Automated Essay Grading: An Evaluation of Four Conceptual Models

Robert Williams, School of Information Systems, Curtin University of Technology

Automated essay grading has been proposed for over thirty years. Only recently have practical implementations been constructed and tested. This paper describes the theoretical models for four implemented system described in the literature, and evaluates their strengths and weaknesses. All four models make use of comparisons with one or many model answer documents that have been previously assessed by human markers. One hybrid system that makes use of some linguistic features, combined with document characteristics, is shown to be a practical solution at present. Another system that makes use of primarily linguistics features is also shown to be effective. An implementation that ignores linguistic and document features, and operates on the "bag of words" approach, is then discussed. Finally an approach using text categorisation techniques is considered.

INTRODUCTION

Teaching staff around the world are faced with a perpetually recurring problem: how do they minimise the amount of time spent on the relatively monotonous tasks associated with grading their students' essays. With the advent of large student numbers, often counted in thousands in first year common core units, the grading load has become both time consuming and costly. A system that can automate the tasks is currently just a dream for most staff.

One of the earliest mentions of computer grading of essays in the literature was in an article by Page in which he described Project Essay Grade (PEG). (Page, 1966). Various aspects of students' essays, such as proportion of words on a common word list acting as a proxy for diction, and the proportion of prepositions acting as a proxy for sentence complexity, were measured. A multiple regression technique was then used to predict the human rater's score, based on these measures. We discuss the latest version of PEG later in this article.

Page made a distinction, which is still relevant today, between grading for content and grading for style.

> ""Content" refers loosely to what the essay says, and "style" refers to syntax and mechanics and diction and other aspects of the *way* it is said."
> (Page, 1966: 240)

This dichotomy gives us the basis for classifying the systems that have been developed : do they grade primarily for subject matter, or for linguistic style. And, do we measure proxies for these dimensions (rating simulation), or do we measure the actual dimensions (master analysis). Figure 1 shows the resulting four categories.

	I Content	II Style
A. Rating Simulation	I(A)	II(A)
B. Master Analysis	I(B)	II(B)

Figure 1: Possible Dimensions of Essay Grading
(Source: Page, 1966: 240)

There are inherent problems to be overcome if automated grading of text is to become a reality. Student essays addressing a particular topic can theoretically be expressed in possibly thousands of forms, using different combinations of words and sentences. Simply checking for the occurrence of some key words does not allow for a very accurate assessment of the work, nor does it allow for the richness and diversity that English allows for expression of ideas. Many words have thirty to forty entries in a thesaurus, and generally many of them are interchangeable in a particular and given context, so checking for the occurrence of key words is not an acceptable approach.

CONCEPTUAL MODELS FOR AUTOMATED ESSAY GRADING

The first model, Project Essay Grade (PEG), is one of the earliest and longest-lived implementations of automated essay grading. It has been developed by Page and colleagues, and primarily relies on linguistic features of the essay documents.

The second model, E_RATER, is one developed by Burstein et al at the Educational Testing Service (ETS) in the US, which has been implemented to the prototype stage for evaluation. This model uses a hybrid approach of combining linguistic features, derived by using Natural Language Processing (NLP) techniques, with other document structure features.

The third model, the LSA model, makes use of Latent Semantic Analysis (LSA) and the "bag of words" approach, and has been developed and evaluated by Landauer et al at the University of Colorado at Boulder. It ignores document linguistic and structure features.

The fourth model, which uses text categorisation techniques, identified in this paper as TCT, has been developed by Larkey at the University of Massachusetts. It uses a combination of modified key words and linguistic features.

PEG

Description

The idea behind PEG is to help reduce the enormous essay grading load in large educational testing programs, such as the SAT. When multiple graders are used, problems arise with consistency of grading. A larger number of judges are likely to produce a true rating for an essay.

A sample of the essays to be graded is selected and marked by a number of human judges. Various linguistic features of these essays are then measured. A multiple regression equation is then developed from these measures. This equation is then used, along with the appropriate measures from each student essay to be graded, to predict the average score that a human judge would assign.

PEG has its origins in work begun in the 1960's by Page and his colleagues (Page, 1966).

> "…we coined two explanatory terms: *Trins* were the in*trinsic* variables of interest – fluency, diction, grammar, punctuation, and many others. We had no direct measures of these, so began with substitutes: *Proxes* were ap*prox*imations, or possible correlates, of these trins. All the computer variables (the actual counts in the essays) were proxes. For example, the trin of fluency was correlated with the prox of the number of words."
> (Page, 1994, p 130)

The multiple regression techniques are then used to compute, from the proxes, an equation to predict a score for each student essay. In the research reported in Page (1994), the goal was to identify those variables which would prove effective in predicting human rater's scores. Various software products, including a grammar checker, a program to identify words and sentences, software dictionary, a part-of-speech tagger, and a parser were used to gather data about many proxes.

Evaluation

Details of most of the predictive variables are not given in Page's work. However, amongst the variables found useful in the equation were the fourth root of the number of words, sentence length, and a measure of punctuation. One set of results, based upon a regression equation with twenty-six variables, showed correlations between PEG predicted scores and human rater scores varying between 0.389 and 0.743.

E_RATER

Description

E-rater uses a combination of statistical and NLP techniques to extract linguistic features of the essays to be graded. As in all the conceptual models discussed in this paper, e-rater student essays are evaluated against a benchmark set of human graded essays. E-rater has modules that extract essay vocabulary content, discourse structure information and syntactic information. Multiple linear regression techniques are then used to predict a score for the essay, based upon the features extracted. For each new essay question, the system is run to extract characteristic features from human scored essay responses. Fifty seven features of the benchmark essays, based upon six score points in an ETS scoring guide for manual grading, are initially used to build the regression model. Using stepwise regression techniques, the significant predictor variables are determined. The values derived for these variables from the student essays are then substituted into the particular regression equation to obtain the predicted score.

One of the scoring guide criteria is essay syntactic variety. After parsing the essay with an NLP tool, the parse trees are analysed to determine clause or verb types that the essay writer used. Ratios are then calculated for each syntactic type on a per essay and per sentence basis.

Another scoring guide criteria relates to having well-developed arguments in the essay. Discourse analysis techniques are used to examine the essay for discourse units by

looking for surface cue words and non-lexical cues. These cues are then used to break the essay up into partitions based upon individual content arguments.

The system also compares the topical content of an essay with those of the reference texts by looking at word usage.

Evaluation

The system has been evaluated by Burstein et el (1998) and has found that it can achieve a level of agreement with human raters of between 87% and 94%, which is claimed to be comparable with that found amongst human raters. For one test essay question the following predictive feature variables were found to be significant.

1. Argument content score
2. Essay word frequency content score
3. Total argument development words/phrases
4. Total pronouns beginning arguments
5. Total complement clauses beginning arguments
6. Total summary words beginning arguments
7. Total detail words beginning arguments
8. Total rhetorical words developing arguments
9. Subjunctive modal verbs

The LSA model

Description

LSA represents documents and their word contents in a large two dimensional matrix semantic space. Using a matrix algebra technique known as Singular Value Decomposition (SVD), new relationships between words and documents are uncovered, and existing relationships are modified to more accurately represent their true significance.

The words and their contexts are represented by a matrix. Each word being considered for the analysis is represented as a row of a matrix, and the columns of the matrix represent the sentences, paragraphs, or other subdivisions of the contexts in which the words occur. The cells contain the frequencies of the words in each context.

The SVD is then applied to the matrix. SVD breaks the original matrix into three component matrices, that, when matrix multiplied, reproduce the original matrix. Using a reduced dimension of these three matrices in which the word-context associations can be represented, new relationships between words and contexts are induced when reconstructing a close approximation to the original matrix from the reduced dimension component SVD matrices. These new relationships are made manifest, whereas prior to the SVD, they were hidden or latent.

To grade an essay, a matrix for the essay document is built, and then transformed by the SVD technique to approximately reproduce the matrix using the reduced dimensional matrices built for the essay topic domain semantic space. The semantic space typically consists of human graded essays. Vectors are then computed from a student's essay data. The vectors for the essay document, and all the documents in the semantic space

are compared, and the mark for the graded essay with the lowest cosine value in relation to the essay to be graded is assigned.

The Intelligent Essay Assessor is a commercial implementation of the LSA approach. Later in this paper we discuss a trial of this system for first year university student essays.

Evaluation

Landauer, et al (1998), report that LSA has been tried with five scoring methods, each varying the manner in which student essays were compared with sample essays. Primarily this had to do with the way cosines between appropriate vectors were computed . For each method an LSA space was constructed based on domain specific material and the student essays. Foltz (1996) also reports that LSA grading performance is about as reliable as human graders. Landauer (1999) reports another test on GMAT essays where the percentages for adjacent agreement with human graders were between 85%-91%.

The Text Categorisation Technique (TCT)

Description

Larkey (1998) implemented an automated essay grading approach based on text categorisation techniques, text complexity features, and linear regression methods. The Information Retrieval literature discusses techniques for classifying documents as to their appropriateness of content for given document retrieval queries (van Rijsbergen, 1979). Larkey's approach

> ".. is to train binary classifiers to distinguish "good" from "bad" essays, and use the scores output by the classifiers to rank essays and assign grades to them." (Larkey, 1998, p90)

The technique firstly makes use of Bayesian independent classifiers (Maron, 1961) to assign probabilities to documents estimating the likelihood that they belong to a specified category of documents. The technique relies on an analysis of the occurrence of certain words in the documents. Secondly, a k-nearest neighbour technique is used to find the k essays closest to the student essay, where k is determined through training the system on a sample of human graded essays. The Inquery retrieval system (Callan et al, 1995) was used for this. Finally, eleven text complexity features are used, such as the number of characters in the document, the number of different words in the document, the fourth root of the number of words in the document (see also the discussion on PEG above), and the average sentence length.

Larkey conducted a number of regression trials, using different combinations of components. He also used a number of essay sets, including essays on social studies (soc), where content was the primary interest, and essays on general opinion (G1), where style was the main criteria for assessment. The results presented here are for these two essay sets only.

Evaluation

When all the criteria for assessment were used the proportion of essays graded exactly the same as human graders was 0.60 and scores adjacent (a score one grade on either side) was 1.00. For the general opinion essays the corresponding figures were 0.55 and 0.97. The system performed remarkably well.

DISCUSSION

We are now in a position to characterise these essay grading techniques according to the classification postulated by Page.

PEG focuses on simple linguistic features, focusing on style, and can be categorised as II(A). E_RATER focuses on linguistic features and document structures, and is thus performing a Master Analysis of style, and falls in the category II(B). The LSA model focuses on the semantics of the essay, but does so using a Rating Simulation, and therefore falls in the I(A) category. The TCT (soc) experiments focused on content in a rating simulation, while the TCT (G1) test focused on style in a rating simulation. Figure 2 summarises these models'classifications.

	I Content	II Style
A. Rating Simulation	LSA, TCT (soc)	PEG, TCT (G1)
B. Master Analysis		E_RATER

Figure 2: Essay Grading Models' Classifications

Figure 3 shows some of the reported performances, in comparison to human graders, of the various models.

Model	Measure	Values	Source
PEG	r	0.389-0743	Page, 1994
E_RATER	%	87-94	Burstein, et al, 1998
LSA	%	85-91	Landauer, 1999
TCT (soc)	r	0.69-0.78	Larkey, 1998
TCT (G1)	r	0.69-0.88	Larkey, 1998

Figure 3: Comparative performance of models

To find the amount of total variation explained by a correlation we take its square (PEG performance thus accounts for between 15% and 55% of the variations between PEG and human ratings, and TCT accounts for between 47% and 77%). It appears then, in terms of comparison with human markers, E_RATER is best, followed by LSA,TCT, and finally PEG.

TRIAL OF THE INTELLIGENT ESSAY ASSESSOR

A team of researchers in the School of Information Systems at Curtin University of Technology trialled the Intelligent Essay Assessor (IEA) during the first semester of 2001.

In March 2001, students enrolled in the unit Information Systems 100 (IS100) were notified that they could receive bonus marks of up to 5 per cent if they took part in the trial by submitting a two to three page essay based on a question taken from their textbook. These essays, in Microsoft Word format, were submitted via email to a special IS100 email address.

In May, 2001 an honours student in the School converted the essays to a standard format, and added student identification. Two hundred formatted essays were then chosen at random to be graded by three human markers. The average grade for these essays was 64.5 These essays, known as the training set, were sent to the USA to be processed by the IEA to form the semantic (knowledge) space, against which the other essays would be graded.

In June 2001 an additional 327 ungraded essays were sent by email to the USA for IEA grading, and the results were received back one week later. The system produced an average grade of 65.53. The accuracy of the IEA was very good, when compared to the human graded average. Figure 4 shows the distribution of grades produced by the IEA.

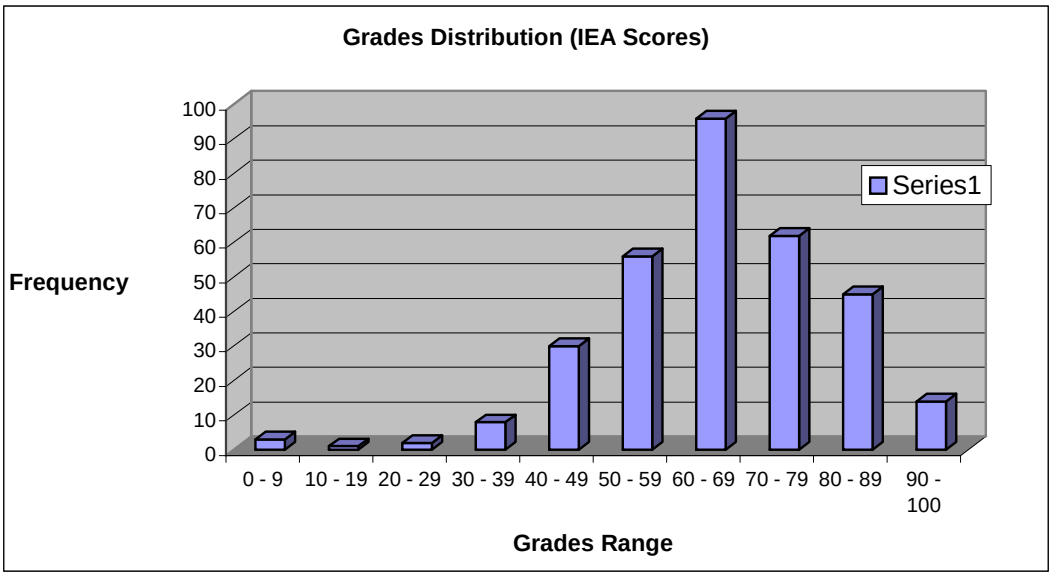

Figure 4: Distribution of grades produced by the IEA

The IEA also detected a number of cases of plagiarism that had escaped the attention of the human graders.

The cost per paper for the automated grading was about A$30, which is high when compared to human grading costs, but economies of scale apply to the IEA, and this cost could be reduced considerably (to about A$5) if more papers were graded against the same semantic space.

The researchers felt that the IEA is suitable when very large numbers of essays are to be graded (eg 2000), but the effort involved in formatting and human grading 200 essays for the semantic space, and the setup costs, are too great when only a few hundred essays are to be graded. The researchers were impressed by the ability of the IEA to detect plagiarism amongst the essays submitted by the students.

CONCLUSION

Automated essay grading is now ready to advance from the research laboratory to the real world educational environment. Current prototype systems, which grade for content, style, or both, can perform equally as well as human graders. Prototype systems only need minor enhancements to move into educational systems worldwide. However, they cannot at present deal with tabular and graphical content in essays. The

administrative resources needed to support these systems are quite substantial. Human judges are still needed to prepare model answers, or to grade samples of student essays before the computer systems complete the task Students also need suitable computer facilities to generate their essays in machine readable form. It is likely that commercial essay grading products will appear in the next ten years, and help ease the grading workload for teachers in a variety of disciplines

REFERENCES

Burstein, J., Kukich, K., Wolff, S., Lu, C., and Chodorow, M. (1998) Enriching Automated Essay Scoring Using Discourse Marking, *Proceedings of the Workshop on Discourse Relations and Discourse Markers, Annual Meeting of the Association of Computational Linguistics,* August, Montreal, Canada.

Callan, J. P., Croft, W. B. and Broglio, J. (1995) TREC and TIPSTER Experiments with INQUERY, *Information Processing and Management,* 327-343.

Foltz, P. W. (1996) Latent Semantic Analysis for Text-Based Research, *Behavior Research Methods, Instruments and Computers,* 28, 197-202.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998) An Introduction to Latent Semantic Analysis, *Discourse Processes,* 25, 259-284.

Landauer, T. K. (1999) Email communication with author, 8th June.

Larkey, L. S. (1998) Automatic Essay Grading Using Text Categorization Techniques, *Proceedings of the Twenty First Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Melbourne, Australia, 90-95.

Maron, M. E. (1961) Automatic Indexing: An experimental Inquiry, *Journal of the Association for Computing Machinery,* 8, 404-417.

Page, E. B. (1966) The Imminence of Grading Essays by Computer, *Phi Delta Kappan,* January, 238-243.

Page, E. B. (1994) Computer Grading of Student Prose, Using Modern Concepts and Software, *Journal of Experimental Education,* 62, 127-142.

Page, E.B. and Petersen, N.S. (1995) The Computer Moves into Essay Grading, *Phi Delta Kappan*, March, 561-565.

Perelman-Hall, D. (1992) A Natural Solution, *Byte*, 17, 2, February, 237-244.

Salton, G. (1988) *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley, Reading, Massachusetts.

van Rijsbergen, C. J. (1979) *Information Retrieval*, 2nd ed., Butterworths, London.