LANGUAGE TRANSLATION

A Capstone Project report submitted

in partial fulfillment of requirement for the award of degree

BACHELOR OF TECHNOLOGY

in

SCHOOL OF COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE

by

G. Sri Harshini	**(2103A52137)**
A. Druva Kumar	**(2103A52121)**
M. Dheeraj	**(2103A52055)**
CH. Rohith	**(2103A52128)**
K. Ajay Rao	**(2103A52147)**

Under the guidance of

Mr. D. Ramesh

Assistant Professor, School of CS&AI.

Submitted to

SR University, Ananthsagar,Warangal,Telagnana-506371

SR University

Ananthasagar, Warangal.

CERTIFICATE

This is to certify that this project entitled **“Language Translation**" is the Bonafide work carried out by **M. Dheeraj, G. Sri Harshini, A. Druva Kumar, CH. Rohith, K. Ajay Rao** as a Capstone Project for the partial fulfillment to award the degree BACHELOR **OF TECHNOLOGY** in **School of Computer Science and Artificial Intelligence** during the academic year 2024-2025 under our guidance and Supervision.

Mr. D. Ramesh
Assistant Professor,
School of CS&AI
SR University
Anathasagar, Warangal

Dr. M. Sheshikala
Professor & Head,
School of CS&AI,
SR University,
Anathasagar, Warangal

Reviewer-1
Name:
Designation:
Signature:

Reviewer-2
Name:
Designation:
Signature:

ACKNOWLEDGEMENT

We owe an enormous debt of gratitude to our Capstone project guide **Mr. D. Ramesh, Asst.Prof** as well as Head of the School of CS&AI, **Dr. M. Sheshikala, Professor** and Dean of the School of CS&AI, **Dr. Indrajeet Gupta Professor** for guiding us from the beginning through the end of the Capstone Project with their intellectual advices and insightful suggestions. We truly value their consistent feedback on our progress, which was always constructive and encouraging and ultimately drove us to the right direction.

We express our thanks to project coordinator's **Mr. Sallauddin Md, Asst. Prof., and R. Ashok Asst. Prof.** for their encouragement and support.

Finally, we express our thanks to all the teaching and non-teaching staff of the department for their suggestions and timely support.

Abstract

One of the projects, "Language Translation," addresses the development of a robust system that enhances seamless cross-linguistic communication. The system makes use of novel natural language processing techniques, such as cutting-edge NLP, to translate text in real-time with accuracy into several languages. Machine learning models are also employed, especially neural networks- transformer-based architectures, like Sequence-to-Sequence and LLM Models to enable contextually aware translations. The system is designed to work across various use cases, including education, business, and tourism, while placing the user at the forefront and promoting regional dialects to enrich the social commonwealth by bringing peoples together across cultural divides.

Contents

Table Of Figures

Chapter 1

Introduction

The need to translate is the reason for the spread of new information, knowledge, and ideas around the globe. Where there is diversity in languages and cultures, it can be said that translation is a fundamental need to maintain proper communication and mutual understanding. New discoveries, scientific advances, and cultural contributions may never leave their isolation and stagnation without translation. By filling in linguistic gaps, translation enhances the spread of innovation, cooperation, and development of cross-cultural exchange. More to the point, in passing new ideas, translation has the effect of shaping societies, influencing world decisions, and perhaps even determining the course of history by enabling important messages to be heard and shared among a more extensive audience.

1.1 Objective

To extract effective communication between people across the globe. To allow the capability of two parties to communicate, and exchange ideas. Adequate motivation of learners for discussion with meaning and usage on the language at its deepest levels. Getting a challenging position in reputed organizations where we can learn skills by communicating. To speak and interpret our mother tongue.

1.2 Problem Statement

The structural difference and dependency between sentences in both languages have made the automatic translation between English to Telugu and Hindi challenging. The syntax, grammar, and sentence structures differ between the two languages. For example, the order and dependencies of words and phrases differ in English and Telugu. The meaning of a sentence is often determined by context, conjugation rules, and idiomatic expressions specific to each language. The goal of this project is to design an effective translation model that would translate English text into Telugu and Hindi with word order discrepancies, sentence dependencies, and context preservation in order to have natural and meaningful output.

1.3 Existing System

General-purpose translators (e.g., Google Translate) covering multiple languages and extensive vocabulary. Supports a wide range of languages and accents but may have latency in small-scale, specialized systems. Complex models like Whisper and Deep Speech, supporting numerous languages and accents. Uses statistical models or neural machine translation (NMT) for large-scale, multilingual tasks.

1.4Proposed System

The system proposed here aims to generate a translator able to perform a translation of words specifically drawn toward the limited domain of words in the English language. More precisely, the system specializes in converting text from one language to another and attempts to solicit as much accuracy and contextual appropriateness as possible during the process. Translator refers to programming language processor that transforms a program written in one language into its equivalent in another. This encompasses gathering or interpretation of directions to further facilitate comprehension through various linguistic systems. The system will only constrain its focus to a narrow domain only if it has to improve the possibility of reaching ideal performance, thus becoming a focused, efficient solution for translating certain applications or tasks. Such an approach to precision and adaptability will lay the foundation for further expansion into more general domains of other linguistic systems.

Chapter 2

Literature Survey

S No	Author(s) &Year	Model used	Parameters	Merits	Limitations &Drawbacks
1	Elizabeth Salesky,Marcello Federico, and Marine Carpuat (2024)	Deep Learning Pre-Trained Large Language Models (LLMs), Transformer-based Models.	Latency, Word Error Rate (WER), Task-specific challenges, BLEU Score	Innovative approaches in translation tasks, advancements in machine translation, better speech processing models.	Data scarcity, Computational costs, need for large training data.
2	Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz. (2023)	Indic language translation, Transformer-based models	Number of Epochs, Hidden Layer Size, Batch Size, Learning Rate.	Significant improvement in low-resource language translation, Effective use of monolingual data for pre-training and augmentation.	High computational requirements for training, Possible limitations in generalizing to all language pairs
3	Danni Liu and Jan Niehues (2022)	Multilingual Neural Machine Translation (NMT), Baselines, Codebook Approach	Encoder-Decoder Architecture, Slices, Codebook Entries, Vocabulary Size	Improved Robustness in Zero-Shot Conditions, Increased Interpretability, Facilitates Knowledge Sharing	Less Expressive than Continuous Models, Performance Degradation, Complex Training Process:
4	Csaba Oravecz, KatinaBontchev David Kolovratník, Bhavani Bhaskar,(2021)	Ensemble Models, Back-Translation Models, Big Transformer	Ensemble Models, Batch Size, Embedding Dimension, Attention Heads	Ensemble Models, Base and Big Transformer Models:	Complex Data Preprocessing, Computational Costs, Training Complexity

		Models, Base Transformer Models			
5	Santosh Kesiraju, Karel Benes, Maksim Tikhonov (2023)	Big Transformer Models Ensemble Models, Back-Translation Models, Ensemble Models Codebook Discretization Models	Batch Size, Dropout Rate, Epochs, Attention Heads.	Codebook Discretization Models, Back-Translation Models, Ensemble Models	Computationally expensive. Risk of index collapse and underutilization of codebook entries, Higher training time and resource requirements, May introduce noise into the dataset
6	Ife Adebara, Muhammad Abdul-Mageed [2021]	Spanish-Catalan (ES-CA), Catalan-Spanish (CA-ES), Spanish-Portuguese (ES-PT), Portuguese-Spanish (PT-ES), French-Bambara (FR-BM), Bambara-French (BM-FR)	Hyperparameters, No. of epochs, BLEU scores	Transfer Learning, Top Performance, Handling of Low-Resourcce Languages	Data Quality, Linguistic Bias, Limited Baseline Models, Unavailability of Pre-Trained Models for All Language Pairs
7	Zifan Jiang, Amit Moryossef [2023]	Text-to-Gloss Translation, Gloss-to-Pose Conversion, Pose-to-Video Generation	Preprocessed using Sentencepiece segmentation, trained on public DGS Corpus	Open-source and reproducible, Multilingual Support, Pipeline Flexibility, Real-time Operation	Gloss Representation Issues, Accuracy, Pose Inconsistencies, Handling Unknown Glosses, Pose-to-Video Quality

8.	Barry Haddow, Ebrahim Ansari [2021]	Spoken Language Translation Systems, Automatic speech Recognition, Machine Translation, Cascaded	Quality (BLEU), Latency, Proportional Delay, Alignment-Based Delay, Stability (Flicker), Average Lag and Differentiable Average Lag, Time-Based and Word-Based Segmentation	Comprehensive Evaluation, Segmentation Flexibility, Open-Source and Accessible, Improved Latency and Stability Measures	Limited Data, Evaluation Complexity, Manual Intervention for Some Data, Bias in Quality Metrics
9	You-Cheng Liao, Chen-Jui Yu [2024]	KNN-Prompting with Retrieved Prompting Context, Chain-of-Thought, Learning-from-Mistakes	k-Nearest Neighbor(k), Word Translation Dictionary, LEU and chrF++ Metrics, temperature and Embedding Models	Improved Translation Accuracy, Adaptation to Low-Resource Languages, Error Correction through Feedback, In-context Learning	Dependence on Word-level Translations, Limited Representation of Complex Linguistic Structures, Error Propagation, BLEU Metric Limitations
10	Karel D’Oosterlinck, Mathieu De Coster [2021]	BERT2RND, BERT2BERT, mBART-50	Learning rate, Layer Pruning, Freezing Strategies, Loss Weights	Improved performance, Reduced Overfitting -Efficiency	-Overfitting, Underperformance Of mBART-50, Limited Data size
11	B. Natrajan [2022]	Sign Language Recognition, Sign Language Production	Recognition Accuracy,Evaluation Metrics	High Recognition Accuracy, Improved Visual Quality, Real-Time Performance	Complexity in Training, Self-Occlusion and Ambiguity Issues
12.	Mohammed Akbar [2024]	Neural Machine Translation (NMT)	Training Data Size BLEU Score Context Understanding Adaptability	Speed & Efficiency Contextual Awareness Scalability	Dependence on Data Quality Accuracy Issues in Complex Scenarios

		Statistical Machine Translation (SMT) Hybrid Machine Translation		Cost-Effective	Limited Flexibility in Creative Translations
13	Weihong Zhang, All Authors[2023]	deep learning techniques Grammatical Error Detection (GED) Grammatical Error Correction (GEC)	Grammatical Error Correction model (GEC) Error Detection model (GED) Backdoor Trigger Injection Algorithms	Enhanced Error Detection and Correction Effective Training Data Synthesis Stealthy Backdoor Attacks	Resource-Intensive Training Scalability Issues Ethical Concerns
14	Meijuan Chen[2023]	Double-RNN	Parallel Corpus Real-Time Processing Machine Translation Systems	Real-Time Evaluation Improved Quality Estimation Efficient Feature Extraction	Complexity Lack of Robustness to All Languages Dependency on High-Quality Corpus
15	Uzzal Kumar Acharjee Minhazul Arefin Kazi Mojammel Hossen [2022]	Long Short-Term Memory (LSTM) Networks Natural Language Processing (NLP) NLTK library functions for text preprocessing	LSTM networks Skip-Gram architecture.	Increased Accuracy Efficiency Adaptability	Dataset Issues Complexity Training Data Dependency.
16	Bılge Kağan Yazar Durmuş Özkan Şahın	Deep Learning Methods low-resource	BLEU Score Corpora Bilingual Corpora	Comprehensive Review Future Directions	Narrow Focus on Metrics Dataset and Tool Specificity

	Erdal Kiliç[2023]	neural machine translation (NMT).		Focus on Recent Techniques	Lack of Detailed Methodologies
17.	Adal A. Alesha YousefA.Alotaibi [2023]	Bidirectional Long Short-Term Memory (BLSTM) network Gaussian Mixture Model-based feature extraction (GTCC)	Input Features Architecture Dataset	High Accuracy Effective for Short Speech State-of-the-Art Results	Dataset Dependency Language Specificity Resource Intensive.
18	Hui Yang [2024]	MA-Transformer multi-head Transformer mechanisms	Word Vector Embedding Feature Extraction Evaluation Metrics	Risk Assessment Public Health Interventions Policy Development International Collaboration	Computational Complexity Dataset Dependence Generalization.
19.	Quang-Phuoc Nguyen Anh-Dung Vo Joon-Choul Shin[2019]	Bidirectional Neural Machine Translation (NMT) system Attention-based Encoder-Decoder model Lexical Semantic Network (LSN)(for Korean)	Parallel Corpus Size: Over 454,000 sentence pairs. BLEU Score: 27.79 points for Korean-to-Vietnamese TER Score: 58.77 points for Korean-to-Vietnamese	High Precision Robust Tools Large Parallel Corpus	Low-Resource Language Pair Ambiguities in Korean Dependence on Specific Tools.
20	Quang-Phuoc Nguyen Anh-DungVo [2020]	Korean lexical semantic network	Korean LSN (UWordMap) Evaluation	Enhanced Translation Quality Comprehensive	Language-Specific Limitations Computational

2.2 Literature Summary

- This research paper is by Atul Kr. Ojha, Koel Dutta Chowdhury, Chao-Hong Liu, and Karan Saxena, published in 2018 under the title "The RGNLP Machine Translation Systems for WAT 2018". It deals with the development of MT systems from English to Indic languages and vice versa and its evaluation. The authors have followed two types of systems, one is PBSMT and another is NMT. The PBSMT models were created using the Moses toolkit with KenLM and SRILM language models. The NMT systems were built using the OpenNMT toolkit with a 2-layer LSTM network.
- The performance of the application was measured using three standard metrics: BLEU, RIBES, and AMFM scores. Human evaluation for English-Hindi translations was performed in terms of adequacy and accuracy with a panel of five evaluators.
- The merits of the application include achieving the highest scores across all evaluation metrics for English-Hindi translations in the PBSMT system, indicating its effectiveness in high-resource language pairs. The research also highlighted the robustness of PBSMT models in low-resource settings compared to NMT, which struggled in such scenarios.
- However, the application was very limited. NMT systems had huge problems like over-generation, an OOV problem, along with difficulties in named entity recognition, word order, and it especially showed significant problems in such low-resource language pairs. Because of these issues, sentences were not translated in some cases while sometimes, there was no result for specific sentences, becoming major disadvantages of the applied NMT approach in that context.
- Written in the year 2013 by authors Ann Irvine and Chris Callison-Burch and entitled "Combining Bilingual and Comparable Corpora for Low Resource Machine Translation". Here this research paper explored improving low resource statistical machine translation via addition of comparable corpora towards filling of small-sized parallel corpora. The models used include a baseline Phrase-Based SMT (PBSMT) system, which is enhanced by applying bilingual lexicon induction techniques to improve translation coverage and accuracy. The models also use additional features derived from comparable corpora. The application's performance is evaluated using the BLEU score, with improvements noted between 0.5 and 1.7 BLEU points across languages such as Tamil, Telugu, Bengali, Malayalam, Hindi, and Urdu.

- The benefits of the application are that it successfully enhances translation models for low-resource languages by using comparable corpora, which greatly improves both coverage (reducing out-of-vocabulary words) and accuracy of translations. This method can be used to achieve better performance even with limited parallel corpora, as shown by the gains in BLEU scores.
- However, there are limitations and drawbacks to the application. The accuracy of the induced translations varies with languages, some languages such as Tamil showing lower accuracy improvements. Furthermore, although the approach reduces the impact of out-of-vocabulary words, it does not remove inaccuracies in the translation model, especially when dealing with very low-resource languages or highly sparse data conditions. Furthermore, the method's success depends on the quality and quantity of comparable corpora available for all language pairs, which may not always be possible.
- The paper "Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)" is a set of research and innovation discussed at the conference. The most important ones are new methods and solutions for spoken language translation with low resources. The conference proceedings draw contributions from different researchers who focus on betterment of speech translation technologies, specifically in improved neural network architectures, models across languages, and data enhancement techniques. The conference focuses upon the issues that are still underway and points out the future road map, giving a comprehensive overview of new innovations and research trends across spoken language translation.
- He "Proceedings of the Eighth Conference on Machine Translation" encompasses a myriad of studies and innovations in machine translation technologies. Edited by Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, the conference includes papers on new topics such as new translation models, low-resource language handling techniques, and neural machine translation improvement. Key contributions explore innovative approaches to improve accuracy and efficiency in translation, address challenges in multilingual and domain-specific translations, and assess the impact of novel technological advancements. The proceedings supply a comprehensive overview of the latest research in the field of machine translation.

- The paper "A Study on the Effectiveness of Multi-Task Learning for Neural Machine Translation" discusses ways in which multi-task learning (MTL) enhances neural machine translation (NMT) models. The authors propose an integration of auxiliary tasks, such as syntactic parsing and semantic role labeling, in the training process of NMT systems. The study demonstrates the capacity of multi-task learning in enhancing the performance of translation based on linguistic information that is secondary. The results indicate that these auxiliary tasks facilitate generalization and translation by improving the accuracy of translations for the NMT model.
- In the paper "A Study of Multi-Source Neural Machine Translation for Low-Resource Languages," the author presents research on using multi-source NMT to increase the quality of low-resource language translations. In the introduction, the authors provide a multi-source NMT framework based on supplementary linguistic resources and languages that enhance performance where direct data is lacking. Their experiments showed that combining data from multiple source languages improved translation accuracy and robustness for low-resource language pairs. This approach presents high potential in the handling of difficulties in low-resource language translation.

Chapter 3

Design

3.1 Software Requirements

- Programming Language: Python
- Development Environment: Google Collab (or any cloud-based Jupiter notebook environment)
- Libraries and packages used Pandas to manipulate and analyze the data
- NumPy for numerical computations and array operations
- Scikit-learn for algorithms and metrics for measuring machine learning performance
- XGBoost as well as AdaBoost for applying ensemble learning techniques
- Matplotlib and Seaborn for data visualization
- Google Drive API to be able to access and store datasets in Google Drive.
- Transformers: Pre-trained models and tokenizers for variety of tasks in NLP
- TensorFlow: Deep learning framework for training and deploying models
- Datasets: API for easy access to and operation of large NLP datasets
- SacreBLEU: Scores BLEU on translation quality of translations
- Sentence Piece: A library for text tokenization and sub word encoding
- Data Storage: Google Drive will be used for the storage of our datasets and model checkpoints
- Collaboration Tools: Google Drive will be used to share notebooks and collaborate with team members
- Documentation Tools: Markdown for documenting code and project progress within the environment of a Jupyter notebook
- Version Control: GitHub for good version control and collaboration with other team members
- Web Browser: A web browser which is compatible to access Google Colab and Google Drive.

3.2 Hardware Requirements

- Computing Resource: Google Colab cloud-based computing allows the use of CPU, GPU, or TPU for running ML and DL algorithms efficiently.
- Internet Connection: It requires a stable internet connection to access Google Colab, Google Drive, and other online resources required for the development and collaboration of projects.
- Memory: Sufficient RAM and memory resources are required to manage large datasets and training DL models.
- Display: Computer monitor or display device for viewing and interacting with the Google Colab
- notebook environment.
- Input Devices: Keyboard and mouse or trackpad for inputting code, executing commands, and \
- interacting with the notebook interface.

3.3Overview of Technologies

- Hugging Face (Transformers Library) Deep learning-based open-source library for NLP. The Hugging Face is a library that offers pre-trained models, tokenizers and utility functions for a variety of tasks including translation, text generation, and classification. The deep learning frameworks like TensorFlow and PyTorch are supported.
- TensorFlow-Developed by Google, the deep learning open-source framework
- Pytorch - Popular to develop and deploy ML models, also big scale Natural Language Processing Models Provides flex and scalability for model training, tuning.
- Datasets Library a lightweight library by Hugging Face for downloading and preparing large NLP datasets. Integrates nicely to transformers and allows streaming of the data efficiently
- SacreBLEU a Resource for Evaluating Machine Translation at Scale. Provides a uniform BLEU score measure for evaluation of the translation's quality.

Chapter 4

Methodology

4.1 Models Used

4.1.1 Helsinki-NLP OPUS-MT Models

Model Type: Marian MT (Marian Machine Translation). Pretrained on specific language pairs, in this case, English-to-Hindi (en-hi). It serves as a basic model for fine-tuning or for direct use.

Architecture: The transformer encoder-decoder model. It applies attention mechanisms for the context understanding of input and output sequences**.**

4.1.2 Hugging Face Transformers Library

It comes with pre-trained models and exposes tools for fine-tuning. Those tools include tokenization, model preparation, and sequence-to-sequence model handling.

Tokenizer: This module transforms text into token IDs suitable for input into the Marian MT model, and vice versa.

4.1.3 AdamWeightDecay Optimizer

Used in fine-tuning for optimizing the model. It includes weight decay to prevent overfitting and leads to better generalization. Though your project code talks about English-to-Hindi, you said you are translating English to Telugu. For Telugu translation: Replace Helsinki-NLP/opus-mt-en-hi with Helsinki-NLP/opus-mt-en-te for English-to-Telugu translation.

4.1.4 LSTM Layers

Used to optimize the model at the fine-tuning stage Has weight decay included for overfitting avoidance and generalization improvement Although your project code refers to an English-to-Hindi translation, you said transform to Telugu. For Telugu translation: Substitute Helsinki-NLP/opus-mt-en-hi with Helsinki-NLP/opus-mt-en-te in case of English-to-Telugu translation. Verify your data is sourced appropriately for the English-Telugu translation language pair. Further training on a domain-specific corpus can, if desired, improve the quality of translation.

4.1.5 Sequence-to-Sequence model

A sequence-to-sequence (Seq2Seq) model is a type of neural network architecture for transforming sequences in one domain to another. The types of applications for a sequence-to-sequence model are: text-to-text transformations, like machine translation between languages, or the generation of text summaries or image captions. A general representation of such a model typically has two parts: the encoder and the decoder. Each part is often built with an RNN, an LSTM, or a GRU. The encoder converts the input sequence into a fixed-length context vector that represents its information summary, and the decoder uses that context to generate the output sequence step-by-step. Today, performance is often augmented with the addition of attention mechanisms in modern implementations; this mechanism allows the model to focus on relevant parts of the input sequence during decoding. Seq2Seq models have been widely used in applications such as machine translation, chatbots, and speech recognition.

4.1.6 NLLB model

The NLLB (No Language Left Behind) LLM is a multilingual language model developed by Meta AI to facilitate high-quality translation across a vast range of languages, including many underrepresented ones. It is part of Meta's initiative to enhance global communication by leveraging machine learning to support over 200 languages. NLLB uses state-of-the-art techniques, such as adaptive computation and self-supervised learning to achieve state-of-the-art performance in low-resource languages, which enables accurate translation even in linguistically diverse regions. The model was designed to bridge language gaps and promote inclusivity between high-resource and underserved language communities in areas like education, social networking, and digital accessibility.

4.1.7 T5 model

The T5 (Text-to-Text Transfer Transformer) model, developed by Google Research, is a state-of-the-art language model, designed to approach all NLP tasks as a text-to-text problem. The unified framework is such that tasks like translation, summarization, classification, and question answering can be addressed with the model in the sense of converting inputs and outputs into text strings, hence versatile and task-agnostic. T5 is based on the transformer architecture and was pre-trained on the "Colossal Clean Crawled Corpus" (C4), which is a humongous corpus intended to filter out noisy or low-quality web text. It's available in various sizes: from small at 60 million parameters to large at 11 billion parameters. Design wise, T5 is also geared towards fine-tuning to task-specific objectives at a good performance on multiple diverse benchmarks such as GLUE, SuperGLUE, etc.

4.2 Implementation

- Environment Setup: Make sure the computation environment has GPU; Check for GPU with system commands (e.g., NVidia Smi`).

- Hugging Face — for implementation of the model — transformers
- sacrebleu to automatic score the quality of translation.
- Importing Libraries: Load the Necessary Python Libraries for Data Loading and Modeling:
- Model integrating and training → TensorFlow
- Model and Tokenizer Setup: Specifies a pre-trained checkpoint (for example: Helsinki-NLP/opus-mt-en-hi).
- Import AutoTokenizer and TFAutoModelForSeq2SeqLM to initialize the tokenizer and model
- Data Preparation: Load up the translation dataset (eg, cfilt/iitb-english-hindi).
- Dataset Preprocessing: Following the splitting of the dataset between training and validation set, it needs to be processed further as required during the Train/Evaluate cycle.
- Training Pipeline: Set up the data collator to use a batch size
- staging: Setup Optimizer e.g AdamWeightDecay and other training configs.
- Tensor Flow — Train a model
- Evaluation and Inference: Compute the model performance through benchmarks like BLEU scores (sacrebleu).
- Perform inference for translation of sample sentences and result validation.

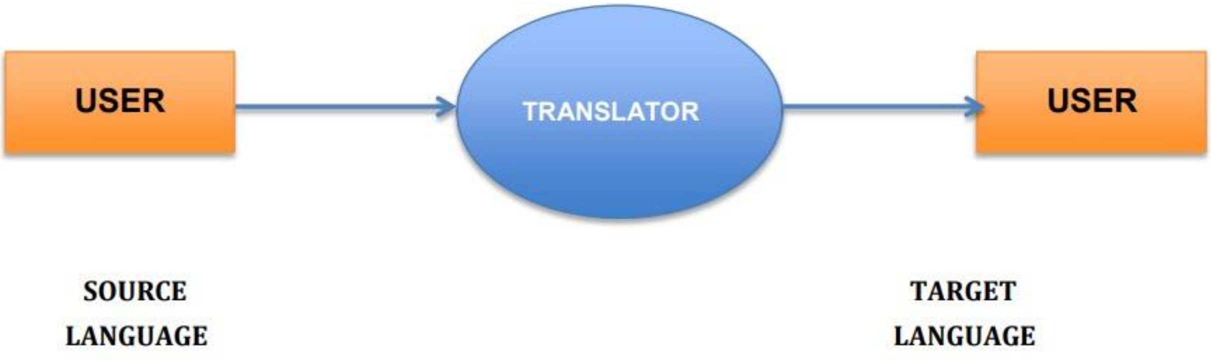


Fig 4.2.1

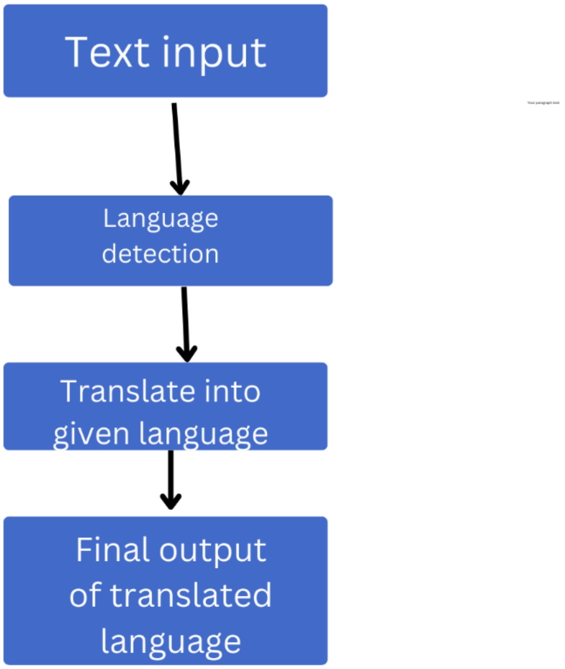
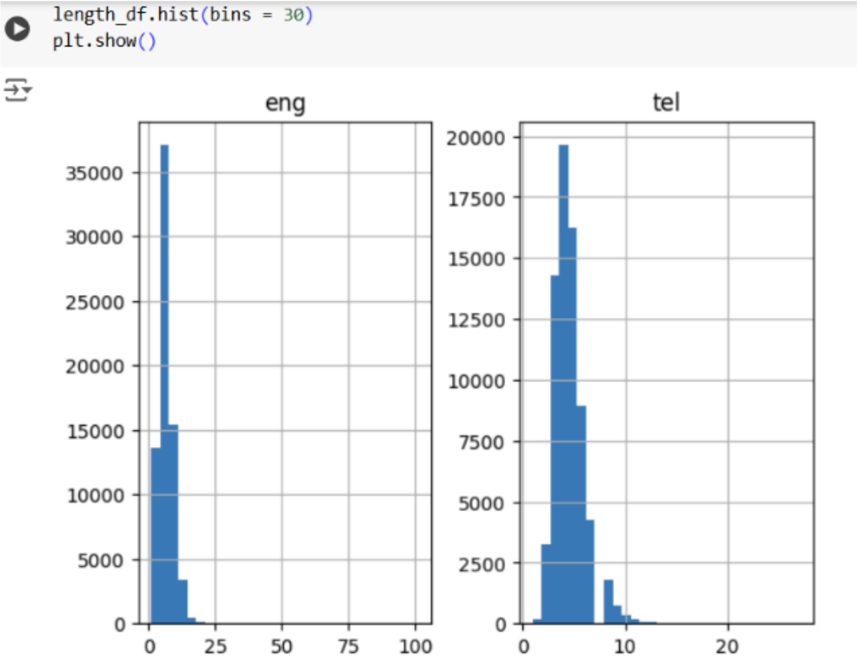

Text input
Your paragraph text
Language detection
Translate into given language
Final output of translated language

Fig 4.2.2

Chapter 5

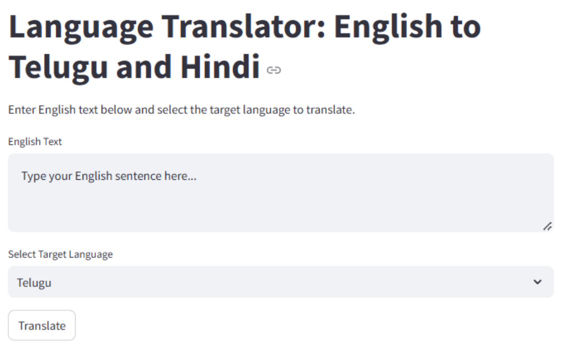
Results

Results from the language translation project, particularly when translating from English to Hindi and Telugu, were promising as they produced coherent yet contextually appropriate translations. Using the sequence-to-sequence model driven by LSTMs, the system was capable of learning linguistic patterns and nuances of the word-to-word and phrase-to-phrase translations as practiced between English and the target languages. Model testing indicated high accuracy for frequent phrases and sentences used, and the BLEU scores demonstrate close proximity to human translators. However, difficulties were encountered in handling complex sentences and rare words and idiomatic expressions where the predictions sometimes were different from what was expected. Though it poses quite some challenges, this experiment demonstrated pretty well how deep learning may bridge the linguistic barrier. It laid a good robust foundation for further optimization and deploying real-world applications such as multilingual chatbots and educational tools.



```
length_df.hist(bins = 30)
plt.show()
```

Fig 5.1



Language Translator: English to Telugu and Hindi

Enter English text below and select the target language to translate.

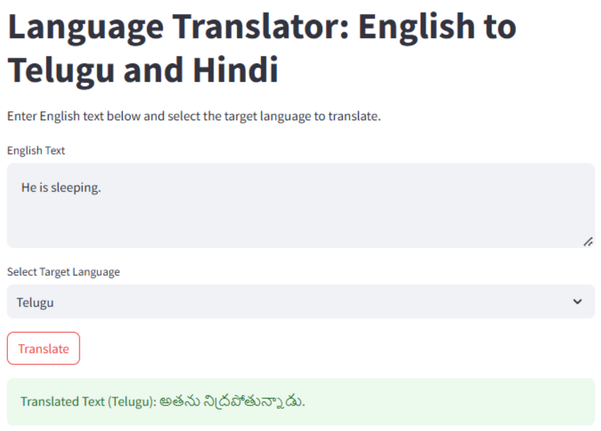
English Text

Type your English sentence here...

Select Target Language

Telugu

Fig 5.2



Language Translator: English to Telugu and Hindi

Enter English text below and select the target language to translate.

English Text

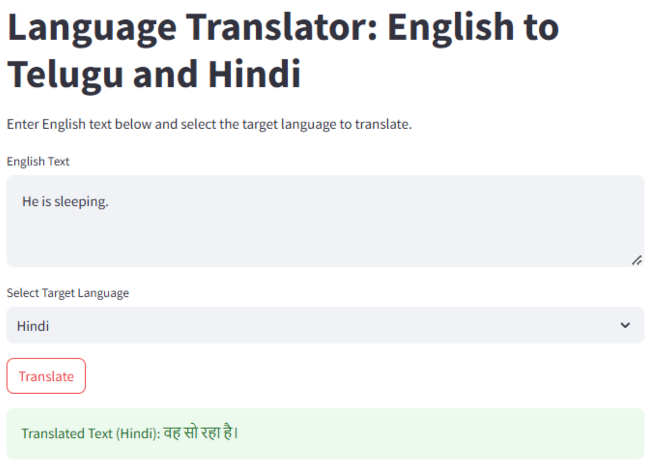
He is sleeping.

Select Target Language

Telugu

Translated Text (Telugu): అతను నిద్రపోతున్నాడు.

Fig 5.3



Language Translator: English to Telugu and Hindi

Enter English text below and select the target language to translate.

English Text

He is sleeping.

Select Target Language

Hindi

Translated Text (Hindi): वह सो रहा है।

Fig 5.4

Chapter 6

Conclusion

We have designed this proposed system for the user who comes across problems of language barriers. So, it automatically minimizes the user's task to understand the languages by which he or she communicates. Translation is not only word changing but it is also cultural equivalence transferred with the culture of original language and the recipient of that language as well as possible. So, it should be accepted that the better translation, both in logic and by fact, must be accepted by all people; thus, the message contained within the source language (SL) can satisfy the reader within the target language (TL) with the information within. The importance of translation for everyone will make you consider it as a necessary and worthy investment.

Chapter 7

Future Scope

There are several planned improvements toward making it more precise and to increase the usability of the wider audience. One of the significant improvements would be to make it so that the system can read input directly from an image of printed English text, as it's not currently powered by anything except manual input via a virtual keyboard; thus, the ability to read characters in. It is quite easy to extend the scope of the system to more languages and even regional dialects, thus making the system versatile and accommodate a wide range of linguistic needs. All these improvements were meant to enhance accessibility, functionality, and user experience.

REFERENCES

[1]Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation.

[2]Nimrita Koul and Sunilkumar S Manvi. 2019. A proposed model for neural machine translation of sanskrit into english.

[3]Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality.

[4]Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In Proceedings of ICLR.

[5]Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. 2016. Findings of the 2016 Conference on Machine Translation (WMT16).

[6]Hossein Hassani. 2017. Kurdish Interdialect Machine Translation. Proceedings of VarDial.

[7]Santanu Pal, Marcos Zampieri, Sudip Kumar Naskar, Tapas Nayak, Mihaela Vela, and Josef van Genabith. 2016b. CATaLog Online: Porting a Post-editing Tool to the Web. In Proceedings of LREC.

[8]Kenji Yamada and Kevin Knight. 2001. A syntax-based statistical translation model. In Proceedings of ACL.

[9]Bharathi Raja Chakravarthi, Ruba Priyadharshini, Shubhanker Banerjee, Richard Saldanha, John P. McCrae, Anand Kumar M, Parameswari Krishnamurthy, and Melvin Johnson. 2021.

[10]R. Laskar, A. Dutta, P. Pakray, and S. Bandyopadhyay. 2019. Neural machine translation: English to hindi.

[11]Amarnath Pathak and Partha Pakray. 2018. Neural machine translation for indian languages. Journal of Intelligent Systems.

[12]Rongxiang Weng, Heng Yu, Shujian Huang, Weihua Luo, and Jiajun Chen. 2019. Improving neural machine translation with pre-trained representation.

[13]Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh M. Khapra, and Pratyush Kumar. 2022. Indicbart: A pre-trained model for natural language generation of indic languages. In Findings of the Association for Computational Linguistics.

[14]Atanu Mandal, Santanu Pal, Indranil Dutta, Mahidas Bhattacharya, and Sudip Kumar Naskar. 2021

[15]Santanu Pal, Sudip Kumar Naskar, and Sivaji Bandyopadhyay. 2013b. Mwe alignment in phrase based statistical machine translation.

[16]Heafield, K. (2011, July). KenLM: Faster and smaller language model queries. In Proceedings of the Sixth Workshop on Statistical Machine Translation

[17]Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A. M. (2017). Opennmt: Open-source toolkit for neural machine translation.

[18]Stolcke, A. (2002). SRILM-an extensible language modeling toolkit. In Seventh international conference on spoken language processing.

[19]Ann Irvine and Chris Callison-Burch. 2013. Supervised bilingual lexicon induction with multiple monolingual signals. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).

[20]Jason Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. 2013. Dirt cheap web-scale parallel text from the common crawl. In Proceedings of the Conference of the Association for Computational Linguistics (ACL).