Detecting Word Issues in Course Transcripts: Progress Report

Team: The Palindromers
Theme: Free Topics
Date: Nov 15, 2021

1) Which tasks have been completed?

- Started exploring pre-trained neural language models such as BERT, BERT variants, ELMo, GPT and the likes. We are planning to use ALBERT for this task, because it is significantly lighter on resource requirements compared to BERT
- In case we train our own model instead of using a pre-trained neural language model, determined an appropriate corpus for training a bigram language model - Wikipedia pages in bulk
- For evaluation, created sample testing documents to evaluate performance with and without word issues
- Created a script to iterate through course transcripts and store in a format that can be fed into a language model

2) Which tasks are pending?

- Explore ALBERT and learn how to fine-tune it to fit our project goal
- Build a python application based on ALBERT
- Use sample documents to test our project code and analyze it
- Train bigram language model in case pre-trained neural language model is not a viable option from resource or implementation point of view
- Write final script to utilize the trained model to parse through the actual course transcripts
- Perform manual review of random sampling of course transcripts to determine effectiveness, tweak model parameters and thresholds if needed

3) Are you facing any challenges?

- Pre-trained neural networks like BERT are the current state-of-the-art but they are resource intensive
- Understanding / unpacking how to utilize ALBERT
- Determining how best to output the exact place in the transcript that a flagged word is from, such that it is easy for human to know where in the transcript to find the erroneous words to be fixed (instead of having to search the whole transcript)