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

April 4, 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 created a new, easier version of the sentiment analysis task as a binary classification task. This version of the task was meant to prevent the issue where models tend towards predicting the majority label for sentiment inputs. However, performing this modification required removing roughly 45% of the inputs from a 10,000 element corpus.

I also trained models using the bert-base-multilingual-uncased and xlm-roberta-base baseline models and evaluated their performance on the existing tasks.

3 Meeting Your Milestone

I achieved my goal of finding a statistically significant performance difference from the mBERT baseline, my first objective at my previous milestone (the xlm-roberta-base model). However, this was due to an exploration of the space of existing pretrained models, rather than due to fine-tuning a model.

I did not achieve my goal of creating an easier sentiment task. Despite removing the neutral class (the hardest input class), the trained models still tend towards a majority vote. It may be that this is due to a lack of training data available for this task.

4 Surprises

There were no particularly surprising results since the last milestone.

5 Looking Ahead

I plan to further explore the xlm-roberta-base model to see if further improvements can be made on any tasks. I also plan to start writing the code to compute dataset difficulty curves, in which

we can rank inputs of a dataset by how well an ensemble of models performs on them.

6 Revisions to Your Future Milestones

The primary focus of my project until the Meeting of the Minds will be on constructing an evaluation that differentiates different corpora of synthetic code-switched data by finetuning capability.

7 Resources Needed

No further resources are needed for this project at this time.