Development of Physics Text Corpora for Latent Semantic Analysis

Donald R. Franceschetti (dfrncsch@memphis.edu)
Department of Physics, University of Memphis, CAMPUS BOX 523390
Memphis, TN 38152 USA

Ashish Karnavat (akarnavat@hotmail.com)
Department of Computer Science, University of Memphis, CAMPUS BOX 526429
Memphis, TN 38152 USA

Johanna Marineau (jmarinea@memphis.edu)
Department of Psychology, 202 Psychology Building
University of Memphis, Memphis, TN 38152 USA

Genna L. McCallie (jordy911@bellsouth.net)
Department of Psychology, 202 Psychology Building
University of Memphis, Memphis, TN 38152 USA

Brent A. Olde (baolde@memphis.edu)
Department of Psychology, 202 Psychology Building
University of Memphis, Memphis, TN 38152 USA

Blair L. Terry (bterry@memphis.edu)
Department of Psychology, 202 Psychology Building
University of Memphis, Memphis, TN 38152 USA

Arthur C. Graesser (a-graesser@memphis.edu)
Department of Psychology, 202 Psychology Building
University of Memphis, Memphis, TN 38152 USA

Abstract

Student responses to qualitative physics questions were analyzed with latent semantic analysis (LSA), using different text corpora. Physics potentially has a number of distinctive characteristics that are not encountered in many other knowledge domains. Physics texts exist at a variety of levels and typically involve an integrated presentation of text, figures and equations. We explore the adequacy of several text corpora and report results on vector lengths and correlations between key terms in elementary mechanics. The results suggest that a carefully constructed smaller corpus may provide a more accurate representation of fundamental physical concepts than a much larger one.

Introduction

The physics classroom has often served as a laboratory for cognitive science. Studies of students learning or failing to learn physics have influenced notions of conceptual change, question answering and tutoring strategy (Albacete & VanLehn, 2000; Van Heuvelen, 1991). The physics teaching community is now aware that conventional teaching methods often fail to make any significant change in the student's understanding of the physical world. While students in the more technical introductory courses might develop the ability to recognize certain problem templates and to manipulate equations, and those in "conceptual" physics courses learn enough set answers to pass multiple choice exams, there is ample evidence that many students retain the same misconceptions about the nature of everyday phenomena with which they began the formal study of physics (Ploetzner & VanLehn, 1997).

The study of physics allegedly places rather different demands on the student than other academic work, as is readily apparent in the texts that are used. The goal of "understanding physics" or "thinking like a physicist", to which most instructors aspire, involves a combination of declarative and procedural knowledge in which the procedural component figures far more significantly than in, for example, a survey of history or introductory computer literacy. Language is used somewhat differently in physics than in other scientific fields. While biology and chemistry resort to Greco-Roman or Germanic word forming conventions to

introduce new words with precise meanings, physics more often than not takes words from ordinary language, like force and momentum, and restricts their meaning to a single sense. In most modern physics texts (such as Hewitt, 1998), there are multiple photographs or simple sketches on every page, and much of the text is directly organized around these figures. Much of the exposition in conceptual physics courses includes questions and answers that may be separated by text. Physics texts often devote considerable space to the historical evolution of physical concepts, the cultural context of physics, and its social impact. Some authors also devote appreciable space to discussing discarded theories and chains of reasoning that lead to incorrect conclusions. Thus, a significant fraction of the text found in a physics text may, in fact, exemplify incorrect thinking.

Our group has been developing a corpus of texts about physics that will eventually be used in an intelligent tutoring system on conceptual physics. The text corpus is needed to build a latent semantic analysis (LSA) space, which will be used to process the meaning of student answers in ordinary language. This paper is concerned particularly with the best strategy to construct such a corpus. A naive approach would be to gather a number of physics texts, and combine them into one corpus. However, there are unusual challenges taking this approach. What should be done about the diagrams in the text? What about the text that was written to illustrate incorrect reasoning? Does the inclusion of texts at different levels strengthen or dilute the accuracy with which physics concepts are represented in the LSA space? In short, how much special preparation of the corpus is needed, if it is to provide a reliable representation of the physics that students are expected to learn?

Latent Semantic Analysis

LSA has recently been proposed as a statistical representation of a large body of world knowledge (Kintsch, 1998; Landauer & Dumais, 1997). LSA provides the foundation for grading essays, even essays that are not well formed grammatically, semantically, and rhetorically; in fact, LSA-based essay graders assign grades to assays as reliably as experts in composition (Foltz, Gilliam, & Kendall, 2000). LSA has been used to evaluate the quality of student contributions in interactive dialogs between college students and AutoTutor, a tutoring system in the domain of computer literacy; the LSA module evaluates the quality of student answers to questions almost as reliably as graduate student research assistants (Graesser, P. Wiemer-Hastings, K. Wiemer-Hastings, Harter, Person, & TRG, 2000; P. Wiemer-Hastings, K. Wiemer-Hastings, Graesser, & TRG, 1999). Given these successes in using LSA to evaluate the quality of student essays and contributions in tutoring systems, on a variety of topics, we were interested in exploring how LSA would fare in the domain of qualitative physics.

LSA is a mathematical technique in which the information contained in the co-occurrences of words in a body of text is compressed into a set of vectors in N-dimensional space. The input to LSA is word co-occurrence matrix M, where the individual elements Mij is the number of times that the ith word occurs in the jth document. A document is an arbitrarily defined unit, but normally is a sentence, paragraph, or section in a text. The rows and columns of the matrix are then subjected to mathematical transformations that take into account the frequency of word use in the document (Berry, Dumais, & O'Brien, 1995; Landauer, Foltz, & Laham, 1998). Using the mathematical process of singular value decomposition, the matrix is then expressed as the product of three matrices, the second of which contains the singular values on the diagonal. Changing all but the largest N singular values to zero sets the dimensionality N of the vector space representing the text. The matrices are then re-multiplied to produce a matrix of the same dimensions of the original matrix.

At first glance it might seem that by discarding some of the singular values we are discarding information. However, it turns out in practice that the lower dimensional representation better captures the meaning of the text. For instance, there ends up being a positive relationship between the coefficients in the rows corresponding to different words, if the words have similar or associated meanings. The reduced number of dimensions are sufficient for evaluating the conceptual relatedness between any two bags of words. A bag is an unordered set of one or more words. The match (i.e., similarity in meaning, conceptual relatedness) between two bags of words is computed as the geometric cosine (or dot product) between the two associated vectors, with values that normally range from 0 to 1. LSA successfully predicts the coherence of successive sentences in text (Foltz, Kintsch, & Landauer, 1998), the similarity between student answers and ideal answers to questions (Graesser, P. Wiemer-Hastings, et al., 2000; Wiemer-Hastings et al., 1999), and the structural distance between nodes in conceptual graph structures (Graesser, Karnavat, Pomeroy, Wiemer-Hastings, & TRG, 2000). At this point, researchers are exploring the strengths and limitations of LSA in representing world knowledge.

Constructing an LSA Corpus That Knows About Physics

We have assembled several different physics corpora to test the effect of the content of the subject matter on the

quality of the LSA solutions. The documents in the texts were classified into different rhetorical categories, such as exposition, example problems, historico-cultural material, incorrect reasoning, and so on. The fundamental research question is whether the inclusion of different texts and categories of content have an impact on the representation of core concepts in the mechanics portion of a conceptual physics course. All the corpora include text materials from the mechanics portion of Paul Hewitt's *Conceptual Physics* (1998), a text that is widely used in conceptual physics courses at the college level; these were used with permission from the publisher. The "Omnibus" corpus included chapters 2-9 of the Hewitt book plus six volumes of a comprehensive text aimed at students in technical or life science majors, two advanced texts in electromagnetism, and another physics text that was available electronically. The "Large" corpus was constructed from the former by deleting the three latter texts. A "Small" corpus further deleted the texts that did not cover mechanics. A "Restricted Small" corpus further deleted any text identified as primarily historico-cultural or involving misconceptions. In the "Restricted Hewitt" corpus, we included only those texts from Hewitt in the restricted small corpus. Each of the corpora was thus a proper subset of the preceding one, with the goal of further refining or sanitizing the text corpus to handle the core concepts in mechanics. The time needed to "restrict" a text was minimal once the text was converted to electronic form.

Vector Lengths and Similarity

Kintsch (1998) proposed that the length of the vectors representing key terms provides a measure of the extent to which the LSA has captured the meaning (or importance, centrality) of the word with respect to the subject matter. The vector length increases to the extent that the set of values in the vector deviate from zero. Words like *force*, *momentum* and *gravitation* should have reasonably large vector lengths in any corpus that represented basic physics concepts well.

LSA spaces of 100, 200, 300, 400, and 500 dimensions were created for each of the above five corpora. Each text paragraph was treated as a document. Figure captions were eliminated. Questions & answers were lumped in the same document. Based on the vector lengths computed for the key mechanics words listed in Table 1, it was decided that little improvement would be achieved by going beyond 500 dimensions. The restricted Hewitt corpus was so small that only a 400 dimensional representation could be obtained. The vector lengths for selected physics words in a 300 dimensional space are shown in Table 1.

A number of conclusions can readily be drawn from Table 1. There is a general correlation between vector lengths of the first two corpora and between those of the two smallest corpora. When we eliminated the material not pertinent to mechanics as presently understood, some vectors ended up increasing in length. For example, the *impulse* concept, which occurs only in mechanics, had a significantly larger vector length in the smaller corpora than in the larger corpora. The same can be said for *tension*, which is the force transmitted by a rope or cable, and is useful only in mechanics. Even a concept like *energy*, which pervades all areas of physics, appears to be more crisply represented in the smaller corpus.

Table 1. Vector Lengths for Physics Words.

Word	Omnibus	Large	Small	R-small	R-Hewitt
Gravity	.288	.281	.262	.242	.240
Gravitational	.256	.250	.223	.256	.283
Mass	.300	.293	.269	.239	.288
Acceleration	.300	.296	.266	.270	.284
Force	.186	.179	.155	.128	.153
Momentum	.267	.263	.258	.283	.288
Energy	.222	.219	.228	.238	.313
Impulse	.367	.371	.400	.432	.466
Friction	.320	.314	.301	.313	.375
Velocity	.252	.250	.240	.236	.291
Vector	.285	.305	.292	.382	.455
Potential	.323	.328	.361	.386	.464
Tension	.266	.271	.302	.390	.475
Kinetic	.298	.294	.301	.312	.422
Normal	.315	.314	.352	.414	.373
Newton	.347	.242	.211	.206	.265
Aristotle	.309	.309	.318	.409	.436
Galileo	.324	.326	.325	.338	.355
Newtonian	.242	.223	.221	.339	.000

The names of the key physicists, *Galileo* and *Newton*, along with that of *Aristotle*, whose notions of physics are now largely discarded, were also included in our study of vector lengths. Interestingly, our efforts to eliminate material of only historical value in the two restricted corpora did not eliminate a rather well represented *Aristotle*. LSA did pick up stylistic characteristics of individual authors.

The similarity between concepts in LSA is represented by the cosine values in corresponding vectors. We computed the cosines between the physics terms in Table 1 and these appear in Table 2. The greatest similarity appeared for *kinetic energy*, which is in effect a composite word and for *impulse-momentum*, which would appear as a composite in the "impulse-momentum theorem" and the exposition of it, in that impulse equals the net change of momentum in a collision. We note that the similarities between *mass* and *acceleration* and between *force* and *acceleration*, which would be expected in any exposition based on Newton's second law (the net force on an object equals

the mass times its acceleration). The similarity scores are appreciably more apparent in the smaller corpora with the irrelevant text removed.

Table 2. Largest magnitude cosines between key physics terms. (Corpora titles are abbreviated)

Correlation	Om	Lge	Sm	R-Sm	R-H
Gravitational force	.084	.083	.093	.146	.029
Gravitational potential	.058	.091	.097	.107	.032
Force acceleration	.006	.009	.009	.048	.087
Mass acceleration	.033	.035	.044	.070	.066
Normal force	.080	.084	.125	.096	.043
Mass momentum	.010	.013	.028	.040	.077
Impulse momentum	.182	.187	.176	.196	.148
Kinetic energy	.209	.228	.265	.267	.267
Tension friction	.052	.052	.020	.066	.001
Vector Velocity	.052	.055	.053	.065	.059
Kinetic friction	.081	.083	.020	.066	.026

Summary

We have developed a number of alternative physics text corpora for use in the evaluation of student answers to physics questions. Comparisons of word length and word similarity suggest that both the elimination of material from other areas of physics and other levels of exposition, as well as the elimination of material not dealing with the exposition of the physical concepts, allows an improved representation of core physics terms and the relationships between them, even with a rather small corpus. However, this conclusion is currently being tested on a large body of student and expert answers to physics questions. The preliminary results suggest that although vector lengths increase for individual words with a refined selection of texts, it is a large corpus that works best when the entire sentence is used to evaluate the match of student and expert answers. In other words, individual words may have a crisper representation when a smaller, well-defined text is used but when analyzing an answer formed around the integration of several complex concepts, a broader selection of texts is more beneficial. Furthermore, it is our contention that a regression could be used to capitalize on the unique information provided by both types of LSA spaces. In future work, we will examine the feasibility of adding picture descriptions in natural language to the corpus and alternative treatments of equations and composite words.

Acknowledgments

This research was supported by Grant N00014-00-1-0600 from the Cognitive Science Division of the Office of Naval Research.

References

Albacete, P. L., & VanLehn, K. A. (2000), Evaluating the effectiveness of a cognitive tutor for fundamental physics concepts. *Proceedings of the22nd Annual Conference of the Cognitive Science Society* (pp. 25-30). Mahwah, NJ: Lawrence Erlbaum.

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.

Graesser, A. C., Karnavat, A., Pomeroy, A., Wiemer-Hastings, K., & TRG (2000), Latent semantic analysis captures causal, goal-oriented, and taxonomic structures. *Proceedings of the22nd Annual Conference of the Cognitive Science Society* (pp. 184-189) Mahwah, NJ: Erlbaum.

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.

Hewitt, P. G. (1998) *Conceptual physics* (Ed. 8). Reading, MA: Addison Wesley Longman.

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.

Ploetzner, R., & VanLehn, K. (1997). *Cognition & Instruction, 15*, 169-205.

Van Heuvelen, A. (1991). Learning to think like a physicist: A review or research-based instructional strategies, *American Journal of Physics, 59*, 891-897.

Wiemer-Hastings, P., Wiemer-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.