Evaluating language knowledge of ELL students

TEAM 8
Kyuri Kim, Sanchit Vijay

Problem Statement

The chosen problem revolves around evaluating the language proficiency of English Language Learners (ELLs) through their written essays. This issue holds substantial importance as it directly contributes to the educational advancement of ELL students by facilitating more precise feedback and aligning learning tasks appropriately. Additionally, it addresses a crucial need to streamline the grading process for teachers. The overarching goal of this project is to enhance existing tools by creating various Natural Language Processing models. These models aim to furnish ELLs with more accurate feedback, streamline the grading process for teachers, and ultimately empower ELLs to elevate their English language proficiency through contextually relevant learning tasks.

Dataset

https://www.kaggle.com/competitions/feedback-prize-english-language-learning/data
We will use the ELLIPSE corpus provided by Vanderbilt University, which contains argumentative essays written by 8th-12th grade ELLs. The dataset comprises essays that are argumentative in nature, likely reflecting the students' ability to form and express opinions, reason through arguments, and use language to persuade.
These essays have been scored on several analytic measures, which are:

- Cohesion: The ability to use connectives and transitions to link ideas and make the writing flow smoothly.
- Syntax: The complexity and correctness of sentence structure.
- Vocabulary: The range and appropriateness of word choice.
- Phraseology: The use of common expressions, idioms, and collocations.
- Grammar: The accuracy and range of grammatical constructions.
- Conventions: Adherence to writing standards such as spelling, punctuation, and capitalization.

Each of the analytic measures is assigned a score ranging from 1.0 to 5.0 in increments of 0.5, indicating the level of proficiency in that specific aspect of writing.

Methodology

Given the goal of the project, a combination of classical and modern NLP methods may be utilized. We will begin with a classical model, logistic regression, to establish a baseline. As we progress, we will shift towards more complex models like LSTM, BERT-based models, and Transformers based networks.
To implement these models, we will use spaCy, Gensim, NLTK, scikit-learn, pandas, Matplotlib, Pytorch, Pytorch-NLP, Keras and Tensorflow.

Tasks

To start the project, we will start with data preprocessing, which includes tokenization, text normalization, stopword removal, stemming/lemmatization, and POS tagging. Then, we will work on feature extraction by doing a bag of words, TF-IDF, word embedding(Word2Vec), and syntax trees. After these steps are

done, we will move onto model building, starting with classical models, then to more complex models such as neural networks and customized models.

Evaluation

The performance of the NLP models will be evaluated using various metrics such as MCRMSE(mean column wise root mean squared error), R-squared, MAE(mean absolute error). Moreover, for classical tasks evaluation, accuracy, precision, recall, and F1-score will be used.

Schedule

The project is expected to be completed in the next four weeks. The schedule is as follows:
Week 1: Data preparation and preprocessing
Week 2: Feature extraction
Week 3: Model development and optimization & Model evaluation and comparison
Week 4: App preparation and report writing