Detecting Word Issues in Course Transcripts

Team: The Palindromers
Theme: Free Topics
Date: Oct 24, 2021

Team Composition

First Name	Last Name	NetID	Email	Role
Catherine	Parker	cph7	cph7@illinois.edu	Team member
Danielle	Richmond	dcr4	dcr4@illinois.edu	Team captain
Scott	Downey	scottmd3	scottmd3@illinois.edu	Team member
Zixiang	Li	zixiang9	zixiang9@illinois.edu	Team member
Jharna	Aggarwal	jharnaa2	jharnaa2@illinois.edu	Team member

Project Description

For those with hearing impairment or for whom English is a second language, the transcripts provided for Coursera videos are a vital component of learning the material. Thus when there are incorrect words at key points in explanations it can make learning the material more difficult. Our goal is to build a method for detecting and flagging words that are potentially incorrect so that human auditors can more easily fix them.

We will use Python as our coding language for this project. Dr. Zhai has offered to provide us with the video transcripts for CS410. We will then build a system with two main components:

- A bigram language model built using a larger corpus to determine the likelihood that two words would actually be next to each other.
- A unigram mixed language model built using the course content and a mixture of a background language model to determine how likely it is for a particular word to show up in the course transcript collection at all.

Both of these will produce scores and if their combined score is above a certain threshold then we will add the following to a list: the course transcript's name, the word in question, and the word's location in the body of text. In order to evaluate our work we will manually review a random sample of both flagged and non-flagged course transcripts to confirm that the majority of flagged transcripts due in fact have words to be corrected and that the majority of non-flagged transcripts do not have any word issues. We will also explore using pre-trained neural language models to do basic flagging of grammatical errors and typos.

Task & Time Estimates

Time Estimate	Task
12	Explore pre-trained neural language models and determine how best they could integrate with and enhance what our system does
2	Create sample documents for testing with and without word issues
3	Determine appropriate corpus for bigram language model online

15	Build and test bigram language model on sample documents
3	Determine appropriate corpus to use for unigram mixed language model’s background model
20	Build and test unigram mixed language model on sample documents and determine best mixture between course content model and background model
15	Test & determine appropriate threshold for bigram model and unigram models separately and combined
15	Write final script to utilize both models to parse through the actual course transcripts
15	Perform manual review of random sampling of course transcripts to determine effectiveness, tweak model parameters and thresholds if needed
100	