Supporting Content-Based Feedback in On-Line Writing Evaluation with LSA

Peter W. Foltz , Sara Gilliam & Scott Kendall

Published online: 09 Aug 2010.

Submit your article to this journal 🖉

Article views: 81

View related articles 🔍

Citing articles: 41 View citing articles 🗗

Supporting Content-Based Feedback in On-Line Writing Evaluation with LSA

Peter W. Foltz, Sara Gilliam and Scott Kendall
New Mexico State University

ABSTRACT

This paper describes tests of an automated essay grader and critic that uses Latent Semantic Analysis. Several methods which score the quality of the content in essays are described and tested. These methods are compared against human scores for the essays and the results show that LSA can score as accurately as the humans. Finally, we describe the implementation of the essay grader/critic in an undergraduate course. The outcome showed that students could write and revise their essays on-line, resulting in improved essays. Implications are discussed for the use of the technology in undergraduate courses and how it can provide an effective approach to incorporating more writing both in and outside of the classroom.

INTRODUCTION

In the process of writing and revision, students often focus on the task of improving the quality of their writing while not focusing on improving the amount of conceptual information expressed. Because the goal of many essay assignments is to show the depth of knowledge on a particular topic, it is important to be able to provide feedback to students when their essays show that they have not acquired the appropriate amount or level of knowledge. Unfortunately, scoring and commenting essays can be time consuming, often limiting the number of writing assignments that can be given, and there can be a delay of about a week before students receive back comments. By then, the students are no longer thinking about the essay topic and must reinstate much of that knowledge. Thus, it would be useful to provide immediate feedback so that students are still focused on the same topic context when they make revisions to their essays.

In this paper, we describe the development of an automated system for scoring and commenting on the conceptual quality of essays used in an

intermediate level undergraduate class. Next, the effectiveness of the scoring is then compared to a variety of human-derived measures of the quality of the essays in order to evaluate what features of the essays are actually being scored by the computer. Finally, we discuss the implementation and testing of the automated scoring system in an undergraduate class.

Essays have long been used in educational settings for assessment as well as a tool for improving student learning. Essays can more accurately capture a student's current knowledge representation because they require the creation of the answer by the student rather than just a choice of answers provided by an instructor. As a student's knowledge representation changes, so too should his or her essay response. These changes can include expression of more knowledge, change in the expression to match more closely that of an expert, and improvement in overall writing quality. Because writing requires more recall and organization of information than other means of question answering (such as multiple choice or fill-in-the-blank questions with limited vocabulary choices), it also can induce a richer knowledge representation in a student.

Errors that occur on surface features of an essay, such as spelling, punctuation, and grammar, are easily identified by students. Indeed, students are trained early in spelling, punctuation, and grammar and taught to proofread for these errors in their essays. However, identifying missing information, incorrect knowledge, and misconceptions is much more difficult to do in essays. A student's essay may be incomplete because he or she has not acquired sufficient knowledge of the essay topic to fully cover the topic. In this case the student may not know what information would be appropriate to include in the essay. Alternatively, an essay may cover information that is off-topic or incorrect, indicating that a student may believe that he or she has acquired knowledge about the essay topic, but has actually acquired incorrect information. Thus, in order to identify these types of errors, it is necessary to evaluate the conceptual content provided in the essay and compare it against what would be expected for the essay topic.

While grammar and spell checkers can provide feedback on errors in the surface features, assessing knowledge requires very different technology. Computers have been used to train and assess individuals' knowledge, and the information gained through developing these systems has improved theories of learning (e.g., Anderson et al., 1995). However, previous computer-based training systems have been limited to using multiple choice questions, entering numbers, or entering words from a limited vocabulary for assessing students' knowledge representations (Polson & Richardson, 1988; Sleeman &

Brown, 1982; Wenger, 1987; see also, Edelson, 1996; Graesser, 1985). Current advances in technology (particularly in computational linguistics) have allowed for students' knowledge to be assessed based on analyzing essay answers to open-ended questions (e.g., Foltz, 1996; Foltz et al., 1999; Landauer et al., 1998).

Latent Semantic Analysis (LSA) can provide measures of the quality of a student's knowledge representation based on analyses of a corpus of textual information on a domain and a student's essay. The technical details of LSA will not be covered in this paper, but may be found in Deerwester et al. (1990), Landauer and Dumais (1997), Landauer, Foltz and Laham (1998), as well as the introduction to this issue (Landauer & Psotka, 2000).

The basic approach of using LSA for scoring essays is that LSA is first trained on some relevant background material on the topic, such as a textbook, or articles related to the essay topic. This training results in a semantic representation of the knowledge of the topic. Using this representation, pieces of textual information (for example, words, sentences, paragraphs, or essays) can be compared against each other. This comparison provides a measure of the degree of semantic relatedness indicating the extent to which the pieces of text are discussing the topic in the same way. Using this feature, essays, or components of essays can be compared against standard pieces of text in order to determine the extent to which an essay is discussing something in a manner similar to that standard. The degree to which an essay matches that standard can then be converted into a score or comment on that essay.

In scoring essays, a distinction is often made between holistic and analytic or componential scoring. In holistic scoring of essays (e.g., White, 1993), a grader provides a single score for the essay based on an overall impression of the essay. In analytic scoring (e.g., Diederich et al., 1961), the grader provides scores for multiple components of the essays (for example, coherence, punctuation, as well as coverage of individual subtopics covered in the essay). While there is controversy as to which method is most accurate at providing a true measure of the quality of an essay, each approach has different advantages. The holistic method can typically provide a more accurate measure of the overall essay quality, while also easing the burden on the grader. However, the componential scoring method can provide specific feedback to students about what components of the essays were incorrect or needed improvement.

In this study we implemented techniques to automatically provide both a holistic score and a componential measure which gave feedback on missing components within the essays. Automatically providing a holistic score can

give students an idea of the overall quality of their essays. This can be useful to help students know how their essay stands relative to the teacher's expectations for the essay topic, but does not let them know what they would need to do to improve the essay. Thus, for automated training the componential scoring method can be additionally used to inform students where to look for additional information in their texts or notes, or to get them to think more completely about the missing concepts before writing revisions.

TESTS OF THE EFFECTIVENESS OF AN AUTOMATED GRADER

The Essay Topic

An essay topic used in a undergraduate psycholinguistics course was chosen as the target for developing the essay grader. The topic was: *Describe McClelland and Rumelhart's Interactive Activation model and how it accounts for the Word Superiority Effect.* In previous years, students in the class had shown great difficulty in mastering the concepts covered in the topic area as well as relating the model to how it could account for the Word Superiority Effect. Because this was a central concept that re-occurred in the course, it was important that students mastered the topic before they would be able to learn other information covered in the course. Thus, the topic was appropriate for evaluating students' mastery of the concepts.

There were three goals in the development and testing of the essay grader. The first was to evaluate the effectiveness of LSA for accurately scoring essays on the topic and for identifying missing subtopics in the essays. The second was to evaluate the role of the quality of writing and the quality of content expressed in the essays and the extent to which each affected human graders' scores on the essays. The third was to evaluate the essay grader in two psycholinguistics courses in order to determine the effectiveness of applying LSA as a Web-based training system.

Human Scoring of the Essays

Forty-one essays based on the question: *Describe McClelland and Rumelhart's Interactive Activation model and how it accounts for the Word Superiority Effect* were used for the initial training of the essay scoring system. These essays were collected from an in-class essay exam from an undergraduate psycholinguistics class at New Mexico State University. The average length of the essays was 106 words with a range of 22 to 227 words. The essays were initially graded

holistically by the instructor and two graduate assistants. The mean correlation among the three graders was 0.73, with a range from 0.69 to 0.75.

One criticism of LSA-based scoring has been that LSA scores are based on the semantic content only and ignore other factors attributed to the quality of writing, such as grammar, mechanics and the flow of arguments. The criticism further states that without considering these factors, LSA cannot accurately measure the true quality of an essay. Therefore, in this study we examined the relationship between scores for the content of the information in the essays versus scores for factors involving the quality of writing. This examination permits us to determine how related measures of the quality of writing are to measures of the quality of content expressed in the essays. Evidence for a strong relationship between the two should indicate that scoring for the expression of the content can accurately account for the overall quality of an essay.

To measure the quality of the writing in the 41 essays, four English graduate students with no knowledge of the topic were recruited to serve as graders. Additionally two psychology graduate students with knowledge of the topic also served as graders. The graders were instructed to grade each essay, providing two scores. The first score was for the quality of writing reflecting such factors as, grammar, syntax, flow of arguments, mechanics. Because the English graduate students had no knowledge of the topic, their scores for the quality of writing should not be biased by a student's expression of content in the essay. The second score the graders provided was a measure of the correct semantic content in the essays, reflecting whether a student was using the correct words for the essay topic. The English graders were told to just provide their best guess for this score since they were unfamiliar with the topic.

The results of the grading for these two criteria indicated that there was a strong relationship between the measures of the quality of the writing and the quality of the expression of semantic content in the essays. The average correlation across the six graders was 0.76 with a range of 0.53 to 0.91. Note that this correlation is actually greater than the average agreement between graders for providing the holistic scores. This result indicates that quality of writing and quality of semantic content tend to vary together, or at least graders tend to score essays in this consistent way. The relationship between these two measures can also be interpreted as capturing the effects of students' general abilities or *g*.

As a second measure of the relationship between quality of writing and semantic content, the scores for the quality of writing provided by the English graduate students were compared to the holistic scores initially provided by the

psychology graduate assistants and professor. The correlation for the scores for quality of writing with the average of the holistic scores was 0.63. While this is slightly below the correlation between scores of the holistic graders (0.76), it does indicate that an independent measure of the quality of writing covaries with the measure of the overall quality of the essays. Taken together, the above results indicate a strong relationship between quality of writing and quality of expression of content. Although LSA may assess primarily the quality of content, because the two are so related, it does accurately account for quality of writing indirectly.

Holistic Scoring of Essays with LSA

In order to score essays, a semantic space representing the typical knowledge relevant to the essay topic must first be created. With this space, essays and components of essays can be compared to each other in order to derive measures of their semantic similarity. The LSA space was created using the first four chapters from the textbook used for the course (*The Psychology of Language,* Carroll, 1994). The SVD analysis was performed on the 1909 sentences from those chapters and was made up of 3952 unique terms.[1] All analyses of essay scoring were performed using 300 dimensions.

Several LSA-derived measures of the quality of the essays were tested to evaluate their effectiveness. The holistic grades by the three psychologists were averaged to create a single reference grade for each essay. By using the average of the three scores, this reference grade provides a more accurate measure of the true quality of each essay. Predicted scores from each measure were then correlated to this reference grade. A summary of each measure's correlation to the individual graders, the *r*-to-*z* transformed mean[2] of the three graders, the correlation of each measure to the reference grade, and a significance test of this correlation against the average correlation among the three graders (0.73) is shown in Table 1.

As a first measure of essay quality, we can compare the semantic similarity of sentences in each essay to sentences in the original textbook (see Foltz, 1996, for other tests of this approach). If students learned from the textbook, then their writing should reflect more of the semantic information contained in the textbook. For this measure, a score was calculated from the average of

---

1. Typically LSA-based essay scoring has been developed using training documents segmented at the paragraph or larger level. Because only four chapters were available electronically, there was not enough textual information available to perform the analysis at the paragraph level (422 paragraphs). Thus, the document was segmented at the sentence level. Although paragraphs generally provide the best representation for LSA, the sentence level training for this textbook appeared to work appropriately.

2. Since correlations are not distributed in a way that they can be averaged, the *r*-to-*z* transform is used to average the three correlations of LSA to the humans. Each correlation is transformed to a *z*, the three *z*s are then averaged and this average is then converted back to an *r*.

Table 1.   Correlation of LSA Grading Methods to Individual Graders, the r-to-z Mean Correlation of the Three Graders, the Correlation of the Methods to the Reference Grade, and Z Score and Significance Test for Comparing the LSA Correlation to the Reference Grade to the Mean Correlation of the Three Graders.

	Professor	Grad TA1	Grad TA2	Mean of the correlations of LSA to the 3 graders	LSA correlation to Reference grade	Z of LSA corr. to reference grade compared to mean corr. of the three graders (0.73)	p
1. Similarity of sentences from essay to sentences from textbook	0.54	0.51	0.45	0.50	0.56	1.29	0.169
2. Similarity of sentences from essay to sentences in text that teaching assistant considers important	0.58	0.69	0.67	0.65	0.71	0.181	0.822
3. Vector length	0.66	0.69	0.73	0.69	0.76	−0.294	0.715
4. Semantic similarity of essay to other similarly graded essays (holistic)	0.86	0.73	0.69	0.77	0.85	−1.427	0.110
5. Combined methods 3 & 4	0.86	0.78	0.77	0.80	0.89	−2.150	0.027

Note: the mean correlation among the three graders was 0.73 and there were 41 essays.

the cosines between each sentence in an essay and the most similar sentence in the textbook. Thus, if each sentence in an essay were exactly similar to a sentence in the textbook (e.g., plagiarism), the essay would receive a high score. Note: because students took this exam without access to their books, their essays would have to reflect their recall of the text. The correlation between this measure and the reference grade was 0.56, and the r-to-z transformed mean correlation of the measure to the graders was 0.50. While this measure does not indicate whether students are writing sentences that are appropriate to the topic, it does indicate whether they have acquired some of the conceptual knowledge reflected in the text.

The second measure of essay quality also measures the similarity of sentences in the essays to sentences in the textbook, however it also measures whether the sentences were appropriate to the topic (see also Foltz, 1996). After scoring the essays, the two graduate assistants chose ten sentences from the textbook that were "the most important sentences that a student should know if they were to write on the essay topic". Then a score was derived by averaging the cosine similarity score of each these ten sentences to the most similar sentence in the essay. Thus students who wrote essays containing the semantic content that the graduate assistants thought was important would receive the highest scores. The correlation of this measure to the reference grade was 0.71, and the r-to-z transformed mean correlation of the measure to the graders was 0.65. This measure indicates that LSA can successfully compare sentences in essays against identified components in the text and determine the degree to which an essay contains those components.

The third measure of essay quality was the vector length of each essay. The vector length accounts for the quantity of domain relevant information contained in each essay (e.g., Laham, 1998). Vector length takes into account both the number of words used and whether the words were high content (low frequency terms), which typically indicates greater amount of information conveyed. Unlike the other measures described, it does not compare the essays against other texts. The correlation of this measure to the reference grade was 0.76, and the r-to-z transformed mean correlation of the measure to the graders was 0.69. Thus, accounting for the amount of topic relevant information contained in the essay is a good predictor of the quality.

In the fourth measure of essay quality, a score is derived by comparing each essay against a set of pre-scored essays. In this case, with 41 essays, each essay was compared against the remaining 40 essays from this trial. Each essay was assigned a score by first determining the five most semantically similar essays

to it. Then a score was assigned that was the average of the reference grades for those five essays weighted by the cosine of their similarity to the essay (see Laham, 1998). This approach is most similar to a holistic measure because it assesses the quality of an essay as a whole based on how similar it is to other essays that have been holistically scored. Thus, an essay most similar to other essays that received scores around 90 would also receive a score around 90. The correlation of this measure to the reference grade was 0.85, and the *r*-to-*z* transformed mean correlation of the measure to the graders was 0.77. Thus, this measure correlates with the average of the graders as well as the graders correlated with each other.

While the third measure provides a characterization of the quantity of information, the fourth measure characterizes the quality of the essay, as measured against the quality of other pre-scored essays. Thus, these two measures are considered to be assessing different aspects of an essay. Indeed, combining the third and fourth measures using regression to determine the weights allocated to each measure shows that the two measures do assess independent factors in the quality of the essays. Scores for the essays derived by combining these two measures had a correlation of 0.89 with the reference grade, and the *r*-to-*z* transformed mean correlation of the measure to the graders was 0.80. Thus, this final measure correlated with the human graders better than the human graders correlated with each other ($Z = 2.15$, $p < .05$).

In summary, the four measures account well for the quality of the essays. Indeed, the final two measures correlate with graders as well or better than human graders correlated with each other. Nevertheless, each of these measures only provides an overall score for the essay and therefore can only give limited feedback to students about whether something is wrong with their essays. A componential scoring system, however, can be used to identify whether an essay has covered particular subtopics and this information can be provided back to the student.

Componential Scoring

To develop the componential scoring, first a set of components, or subtopics that should be covered in the essay, needed to be identified. The two graduate teaching assistants, who had performed the essay scoring, analyzed the topic and generated a list of seven subtopics that would be critical pieces of information for the topic. Then, for each subtopic, they indicated one to three sample sentences from the textbook or the essays they had graded that would be good examples of each component.

To score the essays with LSA, each sample sentence of a subtopic was compared against all of the sentences in an essay. If one of the sentences in the essay was sufficiently similar to a sample sentence then it was considered that a student had sufficiently covered that particular subtopic (see Foltz, 1996, and Foltz, Kendall & Gilliam, 2000, for additional details on techniques for matching subcomponents). By comparing all the sample sentences against the sentences in an essay, it permits a determination of which subtopics were mentioned and not mentioned in the essay. For any subtopic not covered, feedback can then be provided to the student to permit him or her to know how to revise the essay.

As a measure of the effectiveness of this approach to detecting subtopics, the number of subtopics identified by LSA for each essay was correlated with the holistic reference grade. In this case, a subtopic was considered covered if any of its sample sentences matched a sentence in the essay with a cosine of 0.4 or greater. The correlation of this measure to the reference grade was 0.78, indicating that the number of subtopics covered in an essay is a good indicator of the quality of the essay. In order to determine whether LSA was able to correctly identify which of the subtopics were covered or not covered, the two graduate assistants re-graded the initial 41 essays, providing a score for each subtopic on a scale of 1 to 5, with 5 representing that the subtopic was completely covered. A score for each subtopic was then derived using LSA by returning the average cosine of the three most semantically similar sentences in an essay to any of the sample sentences for the subtopic. By averaging across three sentences, the measure takes into account whether a student has written multiple sentences on a particular subtopic.

Table 2 summarizes the correlation for each subtopic between the two graduate assistants and between the LSA measure and the average subtopic score given by the graduate assistants. The results indicate that the correlations for

Table 2. Correlations of Ratings of Coverage of the Seven Subtopics.

	Topic						
	1	2	3	4	5	6	7
Human-human	0.71	0.88	0.74	0.67	0.59	0.76	0.41
LSA-average of humans	0.09	0.71	0.63	0.37	0.36	0.36	0.18
Z	3.474**	2.130*	0.911	1.841*	1.311*	2.700**	1.106

$* p < .05 ** p < .01$

the human-human ratings were significantly better for five of the 7 subtopics than LSA's correlations with the humans. While LSA's correlations are not as strong as the human-human correlations, they do indicate that the measure does detect the presence of subtopics moderately well in that the correlations are in the right direction. For subtopics one and seven, in which LSA performed fairly poorly, there was only one sample sentence on which to match for each of those subtopics. Thus, there may not have been enough examples of how a student may correctly write on that subtopic for LSA to be able to match. This suggests that the measure would likely be more effective if one were to generate additional sample sentences for each subtopic.

BUILDING AN AUTOMATIC ESSAY GRADER/CRITIC FOR THE CLASSROOM

The above results indicate that the technique can provide accurate overall scores on essays as well as a moderately accurate determination of which subtopics are sufficiently covered within each essay. These techniques were then used to develop a Web-based automatic essay grader/critic for this essay topic. The system permitted students to write their essays on a Web page and then submit them. Within 15 seconds, they would receive feedback with an overall score for their essays along with comments on subtopics that were missing from their essays. Students could then revise their essays and resubmit them.

The overall score for a student's essay was determined by using the combined holistic and vector length measure using the scores from the 41 essays that had been previously collected and tested. To provide comments to the students, a question or comment was created for each subtopic. When a student's essay did not match against a subtopic, a question or comment could be returned to the student. For example, if there appeared to be no sentences in an essay discussing the fact that processing in the Interactive Activation Model occurs at three levels, or that there was no mention of the three levels, then the question: *The processing occurs at three levels. What are they?* was returned to the student. Links were also provided to relevant pages of the professor's course notes, so that students could look up additional information as they were writing their essays. The computer logged all interactions with the students, keeping track of their essays, the scores and comments given on the essays and the number of revisions a student did. An example of the feedback

on an essay is shown in Figure 1. A demonstration of the grader may also be seen at: *http://psych.nmsu.edu/essay.*

The grader was then tested on undergraduates in two psycholinguistics courses. The students were given the assignment as a required take-home assignment. Thus, they could work on their essays at any time. They were allowed to use reference materials such as books and class notes in the writing of their essays. Students were instructed to write their essay on the Web page, submit it, and then, based on the comments and grade, they could revise and resubmit the essay. They were permitted to revise their essays as many times as they wanted and were told that they would receive the maximum grade that the grader gave them. In addition, students were told that if they
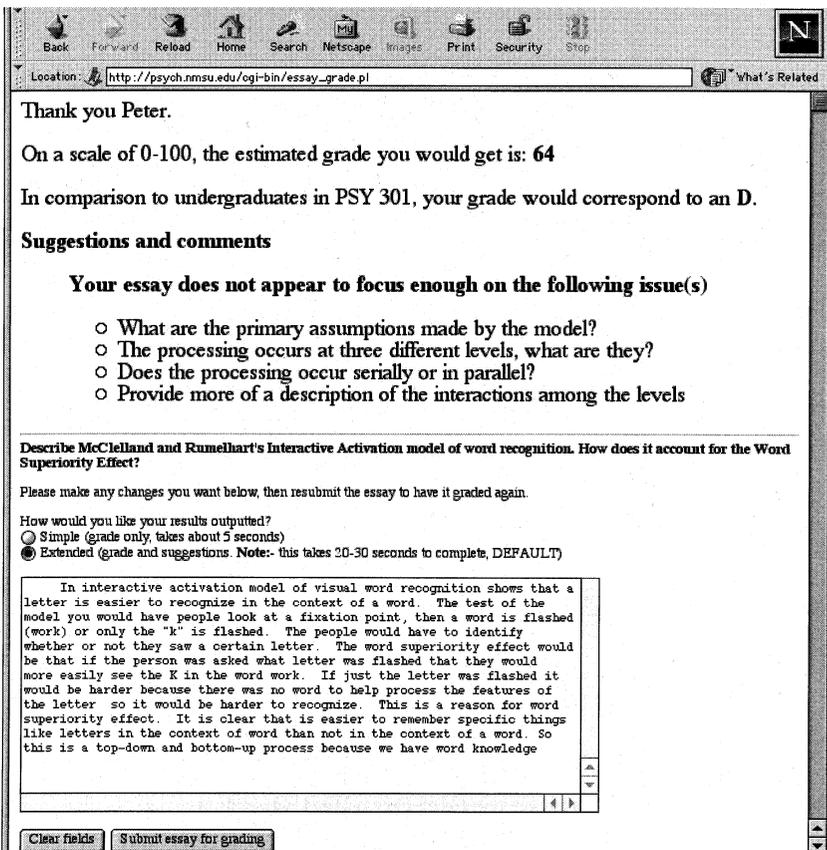

Fig. 1.   Example feedback Web page from the essay.

were unhappy with the grade given by the computer, they could turn in their essay to the professor to be graded by hand. No student exercised this option. This result could be interpreted in several ways: students trusted or liked the score provided by the computer, or possibly students thought that the professor would grade more strictly.

Over the two courses, a total of 40 students used the grader. All 40 revised their essays at least once. The average number of revisions was 3, with a range from 1 to 12. The average grade given by the grader (out of 100 possible points) on the first revision was an 85, while the average grade on the last revision was a 92. The largest improvement from first to last revision was 33 points.

The fact that students scored an average of 85 on their essay was initially surprising considering that students typically have difficulty with the topic. After the assignment, students were asked about why they may have done so well on their first essay. Several students noted that because they were unsure about how the grader would score their essays, they spent a lot of time working on the first essay they submitted. Thus, an indirect effect of using the technology was that students may have worked harder because they were unsure of the expectations of the computer as compared to the expectations of the professor!

Because the computer provided an overall score for the essays, it is important to verify whether the computer was accurately scoring the essays that the students submitted. It should be noted that to score the essays, the computer was basing its grade on the 41 original essays which had been collected in an in-class exam. Because the students who used the Web-based grader were writing their essays as a take-home assignment, they might have written very different essays which would not have been scored as accurately. To test this, 25 of the submitted essays were randomly selected and were graded by one of the graduate assistants who was blind to LSA's grading. The correlation of her grades to the grades given by LSA was 0.75, well within the range of the human-human correlations for the original 41 essays. This result indicates that LSA was scoring these new essays as accurately as it had scored the original essays.

After the assignment, students in the class were given a survey asking for their reaction to the system. Students were asked *Would you use this system if it were available?* 30 students reported that they would definitely use such a system, 9 said they would probably use such a system, and one student said that he would probably not use the system. The one student who said he would not use the system wrote further in his comments that he did not need such a

system since he could already write well. Additional comments on the survey were very positive with students wishing to have such a system for their other courses and for future essay topics in the class. Several of the students further reported that the comments/questions on the subtopics helped them determine what additional information they needed to study. Overall, students liked the idea of receiving immediate feedback on their essays and felt like it was a very useful in helping them identify problems in their essays.

DISCUSSION

Overall, the results from this study show that LSA-based measures can be used to provide accurate judgements of the quality of essays and can be applied in classroom settings for helping students write essays. Although some of the measures tested in this study have been tested in other studies (Foltz, 1996; Foltz et al., 1999; Laham, 1998) these results extend and refine the basic findings by showing that LSA-based grading can be used in a variety of domains and for a wide range of student ability levels.

Evaluating Essays with LSA

One of the goals of this research was to show that we can use methods to derive overall scores for essays as well as scores for the quality of the sub-components of essays. In tests of the measures which provided a single score for each essay we showed that there are a variety of approaches to measuring the quality of an essay relative to a standard, such as sentences in the original text, sentences that a graduate assistant thinks are important, or essays that have been scored previously. The best of these approaches can score the essays as accurately as the human graders for the same class.

For evaluating the components of the essays, one method was shown which compares sentences in essays against sentences that are good examples of particular subtopic knowledge. While the measure did not correlate as well with human graders as the human graders correlated with each other, it did show that it could account for the number of subtopics in an essay and still was fairly accurate for determining which subtopics were present or missing.

One factor involved in properly identifying the subtopics is to provide enough appropriate sample sentences of that subtopic. For this study, more sample sentences may have needed to have been provided thus so some subtopics may not have been identified appropriately. Second, LSA generally

works better at matching larger units of text since some sentences may have not contained enough content to permit them to match. Combining sentences to create a more complete example of the required subtopic may provide better measures for each subtopic. Finally, using a fixed cutoff value in which the computer only returned a comment if no sentence matched with a cosine of 0.4 or greater may not have been optimal. It may be that certain subtopics need more strict criteria for matching, while others could use looser criteria. Further tests of ways of improving the identification of subtopics are currently underway (Foltz et al., in preparation).

By independently measuring the quality of writing and the quality of content, we have shown that the two are strongly linked. Therefore, one can evaluate the quality of content and still account for much of the variance in the quality of the writing. This result indicates that students who do have more knowledge about the domain tend to have better quality of writing, while those who do not, show deficits in their writing quality. Although we can speculate about the many possible ways these two factors are linked, in this case, it is just important to show that measuring the content of knowledge is sufficient for characterizing the overall quality of an essay.

Issues for Developing Automatic Essay Critics

The classroom test of the essay grader/critic was successful at showing that students could improve their essays through interactive iterations with the computer. Combining both an overall score and comments on missing subtopics appeared to provide the appropriate feedback to allow students to write better essays. Nevertheless, no control group was used in this test. Therefore, we do not know what features of the interface permitted students to improve their essays, or even whether it was just the fact that they had to write (i.e., do more cycles of revising) that caused the improvement (see Kintsch et al., this issue, for a similar study which showed success when comparing an essay grader against other forms of writing practice).

It is important still to determine how the different forms of feedback from an essay critic can help students. Providing information about missing pieces of information may help students become more aware of their own level of knowledge as well as helping them know where to look in a textbook or other sources for information. Thus, these techniques can potentially help improve students' knowledge. Additionally, providing feedback on content may still help students improve the quality of their writing. Since students are able to receive feedback and make revisions in a very short time, they receive much

more writing practice. Thus, an open issue is the extent to which using such a system can actually improve writing skill. Further tests are currently under-way which compare writing with different forms of feedback and no feedback in order to determine what components of the essay grader/critic aid in improving writing and students' knowledge.

Developing an automatic grader/critic still requires a fair bit of initial work. A textbook, or some relevant background training material in electronic form, must first be acquired in order to train LSA. Second, a teacher must then indi-cate some form of standard by which to judge the essays. This standard could include grading a subset of essays, indicating relevant passages in the text-book, or dividing an essay topic into subtopics and identifying appropriate sample sentences and appropriate feedback for each subtopic. The feedback given to the students may then need to be fine-tuned so that it is given at the appropriate times. However, as more of these graders are developed, many of these tasks can be automated or streamlined.

The essay grader/critic as a Web-based application was simple to integrate into an undergraduate class. Because the students were allowed to work on the assignment in their own time, there was no need to have a large classroom of computers. Students could use any computer on campus, or at home to com-plete their work. There were also no reports of problems in using the system. Since most students had used the Web before, they had little difficulty enter-ing essays, making revisions, and understanding the feedback provided.

Overall, the results of this study show that LSA is an effective tool for scor-ing and commenting essays. The scores it provided proved to be as accurate as those of the human graders, yet it was able to provide almost instantaneous feedback to students, thereby permitting them to make multiple revisions in a short period of time. Because the technology can work in any content-based area in which we have electronic texts, it has the potential to be applied to a wide range of domains. Thus, this technology is well suited for use in under-graduate courses and provides an effective approach to incorporating more writing both in and outside of the classroom.

REFERENCES

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences, 4*, 167–207.
Carroll, D. (1994). *Psychology of language*. Pacific Grove, CA: Brooks/Cole.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, *41* (6), 391–407.

Diederich, P. B., French, J. W., & Carlton, S. T. (1961). *Factors in judgements of writing ability* (Research Bulletin RB 61–15). Princeton, NJ: Educational Testing Services.

Edelson, D. C. (1996). Learning from cases and questions: The Socratic case-based teaching architecture. *The Journal of the Learning Sciences, 5*(4), 357–410.

Foltz, P. W. (1996). Latent Semantic Analysis for text-based research. *Behavior Research Methods, Instruments and Computers, 28,* 197–202.

Foltz, P. W., Kendall, S., & Gilliam, S. (in preparation). Technical issues for developing automated essay scoring and commenting.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Education Journal, 1* (2).

Foltz, P. W., Kendall, S., & Cilliam, S. (in preparation). Technical aspects of scoring essays with Latent Semantic Analysis. Manuscript under preparation.

Graesser, A. C. (1985). An introduction to the study of questioning. In A.C Graesser and J.B. Black (Eds.), *The psychology of questions.* (pp. 1–14). Hillsdale, NJ: LEA.

Kintsch, E, Steinhart, D., Stahl, G, Matthews, C, & Lamb, R. (2000). Developing summarization skills through the use of LSA-based feedback. *Interactive Learning Environments*, *8*(2), pp. 87–109.

Laham, R. D. (1998). *Automated holistic scoring of the quality of content in directed Student essays through Latent Semantic Analysis.* Unpublished Master's thesis. University of Colorado, Boulder.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 211–240.

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

Landauer, T. K., & Psotka, J. (2000). Simulating text understanding for educational applications with Latent Semantic Analysis: Introduction to LSA, *Interactive Learning Environments*, *8*(2), pp. 73–86.

Polson, M. C., & Richardson, J. J. (1988). *Foundations of intelligent tutoring systems.* Hillsdale, NJ: LEA.

Sleeman, D., & Brown, J. S. (1982). *Intelligent tutoring systems.* NY: Academic Press.

Wenger, E. (1987). *Artificial intelligence and tutoring systems.* Los Altos, CA: Morgan Kauffman.

White, E. M. (1993). Holistic scoring: Past triumphs, future challenges. In M. M. Williamson & B. A. Huot (Eds), *Validation of holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 79–108). Cresskill, NJ: Hampton Press.

Wolfe, M. B., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). Learning from text: Matching readers and texts by Latent Semantic Analysis. *Discourse Processes, 25*, 309–336.