Automatic Essay Evaluation

A Project Report

submitted by

Sai Chaitanya Banala(12EC25)

Chetan Giridhar Vashisht (12EC31)

Sharang Kulkarni (12EC85)

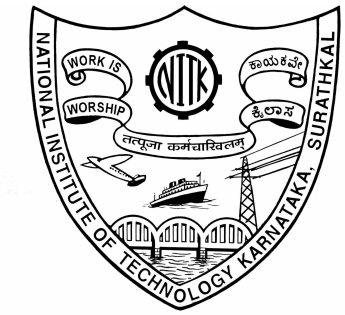
under the guidance of

Dr. Raghavendra Bobbi

in partial fulfilment of the requirements

for the award of the degree of

BACHELOR OF TECHNOLOGY



DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA

SURATHKAL, MANGALORE - 575025

March 18, 2016

ABSTRACT

Manual grading of students' essays is a time-consuming, labour-intensive and expensive activity for educational institutions. It is nevertheless necessary since essays are considered to be the most useful tool to assess learning outcomes. Automated essay evaluation represents a practical solution to this task. In this project we evaluate essays on a content based approach coupled with statistical substitutes using various algorithms.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1
Introduction

Essays are an important expression of academic achievement, but they are expensive and time consuming for states to grade them by hand. So, we are frequently limited to multiple-choice standardized tests. We believe that automated scoring systems can yield fast, effective and affordable solutions that would allow states to introduce essays and other sophisticated testing tools. We believe that you can help us pave the way towards a breakthrough. This project

- challenges us of automated student assessment systems to demonstrate their current capabilities.
- compare the efficacy and cost of automated scoring to that of human graders.
- reveal product capabilities and motivates people to adopt them.

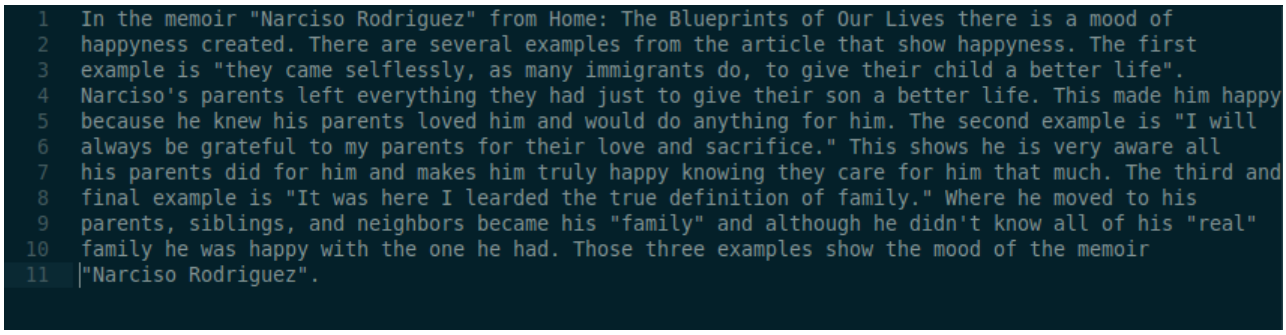


Figure 1.1: A typical essay

1.1 Previous work

Measured relevance of search results using basic NLP techniques and statistical features.

1.2 Motivation

Our previous project on Search Engine Relevance using NLP techniques and statistical features motivated us into taking up a project in the same domain when we learnt that one of the key roadblocks to advancing school-based curricula focused on critical thinking and analytical skills is the expense associated with scoring tests to measure those abilities. For example, tests that require essays and other constructed responses are useful tools, but they typically are hand scored, commanding considerable time and expense from public agencies. So, because of those costs, standardized examinations have increasingly been limited to using bubble tests that deny us opportunities to challenge our students with more sophisticated measures of ability.

CHAPTER 2
Feature Extraction

Since the essays are all rated on a different scale, the training file is first split into eight different files (split according to the essay set). Then, the features are extracted on each of the files and stored to use in the future. Instead of extracting the same features repeatedly, we append the new features extracted to the existing file. (extracting the features is costly as it takes 1.5 seconds per essay and there are over ten thousand essays)

The evaluation metric for the project is the quadratic weighing kappa. This metric penalises the square of the difference between the scores of the rater and the prediction.

Statistical features, grammatical features, argument based features and miscellaneous features are the broad types of features extracted.

2.1 Statistical features

These statistical substitutes are used to judge how well the students write their essays. This works as there is a very strong correlation between a person with good english and person who writes good essays

- Number of sentences
- Number of words
- Number of unique words
- Number of long words
- Number of punctuations used (commas, brackets, quotes)

2.2 Grammatical features

: The wealth of the grammar used by the author is explored using grammatical features. These give a good idea about proficiency of the user. Generally a person with good grammar writes good essays.

- Number of spelling errors (using enchant)
- Parts of speech tagging, number of nouns, verbs, adverbs and adjectives

2.3 Linguistic model

Apart from the statistical features of the text of each essay (such as the total number of words), linguistic features enable us to evaluate essays based on their argument strength. Various features of the essay, such as the Semantic Frames, Transitional Phrases, co-reference, adherence to the prompts topic, take model closer to duplicating human insight while grading essays.

2.3.1 Transitional phrases

14 transitional categories identified and the ngrams of the words are classified into these categories. The total count of each category is taken into account. The categories are:

- Addition: also, again, besides, similarly
- Consequence: accordingly, as a result
- Contrast: accordingly, otherwise
- Direction: here, there
- Diversion: by the way, incidentally
- Emphasis: above all, chiefly
- Exception: other than, outside of
- Exemplifying: chiefly, for instance
- Generalizing: as a rule, as usual
- Illustration: for example, for instance
- Similarity: comparatively, coupled with
- Restatement: in essence, in other words
- Sequence: at first, secondly
- Summarizing: after all, alas

There are 149 phrases that have been accounted and placed into the above categories.

2.3.2 Content based

: First the introductory line and the concluding lines are analysed to identify the side of the argument the author is referring to. The rest of the paragraph is analysed and supporting and opposing views are counted. The analysis is on a comparison basis with a vocabulary for each argument that has been written by us.

2.3.3 Co-reference features

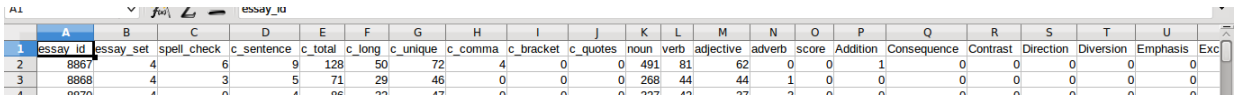
: A strong argument must be cohesive so that the reader can understand what is being argued. While the transitional phrases already capture one aspect of this, they cannot capture when transitions are made via repeated mentions of the same entities in different sentences. Various prompt features are collected.

2.3.4 Semantic frames

: For each essay in our data set, we identify each semantic frame occurring in the essay as well as each frame element that participates in it. For example, a semantic frame may describe an event that occurs in a sentence, and the vents frame elements may be the people or objects that participate in the event. For example, given a sentence like "Mary sold the book to John", the task would be to recognize the verb "to sell" as representing the predicate, "Mary" as representing the seller (agent), "the book" as representing the goods (theme), and "John" as representing the recipient. This is an important step towards making sense of the meaning of a sentence. A semantic representation of this sort is at a higher-level of abstraction than a syntax tree. For instance, the sentence "The book was sold by Mary to John" has a different syntactic form, but the same semantic roles.

2.3.5 Prompt Argument

: Measure the strength of the argument from the introduction and conclusion of the paragraph mentioned. This feature is not useful for all the essay sets.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	essay_id	essay_set	spell_check	c_sentence	c_total	c_long	c_unique	c_comma	c_bracket	c_quotes	noun	verb	adjective	adverb	score	Addition	Consequence	Contrast	Direction	Diversion	Emphasis
2	8867	4	6	9	128	50	72	4	0	0	491	81	62	0	0	1	0	0	0	0	0
3	8868	4	3	5	71	29	46	0	0	0	268	44	44	1	0	0	0	0	0	0	0

Figure 2.1: List of the some of the features extracted

CHAPTER 3
Classification

After all the features are stored into different files, we run three machine learning algorithms on the data to obtain different levels of accuracy for each of the essay sets. A parameter sweep is conducted across the entire dataset to obtain the best set of classifiers to use.

The classifiers used include a random forest classifier, naive bayes classifier, a support vector machine and K nearest neigbours. Each classifier is tested on the entire dataset and a parameter sweep is conducted for each one. The best results are shown on the terminal screen.

CHAPTER 4
Conclusions

We proposed a feature-rich approach to the new problem of predicting argument strength scores on student essays. In an evaluation on around 10000 essays selected from the Kaggle, we implement a linguistic based approach and a statistical based approach to compare both the systems.

4.1 Analysis

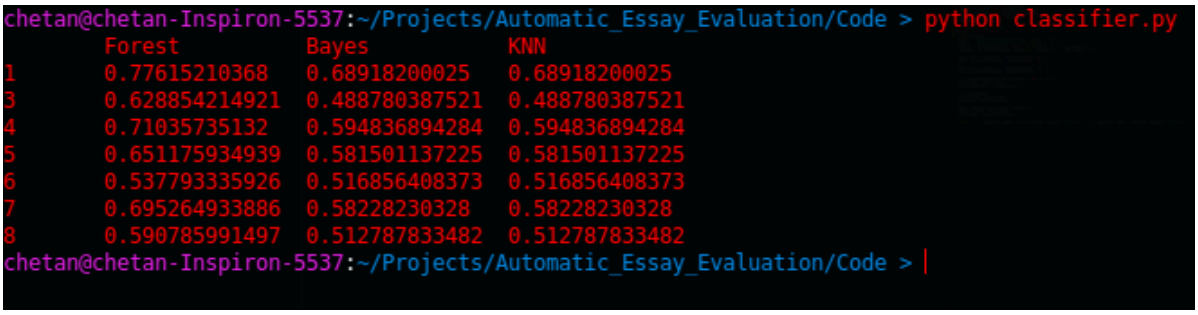


Figure 4.1: Current results with three classifiers

We have currently extracted over 40 features (a mixture of statistical, grammatical and linguistic)and the results with extracted features are documented above. The forest classifier gives the best results among the different machine learning algorithms used. The reported results are after the parameter sweep has been conducted.

REFERENCES

[1] Isaac Persing and Vincent Ng
`Modeling Argument Strength in Student Essays`

[2] Manvi Maha, Mishel Johns, Ashwin Apte
Automatic Essay Grading Using Machine Learning

[3] 2002. MALLET: A Machine Learning for Language Toolkit.
http://mallet.cs.umass.edu.

[4] Transitional Phrases, study guides and Strategies
http://www.studygs.net/wrtstr6.htm

[5] Derrick Higgins, Jill Burstein, Daniel Marcu, and Claudia Gentile.
2004. Evaluating multiple aspects of coherence in student essays

[6] Dipanjan Das, Nathan Schneider, Desai Chen, and Noah A. Smith. 2010.
Probabilistic frame-semantic parsing.

[7] Dipanjan Das, Sam Thomson, Meghana Kshirsagar, Andr F. T. Martins, Nathan Schneider, Desai Chen, and Noah Smith.
Semafor, a frame semantic parser, http://www.cs.cmu.edu/ ark/SEMAFOR/

[8] Kaggle(2012)
https://www.kaggle.com/c/asap-aes

[9] Bird, Steven, Edward Loper, Ewan Klein
Natural Language Processing with Python, O'Reilly Media Inc

[10] Pedregosa, F.Weiss, R. & Brucher
Scikit-Learn: Machine Learning in python

[11] Kelly, Ryan
http://packages.python.org/pyenchant