Natural Language Processing (UCS664)

Machine Translation English to Hindi Project

Submitted To:

Dr. Jasmeet Singh

Submitted By:

Maulik Gupta 102103294

Rimjhim Mittal 102103430

Vartika Gautam 102103397

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Computer Science and Engineering Department

Thapar Institute of Engineering and Technology, Patiala

Jan - May 2024

NLP Application: Machine Translation English to Hindi

This project focuses on machine translation, specifically translating text from English to Hindi. Machine translation is a subfield of natural language processing (NLP) that involves the automatic conversion of text from one language to another while preserving the original meaning. The goal here is to build a model that can accurately translate English sentences into Hindi.

Dataset: IIT Bombay English-Hindi Translation Dataset

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

The dataset used in this project is a parallel corpus consisting of English-Hindi sentence pairs. This means that for each English sentence, there is a corresponding Hindi translation. The data is loaded from a CSV file named hindi_english_parallel.csv. It is essential that the dataset is clean and well-aligned to ensure the model learns the correct mappings between the two languages. In this case, the dataset is split into training and testing sets, with 20% of the data allocated for testing.

Transformer Model: Helsinki-NLP/opus-mt-en-hi

Link: https://huggingface.co/Helsinki-NLP/opus-mt-en-hi

The Helsinki-NLP/opus-mt-en-hi model from Hugging Face is a Transformer-based model specifically designed for translating text from English to Hindi. Transformers have revolutionized natural language processing (NLP) because they handle sequential data so well. This model, built on the Transformer architecture, excels in machine translation.

1. **Architecture:** The model uses the Transformer architecture, which consists of encoder and decoder layers. The encoder processes the input English text, and the decoder generates the Hindi translation. This setup helps the model capture complex patterns in the text and produce accurate translations.
2. **Pre-Training:** Before being fine-tuned, the model is pre-trained on a large corpus of English and Hindi text. During this phase, it learns the semantics and syntax of both languages, setting the stage for high-quality translations.
3. **Model Fine-Tuning:** Fine-tuning further refines the model. In this case, it was fine-tuned on the IIT Bombay English-Hindi Translation Dataset. Fine-tuning involves tweaking the model's parameters, like the weights in its neural network layers, based on the errors between its predicted translations and the actual translations in the training data.
4. **Vocabulary:** The model's vocabulary covers a wide range of English and Hindi words, enabling it to handle various topics and domains effectively.
5. **Performance:** This model is known for its impressive performance in translating English to Hindi. It produces translations that are both fluent and accurate. However, like any

machine translation system, its effectiveness can vary based on the complexity of the input text and the quality of the training data.

The Helsinki-NLP/opus-mt-en-hi model is a powerful tool for English-to-Hindi translation tasks. It offers state-of-the-art performance and generates accurate translations across a wide range of subjects and contexts.

Result

Blue Score: 0.0085778022184968

```
input_text='sir give full marks, please (:'
input_ids=tokenizer(input_text,return_tensors='tf').input_ids
output=model.generate(input_ids)
tokenizer.decode(output[0],skip_special_tokens=True)
```

```
'सर सर पूरा निशान दे, कृपया भरें निशान दे दो, कृपया (G: 09:'
```