IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Volume 11, PP 22-26

Formative Assessment with Evaluation & Scoring Tool - A-TEST

Jyothi Arun[1], Rovina D'Britto[2], Sachin Gavhane[3], Snigdha Wasnik[4], Yogita Shelar[5]

[1]*(Atharva College of Engineering/ Mumbai University, India)*
[2]*(Atharva College of Engineering/ Mumbai University, India)*
[3]*(Atharva College of Engineering/ Mumbai University, India)*
[4]*(Atharva College of Engineering/ Mumbai University, India)*
[5]*(Atharva College of Engineering/ Mumbai University, India)*

Abstract: *In large classrooms with limited teacher time, there is a need for automatic evaluation of text answers and for real-time, personalized feedback as students learn. Most automatic essay scoring systems focus on providing scores that are comparable to human scores. While this helps save time in large scale or ongoing assessments, it does not provide real time help to learners. Moreover, training the system generally requires the help of experts in machine learning and NLP.*

We discuss the design of A-TEST, a text evaluation and scoring tool that can learn text answers and the scoring methodology from manually scored answers and also directly from teacher input using LSA, NLP and SVD techniques. In addition, A-TEST provides the necessary information to our web tutoring system thus enabling it to provide real-time feedback during the formative assessment process.

Keywords *- Essay scoring, Latent Semantic Analysis (LSA), SVD, word by context matrix, Natural Language Process, NLP, AES, formative assessment*

I. INTRODUCTION

The advancement in internet technologies has increased the reach of web based assessment and tutoring to tens of thousands of users. However in most systems the process of evaluation and grading is limited to objective questions which are not enough to evaluate a student in more complex tasks. Most web based systems do not automatically evaluate or provide feedback to users for text answers but store them for teacher evaluation. Scoring text answers has challenges as there are many right and wrong answers. Though there are a few systems that score essays with a high degree of accuracy, but they do not provide immediate feedback to the students. This often requires machine learning of a large number of humanly evaluated answers and using similarity functions to score the essays.

Tutoring systems in STEM learning often require a large number of questions that require text answers and it may be difficult to have a large corpus of manually scored answer sheets to learn from. Often for STEM learning, subject matter experts can compile the important list of keywords that they look for in evaluating text answers. The subject matter expert (SME) may also specify the weight-age for evaluation criteria such as content, grammar, spelling etc. This paper describes an text evaluation tool called A-TEST (Amrita Text Evaluation & Scoring Tool). We describe the architecture of the system and evaluate the accuracy of the system using for one essay corpus using both the manually assessed corpus and the teacher entered keywords.

II. BACKGROUND AND RELATED WORKS

A system for automated assessment is consistent for the way it scores essays. The key advantage of the system is the cost and time savings. The system can be shown to grade essays within the range of those given manually. Recently there has been lot of research and studies happening in the area of Automated Essay Scoring. Researchers were successful to a great extent in considering the essays as the most useful tool for assessing, the learning outcomes, ability to recall, organize and integrate ideas, express oneself in writing and the ability for application of data.

The performance of Latent Semantic Analysis in various Information Retrieval tasks apart from automated essay grading was determined by Tuomo Kakkonen, Niko Myller, Erkki Sutinen and Jari Timonen determined [1]. LSA has proved to be one of the most successful methods for content-based essay grading. Depending on the test set, Landauer et al. (1997) and Foltz et al. (2000) have, for example, reported correlations from 0.64 to 0.84 between grades given by two human assessors and correlations from 0.59 to 0.89 between the LSA-based grading system and human graders. This means that LSA-based systems perform as well as the human graders. Kakkonen et al. [2006], [2], [13] used Probabilistic Latent Semantic Analysis (PLSA) [Hofmann, 2001]. PLSA adds a stronger probabilistic model to LSA based on a mixture decomposition derived from the latent class model. This results in a more principled approach which has a solid foundation in statistics.

However, it also has over fitting problems. In fact, the results achieved with PLSA are quite similar of those achieved with LSA.

According to Wolff, Burstein, Li, Rock, and Kaplan [1], [2], the four quality criteria for an automated essay grading system are accuracy, defensibility, coachability and cost-efficiency. For a system to be acceptable, it must deliver on all these criteria. An accurate system is capable of producing reliable grades measured by the correlation between a human grader and the system. In order to be defensible, the grading procedure employed by the system must be traceable and educationally valid. In other words, it should be possible rationally to justify and explain the grading method and the criteria for given grades. Coachability refers to the transparency of the grading method. If the system is based on simple, surface-based methods that ignore content, students could theoretically train themselves to circumvent the system and so obtain higher grades than they deserve. It is also self-evident that an automated grading system must be cost-efficient because its ultimate purpose is to reduce the total costs of assessment.

Jill Burstein, Martin Chodorow, Claudia Leacock [6] proposed the Criterion interface which was developed by showing screen shots and prototypes to teachers and students and eliciting their comments and suggestions. The interface presented one of the larger challenges. A major difficulty was determining how to present a potentially overwhelming amount of feedback information in a manageable format via browser-based software.

Scott Deerwester, Susan T. Dumais*, George W. Furnas, and Thomas K. Landauer [7], proposed the approach to take the advantage of implicit higher-order structure in the association of terms with documents ("semantic structure") in order to improve the detection of relevant documents on the basis of terms found in queries.

A. A-TEST

Spell checkers, word count and such features are important evaluation criteria but only provide feedback on errors in the surface features. LSA is a powerful IR technique that uses statistics and linear algebra to discover underlying "latent" meaning of text and has been successfully used in English language text evaluation and retrieval [8], [9], [10]. To assess knowledge, A-TEST uses Latent Semantic Analysis (LSA) to measure the student's knowledge based on analyses of a set of textual information on the subject domain and the student's answer and then validates the model using similarity measures. A-TEST can learn from either a corpus of manually scored answers using NLP and LSA techniques and from direct entries from the subject matter expert and determine an optimal evaluation and scoring model for each question with a text response. A-TEST looks for significant keywords, combinations of words, spelling, word count, grammar and the closest model that maps the essay to the teacher scores.

For STEM evaluations, teachers tend to give much more weight-age to the actual content than to grammar, spelling or word-count whereas for language essays both the syntactic and semantic aspects are important. A-TEST learns the weights from both the training essays and directly with teacher input. The output of A-TEST are the most important keywords learnt and a scoring model.

III. ENHANCEMENTS TO WEB TUTORING

The information learned from A-Test is used by our web tutoring system called to enhance the learning experience for text answers. When a new user answers a question, the system checks the spelling and grammar using NLP tools and also checks the content using the keywords or group of keywords stored in the system and provides scaffolds based on the errors. The student may tell the system whether she/he agrees with the score or ask for teacher intervention. In the case the student asks for teacher intervention, an alert is send to the teacher. They teacher may either agree or disagree with the automated score and is prompted to enter the reason for this. Therefore the system enables the creation of essay type questions by the subject matter expert and provides personalized feedback to the student along with automatic evaluation. Moreover a mechanism is in place for incremental learning based on teacher and student feedback.

IV. CONTENT EVALUATION WITH A-TEST

An important goal of A-TEST is to find a useful model of the key relationships between terms and documents.

Latent Semantic Analysis is a technique of document retrieval and word similarity method underlying the use of reduced singular value decomposition for the tasks of retrieving original data, usually consisting of a word×document matrix, and breaks it down into linearly independent components with the goal of removing noisy correlations in the original data.

LSA is a complex statistical technique that was initially developed for indexing documents and Information Retrieval [Deerwester et al., 1990]. Nevertheless, it can also be applied to automated essay grading [Haley et al., 2003]. LSA might be described as comprising of the training, test and result phase [Whittingdon

and Hunt, 1999]. In the training phase, weights of the vectors that represent the course materials are calculated using large course materials without manual scoring. The test phase represents each individual student essay along with several transformations including pre-processing, stop-word removal, spelling correction and stemming. The term by matrix is decomposed into the product of three orthogonal matrices using Singular Value Decomposition. One of the matrices is diagonal and its values are the singular eigen values of the original matrix. Dimensionality reduction is done by reducing the rank of diagonal matrix find the more significant terms and relations, remove noise and the new set of matrices result in faster computation with similar accuracy. Finally, once we have the LSA representation of the student text is compared against the LSA model representations and their similarity is computed.

V. A-TEST ARCHITECTURE

The system is being developed for students who will have a quick access to their personal assessment with feedback and the teachers are provided with the concise conceptual model of each student using the Automated Essay Scoring System. The system architecture and programming philosophy are explained in the next sections.

The Figure1 depicted below abstracts the broad level architecture of the proposed A-TEST system. The system consists of mainly two phases: Analysis and Grading phase. In the Analysis phase, series of algorithms are followed such as: pre-processing, dimensionality reduction, weighting schemes, and similarity measure. In the Grading phase, the essays are graded according to the scoring model created in Analysis phase. This system is used to automatically grade the essays and give personalized feedback to the users. Here, it is based on the grading of scoring model which is developed by comparing the manually graded essays with the course materials. The system currently contains the implementation of LSA. The system consists of scoring model which is created from the course materials such as passages from lecture notes, textbooks or pre-graded essays. These passages are selected on the basis of the importance to the essay that has to be graded. The main component of the system includes the Pre-processor, LSA, Grade definition and other feature selection modules. The base forms of words in the input documents as well as the pre-processing stages are functioned. In the Scoring model, Word by Context matrix is processed with LSA and creates reduced representation of WCM. This removes the details which are not required and make system more obvious.

Document vectors are created from manually graded essays to the reduced WCM as a result to determine the similarity between each essay along with trained course materials. Thus, comparison between the individual essay and reduced WCM is done so that the distances between document vector of an essay and every document vector in the reduced WCM are calculated. Then cosine similarity measure is applied between the documents.

VI REGRESSION MODEL FOR SCORING

We discuss the results evaluating a set of essays using A-TEST. The theme of the essays is the effect of computers on people. We tested our model using 385 essays from dataset of 1400 pre-scored essays. The training essays and submitted essay were graded manually by two teachers. The score point of each essay ranged from 1 to 6, where a higher point represented a higher quality. After the LSA, models were created using regression analysis.

An unscored essay was compared to the closest match amongst the keywords learnt during the LSA and SVD process using cosine similarity. This finds the scored essay that best matches the essay to be scored in terms of the keywords but ignores other factors such as word-count and the number of spelling mistakes. Hence the score of the best matched essay needs to be adjusted based on these other factors in the new essay. The following factors were found to be significant by using multiple regression analysis and using the scores of rater1 as the training labels.

	Estimate	Std. Error	t value	Pr(>|t|)
(Spell-errors)^1/4	1.08367	0.054557	19.863	< 2e-16
Difference in Spell-errors between the two essays	-0.44809	0.051643	-8.677	< 2e-16
Word Count of the essay being scored	0.002284	0.00033	6.919	1.94e-11
Rater1 human score of the best matched essay	0.145495	0.032987	4.411	1.34e-05

Table 1: Regression Analysis

Regression analysis is used to predict a continuous dependent variable from independent variables and also to find their relationships. Multiple regression analysis for the model has independent variables such as the content score, count of spelling mistake, word count and their mark and dependent variables as score. For the

model, the adjusted R-squared value was 0.9853, residual standard error 0.5249 with 381 degrees of freedom. The F-statistic: 6436 on 4, with p-value: < 2.2e-1. Using this model, the score was calculated for 385 essays. The percentage agreement (# of agreements/total number of essays*100) between rater1 and our model was 67% while the percentage agreement between the two manual scorer rater1 and rater2 was 63%. And the percentage adjacent agreement (# of rater1 of agreements/total number of essays*100) between rater1 and our model was 99.7% while the percentage adjacent agreement between the two manual scorer rater1 and rater2 was 97.9% which scores better

Kappa value scores (Table 2) indicate moderate agreement between the two manual raters and also moderate agreement between our models and the raters. Kappa (Cohen, 1960; Fleiss, Levin, & Paik, 2003; Spitzer, Cohen,, & Fleiss, 1968) can range values from -1.0 to +1.0. A kappa of 1.0 means that two raters show perfect agreement, a kappa of -1.0 means that they show perfect and consistent disagreement, and a kappa of 0 means that the two raters show no relationship between their ratings. There has to be general agreement between the raters such that the kappa should be at least .60 or .70. The Table 2 shows the quadratic weighted kappa's score that has an agreement of .71 for the model score with rater1. This indicates that, the inter-rater reliability is good enough.

Measure of Agreement	Kappa Value	Quadratic weighted Kappa score	Asymp. Std. Error a	Approx. Tb	Approx. Sig.
Model-score with rater2	0.438	0.672	0.039	12.419	.000
Model-score with rater1	0.443	0.712	0.04	12.508	.000
rater1 with rater2	0.425	0.691	0.038	12.488	.000

Table 2: Measurement of Kappa Score

Though the percentage agreement and the kappa values were slightly better than the corresponding values between rater1 and rater2, these can be improved by additional factors such as grammar, complexity of essay, parts of speech count, punctuations and more advanced NLP features (N-grams, k-nearest neighbors in bag of words).

VI. CONCLUSION

Though there are many scoring engines, our solution is different in that we have developed a generic tool A-TEST that can be easily used by an educator to learn the essays and the scoring materials based on course materials and teacher-graded essays. In addition to learning from a corpus and pre-scored text answers, the system can also learn additional keywords or bi-grams that are directly entered by an educator. Another important difference is that this information can be used to provide real-time feedback to students in the formative evaluation.

The model is useful for large scale formative evaluation of science assessments where learning the significant keyword related to science concepts is important. The formative assessment authoring platform uses this information to provide real-time help to students who attempt text answers. Performance on essays can be improved by incorporating content and advance features that can contribute towards a good prediction. The performance of A_TEST for a sample text answer using a dataset of pre-scored essays was comparable to the accuracy between two humanly scores. Though our prediction model worked as well as the manually evaluated teacher model, additional enhancements such as n-grams, grammar specific errors, and other dimensionality methods such as PLSA and further improve the prediction model and are planned as further work. The proposed system is applicable for essay with raw text. In future the proposed work will be extended towards the grading of essays containing text, tables, mathematical equation etc.

Acknowledgements

This work derives direction and inspiration from the Chancellor of Amrita University, Sri Mata Amritanandamayi Devi.

REFERENCES

[1] Tuomo Kakkonen, Niko Myller, Erkki Sutinen and Jari Timonen, "Comparison of Dimension Reduction Methods for Automated Essay Grading", International Forum of Educational Technology & Society (IFETS), J. (2008).

[2] Rama Adhitia and Ayu Purwarianti., Automated Essay Grading System Using SVM and LSA for Essay Answers in Indonesian.

[3] Michael Flor ,Yoko Futagi, "On using context for automatic correction of non-word misspellings in student essays ",The 7th Workshop on the Innovative Use of NLP for Building Educational Applications, pages 105–115, Montreal, Canada, June 3-8, 2012.

[4] Turney, Peter D. and Michael L. Littman. 2003, "Measuring praise and criticism: inference of semantic orientation from association", ACM Transactions on Information Systems 21: 315-346.

[5] Ryo Nagata, Junichi Kakegawa, and Yukiko Yabuta," A Topic-independent Method for Automatically Scoring Essay Content Rivaling Topic-dependent Methods" 2009 Ninth IEEE International Conference on Advanced Learning Technologies.

[6] Chien-Liang Liu, Wen-Hoar Hsiao, Chia-Hoang L and Hsiao-Cheng Chi," An HMM-Based Algorithm for Content Ranking and Coherence-Feature Extraction", IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS,2012.

[7] T. Miller, "Essay assessment with latent semantic analysis," Department of Computer Science, University of Toronto, Toronto, ON M5S 3G4, Canada , 2002.

[8] Loraksa and R. Peachavanish, "Automatic Thailanguage essay scoring using neural network and latent semantic analysis," in Proceedings of the First Asia International Conference on Modeling & Simulation (AMS'07), 2007, pp. 400–402. "doi:10.1109/AMS.2007.19"

[9] D. T. Haley, P. Thomas, A. D. Roeck, and M. Petre, "Measuring improvement in latent semantic analysis based marking systems: using a computer to mark questions about HTML," in Proceedings of the Ninth Australasian Computing Education Conference (ACE) 2007, pp. 35-52.

[10] Md. Monjurul Islam, A. S. M. Latiful Hoque," Automated Essay Scoring Using Generalized Latent Semantic Analysis" JOURNAL OF COMPUTERS, VOL. 7, NO. 3, MARCH 2012.

[11] M.W. Berry, S.T. Dumais & G.W. O'Brien," Using Linear Algebra for Intelligent Information Retrieval," National Science Foundation under grant Nos. NSF-CDA-9115428 and NSF-ASC-92-03004.

[12] Tuomo kakkonen, Niko Myller, Erkki Sutinen," Applying part of speech enhanced LSA to Automated Essay Grading," Automated Assessment Technologies for Free Text and Programming Assignments by Academy of Finland.