CS689A: Computational Linguistics for Indian Languages

Assignment 2 (75 marks)

Due on: 7th November, 2022, 11:00pm

Choose the corpus file according to your language preference (Bangla, Hindi, Marathi, Tamil, Telugu).

The Python code notebook for Ques 2 and the respective datasets can be found at https://drive.google.com/drive/folders/1VKwQtnSRqWERULev2JwXKj6Y4Ow61EPk?usp=sharing. The BERT models are listed in https://huggingface.co/ai4bharat/indic-bert.

1. Consider the parse file.

   Consider a particular sentence S consisting of n words $w_1 \dots w_n$.

   Corresponding to that are the POS tags $t_1 \dots t_n$.

   For each word w_i, assume the gender, case and number to be g_i, c_i and b_i respectively.

   There are also n dependency relations in the sentence (including the one with the root). Assume a dependency relation to be $\langle w_i, R, w_j \rangle$ which indicates that the word w_i is connected to its head word w_j by the relation R. Correspondingly, the POS tag relation is $\langle t_i, R, t_j \rangle$.

   For each of the questions below, output the list in descending order of frequencies.

   (a) (5 marks) Find the frequencies of POS tags of words.

   (b) (5 marks) List the 50 most frequent words corresponding to each POS tag.

   (c) (5 marks) Find the frequencies of gender, case and number of words separately.

   (d) (5 marks) List the 50 most frequent combinations of gender, case and number as a 3-tuple.

   (e) (5 marks) Find the frequencies of POS tags corresponding to only head words.

   (f) (10 marks) Find the directed POS tag tuples, i.e., $\langle t_i, t_j \rangle$. For each such 2-tuple, list the frequencies separately for each relation R as well as total.

   (g) (5 marks) For each dependency relation R, list the frequencies separately for each 2-tuple $\langle t_i, t_j \rangle$ as well as total.

2. Use the Python code snippet mentioned earlier.

   (a) (20 marks) Fine-tune the pre-trained BERT model for the *UPOS prediction* task. Marks will be given based on how well you can explain the steps of the entire code during the demo.

   (b) (5 marks) List precision, recall, F-score (both micro and macro).

3. (10 marks) The submission MUST contain a README file and a Makefile. The code must have documentation with appropriate comments.

Instructions

Submit the assignment as one zip file `rollno-assignment2.zip` in the course portal (canvas.cse.iitk.ac.in) within the deadline.