CS 4710 | Artificial Intelligence

Webpage Translation with NLLB-200

01

ANUSHA CHOUDHARY

02

The Problem

Machine Translation on Low-Resource Languages

Outside of languages like English, German, Spanish, French, etc., most languages do not have many resource available to use as training data for neural networks. Since neural networks depend on large amounts of data to perform well, machine translation performs poorly on low-resource languages.

Meta AI's NLLB-200

Earlier this year, Meta AI released an unpublished paper + all source code and pre-trained models for a new model named No Language Left Behind (NLLB). It offers multilingual translation on 200 languages and has reported an average 44% increase in BLEU score over previous state-of-the-art models.

Applications of NLLB

Since NLLB is a new model, there are not many real-world applications of the model in use yet (outside of its use in Meta platforms and on Wikipedia). This means the potential of this model is largely unexplored.

A note on Maithili

Maithili is a language with ~34 million speakers in parts of India and Nepal.

Since Maithili is a low-resource language, efficient translation tools do not currently exist for it.

The problem of English to Maithili translation is unique because of the lack of well-performing English to Maithili translation resources online.

03

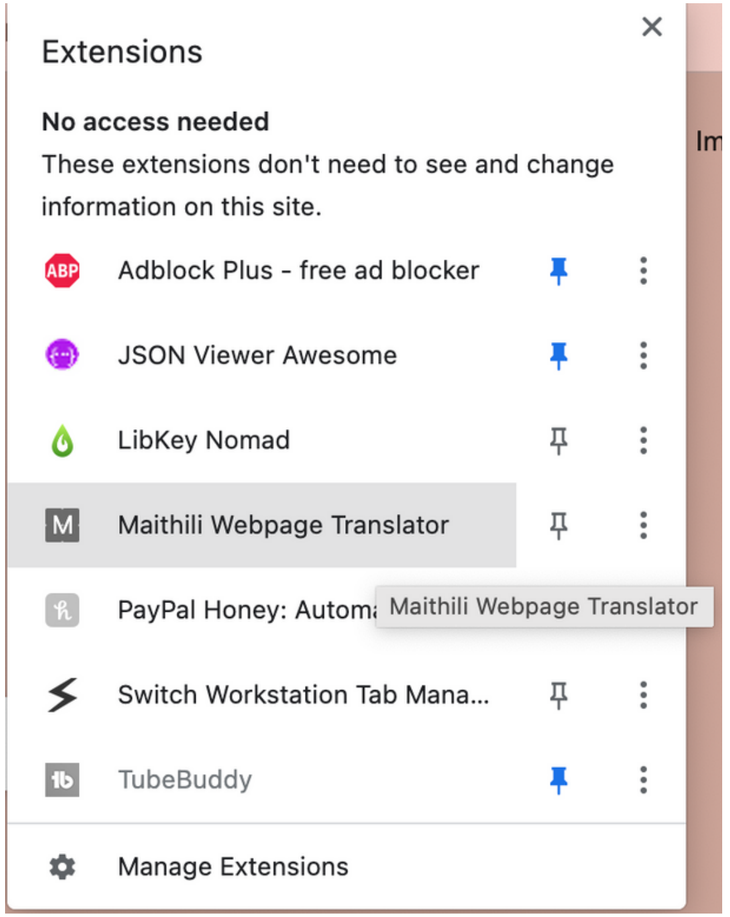
The Approach

An English -> Maithili Webpage Translator

In response to the lack of real-world applications of NLLB-200 and the lack of reliable and accessible English to Maithili translation tools, this project will use Meta AI's NLLB-200 machine translation model to build a Google Chrome extension for translating web pages from English to Maithili.

Resources Used

The project will use the NLLB-200-distilled-600M model from HuggingFace transformers, a Python backend for webscraping and translating, and a JS frontend for the Chrome extension.



04

Progress

Planning

Build a sequence diagram to understand the flow of data from the Chrome extension to the backend and vice-versa.

Proof-of-Concept: NLLB-200

Write a basic python script that uses NLLB-200-distilled-600M for translating from English to Maithili and perform rudimentary human evaluation.

Proof-of-Concept: Chrome extension

Build a Chrome extension with a JS frontend and a Python backend and pass dummy data over HTTP.

Final Implementation

Use a webscraping tool to extract text from the webpage, translate the text using NLLB-200, display translated text on webpage.

Evaluation

Evaluate the model on intelligibility (1-9) and fidelity (0-9). A challenge of this project is to find better evaluation methods for this tool.

Publish to Chrome Web Store

Questions?