A STUDY OF COMPUTER-AIDED GRADING OF WRITTEN TEAM WORK

By

P. Josan & R.V. Fleisig

A Technical Report

Submitted to the Centre for Leadership in Learning

McMaster University

Abstract

The trend towards increased enrolment in higher education has inevitably led to larger classes and increased the stress on educators including the demand on their time and budgets. One of the most significant effects of this is the increased time and effort spent on grading and verification of discrepancies between different graders. A possible aid to deal with this problem is to provide instructors and teaching assistants with computer-aided means of analyzing the written work. The purpose of this study was to explore the use of a computational technique called "Latent Semantic Analysis" (LSA), to aid student authors, their teaching assistants, and instructors in the assessment of written project work in design engineering. Several tests were performed on the course team documents and assignments, to compare LSA's reliability to human graders for accurate and effective grading. LSA produced accurate results and proved to be more reliable than an average grader for the set of documents tested from the engineering course, ENGINEER 1P03 – Introduction to Professional Engineering.

Table of Contents

List of Tables

Introduction

The motivation for this study is derived from recent increases in undergraduate enrolment in the Faculty of Engineering which has raised the pressure on both the educators and the students. New techniques are needed to maintain the quality of education without exceeding the limits of existing resources. The purpose of this study was to investigate a computational method called Latent Semantic Analysis (LSA) and verify its reliability for implementation in engineering courses. LSA is a measure of semantic similarity of written passages and could be used for automated evaluation of student work [1]. This study is one of the efforts that are being directed towards improving the learning in team-based written work throughout the engineering design process, in a first year Engineering course, ENG 1P03 – Introduction to Professional Engineering. On a wider scope, it is hoped that the findings from the study would be helpful in integrating automated assessment technologies in other university courses as well.

Along with the number of students entering the university each year, the demand for additional teaching resources has also increased drastically. ENG 1P03 alone requires about 30 – 40 teaching assistants (TAs) a large part of whose job it is to grade written student work which is a significant component of the course. In large classes of over 800 – 1000 students, the need for frequent feedback on the assignments and project work becomes crucial. Due to the large number of graders, it is hard to maintain consistent grading. As a result, the issue of fair and similar evaluation of all students in an identical manner arises. Instructors also face problems in keeping track of the teaching and grading quality of all the TAs.

A promising solution to deal with all these problems was to provide instructors and teaching assistants with computer aided means of analyzing the written work. Instructors could use such method as a secondary measure and also offer it to the students for indicating the relative quality of their team generated documents before the final submission.

This study aimed to explore LSA's reliability for evaluation purposes, compared to the existing graders and its prospects for integration into the existing ENG 1P03 grading system. LSA measures the cohesiveness and coherence of the text, which are well known to be linked to good writing [2] [3]. However, it does not check the grammar, spelling, punctuation, style or syntax of the passages [4]. It has been proven to be effective in evaluating the quality of academic essays and GMAT test questions under controlled circumstances [5]. The short term goal of this study was to verify and test LSA's effectiveness in accurately assessing the quality of team writing, individual written assignments and engineering outcomes in ENG 1P03.

Research Method

The research involved a series of experiments in which ENG 1P03 team project and individual assignment documents were evaluated using LSA. The results were compared against the human evaluation of those documents done by engineering experts and a statistical correlation was calculated between both sets of evaluations. All experiments included the following three basic steps:

1. Expert Evaluation
2. Assessment by LSA
3. Comparison and Analysis of the Results

Expert Evaluation

Each experiment began with the task of collecting ENG 1P03 student documents from the academic year 2009 – 2010 for testing. TAs' grades were recorded and the documents were ranked according to their assigned grades. If possible, more rankings were obtained from several engineering experts for the same set of documents, to measure inter-rater correlations and keep discrepancies to the minimum. In order to protect the privacy of the students, all the students' identifying information was removed from the documents before they were given to the experts for their evaluation. This step was conducted to rank the test documents in order of their content quality based on the graders' evaluation.

Assessment by LSA

The same documents were converted to text format for inputting into LSA code written in MATLAB. Based on the LSA measures used for the specific experiment, the documents were ranked accordingly. This step, therefore, revealed the LSA's assessment of the student documents.

Comparison and Analysis of the Results

The sets of human rankings were then compared against the LSA's ranking using a statistical method called 'Spearman's rank correlation'. This method measures power of the link or relation that exists between two sets of data [6].

The Spearman's rank correlation formula is given by

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)} \tag{1}$$

where $r_s \in [-1,1]$ is the Spearman rank correlation coefficient, d is the difference in rank between a given a pair, and n is the total number of pairs (i.e., the number of documents used/ranks).

The higher the value of the correlation coefficient, stronger is the link between two data sets, which in this case are the grader and LSA rankings. Reflecting on a previous study done in the field of LSA, a coefficient of around 0.69 was anticipated between the two rankings and this number was used as a threshold for the study [7]. In cases where it was possible to gather more than one set of human rankings, the grader-grader correlation (between all human graders) set the bar and was compared against the grader-LSA correlation.

Experiment 1

The objective of the Experiment 1 is to test LSA's effectiveness and efficiency in assessing the ENG 1P03 individual student assignments.

Assignments Used

For the first experiment, seven scenarios based on Assignment 1 documents were used. Each document was approximately 270 words long. All of these documents were marked by a single TA, which eliminated the existence of any discrepancies in the grading.

Procedure

1. *Expert Evaluation:* The grades assigned to the seven documents were recorded and the documents were ranked in the descending order from the highest grade to the lowest. In order to measure the reliability correlation between different graders, expert evaluation on the same documents was also gathered from four other faculty members. Ranking by these additional graders was conducted in the following manner:
    a. Only the body of the assignment letters was provided for ranking with all the student identifying information taken off for confidentiality purposes.
    b. All graders signed a confidentiality oath letter (shown in Appendix A) to keep all information regarding the study confidential.
    c. Each grader was given seven documents (labeled "Document #") to rank.
    d. In order to make the evaluation as consistent and accurate as possible, assignment question, scenario, requirements, ranking instructions and criteria were given to each grader. Appendix B shows the all this information in the same format as it was provided to the graders. This initial grader (TA) had the same information available for evaluation.
    e. In addition to these, an ideal assignment 'exemplar' was also provided to the graders.
    f. Each grader had a maximum of 40 minutes to rank all the assignments individually.
    g. A list of the document rankings was obtained at the end from each grader.
2. *Assessment by LSA:* This was done using two different metrics: cosine similarity and standard deviation.
3. *Correlation Calculations:* The inter grader and LSA-grader correlations were calculated and compared.

Cosine Similarity Metric

Cosine similarity method measures the cohesiveness/similarity between passages of text [5]. This LSA measure was used to discover semantic ties between the test documents and the 'ideal assignment'. The higher the cosine similarity value, the closer a test document is

to the ideal one, in terms of the overall meaning [5]. This assessment was carried out using the following steps:

- A term-document matrix was constructed using the seven test documents and one Ideal assignment.
- Appendix C shows the steps taken for developing the ideal assignment.
- Single value decomposition (SVD) of the matrix was calculated after applying log-entropy weighting and *K*-reduced.
- Cosine angle similarity values of the test documents were calculated against the ideal assignment.
- The similarity values were arranged in descending order and the documents were ranked accordingly, meaning that document with the highest value was given the first rank.

Standard Deviation Metric

Standard deviation is another LSA technique used to measure coherence between the passages of a single text. The lower the SD value, the higher the semantic coherence and text quality will be [7].

- Each test document was broken down sentence-by-sentence into several passages titled 'Doc (1...7) Passage (1...n)'. Therefore, all documents had their own set of passages, each consisting of only one sentence.
- These Passages were used to construct a 'Passage –Term' matrix.
- Single value decomposition of the matrix was calculated after applying log-entropy weighting and *K*-reduced. The weighting is used to get rid of the unnecessary information or noise in the matrix and works by assigning different weightings to each term-frequency cell depending on their importance in the text. It is calculated by using the following formulas [8] where tf_{ij} is the frequency of term *i* in document *j*, gf_i is the frequency of term *i* across all documents, and *n* is the number of documents.

$$\text{Local Weight: } l_{ij} = log\ (tf_{ij} + 1) \tag{2}$$

$$\text{Global Weight: } g_i = 1 + \frac{\sum_j (p_{ij}\ log\ (p_{ij}))}{log\ (n)} \tag{3}$$

$$\text{Final Weight: } X_{ij} = l_{ij} * g_i \tag{4}$$

$$p_{ij} = \frac{tf_{ij}}{gf_i} \tag{5}$$

- Centroid and standard deviation values were calculated for the passages of each document using the following formulas [7] where $\boldsymbol{S}_{k,i}$ is the *k*th singular vector (passage) representing document *i* and n_i is the number of passages in document *i*.

$$\text{Centroid: } \boldsymbol{c}_i = \frac{\sum_{k=1}^{v_i} \boldsymbol{S}_{k,i}}{n_i} \tag{6}$$

$$\text{Standard Deviation: } \sigma_i = \frac{\sqrt{\sum_{k=1}^{v_i} (\boldsymbol{S}_{k,i} - \boldsymbol{c}_i)^2}}{n_i} \tag{7}$$

- Another set of rankings was then obtained by arranging the documents in the increasing order of their standard deviation values, i.e., the document with lowest standard deviation value was given the first rank.

Observations

The Cosine Similarity and SD values obtained for the seven documents are given in Table 1.

Table 1 Cosine Similarity and SD values

Document #	Cosine Similarity	Standard Deviation
1	8952.4	0.538
2	9613.1	0.459
3	15470.0	0.323
4	17805.0	0.502
5	10525.0	0.341
6	1876.9	0.683
7	9161.6	0.562

The correlation values of LSA's ranking against different graders' are given in Table 2.

Table 2 Grader-LSA correlations

Grader	Cosine Similarity	Standard Deviation
1	0.214	-0.143
2	0.571	0.536
3	0.071	0.536
4	0.429	0.250
5	0.429	0.607

The correlations between different graders are given in Table 3.

Table 3 Inter-grader correlations

vs.	Grader 1	Grader 2	Grader 3	Grader 4	Grader 5
Grader 1		0.214	-0.393	0.286	0.357
Grader 2			0.143	0.857	0.857
Grader 3				-0.071	0.357
Grader 4					0.714

Results

Based on these observations, the following results were obtained:

- Average inter-grader correlation: 0.332
- Average standard deviation grader-LSA correlation: 0.357
- Average cosine similarity grader-LSA correlation: 0.343

If the lowest correlations (i.e., Grader 1 for SD and Grader 3 for Cosine) are ignored then the following results are obtained:

- Average inter-grader correlation: 0.476
- Average standard deviation grader-LSA correlation: 0.482
- Average Cosine Similarity grader – LSA correlation: 0.411

Findings

1. LSA evaluated the individual student assignments with a higher level of reliability than an average grader for the small sample of documents tested. Therefore, it may be effective in accurately assessing student work in ENG 1P03.
2. Assessment of assignments and individual writing with LSA is more consistent than human graders.
3. Cosine similarity method can fail to recognize creative writing by assigning a lower grade to a document that is less similar to the exemplar, but could really be unique and creative. This problem might be solved by using multiple exemplars instead of one.
4. LSA's standard deviation technique proved to be a better measure of overall quality and coherence of individual creative assignments than the cosine method.
5. Cosine similarity method works best for question based writings, where creativity is limited and all possible answers are similar. For example, the short answer questions on engineering tests/exams.

Experiment 2

The objective of Experiment 2 is to test LSA's effectiveness and efficiency in assessing team generated engineering design work.

Assignments Used

To test LSA's ability to evaluate team writing, ENG 1P03 final project reports were used. Each report was over 5000 words long.

Procedure

1. *Expert Evaluation:* It was difficult to gather more sets of expert evaluations on such large documents, so the 45 reports selected for this experiment were marked by the six strongest and most reliable, graduate TAs. Separate tests were run for each TA's graded set of documents to eliminate discrepancies between individuals. The grades for all the documents were recorded and individual sets of rankings were generated for all TAs.
2. *Assessment by LSA:* The project reports were the work of several students put together in teams. Therefore, the best way to asses them using LSA was to measure their coherence between sections, to see if all the authors (team members) had represented or expressed the same ideas/views throughout the report but probably using varying words/synonyms and styles [7]. As described earlier, standard deviation (SD) method is a perfect measure of coherence in writing, The LSA technique used for this experiment. The following approach was taken to carry out the evaluation:
    a. Each report document was broken down section-by-section into several passages titled 'Doc (1...n) Passage (1...n)'. Each TA had his/her own set of passages.
    b. These passages were used to construct a 'Passage-Term' matrix. Again, one matrix for each TA's set of passages.
    c. Single value decomposition of the matrices was calculated after applying log-entropy weighting and *K*-reduced.
    d. Centroid and standard deviation values were calculated for the passages of each document.
    e. A set of rankings was then obtained for each TA's documents by arranging them in the increasing order of their standard deviation values, i.e., the document with lowest standard deviation value was given the first rank.
3. *Correlation Calculations:* The LSA-grader (TA) correlations were calculated and compared. The correlation of 0.69 served as the threshold for this experiment.

Observations

Table 4 shows the standard deviation values obtained for the document sets for each TA.

Table 4 SD values for the document sets of 6 TAs

Document #	TA 1	TA 2	TA 3	TA 4	TA 5	TA 6
1	2.6437	2.1598	1.6561	2.1558	2.6617	2.9198
2	3.9987	2.1583	3.0047	2.7223	1.9455	2.9548
3	1.9282	2.464	3.5156	3.2559	3.2984	2.5659
4	3.8284	2.2996	2.5675	2.5375	2.5134	2.9907
5	5.0703	3.0361	3.0316	2.6873	4.4467	2.1097
6	2.1282	1.272	2.2784	2.9632	3.5084	2.156
7	2.4296	2.7381		1.8034	2.5303	1.6312
8	3.726			1.9302	3.1528	
9	3.5943					

Results

Results are given in Table 5 and yield an average correlation of 0.756.

Table 5 Correlation coefficients for the six TAs (LSA-TA correlations)

TA	Correlation
1	0.717
2	0.786
3	0.543
4	0.810
5	0.679
6	1.000

Findings

1. Getting the average correlation value for this experiment higher than 0.69, means that LSA performed with a higher reliability than the ranges set by the previous studies. This proves that LSA is indeed a good measure for the assessment of team based engineering design work.
2. One interesting observation made during the tests was that LSA's evaluation is not affected by the length of the document. For e.g. the longest documents with the maximum word count never achieved the best ranks and vice-versa. However, the number of passages that each document is split into seemed to have a slight effect on the SD values in some circumstances, consequently affecting the LSA ranking. This is evident from the fact that for three TAs the document with the best and worst ranks had the most and least number of passages, respectively. For the other three, no such relation was observed. Therefore, it could be said that even though it is not direct, there exists some kind of relation between the number of passages for a given document and its SD value.
3. The reports also contained a significant amount of information in picture or table format which did not get inputted with the text into LSA. This could have lead to poor coherence for some documents with plenty of pictures in them, which in turn

affected LSA's rankings. Thus, the correlations are expected to be higher, for text-only documents.

Conclusions

Based on the experiments done in the research, it can be said that LSA may be effective at evaluating the quality of team project documents and class assignments in ENG 1P03. However, it has some limitations and therefore cannot replace human grading in the near future. For now, it will evidently prove to be a great tool for instructors to get supplementary assessment on student work. Future studies with LSA, will investigate applications of LSA to the issues of detecting plagiarism, assessment of individual performance on a team, and team collaboration. It may also be possible to use LSA to determine the level of creative input by students to a project or to analyze year-to-year course improvements.

References

[1] Thomas K. Landauer, Peter W. Foltz, and Darrell Laham, "An introduction to latent semantic analysis," *Discourse Processes*, vol. 25, no. 2 & 3, pp. 259 - 284, 1998.

[2] Peter W. Foltz, Walter Kintsch, and Thomas K. Landauer, "The measurement of textual coherence with latent semantic analysis," *Discourse Processes*, vol. 25, no. 2 & 3, pp. 285 - 307, 1998.

[3] Stephen P. Witte and Lester Faigley, "Coherence, Cohesion, and Writing Quality," *College Composition and Communication*, vol. 32, no. 2, pp. 189 - 204, May 1981.

[4] Tristan Miller, "Essay assessment with latent semantic analysis," *Journal of Educational Computing Research*, vol. 29, no. 4, pp. 495 - 512, 2003.

[5] Peter W. Foltz, Darrell Laham, and Thomas K. Landauer, "The Intelligent Essay Assessor: Applications to Educational Technology," *Interactive Media Electronic Journal of Computer-Enhanced Learning*, vol. 1, no. 2, October 1999.

[6] Barcelona Fields Study Centre. (2009, November) Spearman's Rank Correlation Coefficient. [Online]. http://geographyfieldwork.com/SpearmansRank.htm

[7] Andy Dong, Andrew W. Hill, and Alice M. Agogino, "A Document Analysis Method for Characterizing Design Team Performance," *Journal of Mechanical Design*, vol. 126, no. 3, pp. 378 - 385, May 2004.

[8] Guillermo Jorge-Botana, José A. León, Ricardo Olmos, and Inmaculada Escudero, "Latent Semantic Analysis Parameters for Essay Evaluation using Small-Scale Corpora*," *Journal of Quantitative Linguistics*, vol. 17, no. 1, pp. 1 - 29, February 2010.

Appendix A

McMaster University

Oath of Confidentiality

I understand that as a faculty member/student/employee participating in the study of *Computer Aided Grading of Written Team Work* being conducted by Professor R.V. Fleisig of the Department of Mechanical Engineering, McMaster University, confidential information will be made known to me.

I agree to keep all information collected during this study confidential and will not reveal by speaking, communicating or transmitting this information in written, electronic (disks, tapes, transcripts, email) or any other way it to anyone outside the research team.

Name: _____ Signature: _____

(Print)

Date: _____Witness Signature: _____

Appendix B

The Assignment scenario is as follows:

You are a student in ENGINEER 1X03. Professor Smith, instructor in your ENGINEER 1X03, has shared in lecture with the entire class your first assignment. The Professor has done this without consulting you. You were embarrassed by this event since your name clearly appeared on the work. As a result, you feel uncomfortable coming to class and facing the Professor. To diffuse the situation you choose to write the Professor an email.

Consider carefully what you are going to ask for. **Be very specific. Your request should have a reasonable expectation of resolving the situation to your satisfaction.** However, first you must determine whether or not the Professor had a right to share your work without your permission.

Ranking Instructions: The assignments should be ranked in the order of how well they fulfill the following criteria. All criteria are equally weighted.

1. *Readability:* The assignment sounds very well written i.e. it is easy to read and understand. Contains the necessary details about the incident e.g. Date or section # etc. The letter clearly describes the incident and its effects on the student as stated in the scenario.
2. *Supporting arguments:* Writer provides strong, well-researched arguments about the university's policies regarding student work.
3. *Professionalism:* There is a balance of respect and directness.
4. *Request:* There is a SPECIFIC request made to address the situation at hand.

Appendix C

The process of developing Ideal Assignment 1 document:

Top twelve assignment submissions that got perfect grades were selected. After thorough reading of these assignments, they were all composed together to create one final ideal assignment that met all the requirements in an accurate and efficient form.