CS689A: Computational Linguistics for Indian Languages

Assignment 2 (75 marks)

Due on: 13th March, 11:00pm

Choose the corpus file according to your mother tongue from the *Naamapadam* corpus link https://huggingface.co/datasets/ai4bharat/naamapadam/tree/main/data

You may use `wget` to download these files.

1. (25 marks) Identify the named entities for a set of 25 sentences and mark them in BIO format.

   The classes are: PER (person), LOC (location), ORG (organisation), and MISC (miscellaneous) and O (for others, i.e., non NEs).

   Suppose, the sentence is
   "Subhas Chandra Bose was the Prime Minister of Azad Hind government and fought in India."
   The output is
   B-PER I-PER I-PER O O B-MISC I-MISC O B-ORG I-ORG O O O O B-LOC

   You must create an account at https://bangla.iitk.ac.in/cs689/, complete the task, and submit the manually annotated sentences both at the online portal and *also in the zip file*.

2. (20 marks) Fine-tune IndicBERT and IndicNER model on the Naamapadam corpus of your mother tongue with train 70%, 10% validation, and 20% test set splits as already given in the corpus (each separate file is present). You have to run for at least 20,000 sentences and for at least 3 epochs.

   You may use GPU from Google Colab or Kaggle for this.

   IndicNER model is available at https://huggingface.co/ai4bharat/IndicNER whereas IndicBERT is available at https://huggingface.co/ai4bharat/indic-bert.

   Do a comparison of both the models using *macro-F1* score on the 20% test set. Also, report the training and validation macro-F1 scores on the sets already given in the corpus.

3. (10 marks) Use ChatGPT available at chat.openai.com. Pass the 25 sentences in Question 1 and query the NERs. Submit the output.

4. (10 marks) Test the outputs of Question 2 (both the models) and Question 3 against the manually marked sentences from Question 1 and calculate the precision, recall, and macro-F1-score.

5. (10 marks) What do you learn from this comparison? Submit a detailed report on the values of hyper-parameters you have tuned, specifying their significance, and the optimal values chosen by you. Also, mention the outputs of both models in the report.

Instructions

Submit the assignment as one zip file `rollno-assignment2.zip` in the course portal (https://canvas.cse.iitk.ac.in/) within the deadline. The submission MUST contain a README file and a Makefile. The code must have documentation with appropriate comments.