AUTOMATED ESSAY SCORING SYSTEM

Project Report submitted in partial fulfillment of the requirement for the degree of

Bachelor of Technology
in
Computer Science & Engineering
under the Supervision of

Mr. Suman Saha

By

AKASH GUPTA
111257

To

Jaypee University of Information Technology
Waknaghat, Solan- 173234
Himachal Pradesh

CERTIFICATE

This is to certify that project report entitled: "AUTOMATED ESSAY SCORING SYSTEM", submitted by AKASH GUPTA (111257) in partial fulfillment for the award of degree of Bachelor of Technology in Computer Science & Engineering to Jaypee University of Information Technology, Waknaghat, Solan has been carried out under my supervision.

This work has not been submitted partially or fully to any other University or Institute for the award of this or any other degree or diploma.

Date: 15 May 2015

Supervisor's Name: Mr. Suman Saha

Assistant Professor JUIT, Waknaghat

ACKNOWLEDGEMENTS

It is my pleasure to be indebted to various people, who directly or indirectly influenced my thinking, behavior and act during the course of this project. I express my sincere gratitude to this university for providing me the opportunity to undergo this project and their help whenever required for the project. A special thanks to my final year project supervisor, Mr. Suman Saha, whose stimulating suggestions and encouragement, helped me to get to the thrust of my topic and understanding the importance of the project. I would also like to acknowledge with much appreciation the crucial role of the staff of Computer Laboratory, who provided me with the lab facilities as and when required. Additionally, I appreciate the guidance given by the panels especially during the previous project presentation which made me realize the various dimensions I was probably missing out and hence, they gave away a room for improvement in the project. Again a special thanks to my friends who gave me valuable suggestions regarding the project. Last but not the least, my heartiest appreciation goes to my parents and college for their encouragement and advice that helped me enormously in successful completion of this project.

Date: 15 May 2015

Name: Akash Gupta

TABLE OF CONTENT

Topic	**Page No.**
Certificate	
Acknowledgment	
Table of Content	
List of Figures	
List of Tables/Graphs	
List of Abbreviations	
Abstract	
• Introduction	1-5.
○ History	1.
○ Introduction AES	2.
○ Existing Systems	4.
○ Motivation	6.
• Literature Survey	7-9.
○ Feasibility Study	
○ Assessment Techniques	8.
• Project Design	10.
○ Gantt Chart	
• Implementation	11.
○ Working of AES	12.
○ Applying AES	15.
○ Algorithm	18.
• Verification & Validation	24.
○ Experimental Results	
○ Evaluation Criteria and Performance Management	
○ AES Advantages	
• Future Work	30.
• Conclusion	31.
• References	33.

LIST OF FIGURES

S.No.	**Title**	**Page No.**
1	Assessment Equation	8.
2	Working of AES	11.
3	Application of AES	15.
4	Detailed Working of AES	16.
5	Basic Algorithm	18.
6	Error Detection Algorithm	19.
7	Syntactic Evaluation Algorithm	19.
8	Semantic Algorithm	21.
9	Essay Length Algorithm	22.
10	Final Score Model	23.
11	Experimental Results	24.

LIST OF TABLES/ GRAPHS

S.No.	**Title**	**Page No.**
1	Gantt Chart	10.
2	Experimental Results	24.
3	AES Evaluation and Performance - I	27.
4	AES Evaluation and Performance – II	28.

LIST OF ABBREVIATIONS

- AES – Automated Essay Scoring
- WO – Word Order
- SVA – Subject Verb Agreement
- VU – Verb Usage
- T – Topic Coherence
- TC – Text Coherence
- EL – Essay Length

ABSTRACT

Automatic Essay Grader is a system that scores and evaluates essays. In this project, the focus is on evaluating and scoring essays written by students. The project has two complementary phases, first to detect errors in grammar, usage, semantics, coherence, length etc. and a scorer/grader, an automated essay scoring system. Natural language processing techniques are used to implement out automatic essay grader. Essays are crucial testing tools for assessing academic achievement, integration of ideas and ability to recall, but are expensive and time consuming to grade manually. Therefore, there is a need and a huge potential for such automated evaluation and grading systems. The problem statement is that access to hand scored essay set is provided. It is required to build, train and test scoring engines against a wide variety of essay set. The objective is to implement an open source efficient essay scoring model and then analyze and compare the efficiency and cost of automated scoring to that of human graders.

HISTORY OF AUTOMATED ESSAY SCORING SYSTEM

Computer based assessment began in 1955 when Lindquist developed optical test scoring equipment at the University of Iowa. The idea of AES first came about in 1966 and was advanced by Ellis Page. It took him around two years to come up with working software. His software, called Project Essay Grade (PEG), was later purchased by Measurement, Inc., which continues to develop it. Later, in the 1990s and 2000s, several other companies, such as Educational Testing Service, Pearson, and CTB/McGraw-Hill, started developing their own tools. Some open tools, such as BETSY, also came up. One major use case of these tools was as an automated "second reader" for high stakes tests. A human first scored the test, after which a machine scored it. If the two scored differed by a certain amount, then a third human re-scored the paper to resolve the dispute.

INTRODUCTION

Automated essay scoring is the use of computer programs to assign grades to essays. It is a method of assessment which uses natural language processing. Its objective is to classify a large set of textual entities into a small number of discrete categories, corresponding to the grades. Broadly classifying the scoring, we get the evaluation categories based on syntax (grammar), semantics and essay length. The project devices a scoring model to evaluate each category and compute a final score. The essays used are at first manually evaluated by human evaluators and these scores are used as a benchmark to compare and evaluate the automated essay grader.

Automated Essay Scoring

AES is the art of giving students automatic, iterative, and correct, scores and feedback on their essays and constructed responses.

- Feedback: In AES application like “second reader”, in exams like GMAT, GRE and TOEFL, feedback is very important.
- Iterative: Provides very quick feedback and scoring and there is no upper limit on the number of times the essays can be evaluated by the system.
- Constructed responses: Automated scoring is not just about essays but also grading constructed responses.
- Correct: Automated essay scoring is useful if its scores are nearly close to manual human grading or maybe only marginally inaccurate because its utility goes away if it can't score properly.

The Value of Essays

Essay tests provide a better indication of students’ real achievements in learning. Students are not given ready-made answers. It is expected that they must have command of an ample store of knowledge that enables them to relate facts and principles, to organize them into a coherent and logical progression, and then to do justice to these ideas in written expression. Essays also provide an indication of the nature and quality of students’ thought processes, as well as their ability to argue in support of their conclusions.

An essay examination is relatively easy to prepare but rather tedious and difficult to score accurately. A good objective examination is relatively tedious and difficult to prepare but comparatively easy to score. A conclusion can be made then that computer support for scoring objective tests is widely available, but that essay testing may be preferred for measuring the higher level abilities of students. If essays could also be graded by computers, then the time consuming tasks of human grading could be reduced and efficiencies in grading could be obtained similar to that obtained for objective tests. Computer grading of essays is now possible, and the accuracy of the grading can match that of humans. University students have always been required to write essays for assessment. An essay topic, expected length, and due date are generally specified by the lecturer. The student is then expected to research the topic, think about the issue, and write his/her response. The student has to be careful about plagiarism, and to correctly reference source material. Essays are generally used when the lecturer wants to assess the student’s ability to express and synthesize ideas, which cannot be measured by multiple choice or short answer tests.

Automatically Grading Essays

Essays can now be graded automatically by specialized software. Some of the systems are listed below.

- AutoMark
- Bayesian Essay Test Scoring System
- Conceptual Rater
- Content Analyst
- Educational Testing Service
- Electronic Essay Rater
- Intelligent Essay Assessor
- Intelligent Essay Marking System
- Intellimetric
- Blue Wren Software
- Paperless School Free Text Marking Engine
- Project Essay Grade
- Rx Net Writer
- SA Grader
- Schema Extract Analyze and Report
- Text Categorization Technique

These systems make use of natural language processing technology and statistical techniques to analyze style and content. Most of these systems can perform as well as human markers in the sense that the computer-human score correlations are similar to the human-human correlations on the same essays.

EXISTING SYSTEMS

PEG

Page conducted a large scale study of PEG effectiveness using senior essays from the National Assessment for Educational Progress (NAEP) in 2014. NAEP essays were scored by two human judges and Page recruited six more human judges to score each essay on a six point scale. The human judges achieved a multiple regression correlation of .877 with each other, in comparison PEG achieved a correlation of .869.

The PEG system was sold by Dr. Ellis Batten Page to Measurement Incorporated in 2002. In January 2012, the Hewlett Foundation invited nine major vendors of artificial intelligence (AI) scoring of student essays to participate in the Automated Scoring Assessment Prize (ASAP) competition. AES system scores were correlated with the scores of two professionally trained readers.

In January, the Hewlett Foundation invited Measurement Incorporated (MI) and eight other major vendors of artificial intelligence (AI) scoring of student essays to participate in the Automated Scoring Assessment Prize (ASAP) competition. The competition included essays written to eight different prompts by students in various grade levels. Each essay had been scored by two professionally trained readers. The human readers had agreement indices of .75. PEG achieved the highest agreement index with the human readers at 0.79.

IEA

Landauer, Latham and Foltz in 2000, used linear regression to compare IEA scoring with human graders on 3,926 essays on 15 diverse topics with a resulting correlation of .85. Even better results were achieved with 900 creative narrative essays from the GMAT with a correlation coefficient of .90 which was identical to that of two human graders.

IEA is currently owned by Pearson Knowledge Technologies a subsidiary of Pearson Education. A recent Pearson white paper claims that its Oral Reading Fluency testing system (IEA based) achieved scores that correlate with human scores at 0.98, while the correlation between pairs of human raters was 0.99.

E-Rater

E-rater was used from 1999 through 2006 to score the GMAT. According to Valenti, Neri and Cucchiarelli (2003) the agreement rate between E-rater and human scorers of the GMAT on over 750,000 essays was over 97%.

In 2006 the GMAT switched to IntelliMetric scoring which is based on the BETSY AES. E-rater is owned by ETS and currently is the software that runs the Criterion Online Writing Evaluation service.

BETSY

BETSY is now owned by Vantage Learning and is the software behind the IntelliMetric system. It also powers the My Access writing assessment tool. Since 2007, IntelliMetric has been used to score the GMAT.
According to Valenti, Neri and Cucchiarelli in 2003, BETSY achieved an accuracy rate over 80% on a test involving 462 essays. Dikli reports of 2006, IntelliMetric in a test involving 8th grade student essays achieved an adjacent correlation scoring of .95 with human scorers and .99 with expert human scorers. According to the Vantage Learning IntelliMetric website when using a 6-point scale, two experts will agree with each other within 1 point about 95% of the time. IntelliMetric typically agrees with either expert about 97% to 99% of the time.

MARKIT

Markit was created by Robert Williams and Heinz Dreher of the Curtin University of Technology in Australia. In a study of 20 essays in a business law class Markit scores were compared to the scores assigned by the course instructor. The average human score was 61.75 while the Markit average was 62.35, the correlation between Markit and the human grader was .79. Williams in a study using Markit to score 290 high school student essays found a correlation of .79 with three human graders.

Essays for Sale

Students today have available to them many World Wide Web (Web) sites that can provide an essay for a fee.

- Custom Writing
- CustomEssays.co.uk
- Prime Essay
- Tailored Essays
- Order Papers.com
- OvernightEssay.com

These sites provide essays from databases of pre-written essays, or writers will write custom essays to order. Turnaround time can be as little as three hours. Detection of these bought essays is difficult because we assume that they are not published to the Web and hence cannot be detected by search engines.

MOTIVATION

Essays are crucial testing tools for assessing academic achievement, integration of ideas and ability to recall, but are expensive and time consuming to grade manually.

- **Manual Grading vs Automating Grading**
  Manual grading of essays takes up a significant amount of instructors' valuable time, and hence is an expensive process.
  Automated grading, if proven to match or exceed the reliability of human graders, will significantly reduce the costs being incurred.
- In recent decades, large-scale English language proficiency testing and testing research have seen an increased interest in constructed-response essay-writing items. The TOEFL iBT, for example, includes two constructed-response writing tasks, one of which is an integrative task requiring the test-taker to write in response to information delivered both aurally and in written form. Similarly, the IELTS academic test requires test-takers to write in response to a question that relates to a chart or graph that the test-taker must read and interpret.
- An ideal English language proficiency test should make it possible to differentiate, to the greatest possible extent, levels of performance in those dimensions of performance which are relevant to the kinds of situations in which the examinees will find themselves after being selected on the basis of the test. However, constructed-response writing tasks have both advantages and disadvantages. Unlike multiple-choice items which have a single criterion for correctness, experts often disagree on how to operationalize and score the set of qualities that define excellent writing.
- While the fact that constructed-response essay items require students to generate samples of normative language may make such items a more proximal measure of communicative writing ability, the process of scoring essay items is quite complex. Human raters must be hired and trained to score each of the examinee essays.
- In addition, the use of human raters introduces a new challenge to maintaining the reliability and construct validity of test scores, as raters are bound to differ in their perceptions of candidate performances and their tendencies towards leniency and severity. Raters may also have unconscious biases that are not immediately amenable to correction through training.

FEASIBILITY STUDY

The feasibility study concerns with the consideration made to verify whether the system is fit to be developed in all terms. Once an idea to develop software is put forward the question that arises first will pertain to the feasibility aspects. There are different aspects in the feasibility study:

- **Operational Feasibility:**

There is no difficulty in using the system, since the system will be made available as an open source software and iOS, Android and Windows Application and since apps are a common feature these days. Therefore, it is assumed that no one will face any problem in running the system.

- **Technical Feasibility:**

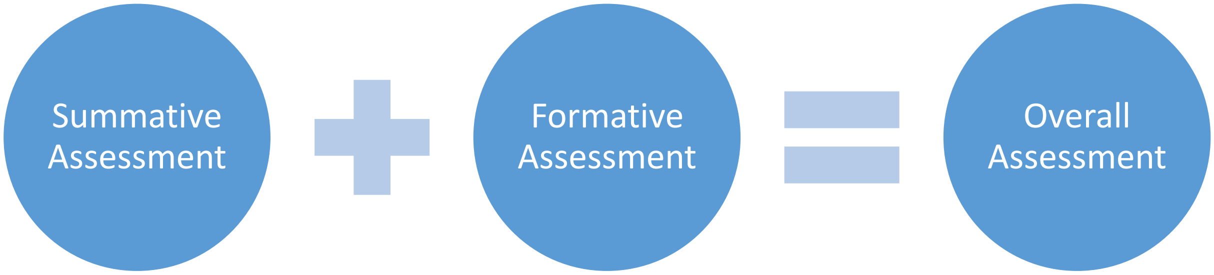
Technical feasibility deals with the study of function, performance, and constraints like resources availability, technology, development risk that may affect the ability to achieve an acceptable system and as we know handling SDK's/IDE's and Tools is quiet easy a task.

- **Economic Feasibility:**

One of the factors, which affect the development of a new system, is the cost it would incur. The proposed system is really cost effective. Hence, very little or no cost has to be incurred to develop the system.

ASSESSMENT TECHNIQUES

Although many instructors enjoy teaching many do not enjoy the effort required in grading student work. Assessment of student learning is an integral part of teaching and also the most time consuming. As Dreher, Reiners and Dreher state, "assessment guides the teaching and learning process by providing reciprocal feedback to both educators and students so that they may improve in their respective tasks".



Assessment can serve two purposes either formative or summative.
Summative Assessment measures a student's learning up to that point in time in a course. Multiple choice tests are often utilized for summative assessment because they can measure a student's knowledge of facts and the course content.
Formative Assessment is used diagnostically to both assist the student and the teacher. Formative assessment provides feedback to the student on their progress and helps the teacher to refine teaching and learning methods to maximize student progress. There are various methods of formative assessment but essay writing is one of the most common.

Summative Assessment

Economic considerations present in education today often dictate large class sizes. Time and effort limitations often necessitate the use of multiple choice exams by instructors. Students experience a great deal of summative assessment but less formative assessment. Blayney and Freeman point out that multiple choice questions are more efficient, especially with vendor provided pre-existing questions, but they "do not test higher order application or provide extensive feedback that students can use to identify their own misunderstandings".

Technology has often been used to assist instructors with grading multiple choice exams, for example the ubiquitous Scantron, which is still in use at many universities. "Such automated assessment can provide a quick, reliable, cost-effective means of assessing large numbers of students" but does not provide formative assessment.

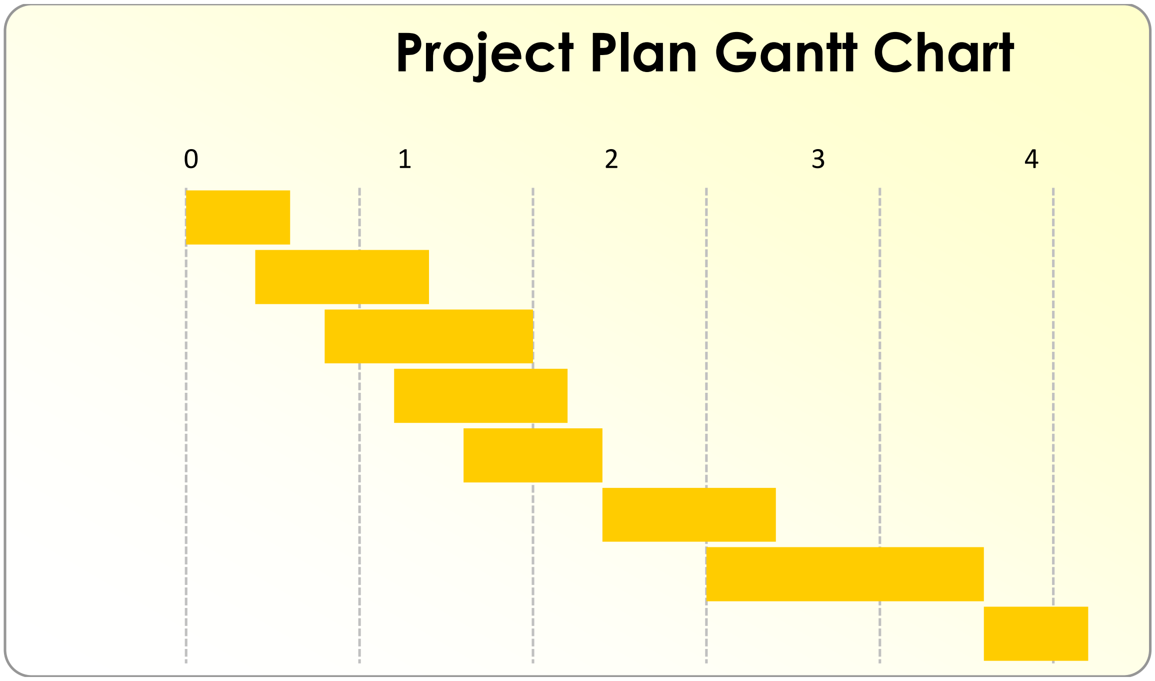
Multiple choice testing while useful for summative assessment mostly addresses surface learning and cannot adequately assess application of knowledge to real life situations. Other limitations of the multiple choice format include lower reliability due to student guessing, lower validity due to inadvertent hints provided by response format, and lower validity due to inability to measure complex construct. Blayney and Freeman found that students retained less when multiple choice questions were graded by a scanner and returned at the next class than when tested on an answer until correct basis.

Formative Assessment

The formative assessment provided by open-ended essay questions is the most productive method of assessing student learning. Formative assessment can collect "detailed information about students' learning status for planning instructional feedback" as well as knowledge and application of concepts. In particular, essay questions are considered by many educators to be the most useful tool for assessing learning outcomes. Open-ended questions require the ability to recall, organize and integrate ideas and the ability to express oneself in writing. Discussion questions "often display wider aspects of students' individuality, personal perspective, and creativity". Essay and discussion questions can also be used to improve students' abilities to solve real world business problems. In today's globalized business environment students must be creative thinkers, problem solvers, planners, decision-makers and able to participate in team activities. Such skills can only be improved through application of the higher levels of Bloom's taxonomy. Bloom's taxonomy has six levels – knowledge, understanding, application, analysis, synthesis and evaluation. Multiple choice questions can assess knowledge and understanding, but the higher levels of application, analysis, synthesis and evaluation require essay and discussion questions. These higher level functions require transfer of theory to practical situations (application), identification of relevant components and logic in the learning material (analysis), combining information to produce new products (synthesis) and making decisions that create an impact on a given application (evaluation). Student success is increased when they are "given challenging, real-world practice assignments with rapid, meaningful feedback".

PROJECT PLAN

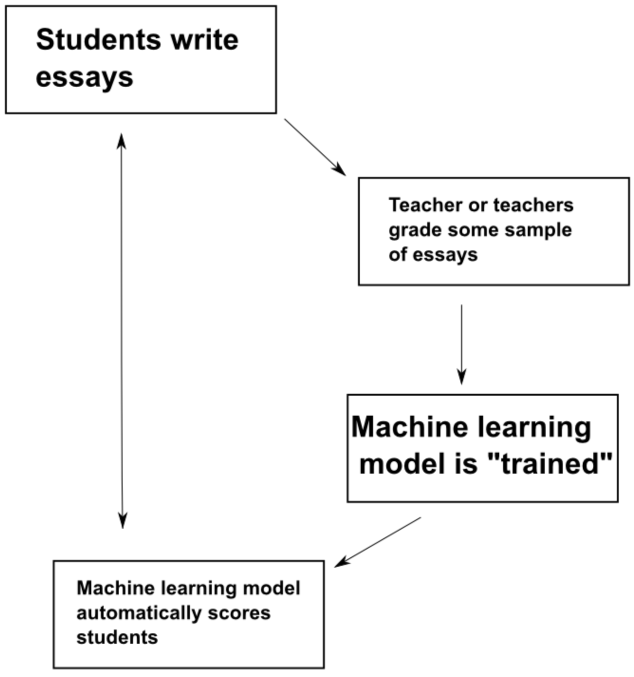
Activity	Start	# of Months
Literature Review	0	½
Existing Systems Research	1/2	1
Module Design	1	3/2
Algorithm Construction	1	3/4
Implementation	2	1
Testing	3	1
Future Work	3	2
Final Report	4	1



IMPLEMENTATION

AES Working

Students first write some essays. Teachers then grade these essays using whatever criteria they want and a machine learning model is created. A machine learning model differs from a machine learning algorithm. A machine learning algorithm is a blank slate that can be trained to do a certain task. To make a bit of a stretch analogy, think of it as a computer brain -- it is capable of learning something, but it doesn't know how to do it yet. Then this algorithm is trained, this computer brain is trained, to score essays. After it has been trained, it gives a machine learning model, which can be used to score more essays. In order for a machine learning model to be created, features first need to be extracted from the text, as a computer cannot directly understand English.


Students write essays
Teacher or teachers grade some sample of essays
Machine learning model is "trained"
Machine learning model automatically scores students

For example, in my current apartment, one feature is that it has 1.5 bathrooms, and another feature is that it has 2 bedrooms. If I was going to build a machine learning model to predict apartment rents, I might pass in these features.

I would then map the features to a certain amount of rent. So, for example, if one apartment has 1.5 bathrooms and 2 bedrooms and costs 1,000 dollars a month in rent, whereas another apartment has 1 bathroom and 1 bedroom and costs 500 dollars a month in rent, a machine could learn that a certain number of bedrooms and a certain number of bathrooms equal a certain amount of rent. So, if we ask it to predict the rent for an apartment with 1 bathroom and 2 bedrooms, it might say 900 dollars.

Let's look at this is the context of essays, using some examples:

Say that I wanted to give a survey today and ask you *why do you want to learn about machine learning?* The responses might look like this:

```
I like solving interesting problems.

What is machine learning?

I'm not sure.

Machine learning predicts everything!
```

Let's say that the survey also asks people to rate their interest on a scale of 0 to 2. So now the responses and associated interest scores are:

Number	**Response**	**Score**
1	I like solving interesting problems.	2
2	What is machine learning?	0
3	I'm not sure.	0
4	Machine learning predicts everything!	2

So, let's say that we get a half-filled-out survey that forgot to include the interest score. All we got was the sentence I really like solving problems. Machine learning is very useful. Now, if we look at this in the context of the other responses, we can infer that the interest of the person is likely a 2/2. But how would a computer do the same thing?
Through features. Some of the features we might extract:

- Presence/absence of the phrase solving problems. (0 if absent, 1 if present)
- Number of sentences.
- Presence/absence of machine learning.
- Average word length.
- Presence/absence of machine.

This is a very simple example, but it gives you a good idea of what features are. Features allow us to represent text, which a machine does not understand, as numbers, which it does understand.

We can then tell a machine learning algorithm, such as a random forest, or a linear regression, that a certain sequence of features means that the teacher gave the student a 2, another sequence of features means that the teacher gave the student a 0, and so on. This trains the algorithm, and gives a model.

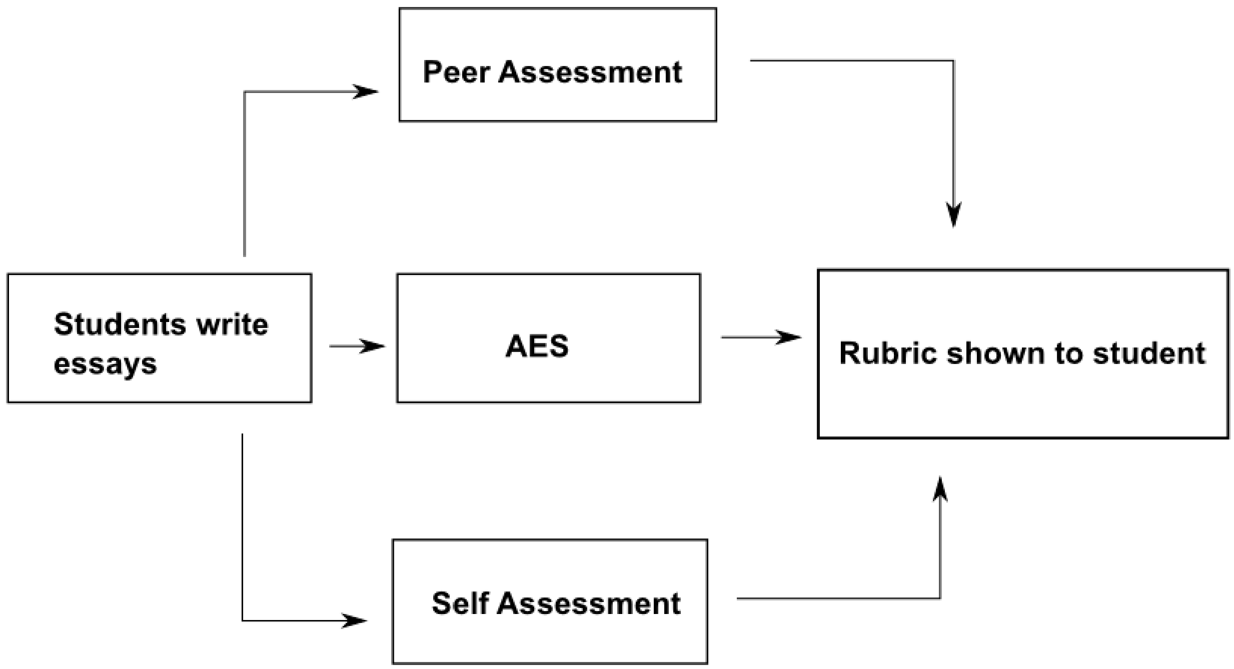
Once the model is created, then it can predict the scores for new essays. Then take a new essay, turn it into a sequence of features, and then ask our model to score it.

As one can see, what the model is trying to do is mimic the human scorer. The model is figuring out how an expert human scorer grades an essay, and then trying to apply that same criteria to other essays. So, it isn't actually a machine judging essays on arbitrary criteria; it is a machine trying to figure out the criteria a human uses to score essays, and then apply those criteria to grade other essays.

APPLYING AES

So, when a student answers a question, it goes to any or all of self, peer, and AES to be scored. Written feedback (from peer assessment), and rubric feedback (from all three assessments) are displayed to the student.

A diagram: Grading essays and Constructing responses



It is completely up to the instructor how each problem is scored, and how the rubric looks. Here is an example rubric:

Topicality

0 points - Student is off topic

1 point - Student stays on topic

Photosynthesis

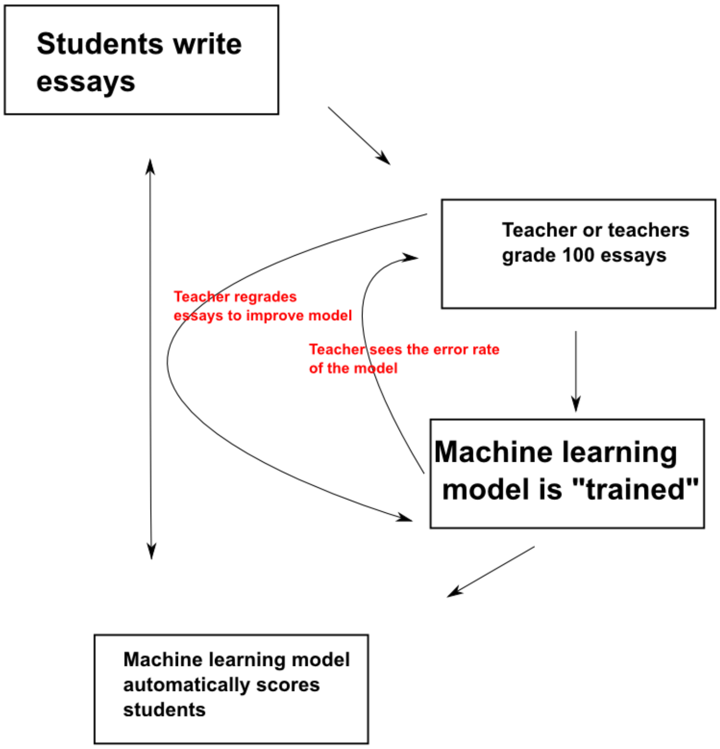
0 points - Incorrectly defines photosynthesis

1 points - Partially correct definition

2 points - Fully correct definition

The AES would tell us how you did on each of the rubric dimensions (which are customizable by the instructor).

Here is specifically how the AES works


Students write essays
Teacher or teachers grade 100 essays
Teacher regrades essays to improve model
Teacher sees the error rate of the model
Machine learning model is "trained"
Machine learning model automatically scores students

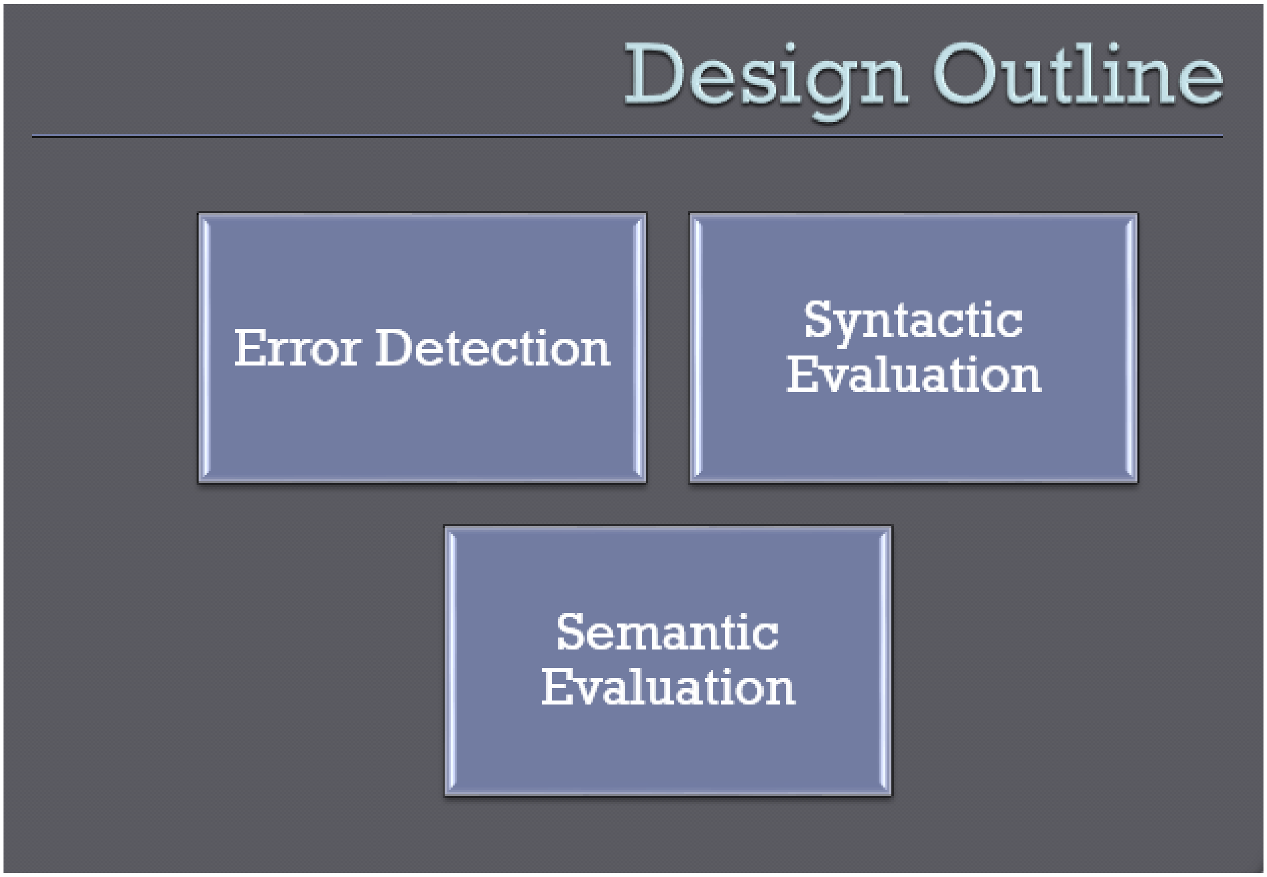
The main difference between this and the generic workflow shown is that it allows teachers to grade essays again that AES has scored poorly. When a machine learning model scores an essay, it doesn't just give you a score; it also gives you a confidence value from 0% - 100% associated with that score. A low confidence indicates that the machine learning model does not know how

to score a given essay well. When a teacher re-scores a paper, it gives the student the correct score, and makes the machine learning model better (it won't make the same mistake twice).
This is called *active learning*.

The AES will give the student feedback on how many points they scored for each category of the rubric.

ALGORITHM

The basic procedure for essay scoring is to start with a training set of essays that have been carefully hand-scored. The program evaluates various syntactic aspects such as word order, subject-verb agreement, verb usage and sentence formation and semantic aspects such as topic and text coherence and the length of an essay. It then constructs a grading model that relates these quantities to the scores that the essays received. The same model is then applied to calculate scores of new essays. Here is a description of my approach to evaluate and score each essay.



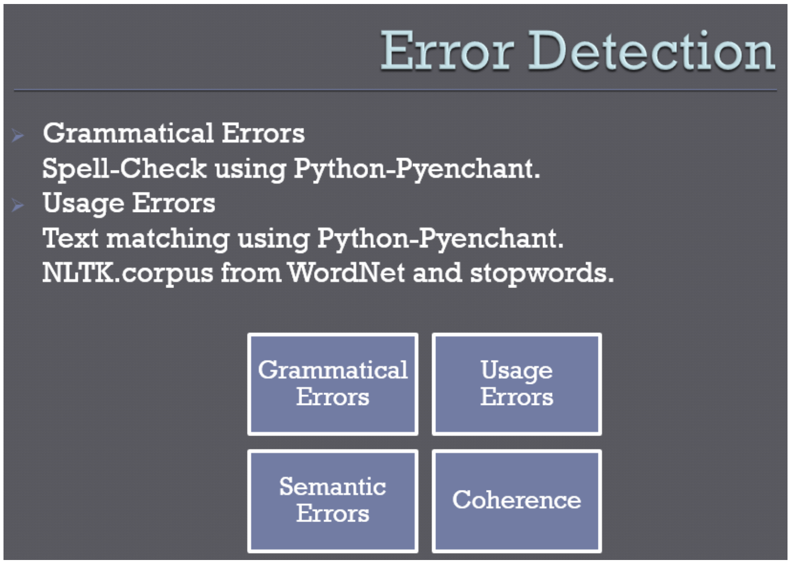
Design Outline

Error Detection

Syntactic Evaluation

Semantic Evaluation

Error Detection



Error Detection

- Grammatical Errors
  Spell-Check using Python-Pyenchant.
- Usage Errors
  Text matching using Python-Pyenchant.
  NLTK.corpus from WordNet and stopwords.

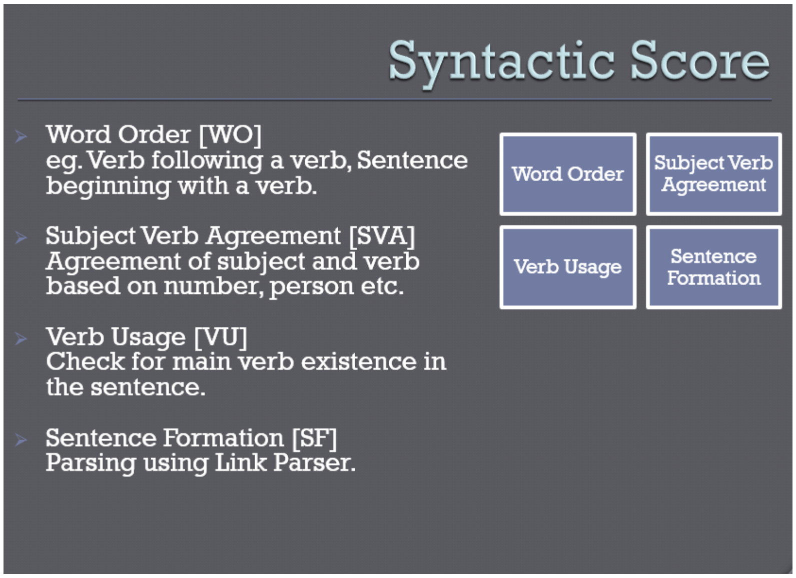
Grammatical Errors

Usage Errors

Semantic Errors

Coherence

Syntactic Evaluation



Syntactic Score

- Word Order [WO]
  eg. Verb following a verb, Sentence beginning with a verb.
- Subject Verb Agreement [SVA]
  Agreement of subject and verb based on number, person etc.
- Verb Usage [VU]
  Check for main verb existence in the sentence.
- Sentence Formation [SF]
  Parsing using Link Parser.

Word Order

Subject Verb Agreement

Verb Usage

Sentence Formation

(a) Order of Words

To facilitate the evaluation of word order in a sentence, speech tagging is used.
To determine errors in the word order, a set of predefined rules generated on the basis of the training set that cause the most common errors e.g. a verb following a verb.
To identify inconsistencies such as sentences beginning with verbs etc. Stanford speech tagger components are used.

(b) Agreement of Subject and Verb

For subject-verb agreement evaluation Stanford speech tagger is used. Firstly, identify the subject and the main verb in the sentence. Check for agreement between the subject and the main verb on the parameters like person, number etc.
To identify attributes for the subject and the main verb, identify the noun phrase and the verb phrase in the sentence, the parts of speech tags would provide the person, number, case etc. of a particular word. A violation of agreement between the person or number of the subject and the main verb is accounted for as an error. Additional exceptional rules such as use of a plural verb with two or more subjects connected by conjunction such as and were employed.

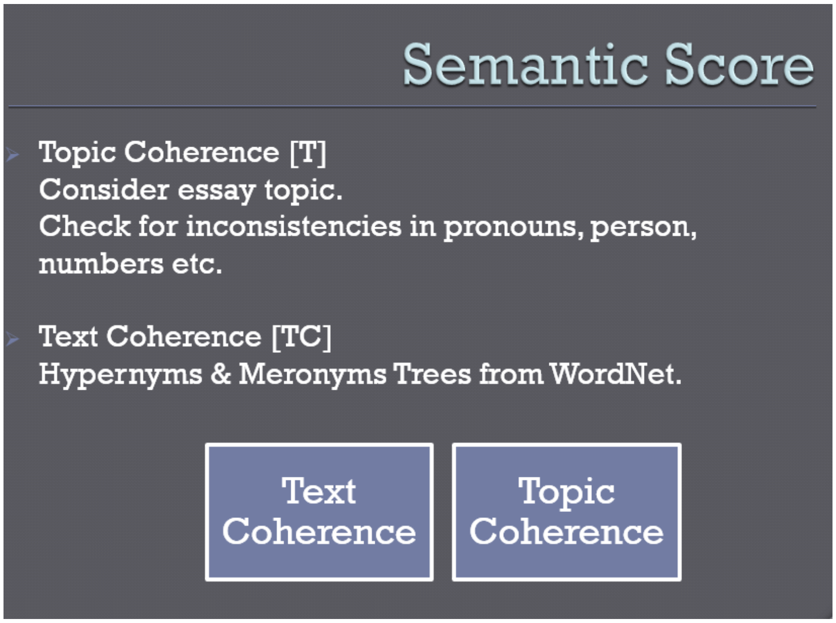
(c) Use of Verb

To evaluate very usage, Stanford speech tagger is used. The existence of a main verb in a sentence is checked for and its absence is marked for errors. The parts of speech tags are used to identify the tense of every verb. Any inconsistencies in the tense of the verbs is marked as an error.

(d) Sentence Formation

The algorithm uses parts of speech tags and the parse trees to perform the evaluation of the grammatical quality of the essay. To tag each word in the essay Stanford speech tagger is used and OpenNLP parser provides the parse tree. Then check the children of root element (TOP) for inconsistencies that is, any element other than a sentence tag (S) as an immediate child of the root is marked as an error.
Check for the existence of a clause introduced by a subordinating conjunction, along with the subordinating conjunction words such as because. All inconsistencies are marked as errors.
The subject and object are identified in every sentence from the Noun Phrase (NP) and Verb Phrase (VP) respectively. Check for existence of verbs in the sentences. Inconsistencies are marked as errors.

Semantic Algorithm



Semantic Score

- Topic Coherence [T]
  Consider essay topic.
  Check for inconsistencies in pronouns, person, numbers etc.
- Text Coherence [TC]
  Hypernyms & Meronyms Trees from WordNet.

Text Coherence

Topic Coherence

(a) Coherence of Text

Evaluate the coherence of the content of the essay. Speech tagging and parsing techniques are used. Considering the topic of the essays (Tell us about yourself and your family), assume that all first person pronouns e.g. I refer to the student writing the essay. Even plural pronouns in first person e.g. we are considered as a group that the user belongs to. So, these pronouns are accepted to be correct. Second person pronouns are considered incorrect e.g. you, given the topic of the essay.

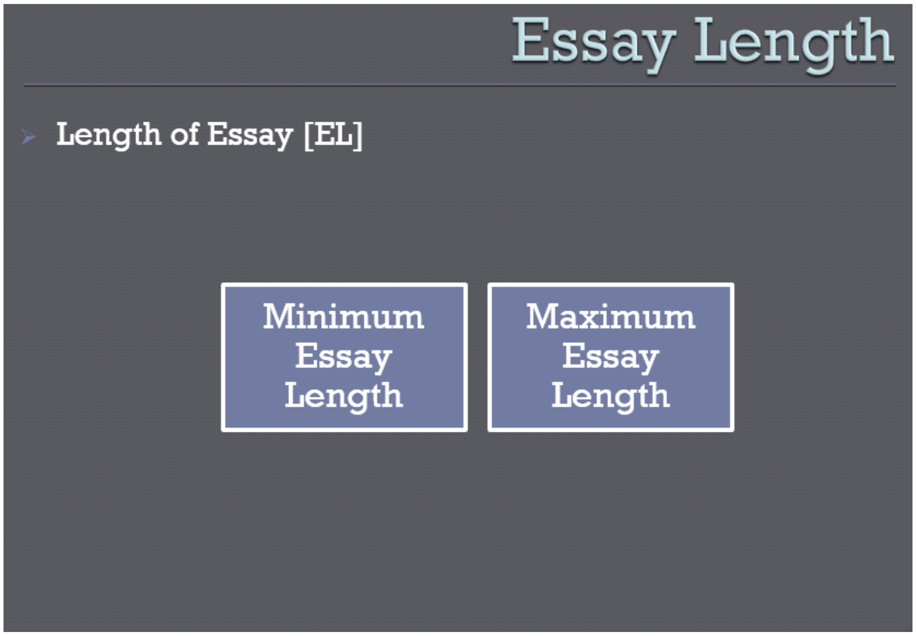
Usage of third person pronouns e.g. they are awarded with a bonus in the score. Approximate that antecedents to be from the previous two sentences. Create a queue of the entities in the previous two sentences and find antecedents based on the gender and number. The scoring pattern adopted is such that absence of an antecedent is marked as an error but an ambiguity of antecedents is marked as an error with an appropriate weight. Device a few exceptional rules such as use of plural pronouns e.g. they when the sentence contains multiple entities and a conjunction. Use a predefined list to identify the gender and number of the entities. Entities such as dog are not assigned a gender because people tend to all the genders for them.

(b) Coherence of Topic

The adherence of the content of the essay to the topic is evaluated here. Speech tagging, parsing and Wordnet is used here. Hypernyms from Wordnet are used to identify words of is-a relation in the essay e.g. the hypernym tree of the word brother would iteratively lead to relative. Check for all the common nouns in the essay for hypernyms that lead to words pertaining to relative, person, family etc.

Meronyms from Wordnet are also used to find words that have a part-of relation with the words in the essay e.g. family has meronyms child, sibling, parent etc. Check for the content words, hypernyms and meronyms and any inconsistencies are marked for errors as deviation from the topic.

Essay length



Essay Length

- Length of Essay [EL]

Minimum Essay Length

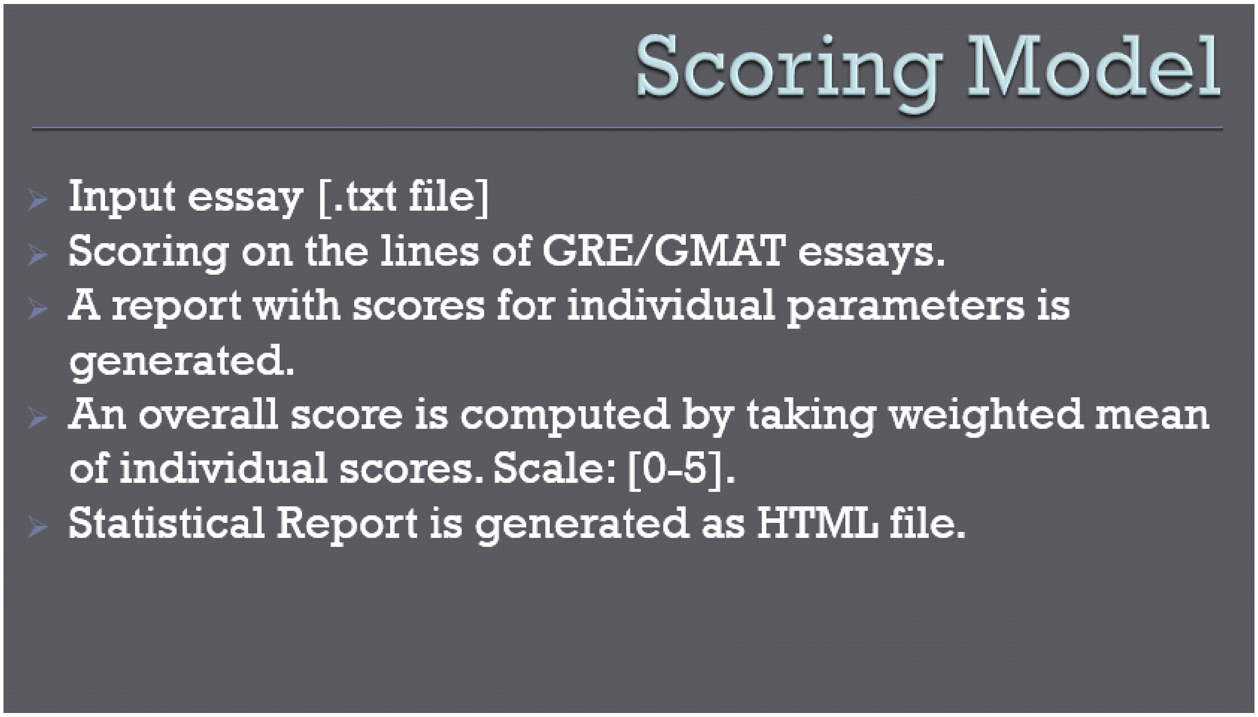
Maximum Essay Length

The essay should essentially have a minimum and a maximum length based on number of sentences. To identify the number of sentences in the essay using speech tagging. Estimate the length of the essay by checking for punctuations. In case of absence of punctuations, the algorithm uses the part of speech tags to count the number of finite verbs, conjunctions such as and, but etc. to estimate the number of sentences.

Calculating Final Score

After estimating the errors in the essay for each category, an error rate is computed based on the number of errors and other appropriate attributes of the essay such as number of sentences, number of verbs etc. based on the category of error.

Then a scoring pattern to award the essay a score closest to the scores awarded by human evaluator in the training set is devised. This model is used to evaluate the essays in the test set and eventually, evaluate the automatic essay grader. Grades are awarded for each category from a scale of 0 to 5 and this expression is used to compute the final score that is awarded to the essay.

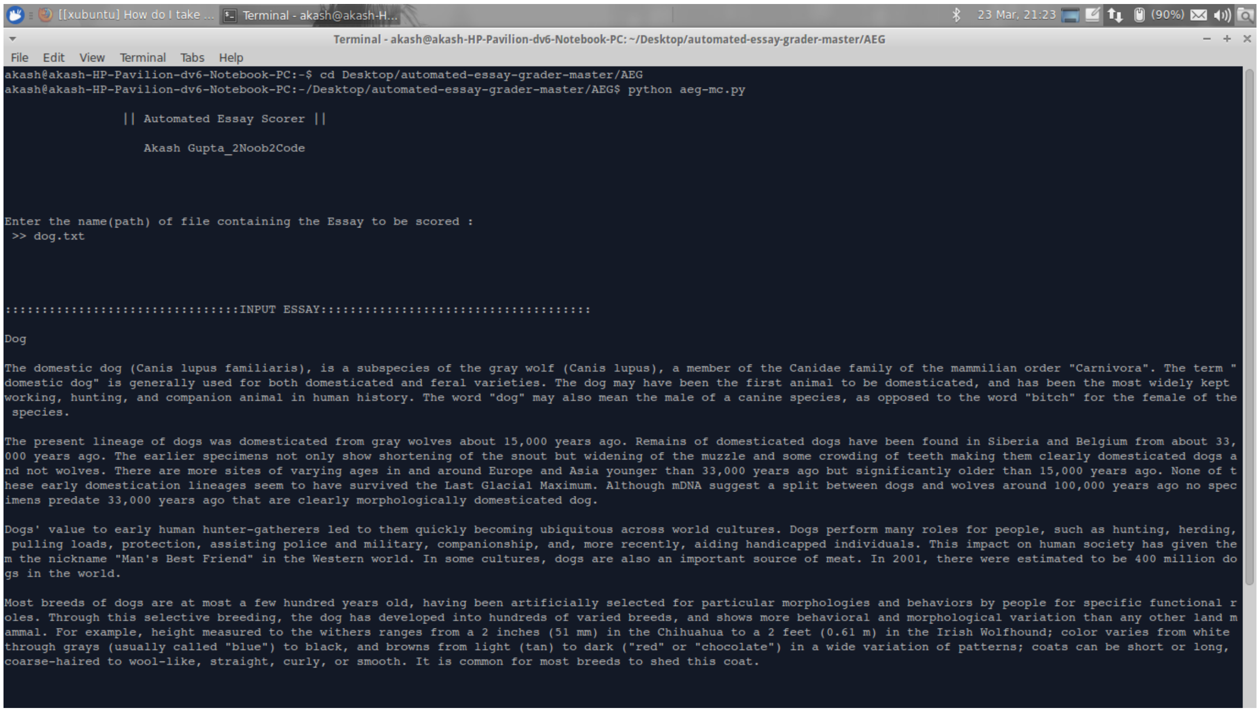


Scoring Model

- Input essay [.txt file]
- Scoring on the lines of GRE/GMAT essays.
- A report with scores for individual parameters is generated.
- An overall score is computed by taking weighted mean of individual scores. Scale: [0-5].
- Statistical Report is generated as HTML file.

Experiment Results

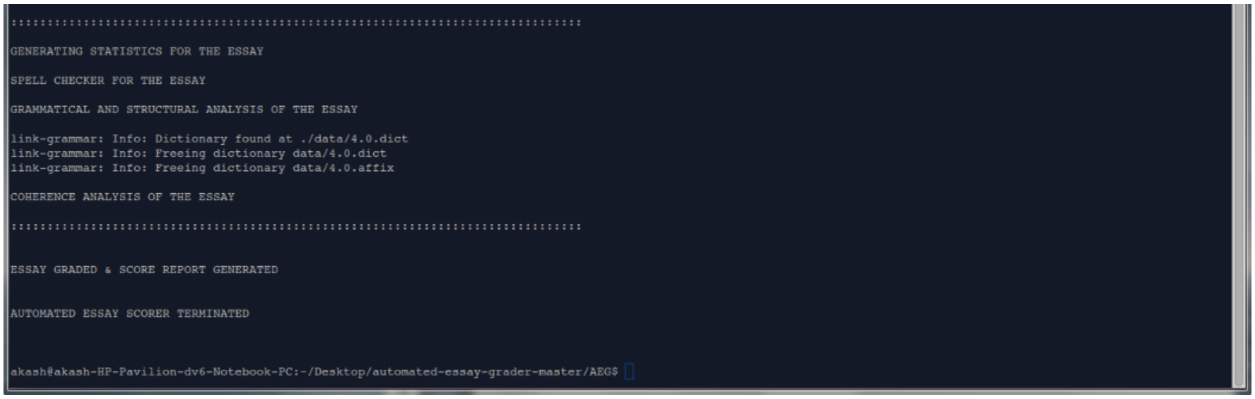
Test Essay



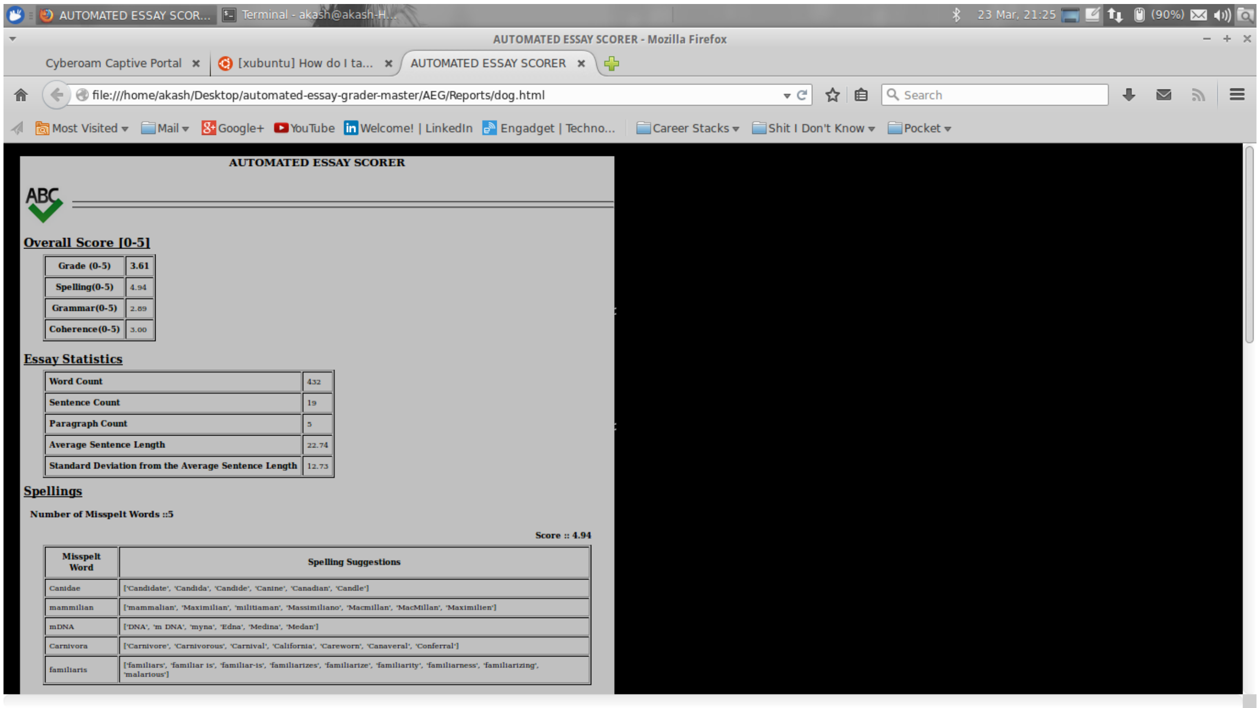
[[xubuntu] How do I take ...
Terminal - akash@akash-H...
23 Mar, 21:23
(90%)
Terminal - akash@akash-HP-Pavilion-dv6-Notebook-PC: ~/Desktop/automated-essay-grader-master/AEG
File
Edit
View
Terminal
Tabs
Help

```
akash@akash-HP-Pavilion-dv6-Notebook-PC:-$ cd Desktop/automated-essay-grader-master/AEG
akash@akash-HP-Pavilion-dv6-Notebook-PC:-/Desktop/automated-essay-grader-master/AEG$ python aeg-mc.py

                || Automated Essay Scorer ||

                   Akash Gupta_2Noob2Code



Enter the name(path) of file containing the Essay to be scored :
 >> dog.txt



::::::::::::::::::::::::::::::::INPUT ESSAY:::::::::::::::::::::::::::::::::::::

Dog

The domestic dog (Canis lupus familiaris), is a subspecies of the gray wolf (Canis lupus), a member of the Canidae family of the mammilian order "Carnivora". The term "
domestic dog" is generally used for both domesticated and feral varieties. The dog may have been the first animal to be domesticated, and has been the most widely kept
working, hunting, and companion animal in human history. The word "dog" may also mean the male of a canine species, as opposed to the word "bitch" for the female of the
 species.

The present lineage of dogs was domesticated from gray wolves about 15,000 years ago. Remains of domesticated dogs have been found in Siberia and Belgium from about 33,
000 years ago. The earlier specimens not only show shortening of the snout but widening of the muzzle and some crowding of teeth making them clearly domesticated dogs a
nd not wolves. There are more sites of varying ages in and around Europe and Asia younger than 33,000 years ago but significantly older than 15,000 years ago. None of t
hese early domestication lineages seem to have survived the Last Glacial Maximum. Although mDNA suggest a split between dogs and wolves around 100,000 years ago no spec
imens predate 33,000 years ago that are clearly morphologically domesticated dog.

Dogs' value to early human hunter-gatherers led to them quickly becoming ubiquitous across world cultures. Dogs perform many roles for people, such as hunting, herding,
 pulling loads, protection, assisting police and military, companionship, and, more recently, aiding handicapped individuals. This impact on human society has given the
m the nickname "Man's Best Friend" in the Western world. In some cultures, dogs are also an important source of meat. In 2001, there were estimated to be 400 million do
gs in the world.

Most breeds of dogs are at most a few hundred years old, having been artificially selected for particular morphologies and behaviors by people for specific functional r
oles. Through this selective breeding, the dog has developed into hundreds of varied breeds, and shows more behavioral and morphological variation than any other land m
ammal. For example, height measured to the withers ranges from a 2 inches (51 mm) in the Chihuahua to a 2 feet (0.61 m) in the Irish Wolfhound; color varies from white
through grays (usually called "blue") to black, and browns from light (tan) to dark ("red" or "chocolate") in a wide variation of patterns; coats can be short or long,
coarse-haired to wool-like, straight, curly, or smooth. It is common for most breeds to shed this coat.
```

Results Obtained



```
:::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::

GENERATING STATISTICS FOR THE ESSAY

SPELL CHECKER FOR THE ESSAY

GRAMMATICAL AND STRUCTURAL ANALYSIS OF THE ESSAY

link-grammar: Info: Dictionary found at ./data/4.0.dict
link-grammar: Info: Freeing dictionary data/4.0.dict
link-grammar: Info: Freeing dictionary data/4.0.affix

COHERENCE ANALYSIS OF THE ESSAY

:::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::


ESSAY GRADED & SCORE REPORT GENERATED


AUTOMATED ESSAY SCORER TERMINATED



akash@akash-HP-Pavilion-dv6-Notebook-PC:-/Desktop/automated-essay-grader-master/AEG$
```



AUTOMATED ESSAY SCORER

Overall Score [0-5]

Grade (0-5)	3.61
Spelling(0-5)	4.94
Grammar(0-5)	2.89
Coherence(0-5)	3.00

Essay Statistics

Word Count	432
Sentence Count	19
Paragraph Count	5
Average Sentence Length	22.74
Standard Deviation from the Average Sentence Length	12.73

Spellings

Number of Misspelt Words ::5

Score :: 4.94

Misspelt Word	Spelling Suggestions
Canidae	['Candidate', 'Candida', 'Candide', 'Canine', 'Canadian', 'Candle']
mammilian	['mammalian', 'Maximilian', 'militiaman', 'Massimiliano', 'Macmillan', 'MacMillan', 'Maximilien']
mDNA	['DNA', 'm DNA', 'myna', 'Edna', 'Medina', 'Medan']
Carnivora	['Carnivore', 'Carnivorous', 'Carnival', 'California', 'Careworn', 'Canaveral', 'Conferral']
familiaris	['familiars', 'familiar is', 'familiar-is', 'familiarizes', 'familiarize', 'familiarity', 'familiarness', 'familiarizing', 'malarious']

EVALUATION CRITERIA & PERFORMANCE MEASUREMENT

Most important characteristics used to measure the effectiveness of an AES system are:-

- Accuracy
- Defensibility
- Coach-ability and
- Cost-Effectiveness

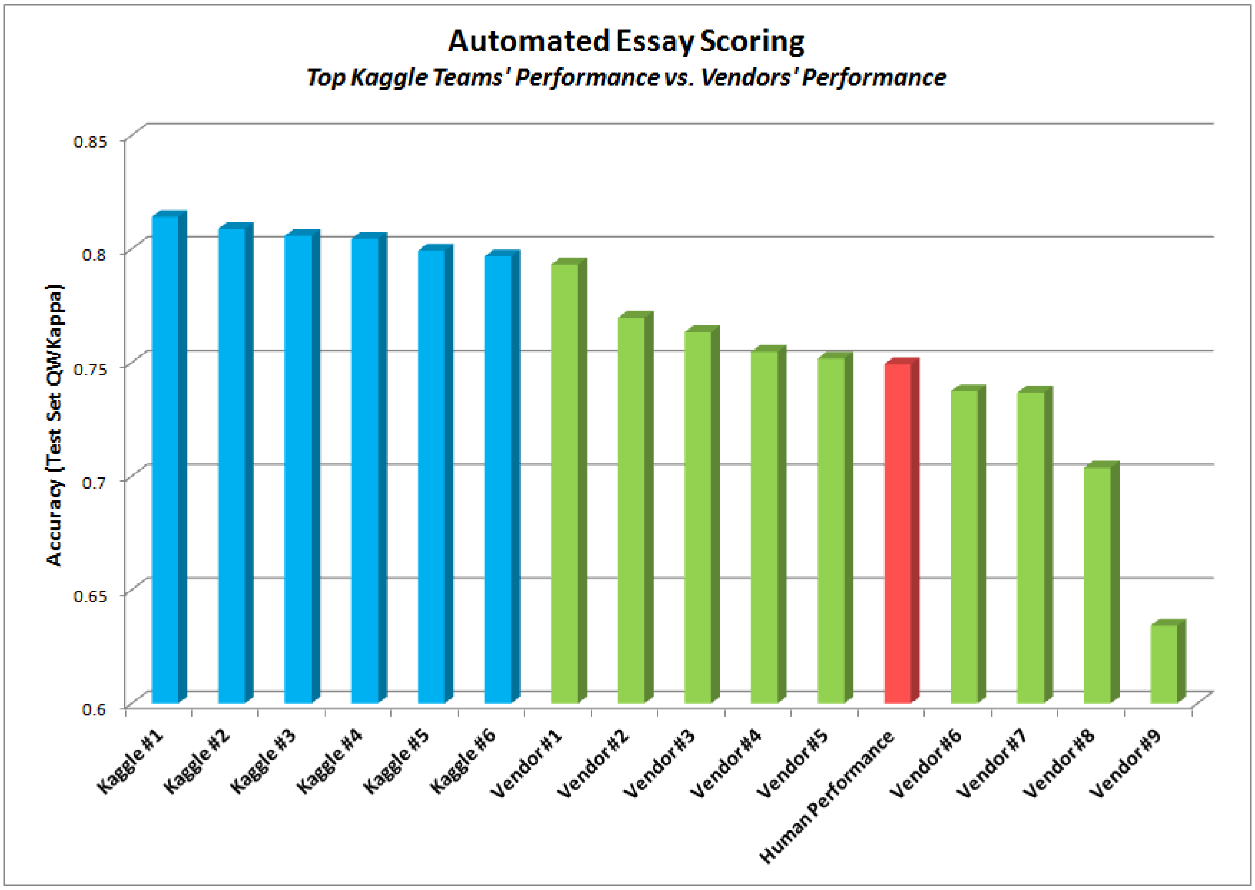
Elaborating Further

- An AES must be accurate when compared to a human grader.
- An assigned grade must be defensible through explanation of grading criteria and comparison to a rubric.
- Coach-ability is not a desired characteristic in AES. A coachable AES is one that is "based on simple, surface based methods that ignore content, students could train themselves to circumvent the system and so obtain higher grades than they deserve".
- One of the criticisms of PEG was that it was coachable simply by writing long essays filled with facts.
- Cost-effectiveness is self-explanatory and is mostly measured through the savings of time and labor for the instructor.
- The vast majority of AES studies have measured AES performance through correlation with a human or multiple human graders. The other two methods used are multiple regression correlation and accuracy of results (error rate).

AES Systems Evaluation

The algorithms of the vendors and the competition participants were evaluated on the same data sets. Competitors and vendors were ranked by quadratic weighted kappa (QWK), which measures how closely the predicted scores from the models matched up with human scores (higher kappa is better).

Summary of the performance with this excellent chart from Christopher Hefele



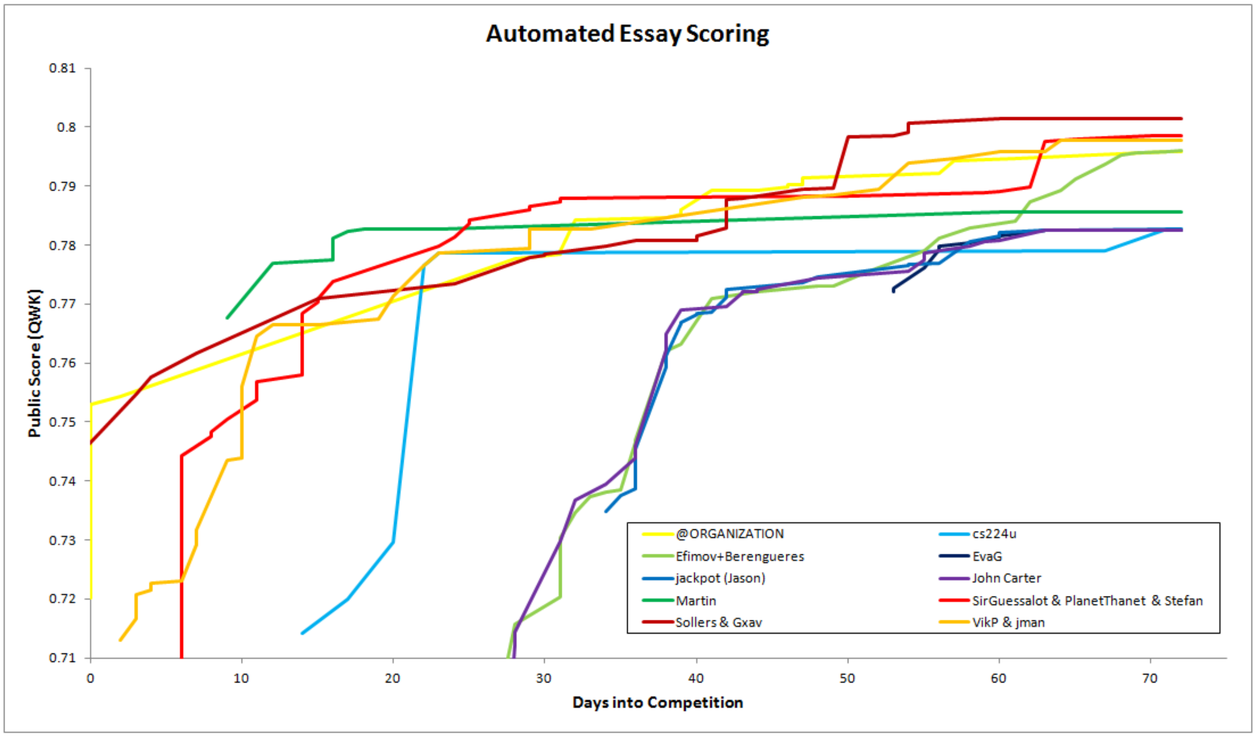
Inference

It can be seen that the top six competition participants did better in terms of accuracy than all of the vendors. As discussed before what accuracy is thought of as the sole metric for AES success, so take this with a bit of salt. The main reason I show this is to illustrate that open competition, with a fair target, can lead to very unexpected results and breakthroughs.

Even the open source solution from CMU that was included in the competition scored a QWK of .7538, good for only 19th place on the final leaderboard, which indicates that it is less about open source than about open information, access, and competition (observant readers may notice that the guy who made the CMU tool is the same guy who called my code "first year graduate student level" and disparaged the edX tool.

As we can see that the best results come about when fresh ideas can be combined with existing knowledge and expertise.

The second chart - also from Christopher Hefele



Inference

Each line is how one of the top competitors performed on the public leaderboard (essentially us testing our algorithms before the final evaluation). Looking at the "VikP & jman" line, brings back some memories of frantic coding and thinking up crazy solutions to increase accuracy. It can be seen how performance changes over time, as algorithms got more and more accurate. But only up to a certain point. The data that was worked with in the competition to train the algorithms was limited. What is seen here is everyone converging on a maximum theoretical accuracy. After this point, there is not much more that can be achieved.

AES ADVANTAGES

- One of the biggest advantages of an AES is that a student receives quick feedback. An AES provides feedback to a student almost instantaneously with submission in most cases. Especially in large classes the speed of feedback is far superior to that of human graders. This is one reason why so many computer science instructors have created homegrown AES systems to grade introductory computer programming courses.
- Grading of programming is tedious and can be handled much more quickly and efficiently by an AES. Quick feedback is vital and often listed as a best practice for teaching and often recognized as being essential in student motivation to grasp and learn.
- In distance education where interaction with an instructor may be minimal or sporadic the use of an AES can not only score essays but tutor the student as well.
- Besides speed of feedback an AES system can also provide consistency in grading, cost-savings for the educational institution, time-savings for the instructor, and reduced error in scoring. Time and date of submission is recorded automatically, and since computers can only be objective no personal bias in grading is possible.
- Another AES benefit is its ability to act as a plagiarism detector. In one study an AES discovered an extremely high rate of plagiarism finding that 98 out of 712 assignments were copied from another student's work.
- Sometimes students become discouraged or belligerent when subjected to criticism even when it is constructive. Students may be more open to such criticism if it is delivered impersonally through an AES.

  An AES also encourages students to revise their work before submitting it for final grading.
- In one study students who used an AES wrote three times as many words on an essay as students who did not use an AES.
- Finally, many Net Generation students enjoy the gamification aspect of utilizing an AES. Students treat the AES system as a video game "in which doing well involves redrafting work to get a higher score".

FUTURE WORK

There is scope for further work in automated essay graders. There could be simple additions such as spell checkers etc. added to this grader.

- **Android, iOS and Windows Application [Students Use Only]**
- **Suggestion Box**

There were many assumptions that were used, worked with in evaluation of topic and text coherence e.g. limiting antecedents to just previous 2 sentences etc. Such assumptions can be resolved to improve the system. The text and topic coherence is quite rigidly bound to the topic in question in this implementation, the tool can be modified to be more generic.

CONCLUSION

Based on the implementation and the experimental results, it is inferred that, some essays that do not use punctuations effectively, make sentence detection harder and this sometimes leads to incorrect parse trees and parts of speech tagging. These can lead to incorrect grading. Some methods employed to evaluate word order, topic coherence such as creating a list of syntax rules to identify incorrect word order and generating a list of words to evaluate topic coherence make the algorithm centric to the training set. So, unless the test set is similar to the training set, there could be inconsistencies in the grading. Many grading methods used have many exceptional rules. Though, it has been tried to incorporate as many exceptional rules as possible there are many rules that are not taken into account like use of singular verbs with sums of money or period of time etc. Specific senses of hypernyms and meronyms provided by Wordnet to grade essays for the provided topic are used. It was experimented with checking all possible senses of the hypernyms and meronyms but that had performance implications.

Lessons from the software

- **Don't forget the goal**
  The goal here isn't to impress people with fancy technology or tell teachers how they should teach. The goal is to maximize student learning and limited teacher resources (time) in a way that is flexible, and under the control of the subject expert (teacher).
- **Scale**
  In a MOOC setting, AES makes sense. It is hard/impossible for a teacher to score thousands of students each week, and writing is a critical component of many courses. But scale can also play a big part in the classroom. Can a teacher grade 10 drafts per student per week? Maybe it makes sense to allow students to score their "intermediate revisions" with AES, improve their writing, and give their key drafts and finished products to a teacher for more detailed feedback.
- **AES is (mostly) best used in combination with other ideas/technologies/concepts**
  In the same vein as the point above, AES is useful in some domains, and can give students accurate scores and rubric feedback. However, AES cannot give detailed feedback like an instructor or peer can. You should evaluate your options and see how you can best use AES. Maybe it works for certain questions. Maybe you can grade tests with AES. Maybe it is good for grading first drafts. Maybe you should combine it with small group discussions or peer scoring. If the tools are built properly, it will be possible to evaluate all these options, and figure out which one, if any, has the most value for students.
- **Put the power in the hands of teachers**
  AES is useless when the power is in the hands of researchers and programmers (although it does make us feel important). The real people who need to shape and implement these technologies are teachers and students, and they need the power to define how the AES looks and works. Maybe a teacher doesn't need to define what features the AES uses, but being able to turn off the AES for certain students might be useful.

- **Give people the information that they need**
  AES is a semi-shadow world to a lot of people, and that may be partially by design. The less we tell people about how things are done, the more valuable and important we become. I am always leery of researchers who take the "non-cooperative expert" stance.
- **Have the algorithm tell people how it is working**
  Algorithms can estimate their own error rates (how many papers they grade correctly vs incorrectly). Giving teachers and students as much information as possible within an AES system is key. If we don't know how something is working, how can we tell if it is doing what we want?
- **It's not all about the algorithm**
  Algorithms are fun and exciting, but learning tools are only useful if they help students, well, learn. The most important thing in this is usability. Can a student quickly digest and use their feedback? Can a teacher quickly create a new problem and deliver it to students? It is actually pretty easy to implement an algorithm. It is hard to put the things in place around it to allow students to succeed. I would even venture to say that once you get a certain level of accuracy in your algorithm, improving usability should become the primary goal.
- **Make everything usable**
  Is the product designed for teachers or for "expert" researchers? Does a user have to manually read a ton of essays into a command line or GUI program (think Microsoft office)? How do students get papers into the system? Everything should be a web-based tool, and students can write papers and receive feedback entirely through a web interface. Teachers can create problems that use AES in a few clicks, and can grade student papers through a web interface. This isn't the end all be all of ways to approach this, but more user friendly is better.
- **Grading isn't all about essays**
  Can we grade uploaded videos? How about pictures or songs? This can be done with peer and teacher grading, but AES needs to be extended to work with alternative media as technology advances.

REFERENCES

Research Paper:

- Stephen P. Balfour, "Automated Essay Scoring and Calibrated Peer Review", Ph.D. Texas University, 2013.
- Pavan Reddy and Girish Jambagi,"Automated Essay Grading – Natural Language Processing", University of Illinois at Chicago, USA, 2013.
- Yigal Attali and Jill Burstein,"Automated essay scoring with e-rater R v. 2. The Journal of Technology", Learning and Assessment, 2012.

Technical Report/Article:

- Automated Essay Scoring Research – Where it's been and where it's going.

Web References:

- [OpenNLP parser] http://opennlp.apache.org/
- [Stanford Parser] http://nlp.stanford.edu/software/lex-parser.shtml/
- [Wordnet] http://wordnet.princeton.edu/