Automatic Assessment of the Content of Essays Based on Course Materials

Tuomo Kakkonen, Erkki Sutinen
Department of Computer Science
University of Joensuu
Joensuu, Finland
{tkakkone, sutinen}@cs.joensuu.fi

Abstract **Assessing free-text responses, like essays, is a demanding task even for a human. To reduce the costs of essay grading, methods to automate the assessment process have been developed. Approaches based on surface features and content have been introduced. We present a method for evaluating the content of essays based on Latent Semantic Analysis (LSA), which was originally developed for information retrieval and has successfully been applied for assessment of essays in several applications. We use LSA to compare the conceptual similarity between the essays and selected text passages from the course material covering the essay assignment-specific subject matter. In our experiment the correlation between the scores given by the system and human grader varied from 0.78 to 0.82.**

***Index Terms*— Computer-assisted essay assessment, Latent Semantic Analysis**

I. INTRODUCTION

The need for computer-assisted assessment of learning outcomes is connected to two inter-related factors in today's schooling and education markets. First, teachers need to automate the assessment and evaluation process especially in mass courses. Secondly, a student, particularly when following an online course, may want to assess the degree of his or her own learning process prior to an examination. Evaluation is a broad concept which covers both formal and informal feedback, carried out either explicitly or implicitly. In this paper, we use the term assessment in reference to formal evaluation (i.e. measuring the learning outcomes with a numerical grade).

Many of the computerized evaluation systems do not fully utilize the potential of available technology. Too often, teachers and students have to be satisfied with automated multiple choice questions. Essay assignments, compared with multiple choice and selection tasks, have many advantages [1]. Written responses require students to generate answers which demonstrate higher order thinking skills such as synthesis and analysis.

In this paper, we describe an essay assessment system based on a commonly known information retrieval technique, Latent Semantic Analysis (LSA). In addition to LSA, the features of the essay assessment system to be introduced are as follows. The system preprocesses the essays using a morphological parser. The grade is computed by using both human-graded essays and assignment-representative text from a textbook. We start by giving an overview to the approaches developed to computerize the assessment of essays. Section III describes LSA as method. Section IV introduces the architecture and implementation our essay grading environment. Section V analyzes the usability and accuracy of the method based on experiments. Section VI presents a conclusion of the findings and outlines future opportunities.

II. BACKGROUND

While assessing the quality of students' essays, humans are prone to several types of errors, or "rater effects" [2]. Such errors include the halo effect and stereotyping. The *halo effect* refers to a situation where the assessor's decision is influenced by the earlier impression he/she has of the student rather than on the actual performance in the test that is being assessed. In *stereotyping*, judgements about the quality of the work are based on the impression of the assessor about the group (e.g. ethnic or gender) that the person whose work is being assessed belongs to.

Automatic grading of essays is substantially more demanding and expensive than grading multiple choice or other selection tasks. Automated assessment systems must preserve the benefits of written responses, be able to perform as accurately as human raters, increase the phase of assessment and reduce grading-related costs. In addition to reduced costs, automated assessment of essays can help achieve better accuracy and objectivity [1, 3]. Since computers can grade essays more rapidly than humans, essay writers can get instant feedback. An automated assessment system is not affected by errors caused by lack of consistency, fatigue or bias.

Research to automate the grading of essays has been going on since the 1960's. Several models have been developed. The best known are Project Essay Grade (PEG) [3], e-rater [4] and approaches based on Latent Semantic Analysis [5, 6, 7]. Two ideas common to all these methods is that they assume human ratings to be the best estimate of the true quality of essays and that computers cannot independently determine the quality. These systems must use essays scored by humans for creating assignment-specific scoring models.

The use of computers in such a highly demanding task as assessment of essays has raised several questions and has even generated quite significant opposition [3]. The earliest approaches, especially PEG, were based solely on the surface characteristics of the essay such as the length in words and the numbers of commas [8]. Despite impressive results wide-range acceptance was not achieved in the education community [9]. One of the main concerns was that scoring essays with such simple, indirect measures would have a leveling influence thereby eliminating creativity [10].

To overcome such criticisms more recent studies have focused on developing measures of essay quality which are more direct.

Our approach is based on LSA, which is a corpus-based statistical method. It provides a means of comparing the semantic similarity between a source and target text. LSA has been successfully applied to automated assessment and feedback of free-text responses in several systems; *Intelligent Essay Assessor* [11] and *Select-a-Kibitzer* [7] apply LSA for assessing essays written in English. In *Apex* [6] LSA is applied to essays written in French. Other applications of LSA to educational technology include an intelligent tutoring system for providing help to students [12] and *Summary Street* [13], which is a system for assessing summaries.

III. Latent Semantic Analysis

The basic assumption behind LSA is that there is a close relationship between the meaning of a text and the words in that text. LSA provides a method for determining the similarity of the meaning of words and text passages. The power of LSA lies in the fact that it is able to extract the meaning of words and text passages starting from word co-occurrence data without need of human intervention, for example construction of logical rules. The LSA method is often able to detect the similarity between two texts even when they do not contain common words.

While LSA is far from a complete model of human understanding, it is powerful enough for many applications. Research has shown that LSA is able to simulate learning and several other psycholinguistic phenomena [15]. In addition to information retrieval and essay assessment, LSA has been shown to simulate human learner behavior in several other applications such as providing synonyms and taking multiple choice tests.

We will shortly describe the technical details of the LSA method. A more detailed description of LSA may be found in [14] and [15].

LSA represents words and passages in a ”semantic space”. First, the text is presented as a word-by-context matrix, in which rows stand for unique words and columns for *contexts* where the words occur. A context can be for example a sentence, a paragraph or a whole document. In word-by-context matrix M, each cell M_{ij} contains the number of times the word i occurred in the document j. For example, the size of a matrix built from 100 text passages and 500 words is 500 x 100. Only the words that appear in at least two contexts are represented in the matrix [15]. Standard preprocessing methods in LSA include typical information retrieval techniques stemming, term weighting, and the use of a stopword list. Stopwords are the most commonly occurring words and are not included in the matrix. In term weighting, entries in the word-by-context matrix are transformed so that they better represent the importance of each word. The aim is to give higher values to words that are more important for the content and lower values to those with less importance. In LSA, *log-entropy* weighting, defined in (1), is often used [15].

$$M_{ij} = \frac{\log\left(freq_{ij}+1\right)}{-\sum_{j=1}^{n}\left(\left(\frac{freq_{ij}}{\sum_{l=1}^{n} freq_{il}}\right) * \log\left(\frac{freq_{ij}}{\sum_{l=1}^{n} freq_{il}}\right)\right)} \quad (1)$$

The value of each cell in the word-by-context matrix M is weighted according to (1) by dividing the local weight of the word by its global weight. The local weight of the word is the logarithm of $freq_{ij}$, the word frequency in cell (i,j), plus 1. The global weight is defined by the entropy of the word in all n documents in the document collection.

The essence of LSA is dimension reduction based on the singular value decomposition. Singular value decomposition is a form of factor analysis, which reduces the dimensionality of the original word-by-context matrix and thereby increases the dependency between contexts and words [15]. Singular value decomposition is defined as $X = T_0 S_0 D_0^T$, where X is the weighted word-by-context matrix and T_0 and D_0 are orthonormal matrices representing the words and the contexts. S_0 is a matrix with scaling values. X can be decomposed to the product of the matrices T_0, S_0, D_0^T. Fig. 1 illustrates the idea of dimension reduction.

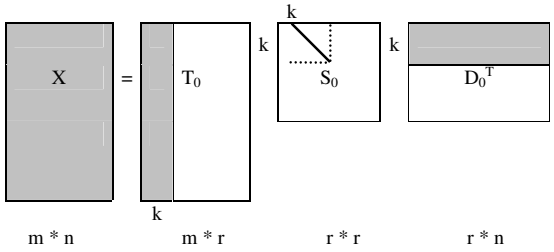

Figure 1. Singular value decomposition.

In Fig. 1, an example of reduction to k dimensions is shown. M and n are the number of words and contexts, respectively, and r the rank of word-by-context matrix X. When the matrices T_0, S_0 and D_0^T are multiplied so that only k largest values of S_0 and corresponding k columns of T_0 and k rows of D_0^T are included, the resulting reduced matrix is the k-dimensional representation of the original word-by-context matrix. In the reduced-dimensional vector space documents are not represented as sets of independent words, but as “continuos values on each of the k orthogonal indexing dimensions” [14].

The aim of the dimension reduction step is to reduce “noise” or unimportant details and to allow the underlying semantic structure to become evident [14]. On the other hand, if the number of dimension is too small, not enough information will be preserved. Since there is no dimensionality that will work in all cases, the selection of number of dimensions must be done empirically. by selecting the one that produces the most accurate results.

IV. Architecture and Implementation

Compared with other assessment methods, LSA has several advantages. First, it focuses on the conceptual content of the essay, not the surface features or content based simply on keyword frequencies. The second

advantage is that LSA-based analysis, compared to for example PEG and e-rater, needs fewer pre-scored essays. PEG and e-rater typically need several hundred essays to be able to form an assignment-specific model [16, 17] whereas LSA is reported to be calibrated with no more than 20 essays [9]. Thirdly, in addition to scoring based on the comparison to the human-scored essays, LSA can base its scores on the comparison between the essays to be graded and domain-representative text (e.g., textbook) [5]. Our approach relies on the latter method.

In our approach, we use parts of relevant learning materials, like parts of a textbook, to train the system with assignment-specific knowledge. The motivation for using the textbook as source for creating the semantic space comes from the assumption that a student's knowledge is usually acquired by reading the course content and that, in turn the student's knowledge can be measured as the degree of semantic similarity between the essay and the parts of the textbook covering the assignment-specific knowledge [18, 19]. Essays with contents that more closely matches the content of the course should be given a higher grade.

The approaches based on the comparison between the essays to be graded and the textbook have been introduced in [5, 6, 9, 11], but have been usually found less accurate than the method based on comparison to pre-scored essays. Our method resembles other methods which combine the use of course content and pre-scored essays. An interesting feature in our approach is that no human intervention in terms of selecting the most important sentences was used when building the semantic space for comparison. Moreover, essays are not segmented into paragraphs or divided into sentences, thus we avoid the problems due to ambiguity of sentence boundary detection [20, 21]. In our approach, the essays to be graded are not directly compared to the pre-scored essays. Pre-scored essays are used for determining the threshold values for grade categories. Fig. 2 shows the details of the grading process of our system.

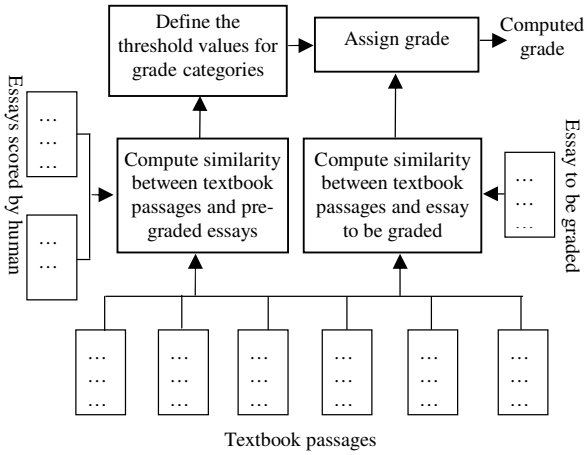

Figure 2. The grading process.

First, we create comparison materials, or a semantic space, based on a relevant textbook by constructing a word-by-context matrix representing the selected textbook passages. Thus LSA derives the meaning of the words from word frequency data, the choice of appropriate corpus for creating the semantic space plays a key role [5].

Our experiments consists of essays and a textbook written in the Finnish language. As we apply LSA to texts written in Finnish, which is a morphologically rich language where words have many different forms and suffixes, the standard way of stemming, stripping the suffixes, cannot be applied. A more sophisticated method is needed due to the grammatical complexity of the Finnish Language. We use *FINCG* [22], Constraint Grammar parser for Finnish, to perform the morphological analysis of the texts.

To compare the similarity of an essay to the textbook passages covering the assignment-specific knowledge, first a query vector of the same form as each of the vectors in the word-by-context matrix is constructed. We compare the query vector $\mathbf{X}$ representing an essay to each text passage $\mathbf{Y}_i$ of the textbook to calculate the similarity score by using the standard LSA similarity measure, the cosine of the angle ($\mathbf{X}$, $\mathbf{Y}_i$), to measure the similarity of the meaning between the documents. The similarity score for the essay is calculated as the sum of the similarity between the essay and each of the textbook passages.

By calculating the similarity scores between each pre-scored essay and the textbook chapters we can define cut-off points for each grade category based on the scores given by the human assessor. Fig. 3 shows an example of cut-off points for one of our test sets, in which essays were graded on a scale from zero to six.

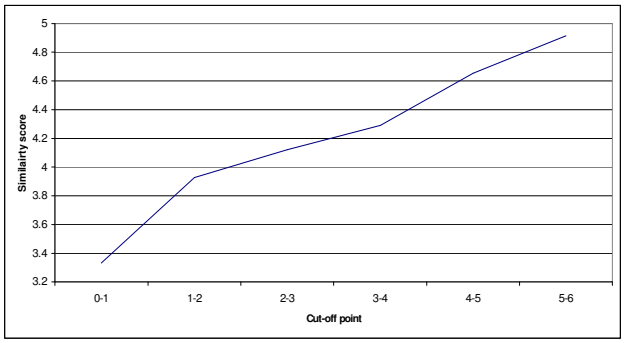

Figure 3. Example of cut-off points for grades.

Now, with grade categories defined, we can determine a grade for each essay to be graded by calculating the similarity score between the essay and the textbook passages. The similarity score is determined with exactly the same method as with the pre-scored essays. After calculating the similarity score for the essay, we simply examine which grade category the essay belongs to according to the defined threshold values.

As mentioned before, the selection of dimensionality plays a key role in LSA and is done experimentally. The aim is to find the optimal number of dimensions so that the underlying relations between words and text passages are found which leads to optimal grading accuracy as compared to grades given by humans. We ran our experiments with all possible dimensions to observe the effects of dimension reduction.

V. Experiments and Results

For the experiments we collected essays from an undergraduate course in education. The assignment was

especially suitable for our method because the assignment was referenced to a certain textbook chapter. For this experiment, 143 essays were collected. The length of the essays varied from 18 to 445 words. The essays were graded by a professor on a scale from zero to six. The grading model was created by using a relevant chapter from the course textbook. The length of the chapter was 2397 words. For the experiments, essays were randomly divided into two separate test sets. In the first experiment, 70 essays were used for defining the threshold values for grade categories and 73 essays were graded. In the second experiment, the grading model was based on 86 essays and 57 essays were graded.

We used a stopword list consisting of 27 words. For our experiments, we used log-entropy weighting and also ran the experiment without using any term weighting. The textbook chapter was divided into paragraphs and sentences for separate experiments. Table 1 shows some of the results of our experiments.

TABLE 1. THE RESULTS OF THE EXPERIMENTS.

Material for creating the scoring model	Weighting scheme	Num. of graded essays	Results		
			Exact	*Exact or adjacent*	*Correlation*
26 paragraphs and 70 essays	LE	73	39.7	83.6	0.78
144 sentences and 70 essays	LE	73	35.6	84.9	0.80
26 paragraphs and 86 essays	LE	57	36.8	77.2	0.81
144 sentences and 86 essays	LE	57	38.6	78.9	0.82
26 paragraphs and 70 essays	N	73	28.8	58.9	0.59
26 paragraphs and 86 essays	N	57	21.1	45.6	0.60

In Table 1, the first column shows the structure of the semantic space and the number of pre-scored essays used. In the column labeled "weighting scheme" LE stands for log-entropy weighting (1) and N for no term weighting. The results of the experiment are shown for the dimensionality that produced the most accurate scores. The proportion of cases where the same score was assigned by the system and human grader and the proportion of the essays where the grade given by the system was at most one point away are shown in columns “exact” and “exact or adjacent”. The last column of Table 1 shows the Spearman rank correlation between the scores given by human and the system.

The results in Table 1 show the Spearman rank correlations which vary from 0.78 to 0.82 between the grades given by the system and the human grader in the optimal dimension when entropy based weighting model was used. Compared to the model not using term weighting or dimension reduction the log-entropy weighting and dimension reduction increased the correlation between the grades given by the professor and the system by 35% to 40%.

Fig. 4 shows the effect of dimension reduction on the correlation on each possible dimension from 2 to 26 on a test run based on 26 textbook chapters.

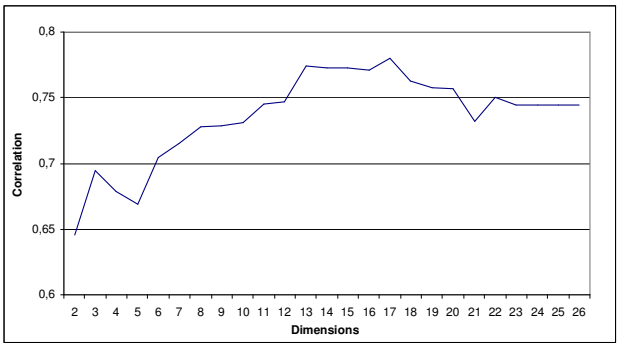

Figure 4. The effect of dimension reduction.

Fig. 4 shows a line graph the Spearman correlation between grades given by the system and the grader on each dimension from 2 to 26 when using the log-entropy weighting. The optimal number of dimensions in the case was 17.

VI. CONCLUSIONS AND FUTURE WORK

We have presented a system for grading automatically essays written in the Finnish language. Our experiments indicate that LSA can successfully be applied to Finnish and that grading based on course materials, such as textbooks, can lead to accurate results. We have also investigated the effects of different weighting methods and stopword lists in our scoring model. The Spearman rank correlation in this experiment (around 0.80) is comparable to the results achieved by the other automated assessment systems based on LSA and to those generally achieved by two human judges. For example Landauer et al. [5], Lemaire & Dessus [6] and Folz et al. [19] have reported inter-rater correlations ranging from 0.64 to 0.84 and correlations 0.59…0.89 between the LSA-based system and human graders, LSA-based usually performing at least as well as the human graders.

Future work involves improving the feedback given by the system. A final grade alone, although useful for assessment purposes, is not very helpful for the students. More detailed feedback is needed. With our current system, we are already able to determine how well the essay writer has covered each section of the textbook. This information could be presented to the students so that they would be able to identify the parts of the course contents where they need more practice. Promising results have been reported using LSA-based systems for providing feedback [6, 7, 19] and providing support for detecting plagiarism [11].

Another development challenge is the selection of dimensionality in LSA. As shown by our experiments, optimally selected dimension yielded significant increase in grading accuracy. An open question is how to automate the selection of dimensionality. Usually the dimensions between 100 and 300 are considered to be optimal in natural language domain [18]. We used only the textbook passages and not any additional text materials to create the semantic space. Our word-by-context matrix used only one chapter from the course textbook. When the chapter was divided into 26 paragraphs, analysis could be repeated with dimensions

from 2 to 26. Using 17 produced the best results.

A limitation of our method is that it only measures the content of the essay and ignores all the other aspects such as syntax and spelling. In addition, human graders also pay more attention to the content than style or mechanics. Foltz et al. [19] discovered that there is a strong relationship between the quality of the writing and the quality of the content of essays, indicating that assessing the content alone can be used to for assessing the overall quality of an essay.

Another claimed limitation of LSA-based assessment systems is that just writing a list of key concepts without a proper essay could cause the system to give an good grade which was undeserved [6]. To preventing this, the system could be augmented with syntax checker. Also, it can be argued that if a writer is able to provide such a list, he or she is likel to possess the required knowledge.

REFERENCES

[1] Chung, G. K. W. K., O'Neil, H. F.: *Methodological Approaches to Online Scoring of Essays*. CSE Technical Report 461. National Center for Research on Evaluation, Los Angeles, USA, 1997.

[2] Rudner, L. M.: Reducing Errors Due to the Use of Judges. *Practical Assessment, Research &Evaluation*, 3(3), 1992.

[3] Page, E. B., Petersen, N. S.: The computer moves into essay grading. *Phi Delta Kappan*, 76(7): 561-565, 1995.

[4] Powers, D. E., Burstein, J., Chodorow, M., Fowles, M. E., & Kukich, K.: *Comparing the validity of automated and human essay scoring (GRE No. 98-08a, ETS RR-00-10).* Princeton, NJ: Educational Testing Service, 2000.

[5] Landauer, T. K., Laham, D., Rehder, B., Schreiner, M. E.: How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. *Proceedings of the 19th annual meeting of the Cognitive Science Society*. Mawhwah, NJ: Erlbaum, 1997.

[6] Lemaire, B., Dessus, P.: A System to Assess the Semantic Content of Student Essays. *Journal of Educational Computing Research*, (24)3, 305-320, 2001.

[7] Wiemer-Hastings, P., Graesser, A. Select-a-Kibitzer: A computer tool that gives meaningful feedback on student compositions. *Interactive Learning Environments*, 149–169, 2000.

[8] Page, E. B.: The imminence of grading essays by computer. *Phi Delta Kappan*, 47(1): 238-243, 1966.

[9] Hearst, M. et al.: The Debate on Automated Essay Grading, IEEE Intelligent Systems, Trends & Controversies feature, 15(5): 22-37, 2000.

[10] Wresch, W.: The Imminence of Grading Essays by Computer – 25 years later. *Computers and composition*, 10(2): 45-58, 1993.

[11] Foltz, P.W., Laham, D., Landauer, T. K.: Automated Essay Scoring: Applications to Educational Technology. *Proceedings of the ED-MEDIA Conference*, Seattle, USA, 1999.

[12] Wiemer-Hastings, P., Wiemer-Hastings, K., Graesser, A.: Approximate natural language understanding for an intelligent tutor. *Proceedings of the 12th International Florida Artificial Intelligence Research Symposium*. Menlo Park, CA, USA, 1999.

[13] Wade-Stein, D., Kintsch, E: *Summary Street: Interactive Computer Support for Writing*. Technical Report from the Institute for Cognitive Science, University of Colorado, USA, 2003.

[14] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R.: Indexing By Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6), 391-407, 1990.

[15] Landauer, T. K, Foltz, P. W., Laham, D. An Introduction to Latent Semantic Analysis. *Discourse Processes*, 25(2&3), 259-284, 1998.

[16] Shermis, M. D., Mzumara, H. R., Olson, J., Harrington, S.: On-line Grading of Student Essays: PEG goes on the World Wide Web. *Assessment & Evaluation in Higher Education*, 26(3), p.247, 2001.

[17] Burstein, J., Marcu, D. Benefits of modularity in an automated scoring system. *Proceedings of the Workshop on Using Toolsets and Architectures to Build NLP Systems, 18th International Conference on Computational Linguistics*, Luxembourg, 2000.

[18] Foltz, P. W., Britt, M. A., Perfetti, C. A. : Reasoning from multiple texts: An automatic analysis of readers' situation models. *Proceedings of the 18th Annual Cognitive Science Conference*. Lawrence Erlbaum, USA, 1996.

[19] Foltz, P. W., Gilliam, S. & Kendall, S.: Supporting Content-based Feedback in Online Writing Evaluation with LSA. *Interactive Learning Environments,*8(2), 111-129, 2000.

[20] Grefenstette, G., Tapanainen, P.: What is a word, what is a sentence? Problems of tokenization. *Proceedings of the Third International Conference on Computational Lexicography*. Budapest, Hungary, 1994.

[21] Mikheev, A.: Tagging sentence boundaries. *Proceedings of the First Meeting of the North American Chapter of the Computational Linguistics*, Seattle, Washington. USA, 2000.

[22] Lingsoft, Inc. http://www.lingsoft.fi/.