Team number 30

AUTOMATIC ESSAY GRADING

- CHANAKYA MALIREDDY
- MRINAL DHAR
- VAIBHAV KUMAR

Problem Statement

- Given an essay, our aim is to evaluate it and give a score that is as close as possible to a score a human grader would give.

Introduction

- Essays are useful tools to assess learning outcomes such as ability to recall, organise, and express one's ideas in writing.

- Many researchers claim that the subjective nature of essay assessment leads to variation in grades awarded by different human assessors, which is perceived by students as a great source of unfairness.

- A system for automated assessment would at least be consistent in the way it scores essays, and enormous cost and time savings could be achieved if the system can be shown to grade essays within the range of those awarded by human assessors.

Datasets

- A dataset consisting of ~13000 essays released by Hewlett and Flora Foundation for a competition hosted on kaggle.com was used.
- Essays were written by school children from grade 7 to grade 10.
- These were divided into 8 sets based on context.
- On average, each essay had about 150-650 words.

Our performance metric

- After evaluating the essays using our system, we compare the score given by the human grader with the one given by our system.
- This is done by calculating the ***average quadratic weighted kappa***.
- It is a robust metric. It also takes into account as baseline the possibility of agreement occurring by chance.
- Therefore, it varies from -1 (complete disagreement between system score and human score) to 1 (complete agreement).
- If there is less agreement between system score and human score than expected, the kappa value is negative. It is 0 when agreement is random.

The project has been divided into three phases….

Phase 1

- The goal of this phase is to evaluate the style of an essay.
- Regardless of the content, the automatic scorer should be able to judge the essay only on the basis of its style.

Approach

- Important features that would contribute to the style of an essay should be identified.
- For this purpose a set of sixteen features were chosen.

Chosen Features

- Word Count
- Long Word Count
- Noun Count
- Verb Count
- Adjective Count
- Adverb Count
- Comma Count
- Quotation mark Count
- Punctuation Count
- Sentence Count
- Bracket Count
- Lexical Diversity
- Spelling Error Count
- Foreign Word Count
- Exclamation Mark Count
- Word Length

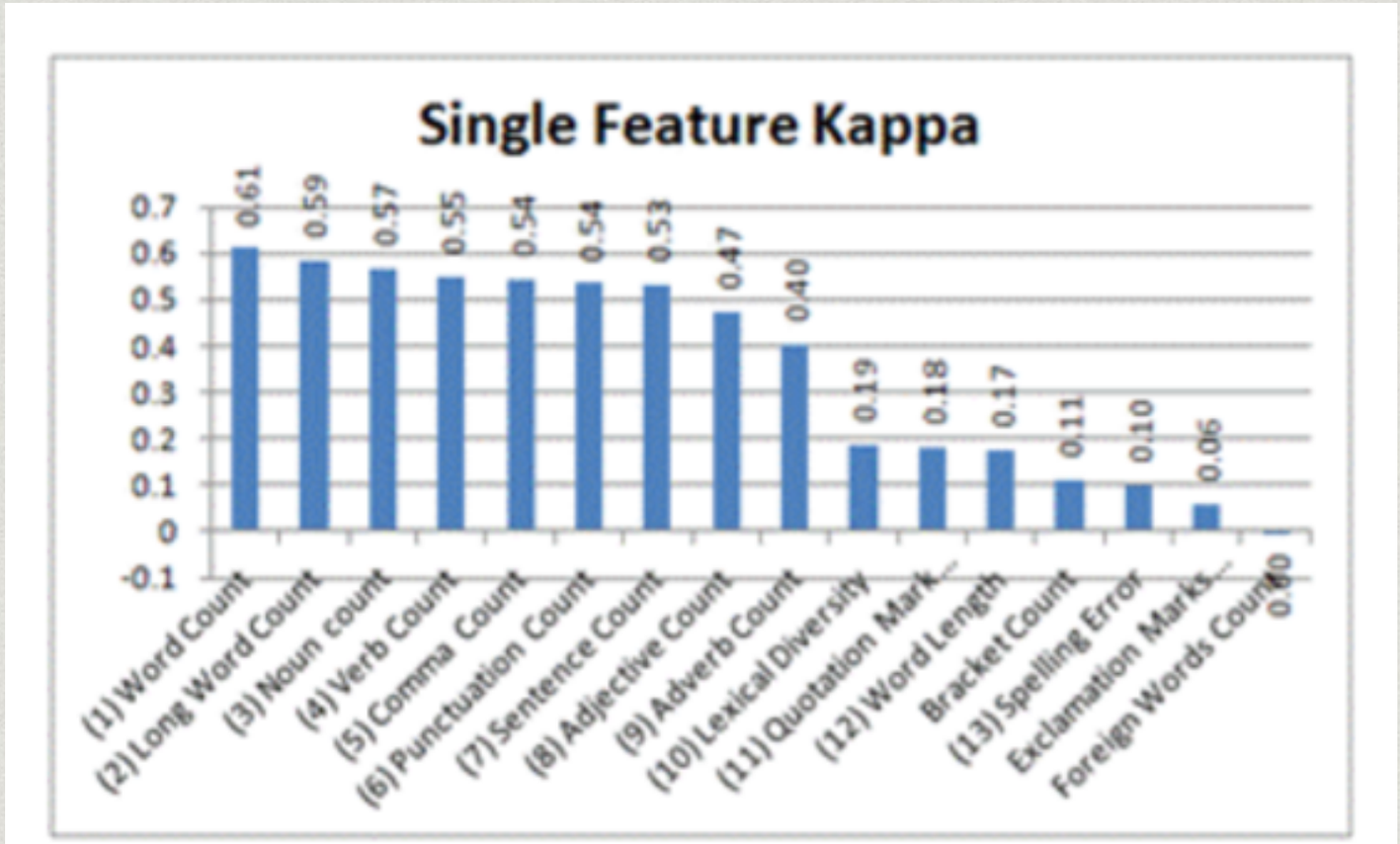
Why these features ?

- These features capture different aspect of the style in which an essay in presented.
- Example :
  - long word count : provides information about how good the vocabulary of the essay is
  - lexical diversity : provides a count of the number of different words used

Learning the model

- Next, we use support vector regression to train our model.
- At first we individually train the model by selecting a feature from the list of features.
- Then, using the quadratic-kappa measure, we calculate the kappa values obtained when each feature was trained individually.

- We then use the forward feature selection approach to choose our final set of features that we would use to train our model.



- Based on the individual feature kappa scores, we first sort it, select each feature incrementally.
- If a new feature which has been added to the set, decreases the overall kappa measure than what was obtained previously, we discard that feature.
- In the end we finalised on 13 features, using which we would train our model.

Limitations

- Grammatical features are not considered.
  - “I am a boy” and “boy I am a” would be evaluated the same.
- Also, the content of the essay is not being accounted for.

Result

```
(venv)sanmanoj:code sankaul$ python run.py -n ../data/SMALL_TRAIN ../models/svm_small ../data_dumps/svm_small -t ../data/SMALL_TEST


================================================

Training the model...
Processing |################################| 100% | 2435 of 2435 | 0 seconds remaining.

Model trained and data dumped successfully.

Loading the test data...
Processing |################################| 100% | 807 of 807 | 0 seconds remaining.

The Average Quadratic Weighted Kappa obtained is:  0.768277571252

================================================
```

Phase 2

- We will try to incorporate grammatical features and structural features of the essay by training a language model to calculate the likelihood of each sentence and sequence of sentences.

Phase 3

- We will try to incorporate content features by considering the prompt of the essay and using techniques such as Latent Semantic Analysis (LSA).

Limitations

- Over reliance on surface features
- Insensitivity to the content of responses
- Insensitivity to the creativity (imagination) of essay takers
- Vulnerability to new types of cheating and test taking strategies

Phase 4 (over ambition)

- Plagiarism Detection.