

Natural Language Processing

Project Proposal

Group – 10

Lokesh Borra - 11758547 | Nikhil Kuchipudi – 11709126

Puneet Puttu - 11691685 | Srilekha Sridasyam - 11610756

https://github.com/nikhil222002/News_Articles_Summarization_and_Translation

Title: News Articles Summarization and Translation

1.Motivation:

The main issue addressed by the news article summarizer and Translator is the challenge of understanding information from different parts of the world and the abundance of availability of information. We aim to make information accessible for everyone in a short and understandable way breaking language barriers.

The Justification for this project lies in the fact that people need to have knowledge about the global news to be knowledgeable and make informed decisions about their investments. Access to information can be made simpler and more efficient using the tool that we plan to produce.

The importance of this challenge is non-negotiable to reduce the spread of misinformation which can spread rapidly especially if people do not have access to information in a way they can comprehend. Without a good translation system, people can miss important updates from around the world. It saves people a lot of time which they can spend productively on other tasks.

2.Significance:

The value this project provides to the field of Natural Language Processing (NLP) cannot be overstated. It greatly pushes the boundaries of domains: Machine Translation and Text Summarization. Future studies can use this project as a benchmark to make further improvements and contributions in this field. It also enables scientists and engineers to broaden their minds and develop products that improve the quality of people's lives.

Possible effects this project will bring to action are reduction in geopolitical tensions, enhanced access to information, efficient usage of time by journalists and students and reduction in spread of misinformation.

Achieving project goals is essential to remove language barriers and information overload on people. Another reason for this is to stop the spread of misinformation as mentioned above, which is the root of many problems around the world.

3.Objectives:

The precise goals of our project are Accurate summarizer, Robust translator, User-friendliness and efficient.

Success can be verified by various methods including feedback mechanism which tells the system how it performed with the summarization and translation. User satisfaction can also be a measure of model performance. Translation accuracy and summarization accuracy are difficult to check by machine, but a human verifies how well it is performing with an article it has never seen before.

4.Features:

Technical characteristics of our project include using Text Summarization and Machine Translation algorithms such as Bidirectional Encoder Representations from Transformers (BERT) model from Transformers are used for most accurate operations. Multilingual translation capability is also a technical characteristic of our project.

Deliverables include a Google Collab notebook with working code that can be run to perform the desired operations. We also deliver a PowerPoint presentation that is useful to understand the project's significance and working. Detailed documentation will also be provided if required in future.

Uniqueness in this project involves additional implementation of translation of the summary or important points into the user requested language. This translation model is integrated with the text summarization. The existing implementation consists of only text summarizing and question answering model.

Milestones in project development are, Data Collection, Data Preprocessing, model Development, Performance Check and Final Model delivery.

Distinctive Elements include contextual understanding and ethical considerations while performing summarization and translation as a minor misunderstanding of information would result in rapid spread of misinformation based on the userbase of the tool. These elements help in the success of a project if they are taken into consideration, and we can make a product with high accuracy.

5.Dataset:

The dataset we chose for our project is [CNN Stories](#) created by Hermann et al in 2015. The dataset consists of around 90k CNN news articles which we try to summarize

based on user selection. The data would be of text datatype where each story is basically a report of a news incident.

The data input for the model would be the story file which is basically a news article. So, there are no direct data cleaning steps that need to be done before working on data. The steps in preparing data to train the model would be tokenization, Lemmatization, Named entity recognition and data augmentation.

6.Visualization:

Workflow of the Project:

1. The Data Collection stage includes collecting data that needs to be given as input to our model. In our case it is the CNN dataset story that is described in the above Dataset section.
2. Data preprocessing involves stripping out any unnecessary data in the input data and then making tokens from the data and then lemmatizing and performing semantic analysis.
3. Then this data is inputted to our models which we imported from transformers module and asks if the user wants the summary of data or important key points from the data.
4. Based on the user selection the model performs semantic analysis and gives the text output to the user.
5. The next step is translating the text into the language the user needs by using machine translation by BERT model.
6. The final translated text is returned to the user, which is the output of our model.

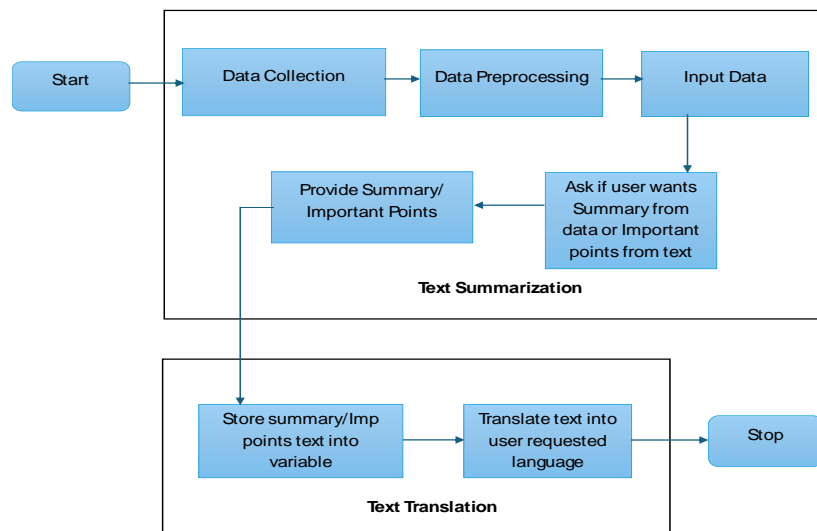


Figure: Workflow of project