Evaluating and Developing Methods of Generating Code-Switched Data
07-400, Spring 2022

Abhishek Vijayakumar,
under Prof. Alan W. Black, Language Technologies Institute
https://inkyubeytor.github.io/category/code-switching.html

February 28, 2022

1 Major Changes

The primary goal of this project has not changed since the previous milestone.

2 What You Have Accomplished Since Your Last Meeting

I made the GLUECoS benchmark more reliable by replacing their fixed train-val-test split with random sampling validation using 80% train, 10% validation, and 10% test datasets. I have implemented this across the token tasks, and I will be implementing this across the sequence tasks (question answering, sentiment analysis, and natural language inference). I have also implemented this for romanized versions of the tasks and trained some models for romanized evaluation.

3 Meeting Your Milestone

My primary goal was to fix the benchmark's variance issues, which I did. Variances on the task performance across instances in my sampling validation are mostly below 1%.

4 Surprises

One surprise is that none of the models finetuned on romanized data performed better than Multilingual BERT, despite mBERT not being trained on romanized Hindi data.

5 Looking Ahead

I plan to "overtrain" some models by increasing the learning rate and training epochs until the models begin to change in performance in some way, allowing us to determine the parameter ranges that can be useful for our model training.

6 Revisions to Your Future Milestones

My current plan is to use the new dataset I acquired this week to attempt to beat mBERT on the romanized tasks. Our goal with this project is primarily to improve performance on romanized tasks, as this is how code-switching appears in real data that we wish to perform our tasks on.

7 Resources Needed

No further resources are needed for this project at this time. I have obtained another dataset of generated Hinglish from another student researcher.