Evaluating and Developing Methods of Generating Code-Switched Data
07-300, Fall 2021

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

January 31, 2022

1 Major Changes

The goals of this project have not changed since the previous milestone. The primary goal of this project is to evaluate the quality of synthetically generated code-switched data with respect to its ability to contribute to language modeling for downstream tasks.

2 What You Have Accomplished Since Your Last Meeting

In addition to the work detailed in the Meeting Your Milestone section, I have done the following.

I am now able to run the GLUECoS evaluation locally for the majority of the Hinglish tasks (the sole exception being machine translation). For several tasks, I have partitioned the labeled validation data provided into a validation set and a test set, allowing us to conduct many preliminary local evaluations before submitting final models to the official benchmark, which uses a test set without public labels.

3 Meeting Your Milestone

The original milestone was to implement 4 distributional evaluations and to acquire datasets. I have implemented the Code-Mixing Index, BLEUScore, BERTScore, and Self-BLEU distributional evaluations. There is currently an error in the Self-BLEU evaluation, but all other evaluations are performing properly. I currently have every dataset I need except for one, which I am unable to gain access to as it cannot be made public. I am currently looking for alternatives to this dataset. If no close alternative is found, I will be using the Hinglishpedia data I collected as a substitute for this dataset.

4 Surprises

There have not been any major surprises in the project since the last bi-weekly meeting.

5 Looking Ahead

Over the next two weeks, I plan to run GLUECoS evaluations over the non-neural generative methods and over the working neural generative method I have. I also plan to try to fix several problems with my current (partially) completed objectives:

- The unknown error in the Self-BLEU evaluation
- The missing dataset that I need to replace the one I do not have access to
- The GLUECoS machine translation task that currently does not run on the machine I have available
- Creating local validation and training datasets for the remaining Hinglish GLUECoS tasks

6 Revisions to Your Future Milestones

I am currently partially done with my next two milestones. If I can complete both of these milestones by my next milestone report, I will shift the entire timeline up by two weeks and add a reach milestone towards the end of the project. As I have several outstanding issues with these future milestones, I will not be preemptively creating an accelerated timeline this week.

7 Resources Needed

No further resources are needed for this project at this time. Deep learning is done on a GTX 1070 on a CMU LTI machine. This is the same system mentioned in previous updated, but I have fixed the issues with the GPU. Datasets have been acquired by scraping or by requesting from authors of previous papers in the field.