CSCE 5290: Natural Language Processing

Increment 1

Project Title

Hindi Literature Translation to English and Summarize the translated literature using Text Summarization method.

Team Members

1. Vinay Chowdary Vemuri
2. Keerthika Kondaveeti
3. Sai Priya Pandeti Vasantha Kumar
4. Sampath Sai Sandeep Gompa

Contents

Goals and Objectives:

1. Motivation

Around the world we have 600 million Hindi speakers and it has critically famous literature. Hindi literature is not very widely read and referred, due to its difficulty to understand grammar, but English is accepted throughout the world. To give Hindi a wide range we want to translate the Hindi literature to English. We also want to summarize the literature so that user can get an understanding about the literature before reading it completely.

2. Significance

The book “Tomb of Sand” (“Rait samadhi”) which won Booker Prize last year was from Hindi literature. Booker prize is a prestigious award given every year in the field of literature for best book in the year. This was possible due to human translation of the book. Not every book author has this luxury to get a human translator. We want to solve this problem of translating the text from Hindi to English and summarize it to make it reach the wide audience around the world.

3. Objectives: The objectives of the project are as follows:

• The primary objective of the project is to translate Hindi literature to English language and ensure that the original meaning, context and style of the literature are present and express correctly in English language.

• The second objective of the project is to summarize the translated text in English using Text summarize algorithm. The Summarization of the text will help users who want to explore Hindi literature can understand the plot, ideas of the literature before deep diving.

• The project aims to promote cultural exchange between the English-speaking communities and Hindi Speaking communities through the translation and summarization of the Hindi literature.

4. Features: The key features of the project at increment 1 stage are as follows:

• **Datasets:** We are using an English – Hindi Parallel corpus dataset for Hindi to English translation open sourced by IIT Bombay researchers. We are using Shakespeare dataset for the text summarization.

• **Translation Process:** In the pre-processing stage we use the tokenization method to tokenize the text and use this data for the next processing steps.

• Next in the feature extraction we use word2vec to represent words as vectors. Word2Vec is used to convert the text into the vector form which will be used in the next steps.

• Next, we train Encoder- Decoder model using our dataset which maps our input Hindi to output English translation using back propagation method to adjust neural network weights and loss minimization.

Increment 1 Report

Related Work (Background)

Machine translation of Hindi to English has a very good background on the research part. IIT Bombay has team of researchers who are working on the building the model of Hindi to English translation for wider application. In this project we are using their dataset corpus. Although we have multiple researches on the similar machine translation but machine translation with text summarization has less research. We have worked on the integration of Machine translation from hindi to english and text summarization. We have attached few of the references of the papers that we considered while researching on the problem statement.

Dataset

We used the IIT Bombay English to Hindi Corpus Dataset and combination of the test dataset generated. The dataset corpus has been developed at the Indian Language Technology IIT Bombay. The Dataset corpus contain the English to Hindi and monolingual Hindi corpus which are collected from various sources.

https://www.kaggle.com/datasets/vaibhavkumar11/hindi-english-parallel-corpus

Detail design of Features

The following features are designed and implemented in the project for Machine translation part.

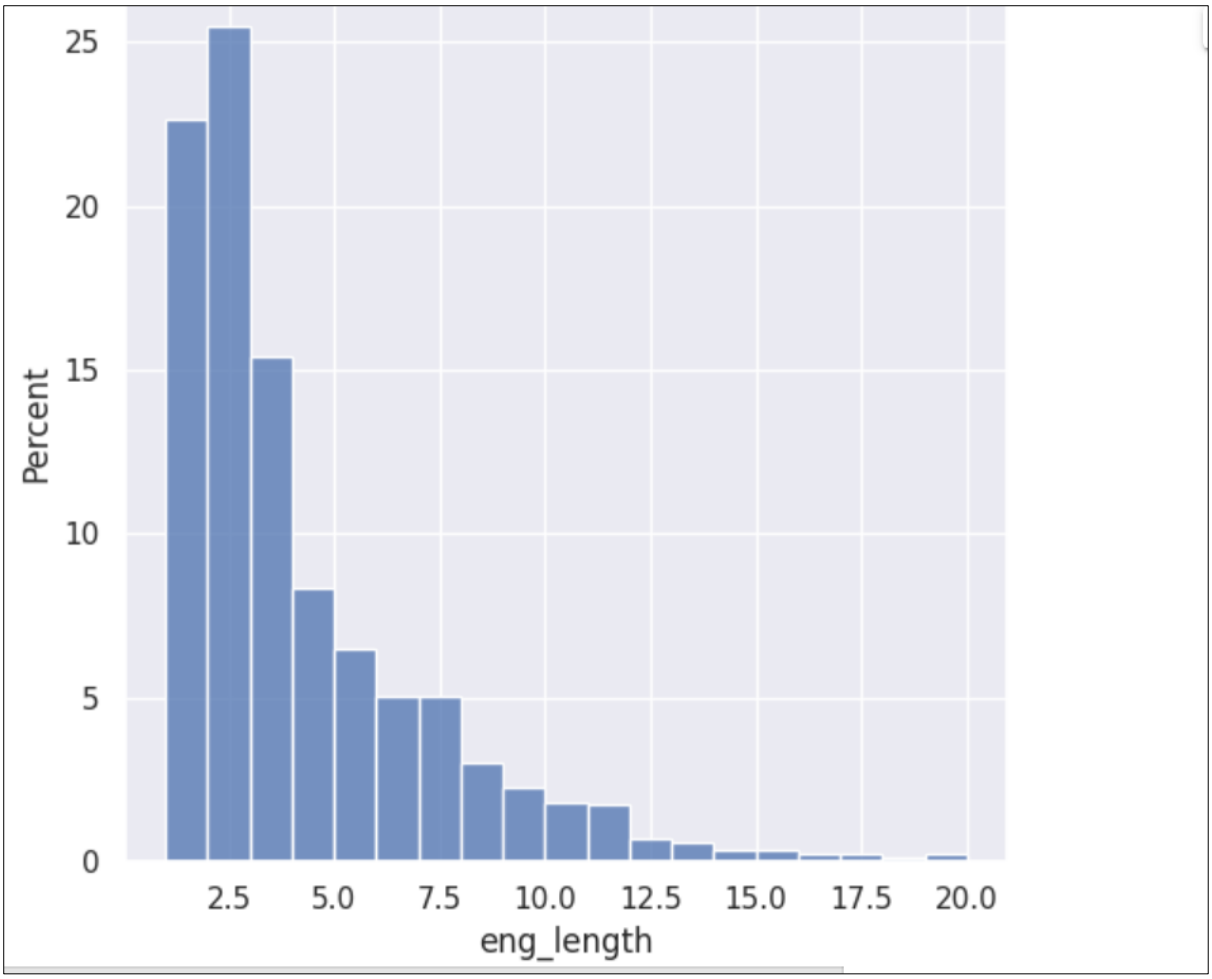
1. Spell correction – This is a pre-processing step where we identify and correct the misspelled words in the data.
2. Removing punctuation – This is a pre-processing step where we remove punctuation to reduce the noise.
3. Preparing Word 2 Vector Dictionary – Here we vectorize the sentences from the given corpus data.
4. Glove Word Embedings – Here we load the pre-trained word embedding in to the memory which are used for the further analysis.
5. Integer encoding for Hindi Vocabulary – Here we perform integer encoding using labeleencoder in the hindi dicitonary
6. Mean square loss Computation – This is a loss function which is used to indicate which part of the sequence should be ignore during the training and evaluation.
7. Seq to Seq tensors flow computation graph with attention mechanism – Here we generate the graphs which contains the placeholders for the input data and the labels. Here attention mechanism is used to weight the importance of each word while generating output.
8. Training and prediction – Training the model and prediction.

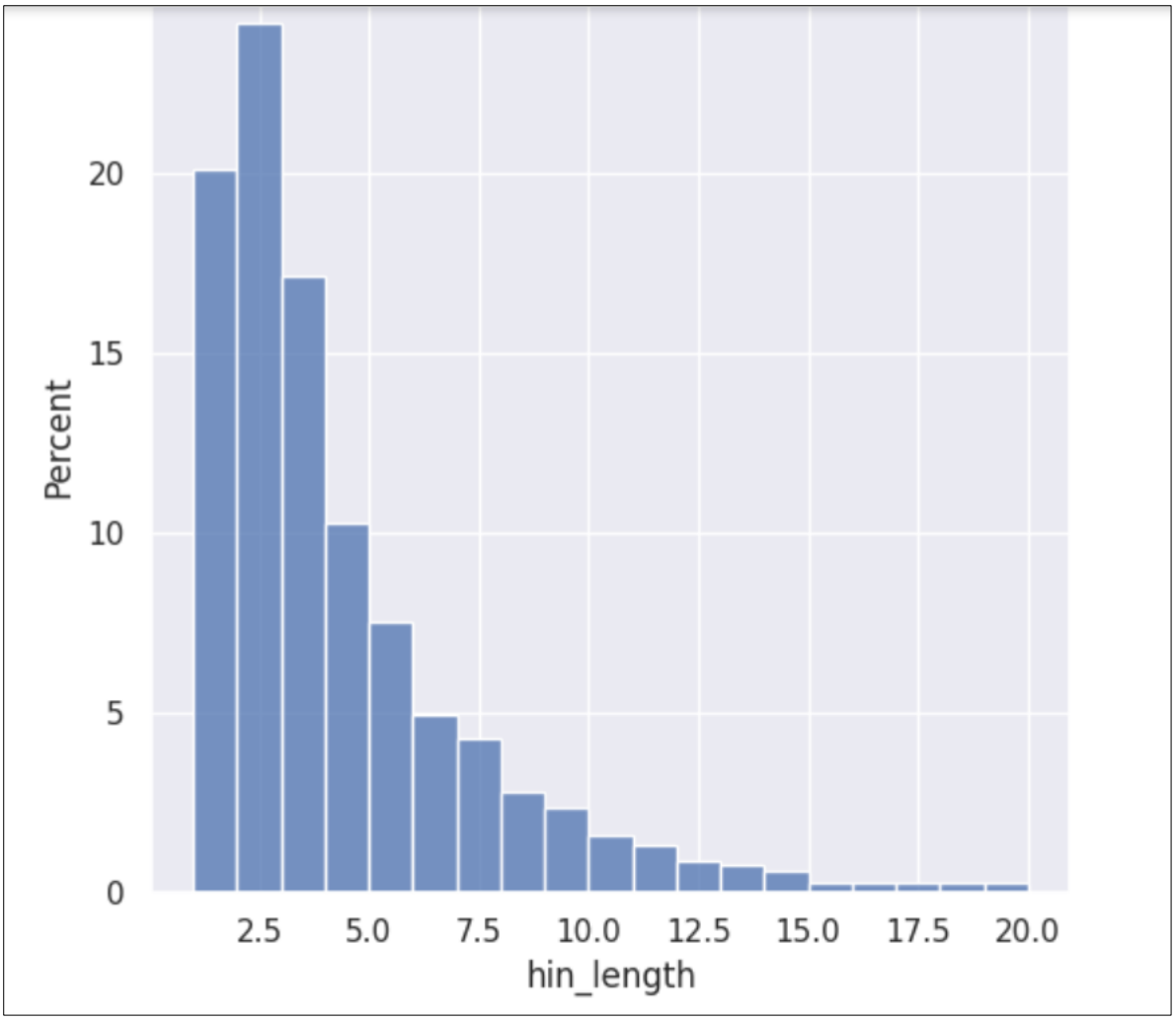
Analysis and Preliminary Results

Hindi to English translation has huge application from literature translation to multiple day to day basis applications. To understand the datasets we tried to make few analysis in the dataset.

We made few analysis on the length of the sentences in Hindi and English

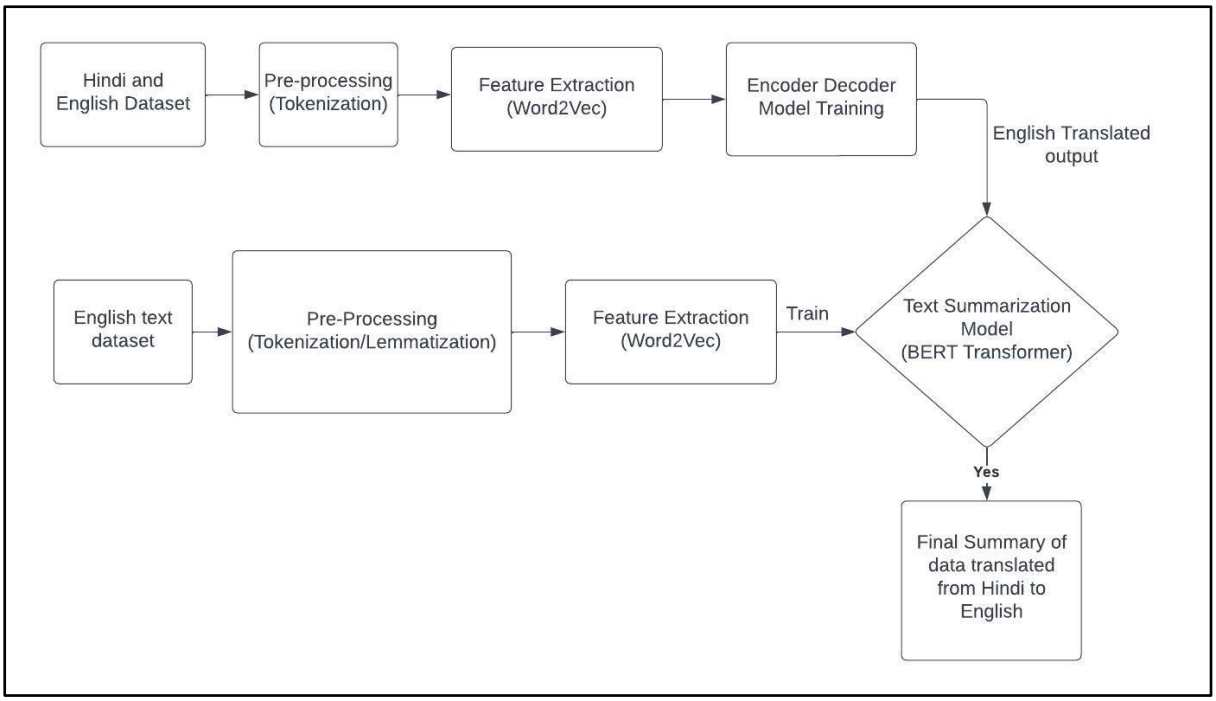
We can see from the graphs that most of the length of sentencens in hindi is of length 2.5 to 5 in both English and Hindi sentences.


25
20
15
10
5
0
Percent
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
eng_length



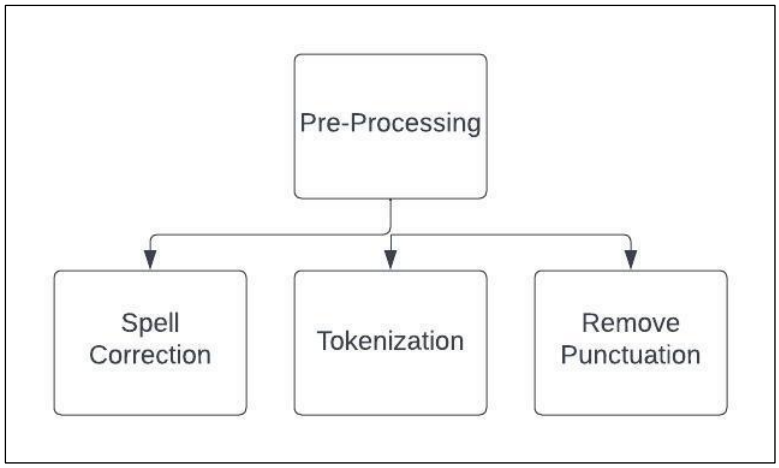
Implementation

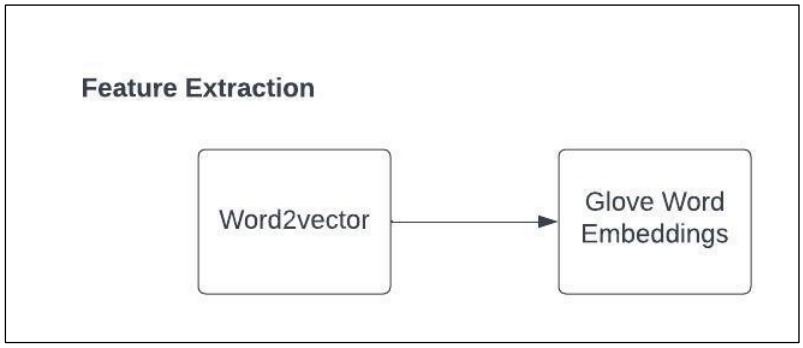
Here the project has two parts, Machine translation of the text from hindi to English and Text Summarization of the text.



In the Increment 1 we worked mainly in the Data Analysis and Machine translation part.

Pre-processing - Here first we use the word_tokenize function from nltk library to tokenize the words and then add the each token in to the word_hind_dic and word_eng_dic. It used a spell checker spell() function to check the spelling of the word. If the last letter in the hindi word has “|”, it removes it and add the word to the dictionary. Here overall we performed the pre-processing of the data where we performed tokenization, spell correction and removed punctuation for the dictionaries for English and Hindi.





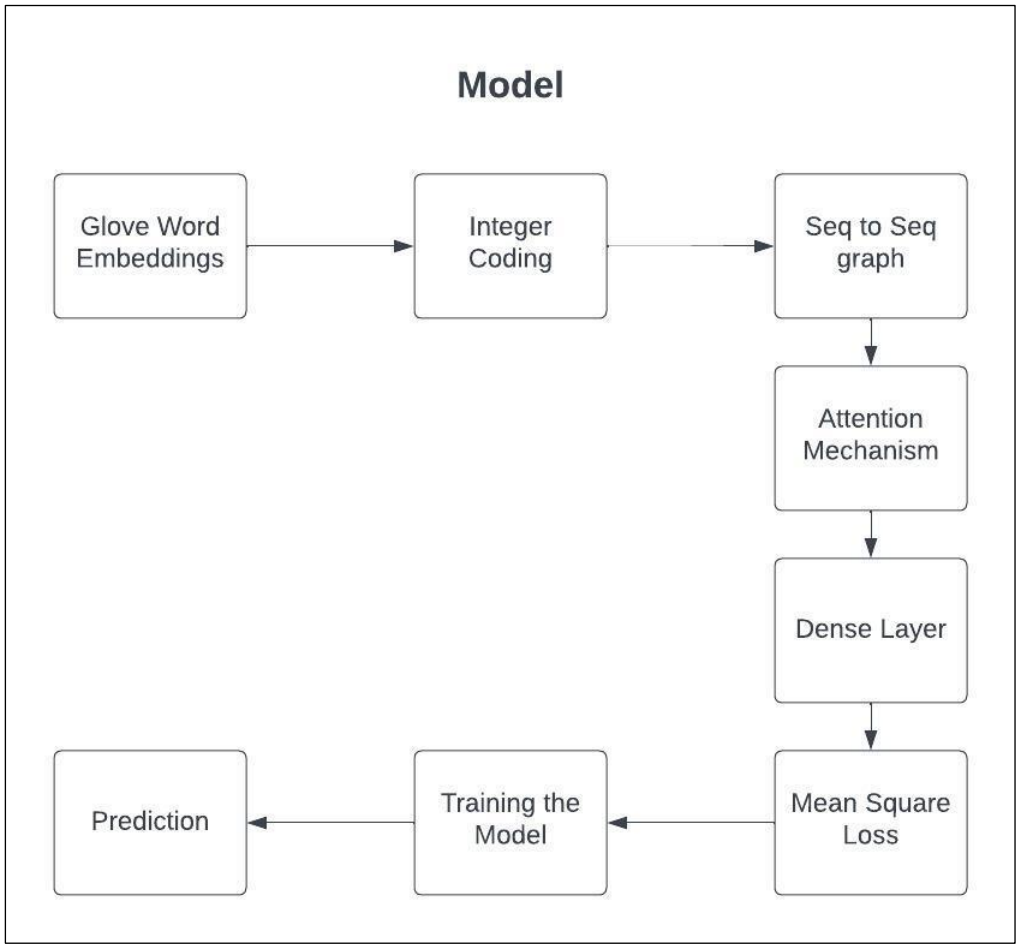
Feature extraction and Model –

- Here we vectorize the dictionary which is corresponding to the to every word.
- Here we made a computational graph in the Tensor flow
- Placeholders - Hindi sentences are represented as inputs in the tensor flow graph. English sentences are represented as labels in the tensorflow graph.
- We have shape variables for the hindi and English depending on the length of the sequence.
- Next we convert hindi vocabulary to embedding size with 50 by embedding look up method.
- Seq to Seq – Here for the Encoder and Decoder user lstm cell with vector size of 30, Encoder takes input from the embedding layer and Decoder take input from the previous dense layer.
- Attention Mechanism – Here the decoder has the permission to utilize the most relevant part of the input sequence with most relevant vectors.

- Dense layer – Here in this step it takes input from the decorder output and map dimension 30 to 50.
- Mean Square loss – Optimizer is Adam optimizer, It is a calculation of root mean square distance between dense layer output and labels output which computed distances between two vector lower value less distance means less loss.

Training and experimentation

- Evaluation – For the evaluation we calculate the Euclidian distance between predicting embedding vector and actual embedding vector.
- We are in the phase of training and error correction.



Project Management

Implementation status report

Work completed:

In the First increment team completed the following.

1. **Description -**

- In the project we have two parts, Machine translation from hindi to English and Text summarization.
- We completed the Pre-processing of the data, Feature extraction and Model building
- We made Data Analysis on the dataset to understand the length and different features of the dataset.

2. **Responsibility (Task, Person)**

- Vinay Chowdary Vemuri - Model Building, Data Analysis, Report
- Keerthika Kondaveeti – Data Analysis and pre-processing
- Sai Priya Pandeti Vasantha Kumar – Feature extraction and Report
- Sampath Sai Sandeep Gompa – Data Analysis and PPT

3. **Contributions (members/percentage)**

    - Vinay Chowdary Vemuri - 25%
    - Keerthika Kondaveeti – 25%
    - Sai Priya Pandeti Vasantha Kumar – 25%
    - Sampath Sai Sandeep Gompa – 25%

Work to be completed

1. **Description**

- In the next submission we will complete the evaluation of the machine translation and testing
- Pre-processing for the text summarization
- Feature extraction
- Model building
- Training and Testing of the model

- Deployment
- Testing with real world data

2. **Responsibility (Task, Person)**

- Vinay Chowdary Vemuri - Evaluation of Machine translation, Model Building for text summarization and evaluation of the mode, Report
- Keerthika Kondaveeti – pre-processing and feature extraction
- Sai Priya Pandeti Vasantha Kumar – Text summarization Model training and Report
- Sampath Sai Sandeep Gompa – Deployment and PPT

3. **Issues/Concerns**

We faced issues with dataset and training the model for the Machine translation. We are trying to debug the training part. Evaluation of the model is also challenging.

References/Bibliography

- https://medium.com/@edloginova/attention-in-nlp-734c6fa9d983
- Dataset - https://www.cfilt.iitb.ac.in/iitb_parallel/
- https://www.kaggle.com/datasets/vaibhavkumar11/hindi-english-parallel-corpus
- https://medium.datadriveninvestor.com/attention-mechanism-encoder-and-decoder-f95d7d7005c8
- https://towardsdatascience.com/extractive-summarization-using-bert-966e912f4142
- https://www.kaggle.com/datasets/kingburrito666/shakespeare-plays
- https://www.kaggle.com/code/shirshmall/english-to-hindi-nlp-text-translation
- https://github.com/shvmshukla/Machine-Translation-Hindi-to-english-
- https://github.com/ArushiSinghal/Neural-Machine-Translation-English-Hindi-for-domain-data/tree/master/Jupyter_Notebbok
- https://arxiv.org/ftp/arxiv/papers/2204/2204.01849.pdf

- https://paperswithcode.com/paper/beyond-reptile-meta-learned-dot-product
- https://paperswithcode.com/paper/a-hybrid-approach-for-hindi-english-machine
- https://paperswithcode.com/paper/neural-machine-translation-by-jointly
- https://paperswithcode.com/paper/attention-is-all-you-need
- https://paperswithcode.com/paper/mlqa-evaluating-cross-lingual-extractive