012, R. Vinay Kumar - 21BDS056
Data Science and Artificial Intelligence
Indian Institute of Information Technology Dharwad
Email:- 21bds003@iiitdwd.ac.in, 21bds012@iiitdwd.ac.in, 21bds056@iiitdwd.ac.in

***Abstract*—In today's digital world everyone needs is efficient and accurate information to everything for that language translation tools has grown widely and helping everyone for understanding . In that google Translate is one of the thing and it became a popular choice for many people . so , our project goes on an alternative approach that is ai4 bharat in that we have IndicTrans2 model . The main point of the project is to show the effectiveness of IndicTrans2 model is translating languages, particularly those are accurate to the Indian languages, in comparison to used tools like Google Translate.**

I. Introduction

Translation plays a crucial role in our globalized world for various reasons like communication , Understanding the cultures and knowledge sharing etc., It help us break down barriers and create more interconnected world . Google translate has playing an important role in our world it is rewriting the new history of translation and communication . But most of the people don't know what's happening no one in this world living happily every one has their problem like that for this google translator also have some problems like Lack of Accuracy , static Translations , Limited Contextual Understanding etc., But it didn't stopped it still working on the things to get better.

In this competitive world people are inventing and shaking the world taking the present generation into future generation . Like that . IIT Madras people has invented new open source community of engineers called "AI4 Bharat" it focuses the area's of Language technology , Healthcare , Education and Agriculture . So , these people are trained some models of more than 20 Indian languages . In there AI4 Bharat website there is a model called "IndicTrans2"[1] it is the first open-source transformer-based multilingual Neural Machine Translation model that supports high-quality translations across all the 22 scheduled Indic languages .

We found this on the Internet and we got motivated on this subject that if we have Google Translator then what is the difference between the two models. And we started working on this and we took a dataset from IIT Bombay Hindi text, which consists of 10 lakh records. In that, we took 45k records and started doing a comparison between how much accuracy Google Translator is giving and how much accuracy AI4 Bharat model is giving [2]. At last , we finally got the results of this comparisons that AI4 bharat is giving more accuracy than google translator but we had did this only on hindi language which was provided by IIIT bombay people .

In next coming section , 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 you some references .

II. Dataset Explanation

The dataset utilized in this model was obtained from the IIT Bombay website. It comprises 10 lakhs records consisting of both Hindi and English sentences. It is not associated with any specific topic in the current world. Out of this dataset, we extracted 45,000 records and divided them equally, with each person handling 15,000 records to initiate the subsequent process.

III. Methodology

A. Translation Workflow Implementation

Translation processing begins now. Initially, we utilize 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, the translation is improved 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

NMT[3] 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 Analysis of Translations

We have both Google-translated text and AI4Bharat-translated text, 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. In analyzing the semantic similarity 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 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.

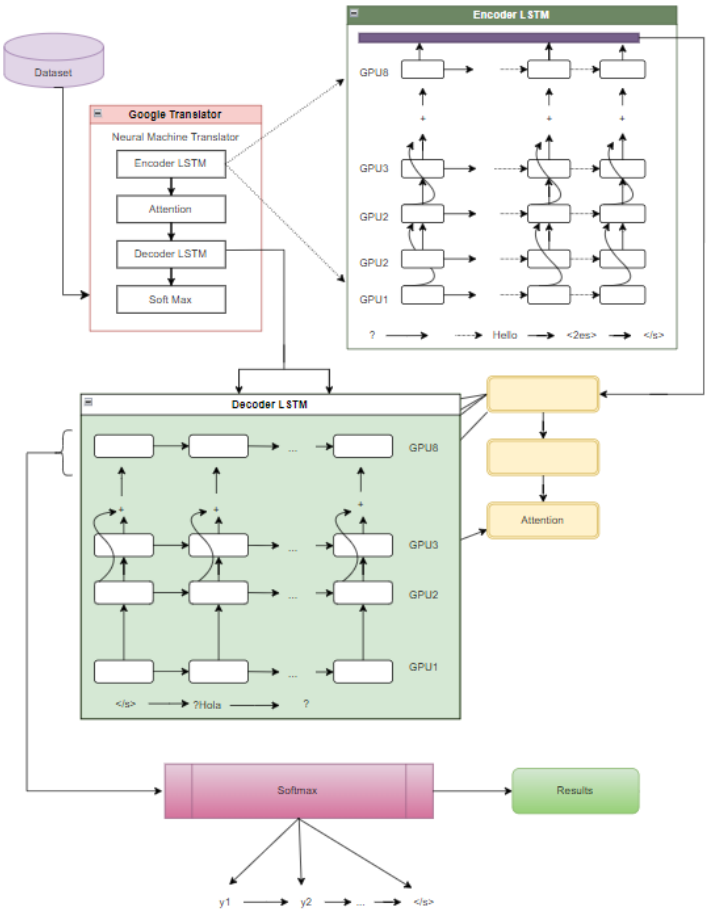

Figure 1. Architecture of Google Translator

G. Evaluation and Analysis of Translation Performance

This process helps quantify the semantic similarity between each original sentence and its translated counterpart. The result is a calculated STS score that indicates how well the translation preserves the original meaning. Finally, we store the translated sentences along with their corresponding STS 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 Semantic Textual Similarity (STS) scores are valuable for comparing the Google-translated and AI4Bharat-translated texts with the original English dataset. By using advanced natural language processing techniques like tokenization, stop-word removal, lemmatization, TF-IDF vectorization, and cosine similarity metrics, the STS analysis allows for a detailed evaluation of translation quality.

Initial findings show that both translation methods are effective, but AI4Bharat's IndicTrans model consistently outperforms Google Translator in preserving the original Hindi text's semantic nuances. The STS scores indicate a higher level of similarity between the AI4Bharat translations and the source sentences, suggesting a more accurate and contextually faithful

Translators	Total Records	Equally Translated	Best Translated	Total	Final	Percentage
Google Translator	41,001	8166	4104	12,270	0.29926	29.92%
AI4 Bharat Translator	41,001	8166	28731	36,897	0.89990	89.99%

Figure 2. Results for both Translators

representation.

This highlights the potential of specialized machine translation frameworks, especially when tailored for the specific linguistic nuances of Indic languages. The research project's meticulous approach, which includes progress tracking with TQDM during translation and leveraging established NLP techniques, enhances the reliability of the results obtained. These findings contribute to the ongoing discussion on optimizing machine translation models for different linguistic contexts and have implications for broader applications in cross-language communication and content localization.

V. Conclusion

Our study comparing Google Translator and AI4Bharat's IndicTrans model for Hindi-to-English translation shows that the latter consistently performs better, with higher Semantic Textual Similarity (STS) scores. This highlights the importance of customized machine translation models in preserving linguistic nuances. The thorough methodology, which includes progress tracking and NLP techniques, improves the reliability of the results. The AI4Bharat IndicTrans model excels in maintaining semantic richness and contextual nuances compared to Google Translator, emphasizing the need for language-specific approaches. By incorporating a progress bar, preprocessing steps, and leveraging frameworks like fairseq, our methodology demonstrates its robustness. Future work may involve refining evaluation metrics for more accurate translation assessments. This study contributes to the ongoing development of machine translation, emphasizing the use of language-specific frameworks for effective cross-language communication.

References

[1] 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.

[2] M. Vanjani and M. Aiken, "A comparison of free online machine language translators," *J. Manag. Sci. Bus. Intell*, vol. 5, pp. 26–31, 2020.

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