Scope/Statement of Work Guidelines

The length of this document should be 1-2 pages. If you choose to detail a full schedule of specific goals (i.e., beyond the calendar of deliverables in the syllabus), then feel free to use more space.

The readers of this document will be the full teaching staff, along with your partner. The intent is to prepare you for future scientific writing you may do in your career. It's good practice to be inclusive and comprehensive in your writing, yet appropriately succinct -- ensure that others could follow your writing, even if they don't have much prior knowledge, but don't bore those who are intimately involved in the project.

Background

Essays are an important expression of academic achievement, but they are expensive and time consuming for states to grade them by hand. Due to this reason, we are frequently limited to multiple-choice standardized tests. Thus, automated scoring systems can provide a potential mitigation to this problem. If developed properly, it can yield fast, effective and affordable solutions that would allow any form of test organizations to introduce essays and other sophisticated testing tools. Another benefit of the automated scoring system is that, with fair training data, it can avoid the discrepancy in scores that come from different human graders and can provide a fair grading of essays. Thus, our team is interested in developing an automated essay scoring system that delivers scores that are close to human expert graders.

Problem Statement

The goal of this project is to build a model and predict the score of provided essays on the scale from integer 1 to 10. To make such a model helpful in reality, we want our score prediction to be as consistent with scores given by humans. In other words, we want our prediction score and human score to be "inter reliable".

We will approach this classification problem using various NLP techniques. We will first start with a Recurrent Neural Network with text embedding as our baseline model, and try pre-trained models as well, such as BERT. Depending on the results of these attempts, we may fine-tune the models, as discussed in our presentation 2 about **How to Fine-Tune BERT for Text Classification? (Sun et.al)**

Some difficulties and challenges:

1. It could be slow to train the model locally, and we need to figure out how to deploy models and data pipelines on the cloud platform to achieve speedup.

2. Grading assay is more subjective and specialized than some other common text tasks, such as translation. We need to figure out how to adapt general pre-trained models to our specific target task.

Our moonshot goal is to develop a model to rank a set of given assays based on predicted score, which could be used for reviewing academic transcript submission.

The performance of the grading model is measured by quadratic weighted kappa error metric. The quadratic weighted kappa error metric of ratings from two raters is defined as follows: Denote two raters as A and B, E_{total} as total number of assays, N as total number of grades (so the grade range from $1, 2, \cdots, N$). Define matrix $O_{N \times N}$ as confusion matrix between A and B, and O_{ij} is number of assays that A grades as i and B grades as j. Define matrix $E_{N \times N}$ as expected matrix, and $E_{ij} = \frac{\sum_{k=1}^{N} O_{kj} \cdot \sum_{l=1}^{N} O_{il}}{E_{total}}$. Notice that $\sum_{i,j} O_{ij} = \sum_{i,j} E_{ij} = E_{total}$. Define weight $w_{ij} = (\frac{i-j}{N-1})^2$, and the metric kappa is defined as $\kappa = 1 - \frac{\sum_{i,j} w_{ij} O_{ij}}{\sum_{i,j} w_{ij} E_{ij}}$. κ normally ranges from 0 to 1, and higher κ means higher agreement between two raters.

We plan to develop a web application as our deliverable of our project, with functionality of retrieve example assays given a score and return predicted score given an assay.

Resources

The data can be found here. The data contain eight essay sets. Each of the sets of essays was generated from a single prompt. Selected essays range from an average length of 150 to 550 words per response. All responses were written by students ranging in grade levels from Grade 7 to Grade 10. All essays were hand graded and were double-scored on the scale from 1 to 10.

The training process requires the usage of GPU, and the web application potentially requires kubernetes, dockers etc.

High-Level Project Stages

Our milestones for this project:

1. Try models from scratch (embedding, RNN, LSTM, etc.) and fine-tuning by Nov 23
2. Try pretrained models (BERT, etc.) and fine-tuning by Nov 30.
3. Discuss with TF and refine the models by Dec 3.
4. Deploy models on web application and test performance by Dec 9.
5. Final presentation on Dec 11.