Executive Summary

Given the growing population of English Language Learners (ELLs) and that writing is one of the most fundamental yet least practiced skills, our model aims to a) evaluate the language proficiency of 8th-12th grade ELLs; and b) provide them with more accurate, detailed scores assessing their language development. This can be used not only to expedite the grading cycle for teachers and standardize the feedback process, but also to provide ELLs and their educators with feedback that helps them narrow down areas where the student's language needs more attention.

We developed a model that evaluates English essays on 6 criteria, namely grammar, vocabulary, syntax, cohesion, conventions, and phraseology. The evaluation measure of the model is chosen to be the MCRMSE, or the mean column-wise root mean squared error, where each column corresponds to one of the 6 criteria being evaluated. Our model scored a MCRMSE of ≈ 0.5, nearly 40% better than our human-baseline model. We also provide an easy-to-use interactive demo targeting the potential end users.

The dataset we used was provided through a competition held by the University of Vanderbilt and The Learning Agency Lab and hosted on Kaggle. The model was trained on 3,911 essays written by non-native students in grades 8 through 12, with teacher-graded scores (labeled training data).

In order to establish a baseline model, each of our team members manually graded 20 essays that are randomly chosen from the training dataset. Our inexpert-grader baseline model got a score of the mean column-wise root mean squared error (MCRMSE) ≈ 0.85. We then used a combination of Machine Learning regression algorithms: a random forest regressor, XGBoost regressor and others, as well as a voting classifier for all of them. Using a grid search to look for the optimal hyper-parameters, we finally obtained a score of MCRMSE ≈ 0.50, indicating that our model performs much better than the untrained eye. In addition, it is computationally inexpensive, allowing its use in schools or centers without the need for intensive infrastructure.

One of the main challenges to build a good model is to extract features that could predict, or at least correlate well with, the language proficiency criteria we are scoring. Through a combination of python packages, we were able to extract features that measure the spelling mistakes, grammatical errors, the use of complex words, and so many more. Since the scores of the 6 analytic measures (i.e., the criteria we are trying to predict, like grammar, etc.) correlate well with one another, we can use the combination of all the extracted features to predict even the criteria with weak correlations to individual features. However, having extracted individual features, we could potentially use them to give more specific feedback to each student on how to improve each of these analytic measures.

The model proves to be of potential use to alleviate the burden of grading on teachers, who are often underpaid and overworked, while giving nearly *instant* feedback to ELLs. Upon improving the model further, it can be used to train inexperienced teachers and help them design exercises better tailored to student's deficiencies.