The Right Threshold Value: What Is the Right Threshold of Cosine Measure When Using Latent Semantic Analysis for Evaluating Student Answers?

Phanni Penumatsa

Department of Computer Science, 373 Dunn Hall
University of Memphis, Memphis, TN 38152-3240, USA
phani_raju@yahoo.com

Matthew Ventura

Department of Psychology, 202 Psychology Building
University of Memphis, Memphis TN 38152, USA
mventura@memphis.edu

Brent A. Olde

Department of Psychology, 202 Psychology Building
University of Memphis, Memphis TN 38152, USA
baolde@memphis.edu

Donald R. Franceschetti

Department of Physics, University of Memphis,
Campus Box 523390, Memphis TN 38152
dfrncsch@memphis.edu

Arthur C. Graesser

Department of Psychology, 202 Psychology Building
University of Memphis, Memphis TN 38152, USA
a-graesser@memphis.edu

and The Tutoring Research Group

Abstract

Auto Tutor is an intelligent tutoring system that holds conversations with learners in natural language. Auto Tutor uses Latent Semantic Analysis (LSA) to match sentences the student generates in response to essay type questions to a set of sentences (expectations) that would appear in a complete and correct response or which reflect common but incorrect understandings of the material (bads). The correctness of student contributions is decided using a threshold value of the LSA cosine between the student answer and the expectations. Our results indicate that the best agreement between LSA matches and the evaluations of subject matter experts can be obtained if the cosine threshold is allowed to be a function of the lengths of both student answer and the expectation being considered.

Introduction

Auto Tutor is a Computer Tutor that simulates natural discourse while executing pedagogically appropriate turns. Auto Tutor engages students in a natural language dialog (Graesser, Person, Harter, &TRG, 2001; Graesser, Van Lehn, Rose, Jordan, Harter, 2001) built around a series of questions in the subject being tutored. Auto Tutor understands student expressions by means of Latent Semantic Analysis (LSA).

LSA is one of the major components in Auto Tutor. It is a statistical corpus-based technique for understanding natural language which represents word, sentences, or paragraphs (generically termed "documents") as vectors in a high dimensional vector space derived from the corpus. The most commonly employed measure of agreement between documents is the cosine of the angle between the corresponding vectors (Kintch, 1998; Landauer and Dumais, 1997; Landauer, Foltz and Latham,1998).

The present study focuses on the appropriate cosine threshold value for declaring agreement in a tutoring situation. In Auto Tutor 2.0 for conceptual physics we declared a match whenever a cosine greater than 0.65 was found between the students answer and the expectation with which it was compared. Had a higher value been used, fewer matches would be found and the students would have been prodded to revise their Answer more often. This would lead to student frustration if the answer were in fact correct but merely phrased differently from the expectation. Further a length effect might be expected. The probability of two longer documents scoring a high cosine match by accident would not necessarily be the same as that for two shorter documents. To determine whether such a length effect could exist we compared the consistency of LSA cosine-based ratings with those of human experts for short, medium and longer length documents.

The following sections reprise LSA and AT separately. We then report results of our document length study.

Latent Semantic Analysis

LSA is a statistical corpus-based text comparison technique that was originally developed for text retrieval. Nowadays it is more often used to capture the content of large bodies of texts (Kintsch, 1998; Landauer & Dumais, 1997; Landauer Foltz and Laham, 1998). LSA has been tested in the grading of essays (Foltz, Gilliam & Kendall, 2000) and found to assign grades consistent with the judgment of experts in composition.

LSA begins with a corpus, a body of documents, generally derived from published texts or reference works. The documents can be individual sentences or paragraphs, or some other convenient unit. From this is constructed a rectangular matrix, with one row for each distinct word in the text and one column for each document. The matrix elements may then be subjected to a mathematical weighting process based on the frequency of occurrence of the words in the document or in the English language as a whole (Berry, Dumais &O'Brien, 1995). The resulting matrix is then subjected to a singular value decomposition (SVD), by which it is expressed as the product of three matrices, the second of which is diagonal with the singular values appearing in decreasing order. For the purposes of latent semantic analysis, all but the N largest diagonal elements are then set equal to zero and the matrices are re-multiplied. The N chosen is typically of the order of a few hundred. This process is thought to eliminate aspects of word use in the text which are incidental to the expression of meaning but to preserve correlations between words that capture meanings expressed in the text.

Once a corpus has been constructed and the corresponding word-document matrix is transformed as outlined above, one can represent any combination of words in the corpus as vector by forming a linear combination of the rows representing the component words. For any pair of word combinations, then, there will be two vectors in the abstract N-dimensional space defined by the (SVD) which meet at an angle, the cosine of which is readily calculated from the vector "dot product," the sum of the pair-wise products of the N components.

LSA Cosine values successfully predict the coherence of successive sentences in a text (Foltz, Kintsch and Landauer, 1998), the similarity between student answers and ideal answers to questions (Graesser, P. Wiemer-Hastings et al, 2000) and the structural distance between nodes in conceptual graph structures (Graesser, Karnavat, Pomeroy, P. Wiemer-Hastings &TRG, 2000). At this point researchers are exploring the strengths and limitations of LSA in representing world knowledge.

LSA Use in Auto Tutor

A thorough description of the Auto Tutor is provided in Graesser et al (1999) and Graesser, Person, Harter and TRG (2001). We provide only a general overview here. Auto Tutor's style of tutoring is modeled after actual human tutoring strategies (Graesser, Pearson and Magliano, 1955). The tutor starts out by asking a question or posing a problem that requires a paragraph length answer. The tutor then works with the student to revise the paragraph until it covers the essential points (expectations) that the tutor deems constitute a correct and complete answer (Olde et al, 2002). Once a question has been satisfactorily answered the tutor poses the next question.

Auto Tutor's general knowledge of its tutoring domain resides in the corpus of texts from which the LSA vector space has been constructed, while the expectations, probable bad answers, and repertoire of dialog moves for each question are contained in separate curriculum scripts. The main dialog moves available to Auto Tutor are hints, pumps and assertions. There are a variety of additional dialog moves in the curriculum script that need not be addressed in the present study. (Olde et al, 2002; Graesser, Person, Harter, 2001)

Auto Tutor matches student responses to the expectations and probable bad answers for each question by calculating the LSA cosine between them. Based on the computed cosines, Auto Tutor selects its next dialog move which might include positive, negative of neutral feed back, pumps for additional information, a prompt for specific words, a hint, assertion, summary, correction or a follow-up question. The smoothness of the mixed initiative dialog in Auto Tutor critically depends on the validity of this cosine matching procedure. This motivated us to look closer at the LSA cosine matches.

Data Collection Methods

Participants

Participants were 24 students from the University of Memphis, Rhodes College and University of Pittsburgh who had previously taken a physics course. Each participant answered 8 problems and the answers were rated by two physics experts in terms of covering expectations of that problem.

Questions and Answers

The data used in the present experiment came from answers to conceptual physics problems. The physics problem answers were graded by LSA, using the space reported in Franceschetti et al. (2001). For each problem, there were sets of expectations to compare to the student answers. We judged the quality of each student's answer by calculating the LSA cosine between each expectation for a problem and all possible combinations of sentences the student gave for this expectation. This gave us a distribution of LSA cosine scores for each expectation. The fact that we considered all combinations of student sentences and taking the maximum LSA values had the intent of biasing assessment of student contributions towards correctness. We then took the highest cosine match for each expectation, and compared such values to expert physics tutors. The cosines matches were then compared with the subjective judgments of two expert physics tutors who read each student answer and for each expectation decided whether the expectation was explicitly present, implicitly present, or absent. The "implicitly present" category was used when the grader felt that it was probable that the student understood the content of the expectation, even though it could not be identified in the written answer. The agreement between LSA and the experts was studied as a function of student answer length and expectation length so as to gauge the impact of these two variables on the agreement.

Results and Discussion

The results of our study are summarized in Figure1 and Figure2. Qualitatively there is a clear correlation of the mean maximum LSA cosine and the experts' opinion that an expectation was covered by an answer. Interestingly, this is so even for those cases in which the expert classified the expectation as "implicitly" present, that is not directly stated or paraphrased by the student. This finding is consistent with the notion that LSA in some sense captures the meanings or ideas expressed in the corpus from which it is constructed.

Quantitatively, the average cosine between expectations and student responses that the experts rated as explicitly expressing them was 0.42. This is surprisingly low compared to the current threshold used in AutoTutor (.65). Additionally, expectation length and sentence length significantly contribute to changing the cosine value under the levels of expert ratings. This suggests that not only should the threshold value be lowered, it should vary according to length of expectation and length of sentence contribution.

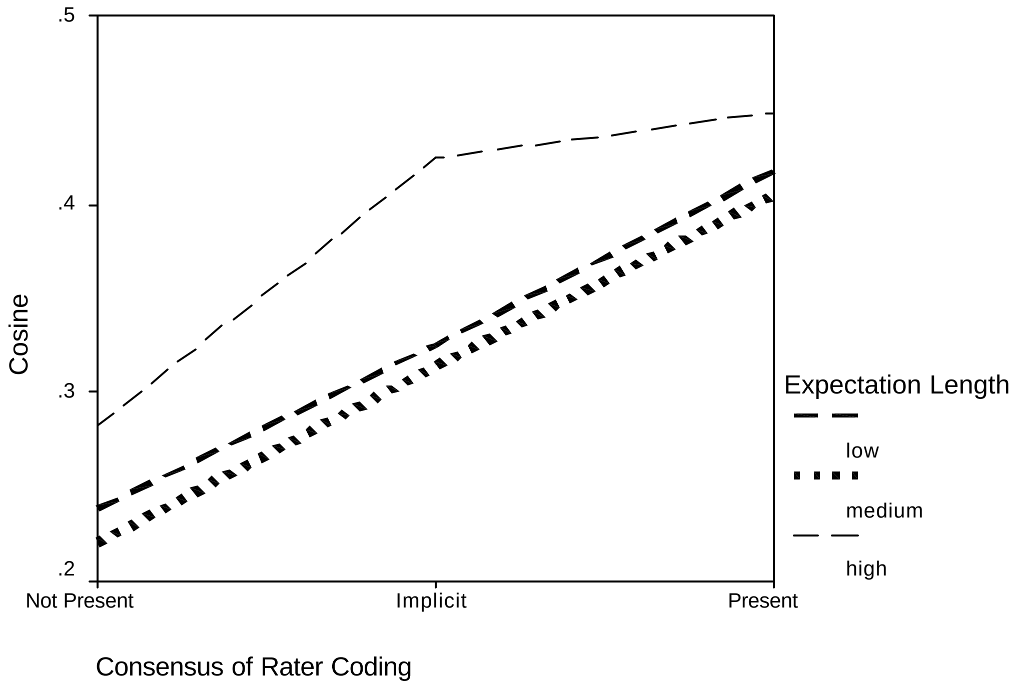

Figure 1.

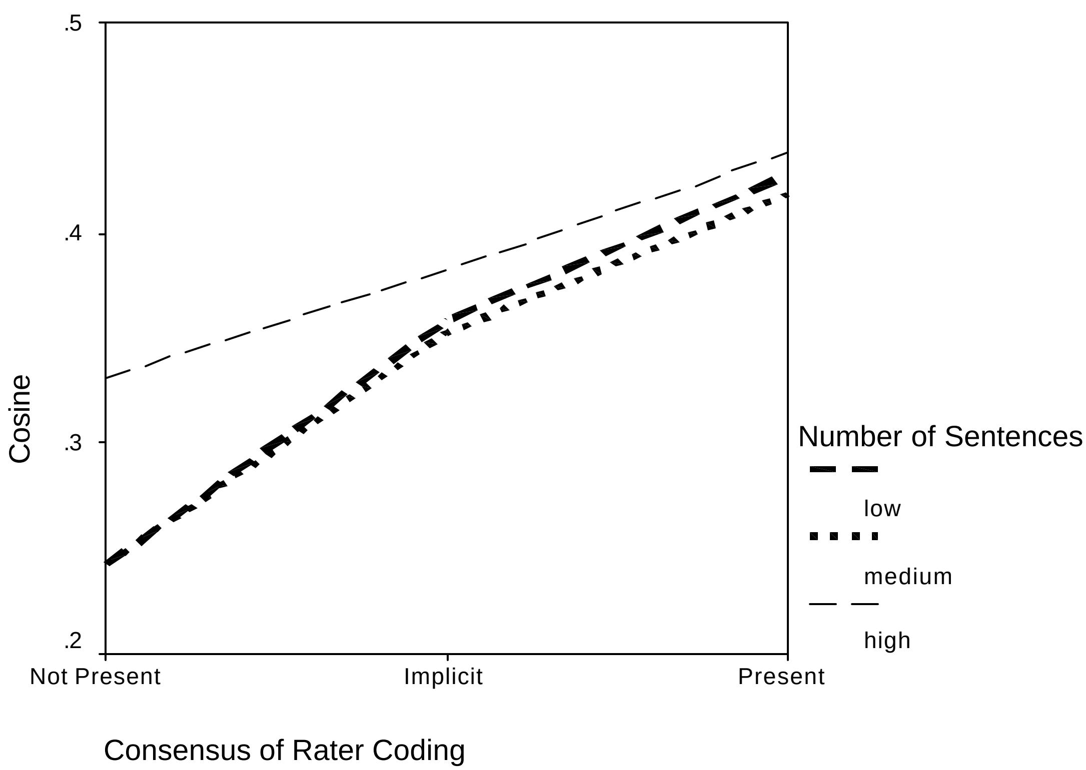

Figure 2

Figures 1 and 2 show the effect of expectation length and student contribution length over maximum cosine value that correlates with expert ratings. We obviously need to adjust the threshold of cosine value according to the expectation length and student length. Currently we are working on a mathematical model that can dynamically adjust the threshold of cosine value according to the length of expectation and length student contribution. We are currently implementing this mathematical model in the next upcoming version of the AutoTutor (AutoTutor 3.0).

Acknowledgments. This research was supported by grants from DoD Multidisciplinary University Research Initiative (MURI) program administered by the Office of Naval Research (N00014-00-1-0600) and the National Science Foundation (SBR 9720314 and REC 0106965). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of ONR or NSF.

References

Berry,M. W., Dumais, S.T., & O'Brien, G.W.(1995), Using Linear algebra for intelligent information retrieval. SIAM Review, 37,573-595.

Foltz, W.,Gilliam, S.,& Kendall,S.(2000). Supporting content based feedback in on-line writing evaluation with LSA. Interactive Learning Environments, 8,111-128.

Foltz, P.W.,Kintsch, W., & Landauer, T. K, (1998). The measurement of textual coherence with Latent semantic analysis . Discourse Processes 25, 285-307

Franceschetti, D. R., Karnavat, A., Marineau, J., Mccallie, G. L., Olde, B., Terry, B. L., Graesser, A (2001), Development of Physics Text Corpora for Latent Semantic Analysis, Procedings of the 23 rd Annual Conference of the Cognitive Science Society (pp. 297-300) Mahwah NJ, Erlbaum.

Graesser,A.C., Karnavat, A., Pomeroy, A., Wiemer-Hastings , K., & TRG (2000), Latent Semantic Analysis captures casual, goal-oriented, an taxonomic structures . Proceedings of the 22 nd Annual Conference of the Cognitive Science Society (pp. 184-189) Mahwah, NJ: Erlbaum

Graesser, A. C., Person, N., Harter, D., & TRG (2001). Teaching tactics and dialog in AutoTutor. *International Journal of Artificial Intelligence in Education*, *12*, 257-279.

Graesser , A.C., N.K.,& Magliano, J.P. (1995). Collaborative dialogue patterns in naturalistic one –to-one tutoring. Applied Cognitive Psychology, 9, 359.1-28.

Graesser, A., VanLehn, K., Rose, C., Jordan. P., and Harter, D. (2001). Intelligent tutoring systems with coversational dialogue. *AI Magazine*, *22*, 39-51.

Graesser, A.C., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., & Person, N., and the TRG (2000). Using latent semantic analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Environments*, *8*, 128-148.

Kintsch, W.(1998). Comprehension : A paradigm for cognition . Cambridge, MA: Cambridge University Press.

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.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. Discourse processes 25, 259-284.

Olde, B. A., Franceschetti, D.R., Karnavat, Graesser, A. C. & the Tutoring Research Group (2002). The right stuff: Do you need to sanitize your corpus when using Latent Semantic Analysis? *Proceedings of the 24th Annual Meeting of the Cognitive Science Society* (pp. 708-713). Mahwah, NJ: Erlbaum.

Wiemer-Hastings P., Graesser, A.C., Hater, D.,and the Tutoring Research group (1998). The foundations and architecture of Autotutor. Proceedings of the 4th International Conference on Intelligent Tutoring Systems (pp. 334-343). Berlin, Germany: Springer-Verlag.

Weimer-Hastings, P., Weimer-Hastings, K., Graesser, A. &TRG (1999). Improving an intelligent tutor's comprehension of students with latent semantic analysis. In S.Lajoie &M.Vivet(Eds.), Artificial intelligence in education (pp. 535-542). Amsterdam: IOS Press.