CSCE 5290: Natural Language Processing

Project Proposal

Project Title

Translation of Thirukkurral from Tamil to English and summarizing the converted text.

Team Members

Group-7

1. Preethi Medepalli
2. Varshitha Inaganti
3. Bhuvaneshwar Reddy Sriyyapu Reddy
4. Vaishnavi Gunna

GitHub link for the proposal:

View 1 · Thirukurral Translation into English and Text Summarization (github.com)

Goals and Objectives:

1. Motivation

Thirukkural is a Tamil Literary work written by Thiruvalluvar. It is known for wisdom, universal truths on various aspects of life. This is often studied in schools and universities as part of Tamil literature. This Book has 3 verses namely Aram (Virtue), Porul (Wealth) and Inbam (Love). Kural means sacred verses, and this has 1330 Kural’s each consisting of seven words. Translating this into English promotes and preserves such books and Tamil Literature and culture.

2. Significance

One of the main aspects of thirukkural is its secular nature. It promotes a set of ethical and moral values that everyone can practice irrespective of their faith and beliefs.
There are three book Aram (Virtue) teaches us on how to lead a virtuous life. In Porul (Wealth), it teaches on how to manage our finances and lead a successful life without any debts. In Inbam (love), it tells us the importance of trust and communication in a relationship. This is a widely respected book and has been translated into various languages such as English, German, French and Russian. By translating Tamil text to English, we make content accessible to a wider audience who might not understand Tamil but are proficient in English. This is particularly important for sharing information, knowledge, or cultural content across linguistic barriers.

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

- The project aim is to translate thirukkural which is in Tamil to English so that each and every individual can read and put all the ethical and moral values into action.
- The main objective is to translate Tamil thirukkural into English while maintaining the meaning, context and ensuring accurate translation happens into English.
- The Other objective is Text summarization. Here after translating the Kural into English, we take that output and summarize it for better understanding. To perform this, we use text summarize algorithm.

4. **Features:** The key features of the project are as follows:

Datasets: We are using a dataset from Kaggle where we have the Tamil-English Thirukkural Dataset. We also have another dataset Tamil to English Conversion from Kaggle.

- **Translation Process:**
- We have gathered a dataset of Tamil-English sentences to train the translation model. Each Tamil sentence will have an English sentence.

- Now, we clean the text by removing punctuation and special characters. We tokenize the text into words using the tokenizer.
- In word embedding, we use algorithms such as word2vec, GloVe or FastText. This process represents the words in dense vectors such that these representations capture semantic relationships and contextual information of words.
- Next, we encode a neural machine translation model, such as encoder-decoder. This consists of an encoder network that encodes the input Tamil sentence into a vector, and a decoder network that generates the output English translation.
- Once the model is trained, we use to translate the Tamil text to English.

- **Text Summarization Method:**

- Initially, the text is broken down into words or tokens using tokenization. Also, lemmatization is used to group the same category of tokens.
- Next the tokens are represented into dense vectors using Word2Vec. Each word is mapped to a high dimensional vector, capturing the semantic relationship between words.
- We train the BERT model on a dataset containing input output pairs, Where the output will be summary of input.
- Once the model is trained, we can summarize the text.
- Finally, once the models are ready we can translate the Tamil to English and the summarize that and give a summary of the text.

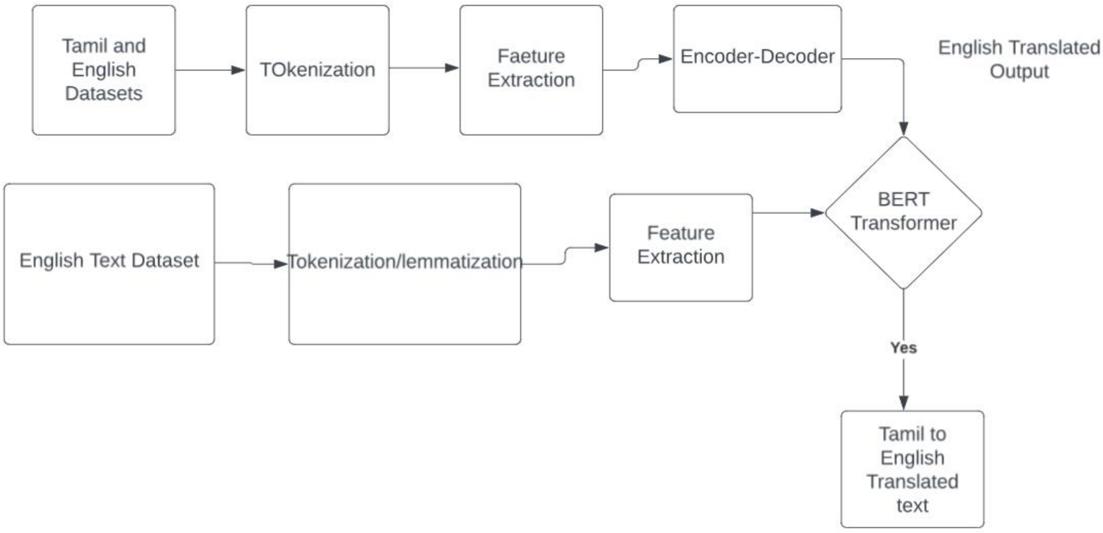
Dataset:

https://www.kaggle.com/code/mpwolke/thirukkural-tamil/input?select=tamil_thirukkural_train.csv

https://github.com/Ishikahooda/Tamil-English-Dataset/tree/master/Dataset

Here, we use these datasets to Translate Tamil to English and summarize the text.

Flow chart:


Tamil and English Datasets
TOkenization
Faeture Extraction
Encoder-Decoder
English Translated Output
English Text Dataset
Tokenization/lemmatization
Feature Extraction
BERT Transformer
Yes
Tamil to English Translated text

References:

I. https://www.kaggle.com/code/mpwolke/thirukkural-tamil

II. https://www.kaggle.com/datasets/bagavathypriya/english-to-tamil-data?resource=download

III. jjasim/Thirukkural-English-Translation-Dataset: Thirukural in English (github.com)

IV. Thirukkural-Tamil-Dataset/data/all_kural.json at master · vijayanandrp/Thirukkural-Tamil-Dataset (github.com)

V. https://www.kaggle.com/datasets/rahulvks/thirukkural