# AUTOMATED ESSAY GRADING CONSIDERING DISSIMILARITY FROM INCORRECT RESPONSES

**Salmerón, L., Arizo, S., García, V., Vidal-Abarca, E., Gilabert, R., Ruzafa, M. & Llacer, E.**
University of Valencia
Valencia / Spain
*ladislao.salmeron@uv.es*

## Abstract

We present a method for grading essays based on Latent Semantic Analysis (LSA), a corpus-based computational method. LSA provides a means of comparing the conceptual similarity between pieces of text without any human intervention. Several systems have successfully used LSA to assess students' free-text responses to a professor's assignment. They usually require LSA to compare students' responses with a 'golden essay', which corresponds to a standard correct response elaborated by the teacher. As the semantic similarity between the student and golden essay increases, so does the grade for the student's essay. A potential problem of these systems is that they do not directly deal with incorrect ideas in the student's essay. Roughly, an incorrect idea will be managed as a low related idea. It will decrease the grade as far as it does not resemble any ideas in the golden essay, but just to the same extent as might do the inclusion of an irrelevant (although not incorrect) sentence. However, in many cases the inclusion of an incorrect idea in a student essay is considered by professors as an index of low comprehension of the topic, which is not necessarily the case with irrelevant ideas. Following this rationale we used LSA to grade student essays, based not only on how much they resemble a golden essay, but also on how much they differ from an essay containing common incorrect ideas. We collected undergraduate class assignment essays from a Psychology and a History class. Professors graded the essays and provided both a golden essay with a standard correct response, and a response including common mistakes to a particular question. Analyses reveal that the automated grade derived from the similarity between student and golden essay could be improved by considering also the dissimilarity between a student and the incorrect essay in one sample (Psychology), but it did not varied in a second sample (History). The discussion points the limitations of the method and outlines possible future improvements.

## Keywords
Latent Semantic Analysis, automated essay grading, free-text responses.

## 1. INTRODUCTION

Current educational methods at Spanish universities seek to maximize student work while reducing the amount of time devoted to traditional classes. In many cases, this new approach consists on increasing the number of student's weekly assignments, such as essays on class topics. Teachers might correct these essays every week, and provide feedback to students based on their performance. Students would then be able to assess the degree of their learning process during the course, and would be able to adapt their study to their goals prior to a final examination. However, this procedure is particularly time-demanding for teachers in mass courses. In order to help teachers on these tasks, researchers have focused on the development of automated essay grading techniques [1]. Traditionally, automated essay grading techniques have stressed the relevance of either syntactic [2] or semantic features [3] of the student's essays. Syntactic approaches are assumed to work better for shorter responses (e.g. one sentence), whereas semantic analysis may be better suited for longer essays (e.g. more than 250 words) [4]. We are concerned with relatively long essays, which are particularly time-demanding for teachers. In this paper we present a method for grading essays based on the semantic features of the responses, which has been applied to a set of two university course assignments.

## 1.1 Latent Semantic Analysis for essay grading

Our method for grading essays is based on Latent Semantic Analysis (LSA), a corpus-based computational method that is indented to extract the meaning of text pieces [5]. Generally speaking, Latent Semantic Analysis (LSA) is an automatic statistical method for representing the meaning of words and test passages. A primary method for using LSA to make predictions as to the grade of a response is to compare some units of a piece of information (sentence, paragraph, or whole essay) and an adjoining unit of text (e.g. golden essay) to determine the degree to which the two are semantically related. The basic idea behind LSA is that the contexts in which words appear and do not appear give sufficient constraints that allow one to estimate the similarity between them. Thus, LSA provides a measure of the similarity between different linguistic units. In fact, LSA permits comparison of semantic similarity between different pieces of textual information such as sentence, paragraphs, summaries as well as essays [6]. There is an emergent body of evidence supporting the reliability of LSA as a tool for evaluating the semantic relatedness between units of discourse, as well as the reliability of LSA when compared to human judgments of documents in terms of similarity. For example, LSA-generated cosines have been tested on a large number of essays over a diverse field of topics obtaining a high correlation with human assessments, as well as with students applying to college who are taking the Test of English as a Foreign Language. Likewise, LSA has been used to determine the coherence of texts. Lastly, others authors have successfully used LSA with verbal protocols and reading strategies and also in a computerized tutor called AutoTutor (a recent revision of these tools is available in [6]).

We will shortly describe the technical details of LSA. A detailed mathematical description of LSA may be found in [7]. LSA represents words and passages in a "semantic space". First, the text is presented as a word-by-documents matrix, in which rows stand for unique words and columns for documents (e.g. paragraph) where the words appear. In word-by document matrix $M$, each cell $M_{ij}$ contains the number of times the word $i$ occurred in the document $j$. 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. The key issue of LSA is dimension reduction based on the singular value decomposition (SVD). SVD is a form of factor analysis, which reduces the dimensionality of the original word-by-context matrix and thereby increases the dependency between documents and words. SVD is defined as $X = T_0 S_0 D_0^T$, where $X$ is the weighted word-by-document matrix and $T_0$ and $D_0$ are orthonormal matrices representing the words and the documents. $S_0$ is a matrix with scaling values. $X$ can be decomposed to the product of the matrices $T_0$, $S_0$, $D_0^T$ . To compare the similarity of an essay to a 'golden essay', first a query vector of the same form as each of the vectors in the word-by-documents matrix is constructed. The query vector $X$ representing an essay is compared to the golden essay $Y$ to calculate the similarity score by using the standard LSA similarity measure, the cosine of the angle $(X, Y)$.

## 1.2 Including the assessment of incorrect ideas

Traditionally, essay grading methods based on LSA compare students' responses with a 'golden essay', which corresponds to a standard correct response elaborated by the teacher, or to a section of a class text-book. As the semantic similarity between the student and golden essay increases, so does the grade for the student's essay. A potential problem of these systems is that they do not directly deal with incorrect ideas in the student's essay. For example, imagine the following ideas included in two student responses to the question "Compare the consequences of acquired and developmental dyslexia":

(1) "DSM-IV indicates several criteria for the diagnosis of dyslexia"

(2) "In developmental dyslexia symptoms disappear naturally without intervention".

In a traditional 'golden essay' system, both ideas might be considered as unrelated, because they would not appear in an ideal correct response. For example, both ideas could have a similar low cosine such as 0.2 with the 'golden essay'. Therefore, the inclusion of these two ideas will lower the student's grade only to the extent that they will decrease the global semantic relation between the student's response and the global essay. However, a professor of a 'Psychology of reading' class may treat both sentences in a student response quite differentially, because the first one is indeed unrelated to the question but correct, whereas the second is related but incorrect. In most cases, a professor would consider the inclusion of such an incorrect idea in a student essay as an index of low understanding of the topic, which is not necessarily the case with irrelevant ideas. Other things been

equal, a professor will surely give a lower grade to the students including an incorrect idea than to those including an irrelevant idea.

The impact of incorrect ideas might not be of relevance in particular contexts in which the 'golden essay' method has been used successfully, such as assignments mainly requiring a summary of a particular text-book section [4]. In that task the student is requested to extract and combine the main ideas of a text, so the errors may come mainly by either the omission of a relevant idea or by the selection and inclusion of an irrelevant one. In both cases, the LSA cosine between the student response and the 'golden essay' will decrease, capturing the student's errors. In real class context, however, class assignments may request students more complex responses that can not be answered referring to a particular part of a text-book. In that context, the possibility that students commit errors by including incorrect responses arises, because they can not rely on a reference text (which certainly only includes correct responses) to base their responses.

Following this rationale we used LSA to grade student essays, based not only on how much they resemble a golden essay, but also on how much they differ from an essay containing common incorrect ideas.

## 2. METHOD AND RESULTS

### 2.1 Materials

Two samples of class assignments were analyzed: one from a class on 'Learning difficulties' from the Psychology school, and another from a class on 'Medieval History' from the History school at the University of Valencia. The Psychology sample included responses from 65 students. Responses length was 3268 characters on average (sd = 1833). The professor responsible for the class grade the responses, which had a mean grade of 2,02 points (sd = 0,83; max = 4; min = 1). The History sample included responses from 40 students. Responses had an average length of 805,42 characters (sd = 252,17). Grades given by the professor had a mean grade of 2,74 points (sd = 1,15; max = 5; min = 0). Professors responsible for each class also provided a 'golden essay' "representing an average correct response to the assignment" and a file containing the most common incorrect ideas from the students.

Two corpora were created based on text-books and professor's class notes. The Psychology (Learning difficulties) corpus included ten chapters from five books on the class topic, drawing a total of 92,777 words. The History (Medieval history) corpus included 13 chapters from four text books, and the professor's class notes, with a total of 76,533 words.

### 2.2 Results

For each sample, we computed the LSA cosines between each student's essay and both the 'golden essay' (LSA_golden) and the 'incorrect ideas file' (LSA_incorrect). This provides an index of the semantic similarity between student's responses and both texts: in the case of the golden essay it can be considered as the student's automatic grade based on semantic overlap; whereas in the case of the incorrect ideas it might be viewed as a negative grade. We then combined these measures in order to have a compound grade following the formula: $LSA\_compound = (LSA\_golden)^2 - (LSA\_incorrect)$. We expected that professors will give more weight in their grades to the inclusion of correct ideas (positive) than to the inclusion of incorrect ideas (negative). Data obtained from these analyses was then correlated to the professor's grade, in order to have an external criterion to validate the results from LSA. It has to be noted that several findings have report the correlation in grade attribution between two human experts to be located between 0,6 and 0,7 (cf. e.g. [8]). That is to say, grades given by two professors to the same sample of responses usually only agree on 60-70% of the cases. Therefore, (ideally) the same results might be expected for the correlation between LSA – professor's grades. Analyses were performed with the corresponding corpus for each class assignments.

For the Psychology sample, Pearson correlation between LSA_grade and Professor_grade was 0,58 (Fig 1). In addition, the correlation between the compound automatic grade (LSA_compound) and Professor_grade was 0,66 (Fig 2). Therefore, taking into consideration the similarity of the student's responses to a group of common incorrect ideas boosted machine-professor grade agreement in 7% .

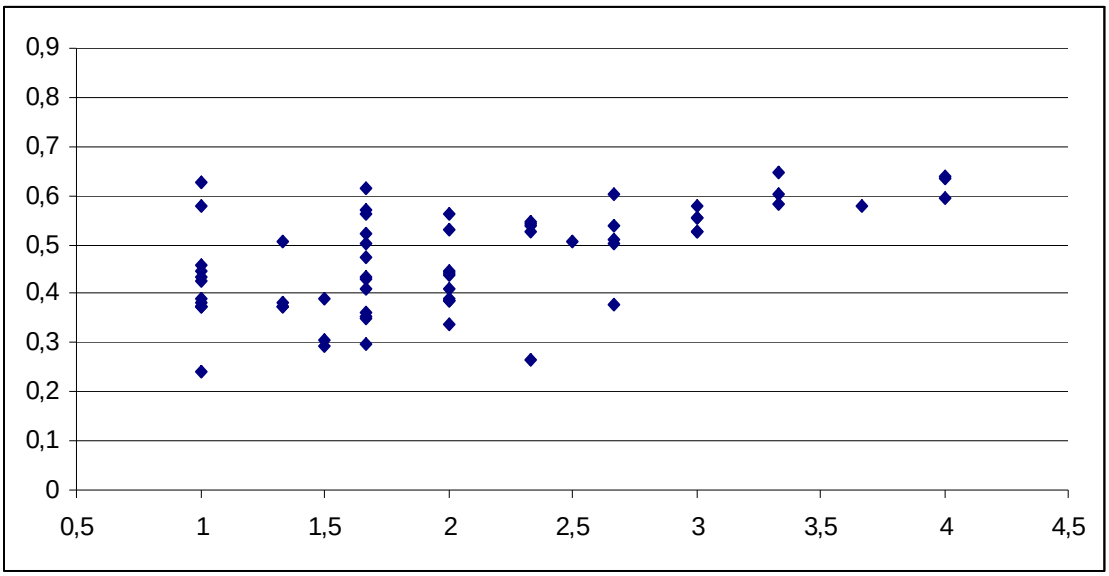


Fig 1, Pearson correlation between professor's grade (min = 1, max = 4) and LSA_golden cosine (min = ~0, max = 1).

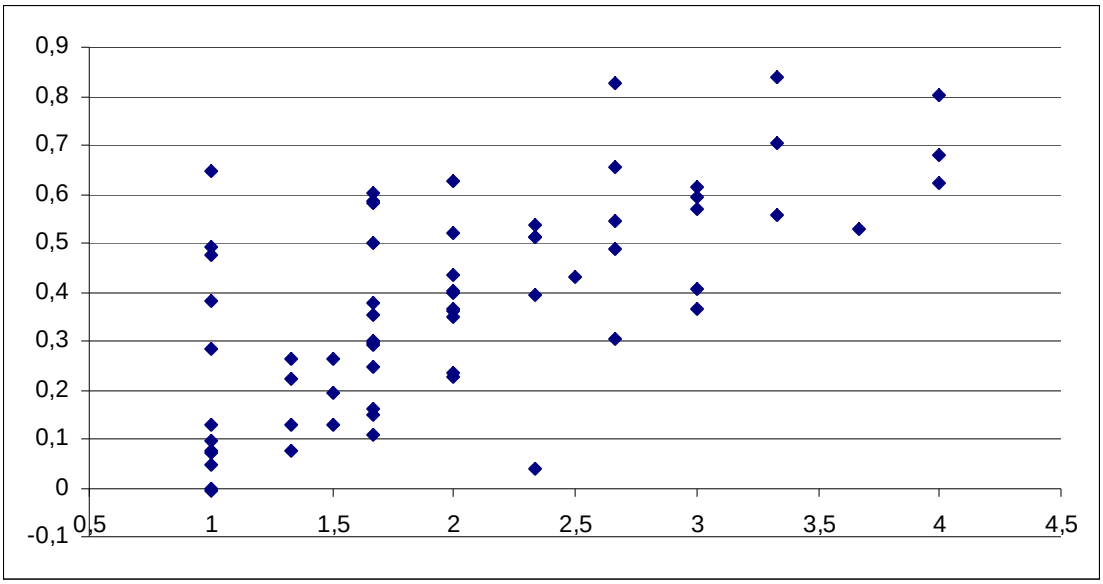


Fig 2, Pearson correlation between professor's grade (min = 1, max = 4) and LSA_compound cosine (min = ~0, max = 1).

For the History sample, Pearson correlation between LSA_grade and Professor_grade was 0,49. Contrary to our expectations and to the data obtained with the Psychology sample, the correlation between the compound automatic grade (LSA_compound) and Professor_grade decreased (r = 0,47).

## 3. DISCUSSION

In the present paper we described a method for automated grading essays is based on LSA. The new method graded student's essays not only on their similarity to a 'golden essay', but also on the dissimilarity of the responses to common incorrect ideas. Analyses from two sample of essays from different disciplines (Psychology and History) revealed mixed results. In one sample (Psychology), considering the dissimilarity to common incorrect ideas boosted machine-professor grade agreement from 0,58 to 0,66, a relevant improvement considering that agreement between two human experts

ranges between 0,6 and 0,7 [8]. However, in a second sample (History), the new method did not improved the traditional 'golden essay' method (r= 0,47 and 0,49 respectively). These results indicate that the assessment of incorrect ideas might have some potential for improving automated essay grading, but a deeper understanding of the assessment process is needed before drawing any strong conclusions.

A qualitative analysis of the files containing the incorrect ideas written by professors may help partially clarify the results. Many of the incorrect ideas for the Psychology assignment included sentences with terms that were not included in the "golden essay". In terms of LSA analyses, that represents a situation in which the golden essay and the incorrect ideas are represented by two distant vectors. In this case, it might be easy to differentiate to which extent a new vector (i.e. student's response) is distant from each of the two key vectors (golden essay and incorrect ideas). By contrast, the incorrect ideas for the History sample follow a different pattern. Many of the incorrect ideas consisted on misnaming a protagonist. The rest of the response could be correct, but the misnaming invalidates the entire response. This type of error is difficult to handle by LSA, because the vector representing the incorrect response will be identical to part of the 'golden essay' except for one term (the name of the protagonist). In this case, both vectors will be really close on the semantic space, making it difficult to discriminate the extent to which a new vector (e.g. student's response) is closer to either the golden essay or the incorrect idea vector. (A student response was graded by the professor as a 0 –the lowest possible score- because he/she referred to the protagonist as Alexander Magnus instead of Charles Magnus. Because the rest of the response was "correct", LSA gave it a high 'golden essay' grade (cosine = 0,75). Indeed, if we drop out this student Pearson correlation between LSA_golden and Professor_grade rises from 0,49 to 0,58). A second type of incorrect ideas in the History sample consisted on a reversal problem: a correct idea was indeed incorrect if the nouns of protagonists / countries were reversed. For example, the common incorrect idea "The Treaty of Verdun gave Charles 'the Bald' Western France, and to Luis 'the Germanic' Eastern France" was the reverse of a correct one included in the 'golden essay': "The Treaty of Verdun gave Charles 'the Bald' Eastern France, and to Luis 'the Germanic' Western France". LSA is virtually incapable to handle this situation, because it will create two identical vectors for the correct and incorrect ideas.

To the extent that LSA does not consider word order or any other syntactic feature, this technique might not be suited to deal with the misnaming and reversal errors committed in the analysis of the incorrect responses. In order to improve this analysis we may rely in additional natural language processing algorithms, such as word-matching procedures. In this sense, a combination of LSA and word-matching algorithms has recently proven to be effective for the analysis of short human responses [9]. Further research will be required to fully understand the potential and generalization of the analysis of incorrect responses for automated essay grading.

## Acknowledgements

Funding for this study was provided from a research grant by the Vicerrectorat de Convergència Europea I Qualitat from the Universitat de València.

## References

[1] Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An Overview of Current Research on Automated Essay Grading. Journal of Information Technology Education, 2, 319-330.
[2] Burstein, J., Leacock, C., & Swartz, R. (2001). Automated evaluation of essay and short answers. In M. Danson (Ed.), Proceedings of the Sixth International Computer Assisted Assessment Conference, Loughborough University, Loughborough, UK.
[3] León, J. A., Olmos, R., Escudero, I., Cañas, J.J. & Salmerón, L. (2006). Assessing Short Summaries With Human Judgments Procedure and Latent Semantic Analysis in Narrative and Expository Texts. Behavior Research Methods, Instruments and Computers 38, 4, 616–627.
[4] Kintsch, E., Steinhart, D., Stahl, G. & LSA research group (2000) Developing Summarization Skills through the Use of LSA-Based Feedback. Interactive Learning Environments, 8(2) , 87-109.
[5] Landauer, T.K. & Dumais, S.T. (1997). A solution to Plato´s problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. Psychological Review, 104, 211-240.
[6] Landauer, T.K., McNamara, D.S., Dennis, S. & Kintsch, W. (2007). Handbook of Latent Semantic Analysis. NJ: LEA.

[7] Martin, D.I., & Berry, M.W. (2007). Mathematical foundations behind Latent Semantic Anallisys. In Landauer, T.K., McNamara, D.S., Dennis, S. & Kintsch, W. (Eds). Handbook of Latent Semantic Analysis. NJ: LEA.
[8] Landauer, T., Psotka, J. (2000): Simulating Text Understanding for Educational Applications with Latent Semantic Analysis: Introduction to LSA. In: Interactive Learning Environments 8 (2), pp. 73-86.
[9] Magliano, J.P.; Millis, K.; Gilliam, S.; Levinstein, I. & Boonthum, C. (2007). Validating the Reading Strategy Assessment Tool (R-SAT). 12th Bienal Conference for research on Learning and instruction. Budapest, Hungary, August 28-Septembre 1, 2007.