Evaluation of essays using incremental training for Maximizing Human-Machine agreement

M.Tech Project Report

Submitted by

Dhananjay Ambekar
Roll No : 123059008

under the guidance of

Prof. D. B. Phatak

Department of Computer Science and Engineering
Indian Institute of Technology, Bombay

October, 2014
Contents

1 Introduction to Topic ................................................................. 1
  1.1 Introduction ................................................................. 1
  1.2 Types of assessments in MOOCs ............................................ 1
    1.2.1 Self Assessment ...................................................... 1
    1.2.2 Peer Assessment .................................................... 2
    1.2.3 AI Grading (or ML Grading) ...................................... 2

2 Automated Essay Scoring (Current Research in AES) .......... 3
  2.1 Tools .............................................................................. 3
    2.1.1 Project Essay Grade (PEG) .......................................... 3
    2.1.2 Intelligent Essay Assessor (IEA) ................................... 4
    2.1.3 Educational Testing Service (ETS-I) ............................ 4
    2.1.4 Bayesian Essay Test Scoring system (BETSY) ................. 4
  2.2 Estimating accuracy against Human grading .................... 5
    2.2.1 Kappa Coefficient (Cohen’s kappa) ............................... 5
    2.2.2 Precision ............................................................... 5
  2.3 Conclusion ...................................................................... 5

3 Process of evaluation .............................................................. 6
  3.1 Training data (Rubrics) ..................................................... 6
  3.2 Feature extraction ........................................................... 8
    3.2.1 Prepossessing of input data ......................................... 8
    3.2.2 n-grams .................................................................. 8
    3.2.3 Synthetic training examples ........................................ 9
    3.2.4 Feature vector .......................................................... 9
  3.3 Training Algorithms .......................................................... 9
    3.3.1 Bayesian Independence Classifiers ............................... 9
    3.3.2 K-nearest Neighbour ................................................ 10
    3.3.3 Gradient Boosting ..................................................... 10
  3.4 Discussion/Conclusion ....................................................... 11

4 Subsequent Work/Plan ............................................................. 12
  4.1 Experiments .................................................................... 12
    4.1.1 Objective ............................................................... 12
    4.1.2 Setup ..................................................................... 12
    4.1.3 Observations ............................................................ 13
5 Problem Formulation

5.1 Problem Statement ............................................................. 14
5.2 Future Work ................................................................. 15
List of Figures

1.1 Process of automatic Evaluation of essays ........................................... 2
3.1 POS Tree representation ........................................................................ 8
3.2 Features in Essay .................................................................................. 10
List of Tables

2.1 AES tools performances ........................................... 4
3.1 POS tags (sample) ..................................................... 8
4.1 Inter grader comparison ............................................ 13
Abstract

Massive Open On-line Courses (MOOCs) have large number of submissions which make automatic evaluation an appropriate choice for assessments. Project aims to Maximize Human-Machine agreement for automatic evaluation of textual summaries or essays in MOOCs leading to near-human accuracy. It also involves study of techniques for automatic evaluation of textual summaries using Machine learning, Natural Language processing & Making use of varied data of MOOCs for incremental training purposes. Propose a method for improving training & evaluation process for greater Human-Machine agreement.
Chapter 1

Introduction to Topic

1.1 Introduction

This research survey was performed with primary focus on following topics/fields:

- Massive open on-line courses (MOOCs)
- Grading essays in MOOCs
- Automated essay grading
- Correlation between Human-Machine grading

Chapter attempts to give detailed description of understandings from the survey in sections following.

MOOCs or Massive Open On-line Courses based on on-line Educational Resources are proving to be one of the most effective ways to offer access to quality education, especially for those residing in far or disadvantaged areas[1]. By offering a course it becomes obligatory to provide grading and evaluation services. Previous generation moocs had specific type of question type limited to numeric, true/false, and multiple choice answers. later came more advanced types such as inputting circuits, mathematical equations or more courses started question types with textual summary or subjective essay answers[2]. With these new generation of questions in MOOCs, came the need for new grading and evaluation strategies.

1.2 Types of assessments in MOOCs

Broadly there are three type of assessment methods[3] in MOOCs which are used to assess and grade student submissions followed by providing score. Following is an introduction to each of this methods[2].

1.2.1 Self Assessment

Self-assessment, is a process whereby students grade assignments or tests based on a teacher’s guidelines. Rubrics are often used to perform self assessment, they help in keeping the process formulated.
1.2.2 Peer Assessment

In peer assessment, a student grades submissions of other students after his completion. A process for calibration may or may not be applied to normalize the given grades.

1.2.3 AI Grading (or ML Grading)

As MOOCs have large number of submission, it is almost impossible for an instructor to manually grade students submissions\[4\]. so a method alternative to earlier mentioned methods was proposed known as AI grading.

AI grading uses machine learning techniques with assistance of language processing tools to grade students submissions particularly of type essay, subjective text. The idea behind this technique is a model is trained with first few (training purpose) submissions being assessed/graded by instructor and then for remaining large portion of the data the AI grading is performed.

Figure 1.1: Process of automatic Evaluation of essays

This literature survey focuses on AI grading type and its peripherals. A detailed study of grading methods, techniques of evaluation of essays and current research in this area are also key objectives for this survey.
Chapter 2

Automated Essay Scoring
(Current Research in AES)

In this chapter summary of study of different techniques is presented. There are number of tools available for essay grading which use different techniques.

Automated essay scoring (AES) is the use of specialized computer programs to assign grades to essays. Essays are considered as one of the best sources by instructors to understand student’s learning of the subject. It’s subjective nature gives human graders freedom of assessing each essay differently, hence any two graders might not grade same set of essays with similar grades. Researchers feel that this gives room for unfairness in the process and consistency is not maintained if multiple graders are grading set of same type of essays. And so automated grading could prove well in maintaining consistency in grading and if trained well, these models can actually do grading as accurate as Human graders with High Correlation between Human and Machine.

Following are some of the tools currently used for essay grading automatically. And in the later section a brief description of measures of correlation is presented.

2.1 Tools

2.1.1 Project Essay Grade (PEG)

PEG is one of the earliest implementation of automated techniques. In PEG essay is graded on the basis of writing quality, taking no account of lexical content hence no Natural Language processing techniques are used (Recent versions may include such features). The process of evaluation is based on concept of proxes. Proxes are variables of interest within essay such as:

1. Count of prepositions
2. Complexity of special words

It uses Regression on training data (graded by instructors) and then based on this, grades number of essays.

Performance: Correlation of 0.87 with human graders.
2.1.2 Intelligent Essay Assessor (IEA)

IEA is another tool which uses Statistical techniques without any language processing. This tool also does not make any use of ordering of words in essay. It represents the textual data using Latent Semantic Analysis (LSA) of documents and their word content in a large two-dimensional matrix. Each word being analysed represents a row of in the matrix, while each column represents the sentences, other subdivisions of the essay in which the word is present. Cells of the matrix contain frequencies of those words. Relationship between matrix and content are established and scoring is performed. **Performance:** Up to 85%-91% agreement with human graders

2.1.3 Educational Testing Service (ETS-I)

ETS was developed by Burstein and Kaplan of the Educational Testing Services. It uses combination of statistical and language processing techniques to extract linguistic features from the essays. The process includes analysis of:

- Discourse structure
- Syntactic structure
- Vocabulary in domain of Essay

Training data is processed by Natural Language Processing tool, a lexicon is created which is used to create grammar rules. this tool requires manual processing of training data which consumes time and cost but graders still say that ETS is worth auto-grading for large number of essay. **Performance:** ETS I has scored Up to 80% accuracy with human graders

2.1.4 Bayesian Essay Test Scoring system (BETSY)

A windows- based BETSY system was mainly used to group essays on scale of few certain predefined points (e.g. extensive, essential, partial, unsatisfactory)[6]. Bernoulli Models (BM) are used to classify text. BETSY can be downloaded and used free of cost. **Performance:** BETSY has scored Up to 80% accuracy with human graders

<table>
<thead>
<tr>
<th>System</th>
<th>Technique</th>
<th>Performance</th>
<th>Accuracy</th>
<th>Correlation</th>
<th>Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>PEG</td>
<td>Statistical</td>
<td></td>
<td></td>
<td>0.87</td>
<td></td>
</tr>
<tr>
<td>IEA</td>
<td>Algebra</td>
<td></td>
<td></td>
<td>85-91</td>
<td></td>
</tr>
<tr>
<td>ETS-I</td>
<td>NLP</td>
<td>93-96</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BETSY</td>
<td>Bayesian/NLP</td>
<td>80</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Automark</td>
<td>NLP/ML</td>
<td>93-96</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C-Rater</td>
<td>NLP/ML</td>
<td>80</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2.1: AES tools performances

Table shows summary of performance of various tools against human grading. Automark and C-Rater are tools used to rank free text responses on specific content. Another important observation is Researchers have been using **Accuracy, Correlation**
and Agreement for estimating relation between human and machine grading. For all
the experiments to be conducted in this project, usage of same measures is planned.

2.2 Estimating accuracy against Human grading

Scoring of assays can be typically done after classifying them into several categories (for
ex. extra-ordinary, good, average, below average). Categorization can be automated using
NLP and machine learning techniques and after this step is done, a comparison of human
and machine grading performance is done by using statistical measures. following is
description of such measures which will be used for comparing performances of various
methods in this project.

2.2.1 Kappa Coefficient (Cohen’s kappa)

This coefficient is statistical measure to calculate agreement between raters for categorical
type of data. This measure was used by Jill Burstein in their experiments where in an
estimate of agreement was to be provided for machine evaluations and human graders for
text essays.

\[
k = \frac{P(a) - P(e)}{1 - P(e)}
\]

Pr(a) is the relative observed agreement among raters
Pr(e) is the hypothetical probability of chance agreement
If the raters are in complete agreement then k is 1.

2.2.2 Precision

This measure is useful here for comparison as we have only two cases of agreement i.e.
the human and machine grading agrees or It does not agree. In their study on essay
evaluation, reference have used precision as follows :

\[
P = \frac{\text{No. of cases of agreement}}{\text{No. of evaluations by Human} + \text{No. of evaluations by Machine}}
\]

2.3 Conclusion

For this project, earlier mentioned measures are sufficient to analyze the results. There
are other measures such as recall and F-measure used to estimate the agreement in text
classification tasks, which will be used if needed.

If implementation is concerned, this project uses edX as base for the experimentation
and techniques are described in later chapters.
Chapter 3

Process of evaluation

The process of evaluation starts with training set of essays which will be used to grade large number of submissions of same type of essays \[5\]. This data set is known as training set.

Evaluation program extracts features from each essay from training data. features are characteristic properties of data. In case of essay, example features can be Number of words, Specific set of words or ratio of uppercase to lowercase words. Many of the system use words and their frequency to classify essays in different classes/categories.

After feature extraction, an algorithm is trained with calculated parameters to score new essays. The score/grade is based on graded essays provided for training. These essays are provided with rubrics (This discussion holds for edX) for evaluation. feature extraction & various algorithms are described below. The feature extraction phase was studied with NLTK Natural language toolkit and EASE libraries for essay evaluation \[7\].

3.1 Training data (Rubrics)

In educational terminology, scoring rubric means "a standard of performance for a defined population". ref Rubrics are the way of communicating expectations from a problem/assignment which student will attempt. Instructors design rubrics around the problem statement so that while evaluating the same a unified criteria can be imposed which leads to fair evaluation as well as It enforces similar evaluation rules on all submissions made by student.

In MOOCs (in general and EdX in specific) evaluation rubrics are defined by Instructor at the time of creating a problem, and are consist of different criterion which will try to understand the submission made by a student and help evaluators to assess and further grade the submission. These rubrics then are used by peer evaluators or even student submitting the problem for self evaluation.

For essay evaluation rubrics such as ‘Idea’ and ‘Content’ can be given which may help users to grade essays with different perspectives. In actual situation when an instructor is manually grading essays he/she may look for more criterion such as number of words or grammatical correctness of essay etc. so while designing automatic evaluation for textual content one should consider all the features required with weight-age to them if needed. In automatic evaluation of essays or text it is expected and required that you provide ample training examples to train in accordance with each criterion.
<rubric>
  <prompt>Write a small essay on Cow. />
  <name>Ideas</name>
  <prompt>Determine if there is a unifying theme or main idea.</prompt>
  <criterion feedback="optional">  
    <!--Following options are defined for above criterion -->
    <option points="0">
      <name>Poor</name>
      <explanation>
        "Difficult for the reader to discern the main idea. Too brief."
      </explanation>
    </option>
    <option points="1">
      <name>Good</name>
      <explanation>
        "Presents a unifying theme & Stays focused on topic."
      </explanation>
    </option>
  </criterion>
</rubric>

XML file sample for Rubrics definition
3.2 Feature extraction

3.2.1 Prepossessing of input data

This step involves removal of non-ASCII characters, then processing punctuation such as period, comma any other marks with inserting a space character between any two of consecutive symbols. In this step essay is truncated with some limit of words, for example essays with more than 2500 words will be truncated to this limit.

3.2.2 n-grams

n-gram is a contiguous sequence of n items from a given sequence of text or speech. Before extracting n grams, a POS tagged tree is created. An example POS (parts of speech) tagged tree is given below.

Sentence:
“Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.”

Tokens:
[Pierre / Vinken / , / 61 / years / old/ , / will / join / the/ ...continued]

A tree is generated using predefined POS tags. List of all the Penn tree tags is available at [10]. Here are few for example:

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNP</td>
<td>noun, proper, singular</td>
</tr>
<tr>
<td>VB</td>
<td>verb, base form</td>
</tr>
<tr>
<td>JJ</td>
<td>adjective or numeral</td>
</tr>
</tbody>
</table>

Table 3.1: POS tags (sample)

![Figure 3.1: POS Tree representation](image)

The implementation of such trees is usually stored in dictionaries. These tagged parts of speech are used to generate n-grams (usually 2-3-4 grams)
3.2.3 Synthetic training examples

Many essay evaluation processes also take account of synonyms of tokens generated. so specified number of synonyms can be added to the feature list so as to accommodate all the different writing styles. Synonyms can be generated using any library such as NLTK(for example).

3.2.4 Feature vector

Following feature details are maintained for each essay in data set. these are specific examples used by NLTK EASE algorithm (Referred through edX docs). Spelling error features:

1. Spelling error features:
   - Average number of spelling errors per essay
   - Average number of spelling errors per character (across all essays)

2. Grammar error features:
   - Counts the number of n-grams that are ”good” (in the good n-gram list).
   - Identifies the positions of n-grams that are ”bad” (not in the good n-gram list).

3. Bag of words:
   - Count tokens that were found in the essay set.
   - Generates two bags (matrices): one for original text tokens and the other for tokens in the spell-checked and stemmed text.

3.3 Training Algorithms

There has been major work in text classification of textual data according to requirements. A document is classified in to multiple classes and categories based on some set of predefined characteristics. here, in essay grading these specific characteristics may refer to features discussed in earlier section. Following subsections contain discussion on methods used for training in grading.

3.3.1 Bayesian Independence Classifiers

In [11] Shermis had put a very precise description of Bayesian independent classifiers in his book. These probabilistic classifier estimate a probability of document being in particular category provided that certain features are present in the document. Bayes theorem is used to derive the probability of document present in each category defined.

Bayesian classifier defines the log probability of essay being in some category as follows:

\[
\log P(C|D) = \log P(C) + \sum_i \begin{cases} 
\log \left( \frac{P(A_i|C)}{P(A_i)} \right) & \text{if } D \text{ contains feature } A_i \\
\log \left( \frac{P(\bar{A}_i|C)}{P(\bar{A}_i)} \right) & \text{if } D \text{ does not contain feature } A_i 
\end{cases}
\]
P(C) is probability that a document belongs to class C, a prior probability. Class C can be a class/category of Good, Average or any class which is defined earlier while process of training.

$P(A_i|C)$ is the conditional probability the document has feature A-i and given that it belongs to class C.

$P(A_i)$ is a prior probability of any document containing feature $A_i$.

$P(\bar{A}_i|C)$ is the conditional probability that document does not feature $A_i$ given that class is C.

$P(\bar{A}_i)$ is the probability that document does not have feature $A_i$.

Based on Lewis’ model [12], this classifiers give 0 or 1 to documents for current feature under examination.

### 3.3.2 K-nearest Neighbour

Another classifier is discussed in [13] uses K-nearest neighbour classifier to classify essays and score them. In k-nearest-neighbor classification k essays in the training collection that are most similar to the test essay are selected. Score is given to essay based on weighted average of k selected essays in test set. Similarity was calculated by querying entire document against test set and ranking scores. k- are the top similar ranking essays chosen for further processing i.e. weighted averaging.

![Figure 3.2: Features in Essay](image)

### 3.3.3 Gradient Boosting

It is a technique for regression problems, which produces a prediction model in the form of an group of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function [14].
3.4 Discussion/Conclusion

This survey was conducted keeping in mind, the process of evaluation of essays using automated techniques. Experiments/Techniques studied in [15, 6] and many other content will be useful in implementing a process of evaluation which can add to improvised correlation between Human-Machine.
Chapter 4

Subsequent Work/Plan

4.1 Experiments

Simple Experiments were performed in order to understand and measure performance of MOOCs AI/Machine grading functionality. Details are as follows.

4.1.1 Objective

Objective was to understand the process of automated evaluation and its components which include Natural language processing, Machine learning and Evaluation Rubrics.

4.1.2 Setup

This evaluation was performed in edX demo course with 10 students. The number of students is less because the main focus of the experiment was to understand the process and measure the correlation in different scenarios.

- edX developer stack instance was used
- edX uses NLTK and EASE for auto-grading text data, which internally applies natural language processing and training algorithms for text classification (namely Gradient Boosting)
- Work-flow is generated from example based assessments and training is performed
- At this point a submission can be auto-graded using classifier trained in previous step

Sample Example based assessment is shown below. In this file (according to edX specification) two sample answers which are already graded by Criteria defined earlier by instructor of the course. It is recommended to provide optimal number of training samples on topic with possible grading-criteria combinations to achieve well trained classifier.
XML file sample for providing Training examples

4.1.3 Observations

Total 10 samples were examined. For experimental convenience samples were graded either good or poor. Table shows the observations obtained after performing grading task. The value 6 in cell infers that their are 6 samples to which algorithm and Human have given same Good score. So Now, If we calculate Cohen’s kappa coefficient-

<table>
<thead>
<tr>
<th>Total 10 samples</th>
<th>edX AI-ML</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Good</td>
</tr>
<tr>
<td>Human assessment</td>
<td>Good</td>
</tr>
<tr>
<td></td>
<td>Poor</td>
</tr>
</tbody>
</table>

Table 4.1: Inter grader comparison

\[ P(a:\text{agreement}): \frac{(6+2)}{10} \]

Human said Good to 6 applicants and Poor to 4 applicants.

Reader B said Good to 8 applicants and Poor to 2 applicants.

Therefore the probability that both of them would say Good randomly is 0.6 * 0.8 = 0.48 and they would say Poor is 0.4 * 0.2 = 0.08

\[ \text{Pr(e:random agreement)}: = 0.48 + 0.08 = 0.56. \]

The inter grader agreement can be estimated as

\[ k = \frac{(0.8-0.56)}{(1-0.56)} = 0.56 \]

13
Chapter 5

Problem Formulation

MOOCs have large number of students from all over the world. It makes the student responses varied & diverse in content as well as discourse of the response. It is fundamental requirement of any course that the assessment procedure should explore student learning in order to give a useful feedback to student helping him to improve the understanding of the subject.

5.1 Problem Statement

- It is observed (through experiments) that due to students being in different geographical/cultural areas, the features contained in responses differ adding diversity to them. (For an example - For Indian students, In an Essay written on Cow, People worship Cow. is an obvious feature.)

- These Features might not be captured while training by picking samples randomly because with respect to large number of students, frequency of these features might be very less and Evaluation Rubrics may or may not capture that content detail.

- While in training phase, another Iteration of training itself can be added to either learn or to add such features

This project aims to address the above mentioned issue by studying the existing Automated evaluation processes and If practicable, suggest an evaluation process for improved assessments. Objective will be to achieve the accuracy as near as human Evaluation.
5.2 Future Work

Following activities include subsequent work to be performed to get more understanding of the problem and techniques surrounded to it.

- Study the process of feature extraction/language processing from textual data in order to gain some knowledge about learning features
- Analyze the performance of different algorithms in automated essay scoring. (Currently edX-ORA2 uses Gradient Boosting for text classification)
- Formulate a process for experimentation of Iterative training and then comparing performances using standard measures.
References


