CS689A: Computational Linguistics for Indian Languages

Assignment 3 (40 marks)

Due on: 17th April, 11:00pm

1. (20 marks) Evaluate three machine translation models:
   (a) Distilled model of NLLB-200 with 600M parameters https://huggingface.co/docs/transformers/en/model_doc/nllb,
   (b) IndicBART https://github.com/AI4Bharat/indic-bart/, https://huggingface.co/ai4bharat/IndicBART,
   (c) ChatGPT, possibly using https://platform.openai.com/docs/guides/text-generation/chat-completions-api to batch translate several sentences.

   Use the Samanantar benchmark dataset available from https://drive.google.com/drive/folders/1hR-8Mc7qQWsZAC-cw-nUqG8_OCqCdq-b as translation datasets.

   If your mother tongue is X, evaluate the following:
   (1) English to X
   (2) X to English
   (3) One Indian language (other than X), say Y, to X
   (4) Y to X

2. (10 marks) Report the BLEU and ROUGE scores (different variants) for all the three models for all the translation tasks on a *random subset* of 1,000 sentences from the `wat2021` test dataset availble from https://drive.google.com/drive/folders/1hR-8Mc7qQWsZAC-cw-nUqG8_OCqCdq-b.

3. (10 marks) What do you learn from this assignment?

Instructions

Submit the assignment as one zip file `rollno-assignment3.zip` in the course portal (https://canvas.cse.iitk.ac.in/) within the deadline. The submission MUST contain a README file. The code must have documentation with appropriate comments.