CSCE 5290: Natural Language Processing

Project Proposal

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

GitHub link for the proposal :

https://github.com/vinayvemuri09/CSCE-5290-NLP-project/blob/main/CSCE%205290%20NLP%20Project%20proposal.pdf

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:

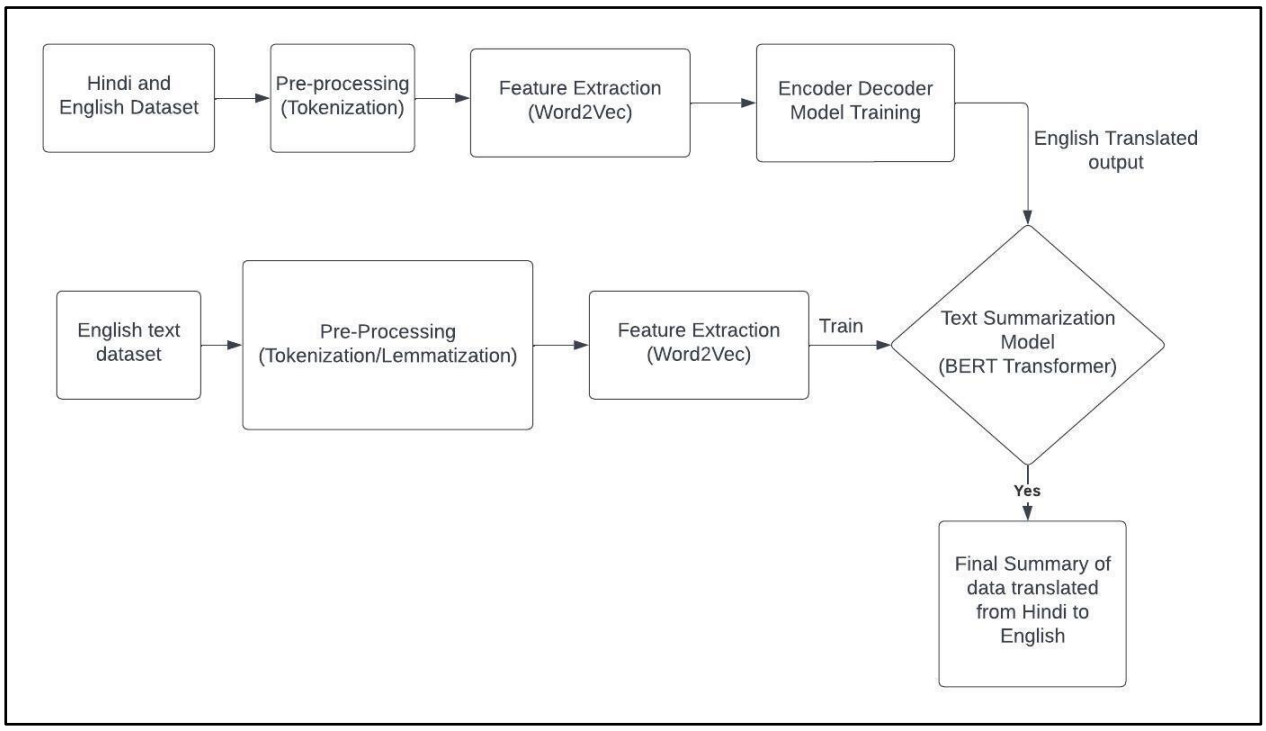
    - The primary objective of the project is to translate Hindi literature to English language and ensure that the original meaning, context and style of the literature are present and express correctly in English language.
    - The second objective of the project is to summarize the translated text in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

- Finally, once the model is trained, we can use the model for the translating the Hindi text to English text.
- **Text Summarization Method:** In the pre-processing stage we use the tokenization and lemmatization methods to tokenize the text and use this data for the next processing. Here we use Shakespeare corpus dataset.
- 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 the BERT model on a dataset of input output pairs where each input is the long paragraphs and output is the short summary of the long paragraphs.
- Finally, once the model is trained, we can use the model to summarize the text.
- **Output:** Once we have the models, we send the Hindi to English translated text to the text summarization model to get the final output of the Hindi literature text summarized in English.
- **Deployment:** We want to deploy the model in the web interface which makes user to interact with the model easily.

Flow chart of the project:



References:

- https://medium.com/@edloginova/attention-in-nlp-734c6fa9d983
- Dataset - https://www.cfilt.iitb.ac.in/iitb_parallel/
- 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