A Survey on the Assessment of Models towards Automated Free-Text Marking Engine

Srujana Inturi
Assistant Professor, Department of CSE, CBIT, Hyderabad

Abstract

Automated free text response grading has been proposed for more than thirty years. Just as of late have practical implementations been built and tried. This paper portrays the theoretical models for four executed system depicted in the writing, and assesses their strengths and weaknesses. Every one of the four models make utilization of correlations with one or numerous model answer documents that have been already evaluated by human markers. One cross breed system that makes utilization of some phonetic features, joined with document qualities, is appeared to be a practical arrangement at display. Another system that makes utilization of essentially linguistics features is additionally appeared to be successful. A usage that overlooks etymological and document features, and works on the "bag of words" approach, is then talked about. At last an approach utilizing text categorization techniques is considered.

Index Terms : Short Answer, Grading System, Question, Conceptual Models, Free-Text Marking Engine

I. INTRODUCTION

Showing staff far and wide are looked with an unendingly repeating issue: how would they limit the measure of time spent on the moderately dull undertakings related with grading their understudies' free text responses. With the coming of substantial understudy numbers, often included thousands in first year normal center units, the grading load has turned out to be both tedious and expensive. A system that can computerize the errands is as of now only a fantasy for generally staff.

One of the most punctual notices of computer grading of free text responses in the writing was in an article by Page in which he depicted Project Essay Grade (PEG). (Page, 1966). Different parts of understudies' essays, for example, extent of words on a typical word list going about as an intermediary for lingual authority, and the extent of relational words going about as an intermediary for sentence multifaceted nature, were estimated. A various relapse method was then used to anticipate the human rater's score, in view of these measures. We talk about the most recent form 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 free text response 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 Free Text Response Grading (Source: Page, 1966: 240)

There are inalienable issues to be overcome if automated grading of text is to end up a reality. Understudy essays tending to a specific point can theoretically be communicated in potentially thousands of structures, utilizing distinctive blends of words and sentences.

Essentially checking for the event of some catchphrases does not take into consideration an exceptionally precise evaluation of the work, nor does it take into consideration the wealth and decent variety that English takes into consideration articulation of thoughts. Numerous words have thirty to forty passages in a thesaurus, and for the most part a significant number of them are compatible in a specific and given context, so checking for the event of watchwords isn't a satisfactory approach.

The existing free text response grading systems(FTGS) as falling into broad themes and time periods, from which our literature review is modeled . Here, each category is an “era” in the field of FTGS , to emphasize the historical organization as shown in figure 2.

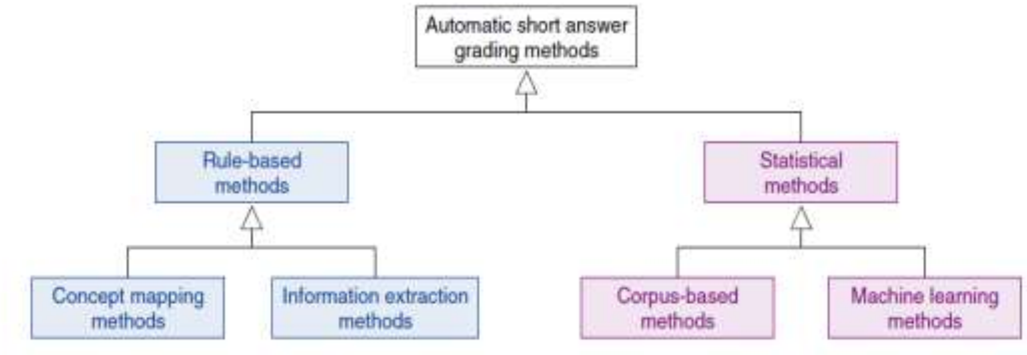

Figure 2: The four method-eras viewed as rule based or statistical methods

II. CONCEPTUAL MODELS FOR AUTOMATED FREE TEXT RESPONSE GRADING

The primary model, Project Essay Grade (PEG), is one of the most punctual and longest-lived implementations of automated free text response grading. It has been produced by Page and partners, and essentially depends on phonetic features of the free text response documents.

The second model, E_RATER, is one created by Burstein et al at the Educational Testing Service (ETS) in the US, which has been executed to the model stage for assessment. This model uses a half breed approach of joining phonetic features, determined by utilizing Natural Language Processing (NLP) techniques, with other document structure features.

The third model, the LSA display, makes utilization of Latent Semantic Analysis (LSA) and the "bag of words" approach, and has been created and assessed via Landauer et al at the University of Colorado at Boulder. It overlooks document semantic and structure features.

The fourth model, which utilizes text categorisation techniques, distinguished in this paper as TCT, has been created by Larkey at the University of Massachusetts. It utilizes a mix of changed catchphrases and semantic features.

PEG

Description

The thought behind PEG is to help decrease the tremendous free text response grading load in extensive instructive testing programs, for example, the SAT. At the point when different graders are utilized, issues emerge with consistency of grading. A bigger number of judges are probably going to create a genuine rating for a free text response.

An example of the free text responses to be graded is chosen and set apart by various human judges. Different semantic features of these free text responses are then estimated. A numerous regression condition is then created from these measures. This condition is then utilized, alongside the fitting measures from every understudy free text response to be graded, to foresee the normal score that a human judge would allocate.

PEG has its starting points in work started in the 1960's by Page and his partners (Page, 1966).

"… we begat two logical terms: Trins were the characteristic factors of intrigue – familiarity, word usage, language structure, accentuation, and numerous others. We had no immediate measures of these, so started with substitutes: Proxes were approximations, or conceivable connects, of these trins. All the computer factors (the real checks in the free text responses) were proxes. For instance, the trin of familiarity was associated with the prox of the quantity of words." (Page, 1994, p 130)

The various regression techniques are then used to figure, from the proxes, a condition to anticipate a score for every understudy free text response. In the exploration revealed in Page (1994), the objective was to recognize those factors which would demonstrate powerful in foreseeing human rater's scores. Different software items, including a language checker, a program to distinguish words and sentences, software lexicon, a grammatical form tagger, and a parser were utilized to accumulate information about numerous 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 mix of factual and NLP techniques to extract etymological features of the free text responses to be graded. As in all the conceptual models discussed in this paper, e-rater student free text responses are evaluated against a benchmark set of human graded free text responses. E-rater has modules that extract free text response vocabulary content, discourse structure data and syntactic data. Multiple linear regression techniques are then used to predict a score for the free text response, based upon the features extracted. For each new free text response question, the system is hurry to extract characteristic features from human scored essay responses. Fifty seven features of the benchmark free text responses, based upon six score focuses in an ETS scoring guide for manual grading, are at first used to construct the regression model. Utilizing stepwise regression techniques, the huge predictor variables are determined. The values derived for these variables from the student free text responses are then substituted into the specific regression equation to get the predicted score.

One of the scoring guide criteria is free text response syntactic variety. After parsing the free text response with a NLP instrument, the parse trees are analyzed to determine clause or verb types that the essay writer used. Proportions are then calculated for each syntactic type on a per free text response and per sentence premise.

Another scoring guide criteria relates to having well-developed arguments in the free text response. Discourse examination techniques are used to examine the free text response for discourse units by searching for surface cue words and non-lexical cues. These cues are then used to break the free text response up into parcels based upon singular content arguments.

The system likewise compares the topical content of an free text response with those of the reference texts by taking a gander at word usage.

Evaluation

The system has been evaluated by Burstein et el (1998) and has discovered 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 among human raters. For one test free text response question the accompanying predictive feature variables were observed to be critical.

1. Argument content score
2. free text response 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 assertion contents in a large two dimensional grid semantic space. Utilizing a framework 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 network. Each word being considered for the examination is represented as a line of a lattice, and the sections of the network represent the sentences, passages, or other subdivisions of the contexts in which the words happen. The cells contain the frequencies of the words in each context.

The SVD is then applied to the network. SVD breaks the first grid into three component matrices, that, when framework multiplied, reproduce the first lattice. Utilizing a reduced dimension of these three matrices in which the word-context affiliations can be represented, new relationships between words and contexts are induced when reconstructing a close estimate to the first framework from the reduced dimension component SVD matrices. These new relationships are made manifest, whereas preceding the SVD, they were hidden or latent.

To grade a free text response, a network for the free text response document is constructed, and then transformed by the SVD technique to approximately reproduce the framework utilizing the reduced dimensional matrices worked for the free text response theme area semantic space. The semantic space ordinarily comprises of human graded free text responses. Vectors are then computed from a student's free text response information. The vectors for the free text response document, and every one of the documents in the semantic space are compared, and the check for the graded free text response with the lowest cosine value in relation to the free text response to be graded is assigned.

The Intelligent free text response Assessor is a commercial implementation of the LSA approach. Later in this paper we examine a trial of this system for first year university student free text responses.

Evaluation

Landauer, et al (1998), report that LSA has been tried with five scoring methods, each varying the manner in which student free text responses were compared with sample free text responses.
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 free text responses. Foltz (1996) also reports that LSA grading performance is about as reliable as human graders. Landauer (1999) reports another test on GMAT free text responses where the percentages for adjacent agreement with human graders were between 85%-91%.

The Text Categorisation Technique (TCT)

Description

Larkey (1998) implemented an automated free text response grading approach based on text categorisation techniques, text complexity features, and linear regression methods. The Information Retrieval literature discusses techniques for characterizing documents as to their appropriateness of content for given document retrieval queries (van Rijsbergen, 1979). The technique initially makes use of Bayesian independent classifiers (Maron, 1961) to dole out probabilities to documents estimating the likelihood that they belong to a specified category of documents. The technique relies on an investigation of the occurrence of certain words in the documents. Secondly, a k-nearest neighbor technique is used to discover the k essays closest to the student free text response, where k is determined through preparing the system on a sample of human graded free text responses. The Inquery retrieval system (Callan et al, 1995) was used for this. At long last, eleven text complexity features are used, for example, the number of characters in the document, the number of different words in the document, the fourth base of the number of words in the document (see additionally the talk on PEG above), and the average sentence length.

Larkey conducted a number of regression trials, utilizing different mixes of components. He likewise used a number of free text response sets, including free text responses on social studies (soc), where content was the essential interest, and free text responses on general feeling (G1), where style was the primary criteria for assessment. The results presented here are for these two free text response sets as it were.

Evaluation

When all the criteria for assessment were used the proportion of free text responses 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 free text responses the corresponding figures were 0.55 and
0.97. The system performed remarkably well.

III. DISCUSSION

We are presently in a situation to characterize these free text response grading techniques as per the grouping postulated by Page. PEG focuses on simple phonetic features, concentrating on style, and can be categorized as II(A). E_RATER focuses on semantic features and document structures, and is along these lines performing a Master Analysis of style, and falls in the category II(B). The LSA model focuses on the semantics of the free text response, yet does so utilizing a Rating Simulation, and therefore falls in the I(A) category. The TCT (soc) experiments focused on content in a rating reenactment, while the TCT (G1) test focused on style in a rating recreation.
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 3: 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 4: 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 FREE TEXT RESPONSE ASSESSOR

A team of researchers in the School of Information Systems at Curtin University of Technology trialed the Intelligent Essay Assessor (IEA) amid the primary semester of 2001. In March 2001, students enrolled in the unit Information Systems 100 (IS100) were notified that they could receive extra characteristics of up to 5 per cent on the off chance that they participated in the trial by presenting an a few page free text response based on a question taken from their textbook. These free text responses, in Microsoft Word arrange, were submitted by means of email to a special IS100 email address.

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

In June 2001 an extra 327 ungraded free text responses 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 precision of the IEA was very great, when compared to the human graded average. Figure 4 demonstrates the conveyance of grades produced by the IEA.

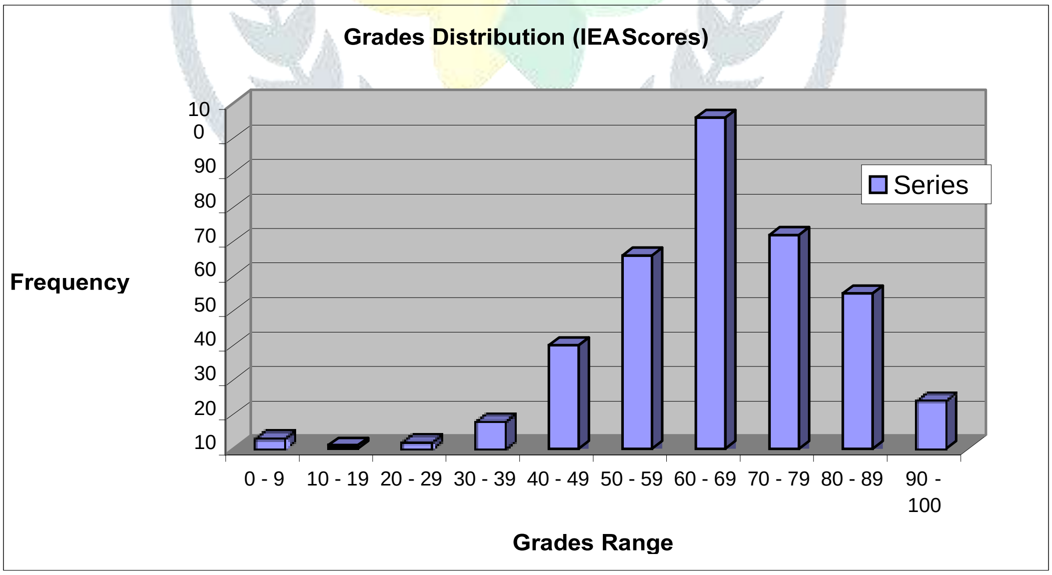

Figure 5: Distribution of grades produced by the IEA

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 free text responses are to be graded (eg 2000), yet the effort involved in arranging and human grading 200 free text responses for the semantic space, and the setup costs, are excessively great when just a few hundred free text responses are to be graded. The researchers were impressed by the capacity of the IEA to detect literary theft among the free text responses submitted by the students.

IV. CONCLUSION

Automated free text response 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 free text responses. 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 free text responses before the computer systems complete the task Students also need suitable computer facilities to generate their free text responses in machine readable form. It is likely that commercial free text response 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 correspondence with creator, eighth 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, second ed., Butterworths, London.