NLP Project Report

on

MACHINE TRANSLATION

by

Devshree Jadeja (20BCP112)

Khushi Shah (20BCP123)

Under the Guidance of

Dr. Santosh Bharti

Submitted to

Computer Science and Engineering,

School of Technology,

Pandit Deendayal Energy University

2023

CERTIFICATE

This is to certify that the project report entitled “Machine Translation,” submitted by Devshree Jadeja and Khushi Shah, has been conducted under the supervision of Dr. Santosh Bharti, Assistant Professor, and is hereby approved for the partial fulfillment of the requirements for the award of the degree of Bachelor of Engineering in the Department of CSE at Pandit Deendayal Energy University, Gandhinagar. This work is original and has not been submitted to any other institution for the award of any degree.

Sign:
Dr. Santosh Bharti
Assistant Professor
CSE Department
School of Technology
Pandit Deendayal Energy University

Sign:
Dr. Santosh Bharti
Assistant Professor
CSE Department
School of Technology
Pandit Deendayal Energy University

DECLARATION

We hereby declare that the seminar report entitled "Machine Translation" is the result of our own work and has been written by us. This report has not utilized any language model or natural language processing artificial intelligence tools for the creation or generation of content, including the literature survey.

The use of any such artificial intelligence-based tools was strictly confined to the polishing of content, spell checking, and grammar correction after the initial draft of the report was completed. No part of this report has been directly sourced from the output of such tools for the final submission.

This declaration is to affirm that the work presented in this report is genuinely conducted by us and to the best of our knowledge, it is original.

Devshree Jadeja
20BCP112

Khushi Shah
20BCP123

CSE Department
School of Technology
Pandit Deendayal Energy University
Gandhinagar

Date: 22 November 2023
Place: PDEU

ABSTRACT

Machine Translation (MT) plays a crucial role in breaking down language barriers and fostering global communication. This project focuses on the development of a state-of-the-art machine translation system for translating English text to Hindi using an encoder-decoder architecture with Long Short-Term Memory (LSTM) networks. The encoder processes the input English sequence, capturing its semantic meaning, while the decoder generates the corresponding Hindi translation. LSTM networks are employed to effectively model the sequential dependencies in the language data, allowing the system to grasp context and handle long-range dependencies. The training dataset comprises a diverse range of English-Hindi sentence pairs, and the model is fine-tuned to optimize translation accuracy. The project contributes to advancing natural language processing techniques, particularly in the domain of neural machine translation, and holds potential applications in facilitating cross-language communication and information exchange.

INDEX

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

In an era characterized by unprecedented global connectivity and communication, the demand for effective language translation systems has grown exponentially. As individuals, businesses, and institutions interact across linguistic boundaries, the need for accurate and efficient multilingual machine translation (MT) systems has become paramount. Machine Translation was the first computer-based application related to natural language, used during World War II. It is a dynamic and evolving field of Natural Language Processing (NLP) which is revolutionizing the way humans communicate across linguistic boundaries. It automates the process of translating text from one language to another, fostering global communication and breaking down language barriers that inhibit the exchange of information and ideas.

The need for effective machine translation has become increasingly prominent in our interconnected world, driven by the surge in global trade, cross-cultural collaboration, and the dissemination of information across diverse linguistic communities. As a result, researchers and practitioners are continually exploring innovative approaches and sophisticated algorithms to enhance the accuracy, fluency, and cultural sensitivity of machine translation systems.

India is a multilingual nation that speaks hundreds of dialects in addition to the eighteen officially recognized languages by the constitution. Despite the fact that less than 3% of Indians can understand English, it is still the primary language of business, education, and administration. Over 400 million people speak Hindi, which is the official language of the nation [5]. Consequently, given the sociological structure of the nation, machine translation becomes even more important in bridging the language gap.

This project embarks on a comprehensive exploration of machine translation, focusing on developing a robust and efficient model for translating English to Hindi. We delve into the intricacies of translating diverse languages, investigating cutting-edge techniques that harness the power of deep learning and neural network architectures. The primary goal is to overcome the challenges posed by linguistic variations, syntactic differences, and cultural nuances to deliver translations that closely capture the meaning and context of the source text.

CHAPTER 2
LITERATURE REVIEW

The expanding importance of neural network-based methods in the particular context of English-to-Hindi translation is discussed in the research [1]. With an emphasis on the rapidly developing subject of neural machine translation, the study investigates how deep learning methods—particularly neural networks—can be applied to improve translation fluency and accuracy. The authors assess the efficacy of their suggested neural machine translation model and look at the particular difficulties faced by the English-to-Hindi language pair. This work offers important insights on the use of cutting-edge technology in enhancing the quality of translation between Hindi and English by shedding light on the complexity of this language pair and outlining a specialized neural approach.

The paper [2] delves into the evolving landscape of machine translation by incorporating multimodal elements. The study combines textual and visual information to improve translation performance, with a focus on the English-to-Hindi translation job. The authors hope to collect more detailed contextual information and enhance the overall translation quality by fusing neural networks with multimodal inputs. By utilizing various informational modalities to handle the complexities of language translation, particularly in the English-to-Hindi domain, this study is essential to the exploration of novel approaches to machine translation.

The study [3] advances the field of machine translation by putting forth a novel hybrid mechanism designed specifically for translating from English to Hindi. The research blends statistical and rule-based methodologies, offering a thorough framework that maximizes the benefits of each technique to enhance translation fluency and accuracy. The hybrid method presents a viable way to improve machine translation systems' effectiveness by tackling the linguistic challenges unique to the English-to-Hindi language combination. The integrative approach taken by this study is noteworthy because it illustrates how different approaches can be combined to provide translation outcomes that are more reliable and efficient.

The research [4] significantly contributes to the development of linguistic resources for English-to-Hindi machine translation. The authors describe the construction of a large parallel corpora, HindEnCorp, with English and Hindi texts aligned. This corpus is a valuable resource for machine translation model evaluation and training, leading to advances in the field. The paper addresses the unique linguistic challenges posed by the English-to-Hindi language pair, while also emphasizing the value of high-quality parallel corpora for efficient machine translation. The initiative of Bojar et al. offers a fundamental resource that keeps influencing research in the creation and assessment of machine translation systems for the Hindi and English languages.

The study [5] makes a noteworthy contribution to the field of machine translation by introducing an innovative system designed specifically for English-to-Hindi translation. The authors emphasize the cooperative role of machine and human intelligence in the translation process as they present a machine-aided method. AnglaHindi is a big step in the right direction toward resolving the syntactic and linguistic quirks that distinguish Hindi and English. The system attempts to overcome the difficulties caused by linguistic variations by fusing human input with automated translation, thus improving the precision and caliber of English-to-Hindi translations.

In the research [6], authors investigated the use of deep learning models in conjunction with neural machine translation techniques to improve translation accuracy and capture minute linguistic details. The study examines the complexities of translating text from English to Hindi and assesses how well neural approaches work. The work highlights the continuous transition to neural-based approaches for better language translation results and offers insightful information about how well-suited sophisticated neural models are for translating between these languages. It also adds important perspectives to the field of machine translation.

The study [7] constitutes a noteworthy advancement in the field of machine translation, particularly in relation to the difficulties posed by Hindi-English language pairs. In order to improve translation accuracy and fluency, the authors suggest a hybrid methodology that combines rule-based and statistical approaches. Through the amalgamation of the advantages of both methodologies, the hybrid model endeavors to alleviate the constraints of distinct approaches and furnish a more resilient resolution. This study shows how hybrid approaches can improve the efficacy and efficiency of machine translation systems by investigating novel approaches to address the linguistic complexities present in Hindi-English translation.

The paper [8] represents a noteworthy investigation in the field of statistical machine translation (SMT) between Hindi and English. The purpose of the study is to determine whether integrating fundamental morphological and syntactic processing methods can enhance the efficiency of English-Hindi SMT systems. The authors hope to improve translation quality and address particular issues presented by the language pair by acknowledging the significance of linguistic structures and morphological features. The results of this study add significantly to our understanding of how to improve the performance of English-Hindi machine translation systems by shedding light on how to supplement conventional statistical methods with straightforward but efficient linguistic processing.

The research [9] addresses the application of deep learning techniques in machine translation, with a focus on Universal Networking Language (UNL) structures. The writers explore the domain of deep learning techniques, making use of their powers to translate and decode text according to the subtleties of UNL structure. By shedding light on how deep learning models might be modified to manage the difficulties involved with UNL, this study advances our knowledge of machine translation in the context of structured languages. The study illuminates novel approaches that have the potential to propel machine translation forward by investigating the amalgamation of deep learning and linguistic structures.

An extensive examination of deep learning's application to machine translation is given in the paper [10]. The writers explore how deep learning methods have developed and how they are being used in different facets of natural language processing, with a focus on machine translation. In order to demonstrate how well deep learning models can capture intricate linguistic patterns, the review covers a broad range of models, such as recurrent neural networks (RNNs), long short-term memory networks (LSTMs), and attention mechanisms. The authors provide a useful overview of the most advanced deep learning techniques in machine translation by combining important developments and difficulties. This greatly advances our knowledge of the field's present state and potential future directions.

CHAPTER 3
METHODOLOGY

The proposed workflow of our methodology involves a series of sequential steps as shown in Fig 3.1. The first step involves the collection of dataset, followed by the preprocessing stage where various text processing techniques are applied. The next step involves the application Encoder-Decoder based LSTM model, which is trained using the preprocessed dataset. Finally, the performance of the trained model is evaluated.

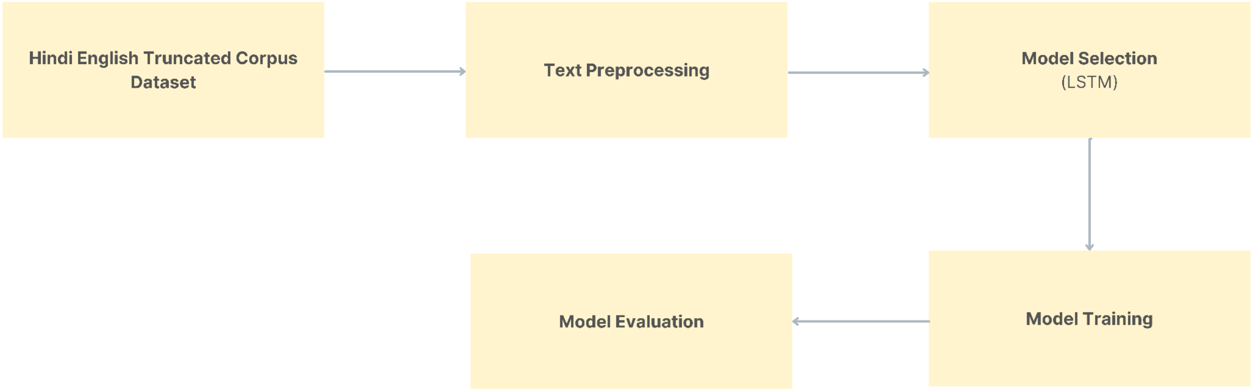


Fig 3.1 : Workflow of proposed methodology

3.1 Dataset

For this project, we have utilized the “Hindi_English_Truncated_Corpus” dataset from Kaggle which is publicly available. This dataset consists of around 1.2 million mirrored Hindi and English sentences, which offers a useful resource for working on natural language processing and machine translation tasks. It comprises three columns namely, source, english sentence and hindi sentence. The following figure 3.2 represents the distribution of English and Hindi sentences based on their token (word) count.

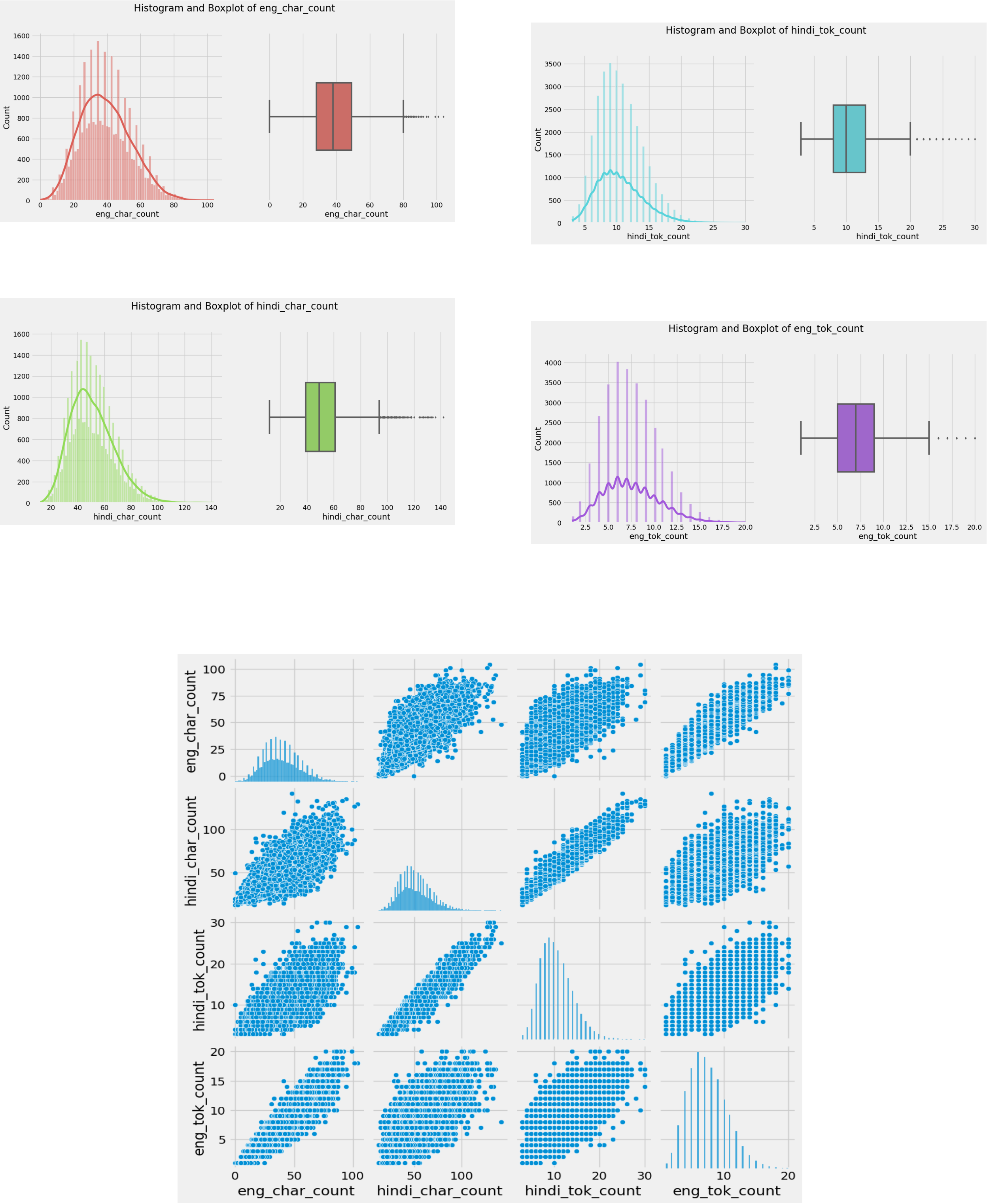


Fig 3.2 : Distribution of sentences wrt token count

3.2 Text Preprocessing

Text preprocessing is a crucial phase in natural language processing that involves cleaning and transforming raw text data into a format suitable for machine learning models. This preliminary step is essential to improve the caliber and efficiency of subsequent NLP tasks. The following preprocessing steps are employed in our project:

Lowercasing is a standard text preprocessing technique that involves converting all characters in a given text to lowercase. It is applied to ensure uniformity in text data, treating uppercase and lowercase letters as equivalent. By performing this, variations in casing are normalized, facilitating accurate word matching, reducing vocabulary size, and improving the generalization of NLP models.

Removing quotes is a text preprocessing step that entails removing quotation marks from a given text. This process is often performed to enhance the cleanliness and simplicity of the text data, particularly when the inclusion of quotes is not essential to the task at hand.

Removing special characters is a text preprocessing step that involves excluding non-alphanumeric symbols, punctuation, and other characters that do not belong to the standard alphanumeric set. The goal of removing special characters is to clean and simplify the text data, making it more suitable for various NLP tasks because they do not contribute significantly to the semantic meaning of the text and can introduce noise in the analysis.

Removing extra spaces is a text preprocessing technique that involves eliminating redundant white spaces from a given text. Extra spaces, including multiple consecutive spaces or leading/trailing spaces, can introduce noise and adversely impact the performance of NLP models. The goal of this preprocessing step is to standardize the spacing within the text, making it cleaner and more conducive to analysis.

Removing numbers is a text preprocessing step in natural language processing (NLP) that involves excluding numerical digits from a given text. The objective of this preprocessing technique is to simplify the text data by eliminating numeric values that may not contribute significantly to the semantic meaning of the text in certain NLP tasks.

Tokenization is a fundamental text processing step that involves breaking down a given text into individual units called tokens. These tokens can be words, subwords, or even characters, depending on the granularity of the tokenization process. We have performed work tokenization for converting both english and hindi sentences into words by splitting them based on space.

Vocabulary creation is one of the most fundamental step in Machine Translation. The coverage of the vocabulary is a major factor that drives the overall accuracy and the quality of the translation. The vocabulary is a comprehensive set of unique tokens encompassing the diverse linguistic elements present in the dataset. For creating the vocabulary for both Hindi and English language we implemented two approaches: word level tokenization.

3.3 Applied Models

3.3.1 LSTM

Long Short-Term Memory (LSTM) is a recurrent neural network (RNN) architecture widely used in Deep Learning. It excels at capturing long-term dependencies, making it ideal for sequence prediction tasks. Because LSTM has feedback connections, as opposed to traditional neural networks, it can process entire data sequences as opposed to just single data points. It can therefore recognize and forecast patterns in sequential data, such as time series, text, and speech, with great effectiveness [11].

The LSTM architecture consists of memory cells and three gates namely, input gate, forget gate, and output gate as shown in Fig 3.2 and Fig 3.3. Input gate determines which parts of the new input should be stored in the cell state, updating and adding relevant information to the cell state for future reference, forget gate helps to decide whether we should keep the information from the previous time step or forget it and output gate determines what information from the current cell state should be passed as the output of the LSTM at a particular time step. It controls the flow of information from the memory cell to the output.

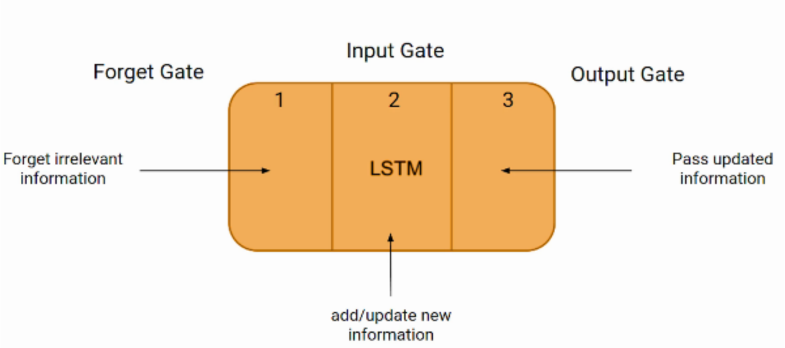
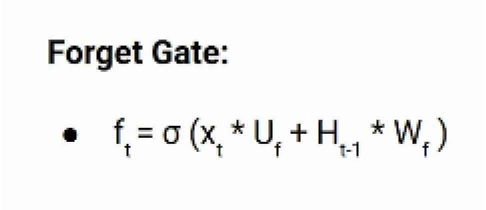
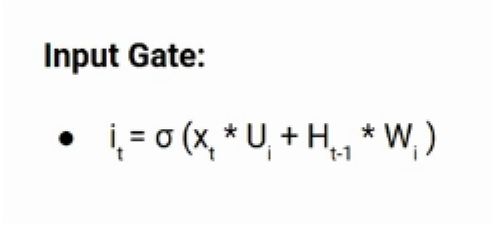


Fig 3.3 : LSTM



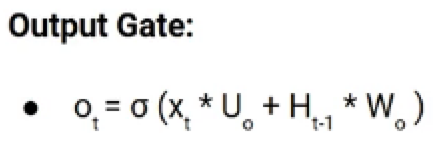
Forget Gate:

- $f_t = \sigma (x_t * U_f + H_{t-1} * W_f)$



Input Gate:

- $i_t = \sigma (x_t * U_i + H_{t-1} * W_i)$



Output Gate:

- $o_t = \sigma (x_t * U_o + H_{t-1} * W_o)$

Where,
Xt: input to the current timestamp
Uf: weight associated with the input
Ht-1: The hidden state of the previous timestamp
Wf: It is the weight matrix associated with the hidden state.

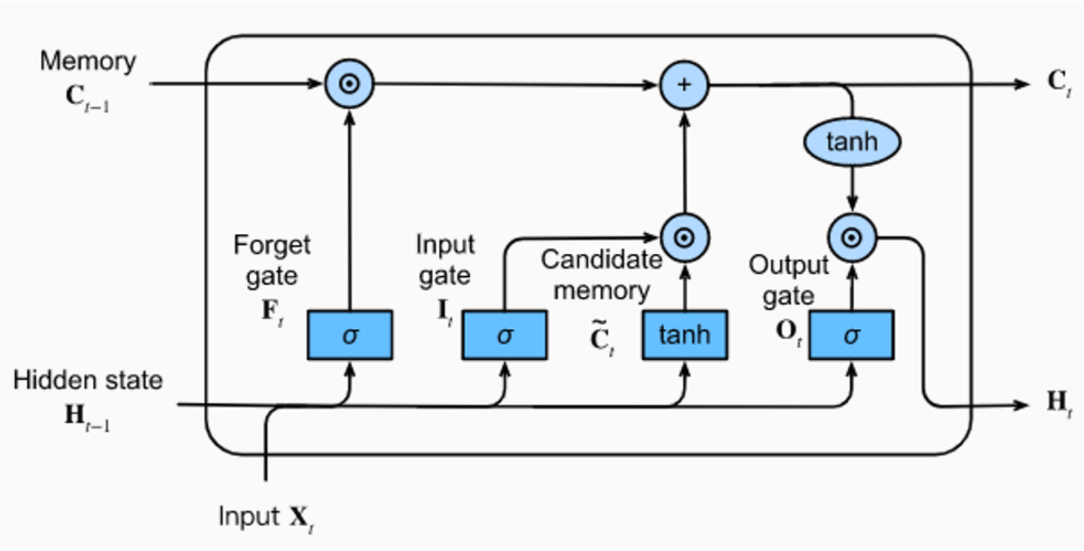


Fig 3.4 : LSTM Architecture

3.3.2 RNN

Recurrent Neural Network (RNN) is a type of Neural Network where the output from the previous step is fed as input to the current step. In traditional neural networks, all the inputs and outputs are independent of each other. Unlike traditional feedforward networks, RNNs possess recurrent connections, allowing them to maintain a hidden state that captures information from previous inputs. The main and most important feature of RNN is its Hidden state, which remembers some information about a sequence. The state is also referred to as the Memory State since it remembers the previous input to the network. This intrinsic memory feature makes RNNs particularly well-suited for tasks involving sequential dependencies such as natural language processing, time series analysis, and machine translation.

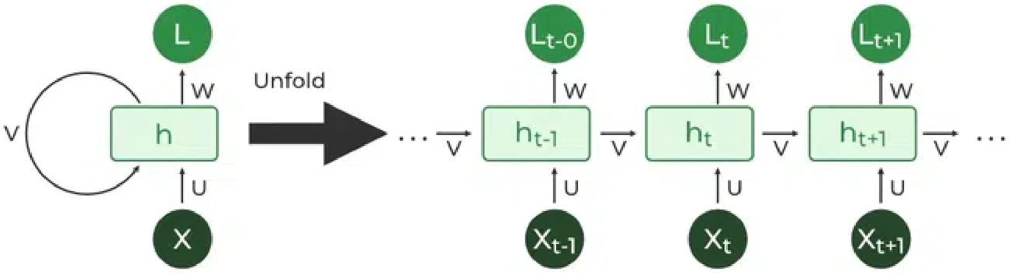


Fig 3.5 : RNN

3.4 Evaluation Metrics

Model evaluation involves assessing the performance of translation systems in converting text from one language to another. Given the unique challenges of translation tasks, several specialized metrics are commonly used for this purpose. We have used BLEU score to evaluate the performance of our system.

3.4.1 BLEU Score

The BLEU (Bilingual Evaluation Understudy) score is a metric commonly used to evaluate the quality of machine-translated text by comparing it to one or more reference translations. BLEU measures the similarity between the machine-generated output and the reference translations based on n-gram precision. The BLEU score ranges from 0 to 1, where 1 indicates perfect agreement between the machine-generated output and the reference translations. Higher BLEU scores generally indicate better translation quality, but it's essential to interpret the scores carefully and consider other evaluation metrics for a more comprehensive assessment.

CHAPTER 4
PROCEDURE

4.1 LSTM

We implemented an Encoder-Decoder LSTM architecture for our machine translation system as shown in Fig 4.1. This architecture is particularly effective for handling sequence-to-sequence tasks like machine translation, where the input and output sequences can have variable lengths.

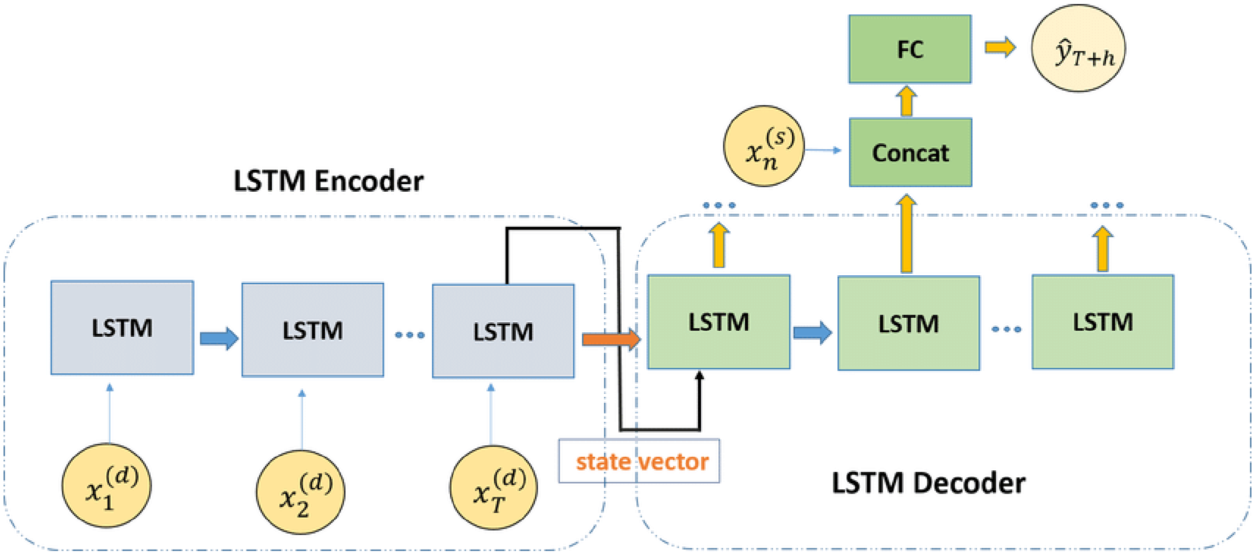


Fig 4.1 : Encoder-Decoder LSTM Architecture

Encoder:
The encoder processes the input sequence, which is the English text. The goal of the encoder is to convert the input sequence into a fixed-size context vector, capturing the semantic meaning of the input. The LSTM cells within the encoder play a crucial role in this process. Each word in the input sequence is fed into the LSTM cells one at a time. The LSTM cells maintain a hidden state that is updated at each time step. The final hidden state of the encoder LSTM captures the information from the entire input sequence and acts as a context vector. This context vector is then passed to the decoder for generating the target sequence.

Decoder:
The decoder takes the context vector from the encoder and generates the output sequence, which is the Hindi translation in this scenario. Like the encoder, the decoder also consists of LSTM cells. However, the decoder LSTM cells have a dual role: they take the context vector from the

encoder as an initial hidden state and generate the output sequence word by word. At each time step, the decoder LSTM cell produces an output word, and this word is used as input for the next time step. This process continues until the entire target sequence is generated.

Training:
During training, the model is fed with pairs of English sentences and their corresponding Hindi translations. The model's parameters, including the weights in the LSTM cells, are adjusted to minimize the difference between the predicted and actual translations. This is typically done using a loss function such as categorical cross entropy

In summary, the Encoder-Decoder LSTM architecture effectively captures the contextual information of the input sequence and uses it to generate a meaningful output sequence.

4.2 RNN

Later, we implemented an Encoder-Decoder RNN architecture for our machine translation system as shown in Fig 4.2. In this model, the encoder processes input sequences and transforms them into a fixed-size context vector, capturing the input's semantic information. This context vector becomes the initial hidden state of the decoder, which generates the output sequence step by step.

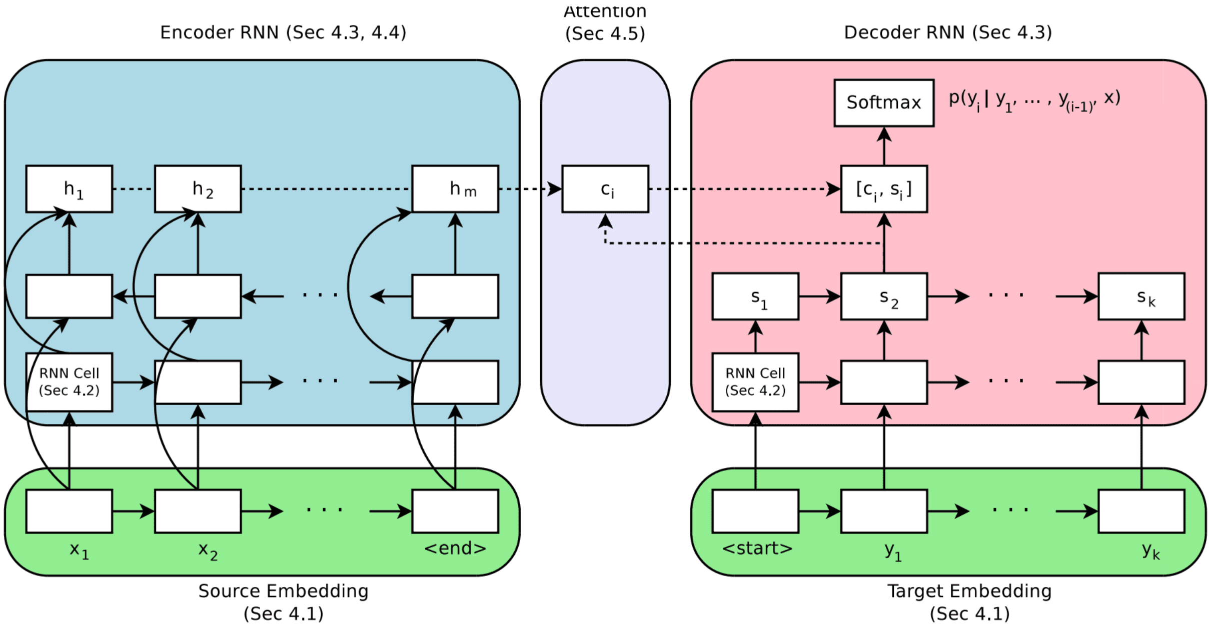


Fig 4.2 : Encoder-Decoder RNN architecture

Encoder:

The encoder of this network is a RNN that outputs some value for every word from the input sentence. For every input word the encoder outputs a vector and a hidden state, and uses the hidden state for the next input word.

Decoder:
The decoder is another RNN that takes the encoder output vector(s) and outputs a sequence of words to create the translation. In the simplest seq2seq decoder we use only the last output of the encoder. This last output is sometimes called the context vector as it encodes context from the entire sequence. This context vector is used as the initial hidden state of the decoder.

Attention:
Attention allows the decoder network to “focus” on a different part of the encoder’s outputs for every step of the decoder’s own outputs. First we calculate a set of attention weights. These will be multiplied by the encoder output vectors to create a weighted combination. The result contains information about that specific part of the input sequence, and thus helps the decoder choose the right output words.

CHAPTER 5
RESULT ANALYSIS

The results of the machine translation project utilizing the encoder-decoder LSTM architecture for English to Hindi demonstrate promising outcomes. The LSTM model, known for its ability to capture sequential dependencies, has effectively learned the complex mappings between English and Hindi sentences. The training and validation loss graph as shown in Fig 5.1, exhibits a desirable downward trend, affirming the model's effective learning.

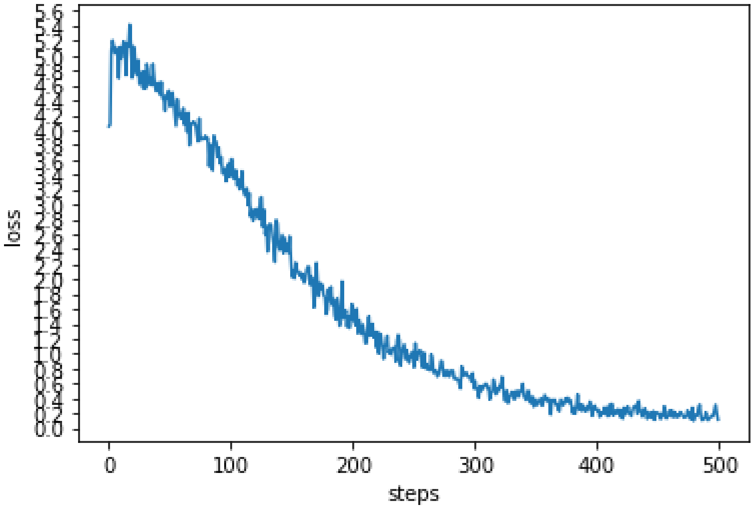
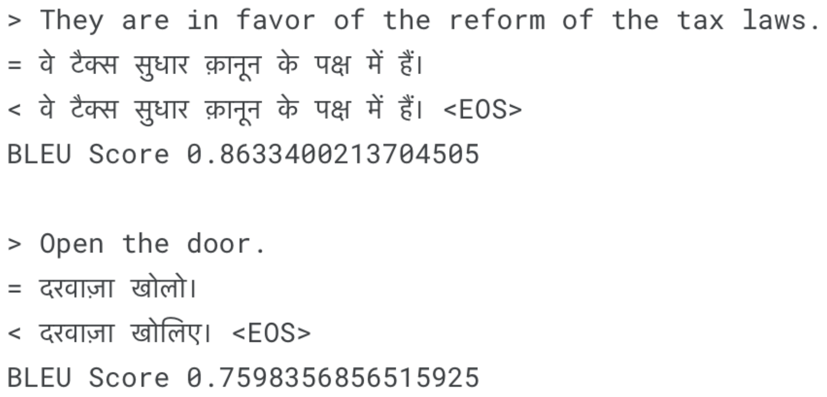


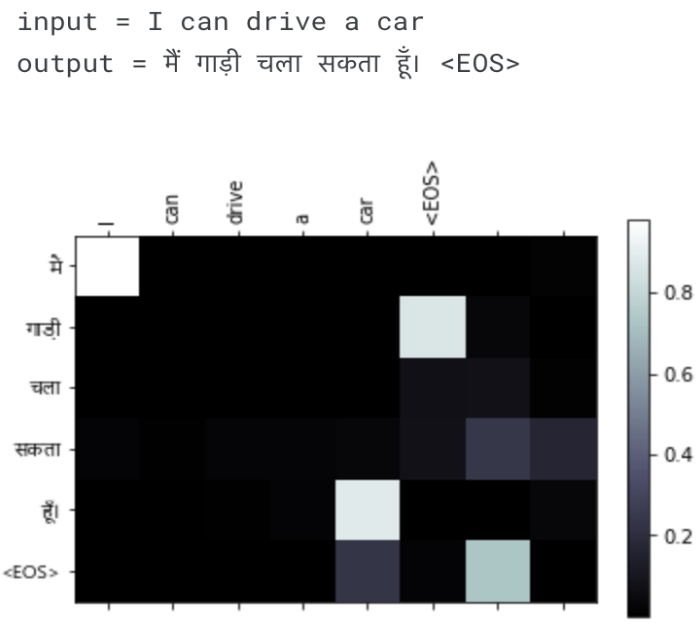
Fig 5.1 : Loss

The BLEU scores of a few examples are represented below.

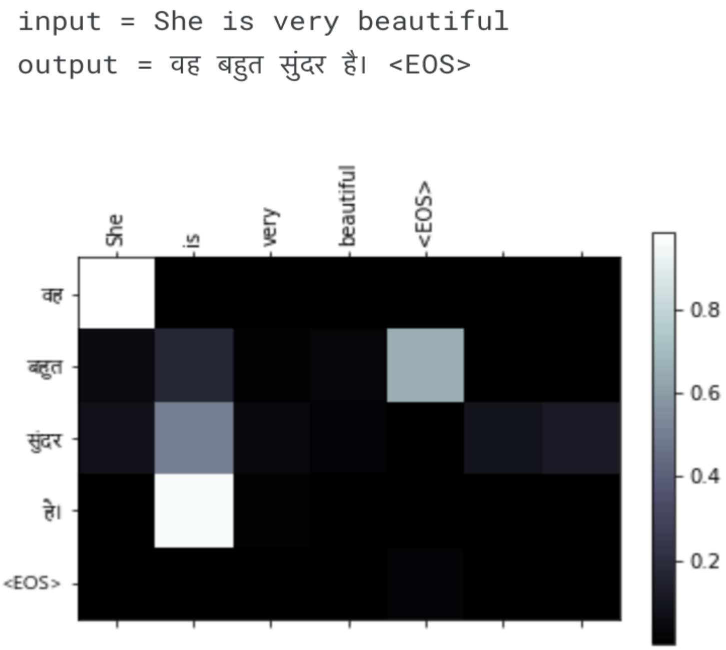


```
> They are in favor of the reform of the tax laws.
= वे टैक्स सुधार क़ानून के पक्ष में हैं।
< वे टैक्स सुधार क़ानून के पक्ष में हैं। <EOS>
BLEU Score 0.8633400213704505

> Open the door.
= दरवाज़ा खोलो।
< दरवाज़ा खोलिए। <EOS>
BLEU Score 0.7598356856515925
```



```
input = I can drive a car
output = मैं गाड़ी चला सकता हूँ। <EOS>
```



```
input = She is very beautiful
output = वह बहुत सुंदर है। <EOS>
```

Fig 5.2 : Word Mapping from English to Hindi

CONCLUSION and FUTURE SCOPE

In the culmination of this machine translation project employing the encoder-decoder architecture, significant strides have been made toward bridging the language barrier between English and Hindi. The applied model, which is based on LSTM networks, has demonstrated good performance in understanding and translating intricate linguistic structures. The system has showcased its ability to produce logical and contextually accurate translations after undergoing thorough training on a carefully selected dataset.

The success of this project underscores the significance of leveraging advanced neural network architectures, such as LSTMs, in natural language processing tasks. However, it is essential to acknowledge the challenges and limitations encountered during the development, including the need for substantial amounts of training data and computational resources.

The future trajectory of this project extends into several promising directions. Expanding the dataset to include a more diverse and extensive set of language pairs would enrich the model's linguistic understanding. Exploring advanced architectures like transformers and attention mechanisms offers potential for further improvement in translation quality and efficiency.

Fine-tuning the model for domain-specific translations and continuous adaptation to evolving language trends are promising prospects. The model's adaptability and user satisfaction may be improved by incorporating user feedback and participation in the training process. Moreover, exploring scalability and multilingual translation capabilities for practical applications continues to be an exciting avenue of research.

In summary, the project has laid a foundation for effective machine translation between English and Hindi, showcasing the potential of deep learning techniques in overcoming language barriers. The continuous evolution of neural network architectures and the availability of more extensive datasets will undoubtedly contribute to further advancements in the field of machine translation and promoting effective communication across diverse linguistic landscapes.

References

[1] Saini, S. and Sahula, V., 2018, March. Neural machine translation for english to hindi. In *2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)* (pp. 1-6). IEEE.

[2] Laskar, S.R., Khilji, A.F.U.R., Pakray, P. and Bandyopadhyay, S., 2020, December. Multimodal neural machine translation for English to Hindi. In *Proceedings of the 7th Workshop on Asian Translation* (pp. 109-113).

[3] Nair, J., Krishnan, K.A. and Deetha, R., 2016, September. An efficient English to Hindi machine translation system using hybrid mechanism. In *2016 international conference on advances in computing, communications and informatics (ICACCI)* (pp. 2109-2113). IEEE.

[4] Bojar, O., Diatka, V., Rychlý, P., Stranák, P., Suchomel, V., Tamchyna, A. and Zeman, D., 2014, May. HindEnCorp-Hindi-English and Hindi-only Corpus for Machine Translation. In *LREC* (pp. 3550-3555).

[5] Sinha, R.M.K. and Jain, A., 2003. AnglaHindi: an English to Hindi machine-aided translation system. In *Proceedings of Machine Translation Summit IX: System Presentations*.

[6] Laskar, S.R., Dutta, A., Pakray, P. and Bandyopadhyay, S., 2019, December. Neural machine translation: English to hindi. In *2019 IEEE conference on information and communication technology* (pp. 1-6). IEEE.

[7] Dhariya, O., Malviya, S. and Tiwary, U.S., 2017, January. A hybrid approach for Hindi-English machine translation. In *2017 international conference on information networking (ICOIN)*(pp. 389-394). IEEE.

[8] Ramanathan, A., Hegde, J., Shah, R., Bhattacharyya, P. and Sasikumar, M., 2008. Simple syntactic and morphological processing can help English-Hindi statistical machine translation. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

[9] Ali, M.N.Y., Rahman, M.L., Chaki, J., Dey, N. and Santosh, K.C., 2021. Machine translation using deep learning for universal networking language based on their structure. *International Journal of Machine Learning and Cybernetics*, *12*(8), pp.2365-2376.

[10] Liu, Y. and Zhang, J., 2018. Deep learning in machine translation. *Deep learning in natural language processing*, pp.147-183.

[11] Saxena, S. (2023, October 25). *What is LSTM? Introduction to Long Short-Term Memory*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/03/introduction-to-long-short-term-memory-lstm/#:~:text=LSTM%20(Long%20Short%2DTerm%20Memory,ideal%20for%20sequence%20prediction%20tasks.