Computerized Evaluation of Essays

Jennifer J. Little

Term Paper

Expert Systems

Fall 2001

13 December 2001
Abstract

A “hot topic” in educational measurement is the area of computerized essay test evaluation. As essays have been incorporated into many of the standardized testing programs, and computerized testing is being implemented in more and more instances, there are many instances of where computers have evaluated the essay portion of standardized tests. In an environment where humans typically evaluate the submitted student writing, essay evaluations have been subject to variance in the level of agreement between readers and individual bias. The computer technology for essay evaluation has been made possible by the use of Naïve Bayes assumptions in text classification, that the probability of occurrence of each word in a document is independent of the probability of occurrence of other words in a document.

There are several computerized essay-scoring systems: (1) Project Essay Grade (PEG), (2) Latent Semantic Analysis (LSA), (3) Electronic Essay Rater (e-rater), (4) Bayesian Essay Test Scoring sYstem Betsy), and others. This paper discusses how these different scoring methods work, where they are currently implemented, and difficulties that have been experienced in their development and implementation.

The idea for classification essays by computer has been around for over thirty years, but the models have progressed from performing surface analysis only, content specific analysis, to a combination of both. They also utilize different statistical techniques, including linear regression models and factor analysis. Further improvements could be made when considering clusters and interactions of these features. These essay grading and classification systems could also be extended for use in the generation of writing diagnostics and instructional feedback.
Introduction

A “hot topic” in educational measurement is the area of computerized essay test evaluation. As essays have been incorporated into many of the standardized testing programs, and computerized testing is being implemented in more and more instances, there are many instances of where computers have evaluated the essay portion of standardized tests. This technology has been made possible by the use of Bayesian techniques in text classification. Bayesian methods have been applied to various areas of text classification, especially in the area of sorting (sorting the resumes of new job applicants into multiple job categories, sorting news articles, sorting spam and other non-relevant e-mail, etc.).

Typically, humans evaluate the submitted student writing. The shortcomings of human readers can include variance in the level of agreement between readers and individual bias. For an increased level of accuracy, multiple readers are required, which can be expensive. To become an essay evaluator, people must undergo strict training and performance monitoring. Page (1994) discusses that the accepted standard number of human evaluators for state and national programs is two, which still yields a very low reliability, a measure of agreement in the score between readers. Often times, a third reader is required to dissolve these discrepancies between the two readers. When human judges are used, it is also unclear that the two independent readers use the same evaluation criteria.

Classification of Essays by Computer

Text classification models are at the core of these computerized essay evaluation systems. There are two basic models that are commonly used, both belonging to the class of Naïve Bayes Models - the multivariate Bernoulli model and the multinomial event model. McCallum and Nigam
(1998) compare and contrast these models in great detail. The Naïve Bayes assumption is that probability of occurrence of each word in a document is independent of the probability of occurrence of other words in a document. The multivariate Bernoulli model represents a document by a vector with binary attributes, accounting for whether or not an item occurs in the document. This model has been shown to perform better at small vocabulary sizes. The multinomial event model represents a document by the set of word occurrences from the document. This model has been shown to perform better at larger vocabulary sizes. The first model has no representation for the frequency of the items, and the second model has no representation for the order in which the words appear.

Bayes Theorem can be stated as follows: 
\[ P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \]. In the context of essay exams, P(A) can be defined as the probability that a feature is included in the essay, and P(B) can be defined as the probability that the examinee belongs to a particular score category. The features of interest Bayesian essay scoring are typically features such as: number of words, average sentence length, frequency of specific words or phrases, and other essay characteristics such the order certain concepts, and the occurrence of specific noun-verb pairs. These features differ based on the essay scoring system used.

There are several computerized essay-scoring systems: (1) Project Essay Grade (PEG), (2) Latent Semantic Analysis (LSA), (3) Electronic Essay Rater (e-rater), (4) Bayesian Essay Test Scoring sYstem BETSY), and others. PEG grades essays based on the quality of writing. LSA evaluates relevant content. E-rater is a hybrid of these two systems, and uses both structural and content analysis. BETSY is also a hybrid system, incorporating both structural and content-based analysis.

The purpose of this research paper is to discuss how these different scoring methods work, where they are currently implemented, and difficulties that have been experienced in their
development and implementation. I will also discuss the direction of future research in regards to essay test scoring systems development and implementation.

**Project Essay Grade (PEG)**

Project Essay Grade was originally developed by Page in 1966, and was developed to evaluate the writing of high school students (in comparison to evaluations performed by a human grader).

PEG is based on the occurrence of features in computing the score of an essay. Essays are evaluated based on essential attributes, such as fluency, diction, grammar, etc. There is no direct measure for these attributes, so the PEG project developed substitutes (“proxes”) for these measures. For example, if the variable of interest is fluency, PEG measures the number of words. PEG measures diction by determining the variance in word length and measures the complexity of sentence structure by counting the number of prepositions, pronouns, etc. These proxes are used in predicting the grade assigned by human judges (in a method similar to multiple regression).

The regression models were developed as follows: a sample of student essays (some with scores, some without) was received and separated into two mutually exclusive groups - a design group and a validation group. The design group is used to construct the model and evaluate the parameters. The constructed model is then “tested” on the validation group. In the regression model, the proxes are the predictor variables and the score is the response variable. By using the proxes, a score could be determined and compared to the score from a human judge. PEG is evaluated by its reliability - proximity to the score from the human judge. For example, when the first paragraph of this paper is examined, the following results are obtained. See Figure 1 below.
The screen above shows that the selected paragraph receives quality scores (in the categories of style, organization, content, etc.) ranging from ~60-80. A weakness of PEG is that it does not judge semantic content in an essay and only measures the quality of writing.

**Latent Semantic Analysis (LSA)**

Latent Semantic Analysis is a machine-learning algorithm designed for indexing documents for information retrieval, introduced by Landauer and Foltz in 1997. LSA is a statistical technique for extracting and representing the similarity of meaning of words and passages by analyzing large bodies of text. LSA represents the meaning of a word as an average of the meaning of all the passages in which it appears; and the meaning of a passage is the average of the meaning of all the words contained in the passage.

Latent Semantic Analysis is used to evaluate the content of a submitted piece of writing by measuring sentence coherence, the similarity of passages, etc. LSA represents the text as a matrix in
which each row stands for a unique word and each column stands for a text passage. The cells of
the matrix contain the frequency in which the word in each row appears in the passages represented
by the columns of the matrix. Single value decomposition (SVD), which is a form of factor analysis,
is used to condense this large matrix, mentioned above, into fewer dimensions for analysis. It is in
this dimensionality reduction step that the inference of the meaning of words and passages is
possible. In this transformation, the cell values become weights that express the word’s importance
in a particular passage. Factor analysis is used to study the patterns of relationship among many
dependent variables, with the goal of discovering something about the nature of the independent
variables that affect them, even though those independent variables were not measured directly.
The independent variables are the meanings of words and passages. The dependent variables, which
can be directly measured, are the frequency of content-specific words and combinations of words.
Landauer, Foltz, and Laham (1997) provide a more detailed description of SVD and provide an
example of the matrix decomposition and dimension reduction steps.

LSA reports measures of near neighbors, matrix comparison, sentence comparison, one-to-
many comparison, and pairwise comparison. Near neighbors are terms (or passages) that are close
in similarity to a particular term (or passage). For example, if the term human computer interface is
considered, a partial list of the near neighbors includes microcomputer, omnipresence, digital, software, and
minicomputer (from the interactive website: http://LSA.colorado.edu/). Matrix comparison
determines the similarity of different passages of text. When the terms above (microcomputer,
onmpresence, digital, software, and apples) are evaluated, the least similarity is of course obtained for
apples in comparison to the rest of the terms. See the Figure 2 below for the results of the matrix
comparison.
The results show that the terms reported by the interactive LSA website are relatively similar to one another (excluding apples). The values in the matrix are cosine coefficients, which are measures of similarity between terms. The cosine of a term with itself is one, as can be seen in the diagonals of the matrix. Similarity measures ≤ 0.40 are considered insignificant; similarity measures ≥ 0.70 are considered significant. The cosine coefficient can be computed for terms, sentences, paragraphs, and entire essays.

Sentence comparison determines the similarity of sequential sentences. One-to-many and pairwise comparisons allow the comparison of multiple passages of text. The difference is that pairwise only examines text segment pairs, while one-to-many examines one body of text to many other bodies of text. For example, if the first paragraph of this paper is examined for pairwise comparison, the results can be seen in Figure 3 below. The results show significant similarity between the sentences.
Prior contextual knowledge must be specified before performing the analysis. The LSA website is limited to certain conceptual bodies of knowledge. For example, before performing the above analyses, specification of the subject area was necessary. I specified "general reading up to first year college". Other bodies of knowledge available include general reading up to 3rd, 6th, 9th and 12th grades, biology, psychology, etc. The LSA website can also be used for educational text selection (selecting text that will enhance learning, based on the text that was submitted) and cross language retrieval (query in one language and receive responses in another language).

Some of the limitations of LSA, similar to the multinomial event model, are that it makes no used of word order or logic. This means that it disregards how the order of words is important in constructing a sentence, and only captures how differences in word choices and passage meanings are related. The relationships inferred by LSA are only relations of similarity (context similarity) and may not be logically defined.

<table>
<thead>
<tr>
<th>Text 1</th>
<th>Text 2</th>
<th>Text 3</th>
<th>Text 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.71</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 3. LSA screen for Matrix Comparison Results** (http://LSA.colorado.edu/)
Electronic Essay Rater (e-rater)

E-rater is currently used by the Educational Testing Service (ETS) and was developed by Jill Burstein and other researchers at ETS. Since 1999, e-rater has been used to score as a second reader on the essays submitted for the GMAT exam.

The essays primarily evaluated by e-rater are for examinees taking the GMAT exam. There are six possible scores for this essay exam, and the essays are of two types: analysis of an argument and analysis of an issue. The e-rater program computes values (frequencies) for 60+ features and submits these feature vectors for stepwise linear regression analysis to compute optimal weights for each of the features. The resulting model of the features and relative weights is the scoring model. There is a separate scoring model creating for each essay topic.

E-rater is based on the occurrence of features in computing the score of an essay. There are three different classes of features: syntactic, rhetorical, and topical content features. The analysis of the syntactic structure of an essay can include the quantity and ratio of simple, compound and complex sentences, types of dependent clauses, use of auxiliary verbs, and other features. The ratios of these structure types per sentence and per essay are also measures of the syntactic structure.

Argument structure in an essay is measured through the use of rhetorical features. Since there is no particular text unit that corresponds to the stages of an argument, the essay reader must focus on cues to identify separate arguments and separate points within an argument. E-rater identifies and quantifies the essay’s use of cue words and other rhetorical features. For example, and words like “possibly” and “perhaps” are often used while developing an argument, words like “this” and “these” signify that the write has not changed topics, and phrases such as “in summary” or “in conclusion” signify the coming of a summary.
E-rater evaluates the topical content of an essay by comparing the patterns of words it contains to those patterns found in manually graded training examples. Two different measures are computed - one ("EssayContent") is based on the vocabulary usage in the essay as a whole, and another ("ArgContent") is based on the vocabulary usage for specific arguments found in the essay. The measurements of EssayContent are based on raw frequent measures according to the multinomial event model; ArgContent is based on weighted values of these frequencies, similar to the procedure used by Latent Semantic Analysis. The analyzed features are used to build a regression model that is used to predict the grade assigned by a human grader. The score reported by e-rater is based the most likely score category, based on the training set of essays.

E-rater currently uses linear statistical analysis techniques in deriving its scoring models. Burstein, et. al. (1998) report that non-linear techniques could be employed (and are currently being investigated) to improve the scoring accuracy. Powers, et. al. (2001) perform a study to challenge e-rater as an effective evaluator and the results show that it is possible to trick these computerized evaluation systems using methods that human evaluators would have picked up. This could be extended beyond e-rater to other computer evaluation systems. The study recommends that human readers be retained to keep e-rater from mis-scoring some essays. The study also recommends the continued (and perhaps widespread) use of online demos to allow external parties to utilize and evaluate the technology, such as the sites of both PEG and LSA.

**Bayesian Essay Test Scoring System (BETSY)**

BETSY was developed by Lawrence Rudner in the Department of Measurement, Statistics and Evaluation at the University of Maryland - College Park. BETSY is a program that classifies text based on trained material.
In BETSY, essays are classified as most likely being appropriate, partial, or inappropriate. The prior probabilities in this instance correspond to the probability of a particular classification and, without any information on the examinee's ability, BETSY assumes equal prior probabilities (an equal initial likelihood of each classification). After examining a feature (presence or absence of desired feature), the probabilities are updated using Bayes Theorem. The posterior probabilities obtained are used as the new prior probabilities and the next feature is examined. This process is repeated until some predetermined stopping criterion - either the Sequential Ratio Probability Test or a minimum posterior probability is reached. An efficient modification to this iteration procedure is rather than examining the entire list of features, only evaluate those features that maximize the expected change in the posterior probability (to provide maximum information).

When the first paragraph of this paper is examined, BETSY is able to classify the subject category in which the text most likely belongs.

![Figure 4. BETSY screen for Text Classification](http://ericae.net/betsy)
BETSY is limited to the four areas of assessment and education (TM), educational management (EA), early childhood (PS), and higher education (HE). In Figure 4 above, it can be seen that the subject that this paragraph most likely belongs to is assessment and education.

**Comparison of Essay Classification Systems**

Wresch (1993) also discusses three more essay evaluation projects and reveals that the technology is very similar to what is being done in PEG, LSA, etc. PEG and e-rater follow the multinomial Bayes model of text classification, taking measures of frequencies of attributes. BETSY follows the multi-variate Bernoulli Bayes model and measures the presence or absence of certain attributes. Rudner describes that the BETSY approach to evaluation of essays by computer has the advantage of combining the best features of PEG, LSA, and e-rater. BETSY can be employed on short essays and is simple to implement. It is similar to LSA and e-rater in that it the more specific terms are given higher weight.

The following table summarizes the essay classification systems described in this paper.

**Table 1. Comparison of Essay Evaluation Systems**

<table>
<thead>
<tr>
<th>Essay Classification System</th>
<th>Evaluation criteria</th>
<th>Scoring mechanism based on...</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two human graders</td>
<td>Range of features including surface and conceptual features</td>
<td>Low inter-judge agreement, Individual judges biased, Multiple judges costly, and Inconsistency in evaluation criteria</td>
<td></td>
</tr>
<tr>
<td>PEG</td>
<td>Surface features</td>
<td>Multiple regression</td>
<td>No measure of conceptual validity</td>
</tr>
<tr>
<td>LSA</td>
<td>Content features (patterns of occurrences)</td>
<td>Factor analysis</td>
<td>No measure of adherence to logical rules or rules of word order</td>
</tr>
</tbody>
</table>
After evaluating the first paragraph of this paper, the different methods produce very different sets of results. The first, PEG, gives only an analysis of the writing style. LSA provides information that is helpful in the structure and use of language in an essay (sentence similarity, choice of words, etc.) BETSY is useful is classifying the subject matter of text based on the use of terms. It can be seen that these tools are all very powerful, and can produce results in a fraction of the time in which a human could produce similar results.

These essay evaluation systems are currently being used in the field of educational measurement. The Intelligent Essay Assessor™ from Knowledge Analysis Technologies (KAT) is a computerized essay evaluation program that uses Latent Semantic Analysis, is used as a topic-specific writing evaluator. KAT was founded in 1998 by researchers at the University of Colorado at Boulder, including Landauer, Foltz and Laham. The Intelligent Essay Assessor™ (IEA) is a web-based service that provides an evaluation and advice on the conceptual concept of written work. Holt, Rinehart and Winston, Prentice Hall, and the U.S. Army use IEA as training material for students.

E-rater is being evaluated for use on the TOEFL exam offered by ETS. The training data for the TOEFL exam will consists of previously submitted TOEFL essays, and the scoring model is obtained in a similar manner. There is variability in the reliability of the reported scores across the different languages, but the results in the 1998 study by Burstein and Chodrow are promising.
All systems have online interactive evaluation systems available. Criterion, an evaluation system using the e-rater technology was unavailable during the time of this writing and a sample evaluation is not available.

**Future Direction of Research**

The idea for classification essays by computer has been around for over thirty years, but as time progresses, the models get more and more advanced. The models have progressed from surface analysis only, to content specific analysis, to a combination of both. They also utilize different statistical techniques, including linear regression models and factor analysis.

In each of the four essays classification systems described here, they all use training sets of essays that were graded by humans. The goal of these scoring models is not to take the human reader out of the loop. The scoring models developed use the judgments made by humans. Burstein et al. (1998) points out that e-rater bases its scoring model on a set of features. Further improvements could be made when considering clusters and interactions of these features. This could be extended to PEG as well, which also employs linear regression techniques in its scoring model. These essay grading and classification systems could also be used in the generation of writing diagnostics and instructional feedback. Landauer, Foltz, and Laham (1997) plan to extend LSA to include more representative samples of both text and spoken language, to extend the range of applications for LSA.
References


2. Bayesian Essay Test Scoring sYstem (BETSY) - [http://ericae.net/betsy](http://ericae.net/betsy)


   [http://LSA.colorado.edu/papers/plato/plato.annote.html](http://LSA.colorado.edu/papers/plato/plato.annote.html)


   [http://LSA.colorado.edu/papers/dp1.LSAintro.pdf](http://LSA.colorado.edu/papers/dp1.LSAintro.pdf)

12. Latent Semantic Analysis @ CU Boulder - [http://LSA.colorado.edu/](http://LSA.colorado.edu/)


   [http://citeseer.nj.nec.bom/mccallum98comparison.html](http://citeseer.nj.nec.bom/mccallum98comparison.html)


   [http://134.68.49.185/pegdemo/ref/J-Exp-E-d-94.htm](http://134.68.49.185/pegdemo/ref/J-Exp-E-d-94.htm)


   [http://134.68.49.185/pegdemo/ref/APA-Inv-95.htm](http://134.68.49.185/pegdemo/ref/APA-Inv-95.htm)


   [http://134.68.49.185/pegdemo/ref/Qual&Qual-b96.htm](http://134.68.49.185/pegdemo/ref/Qual&Qual-b96.htm)

18. Project Essay Grade Information Center - http://134.68.49.185/pegdemo/


   http://corax.cwrl.utexas.edu/cac/archives/v10/10_2_html/10_2_5_Wresch.html