The Application of LSA to the Evaluation of Questionnaire Responses

Dian I. Martin, John C. Martin and Michael W. Berry

Abstract This chapter will discuss research surrounding the development of applications for automated evaluation and generation of instructional feedback based on the analysis of responses to open-ended essay questions. While an effective means of evaluation, examining responses to essay questions requires natural language processing that often demands human input and guidance which can be labor intensive, time consuming, and language dependent. In order to provide this means of evaluation on a larger scale, an automated unsupervised learning approach is necessary that overcomes all these limitations. Latent Semantic Analysis (LSA) is an unsupervised learning system used for deriving and representing the semantic relationships between items in a body of natural language content. It mimics the representation of meaning that is formed by a human who learns linguistic constructs through exposure to natural language over time. The applications described in this chapter leverage LSA as an unsupervised system to learn language and provide a semantic framework that can be used for mapping natural language responses, evaluating the quality of those responses, and identifying relevant instructional feedback based on their semantic content. We will discuss the learning algorithms used to construct the LSA framework of meaning as well as methods to apply that framework for the evaluation and generation of feedback.

Keywords Latent Semantic Analysis (LSA) • Essay grading • Feedback generation • Simulating human raters • Cognitive model • Unsupervised learning • Natural language responses • Open-ended essay questions

D.I. Martin (✉) • J.C. Martin
Small Bear Technologies, Inc., Thorn Hill, TN, USA
e-mail: Dian.Martin@SmallBearTechnologies.com

M.W. Berry
University of Tennessee, Knoxville, TN, USA

M.E. Celebi, K. Aydin (eds.), *Unsupervised Learning Algorithms*,
DOI 10.1007/978-3-319-24211-8_16

1 Introduction

The use of computerized systems for evaluating and understanding open-ended responses to test and survey questions has been researched and discussed for over 50 years in an effort to overcome the fundamental problem of a need for human effort to score them. While closed response questions where a subject picks from a set of pre-defined answers are readily amenable to automated scoring and have shown themselves to be very useful in standardized testing and information surveys, they cannot be used in all situations and do not permit the sort of assessment options that open-ended prompts can provide. The use of essay and short answer questions in testing and surveys persists because of the unique insight they can provide. Advances in computer technology and our understanding of language, thought, and learning have made it possible to produce automated systems for evaluating the unstructured text of open-ended responses that could only have been imagined as little as 20 years ago.

In this chapter we will present one particularly promising machine learning technology used to address open-ended responses, Latent Semantic Analysis (LSA), and its application as a theory of learning, a computational model of human thought, and a powerful text analytics tool. First, we will discuss the nature of open-ended response items or essay questions themselves, covering both their application advantages and the complications involved in their use. We will describe the background of LSA, exploring its application beyond its initial use in information retrieval and its development into a theory and model of learning. Looking toward the subject at hand, we will elaborate on the application of LSA toward essay questions and in the realm of scoring and feedback generation. As an example, we will present a specific case study describing the application of LSA to form an unsupervised learning system used to evaluate open-ended questionnaire responses and generate constructive feedback for the purpose of driver training in a NIH-funded research project. Finally, we will touch on other projects involving LSA for scoring, evaluation, and content analysis, and discuss ongoing research work with LSA.

2 Essays for Evaluation

Humans think, speak, learn, and reason in natural language. The ability to express an answer to a question in one's own words without being prompted by a set of predefined choices has been a measure of understanding and achieved learning probably since the invention of students. While closed response questions in the form of multiple-choice tests have proven their worth in many testing arenas, the essay question still remains as a primary, though problematic, evaluation tool in many areas of learning.

2.1 Open-Ended Responses Provide Unique Evaluation Leverage

One of the key advantages of open-ended response items is that they demonstrate the ability to recall relevant information as well as the synthesis and understanding of concepts rather than the mere recognition of facts [1]. Open-ended responses also allow for creative and divergent thoughts to be expressed as they provide an opportunity for respondents to express ideas or analyze information using their own words. In educational contexts, essays in many ways exhibit a more useful level of knowledge from a student, while providing cues to educators guiding additional instruction. Feedback from essay evaluation can help the students learn certain content as well as the skill of thinking [2]. Additionally, there is a universal need for people to be able to communicate what they know, not just pick it off of a list, and educators need to promote the expressing of knowledge in self-generated prose. Outside of a testing environment, communication is at times perhaps more critical than factual knowledge.

Another consideration favoring open response items is that some types of questions do not lend themselves to closed response prompts. The formation of high-quality closed response questions when they are suitable is still time consuming and expensive. The ability to creatively combine multiple pieces of information and then to recognize and describe their interconnectedness demonstrates a higher order of thinking than can be captured in a closed response evaluation item. We desire to know what people can do, not what they can choose [3].

In short, essay and open-ended response questions provide several advantages for assessment:

- Demonstrate the ability to recall relevant information
- Demonstrate the synthesis and understanding of information
- Allows creative and divergent thoughts to be expressed
- Allows for different type of questions to be posed
- Avoids leading the respondent with pre-formed answers

2.2 The Problem with Essays: Human Raters Don't Scale

Essay questions are not the problem. The problem is essay assessment. Measurement tools for gauging the quality of an open response item need to reliable, valid, and generalizable. Ideally, essay scoring should meet the same level of objectivity as a well-crafted multiple choice test. The standard approach to evaluating open-ended responses has been to use human raters to read them and assign scores. Even though individual human scoring can be unreliable as it is subject to a rater's variability, biases, and inadequacies, human judgment is still considered the standard of quality in the evaluation of essay responses [4]. Human essay assessment however is fraught with issues. It is both expensive and time consuming. Language dependency makes

it difficult or impossible to generalize human assessment for large collections of responses. Finally, such assessment is low in reliability even for relatively small sets of items. Consistency issues increase as the volume of items to be evaluated grows.

2.2.1 Expense

The process of human assessment is more complicated than would appear on its face. It is both time consuming and labor intensive. A human rater can read for comprehension at a certain maximum rate. While repetition may result in some speedup, the rate cannot be increased beyond a certain point. Large volumes of items to be rated will overwhelm the abilities of a human rater and delay the availability of assessment scores. Additionally, using a single human rater can cloud the assessment ratings with questions of subjective bias.

In order to deal both with problems of scale and questions of bias, it is apparent that multiple human raters must be employed. In order to ensure reliable assessment and eliminate questions of bias, more than one expert human needs to grade each essay [4]. This redundancy is expensive both in time and money while doing nothing to increase the throughput of scored items. Even more additional raters are required to deal with larger numbers of items. Using multiple raters at any level introduces both additional costs and problems. First, to facilitate consistent scoring a set of evaluation criteria, a rubric, must be established. The essay items to be evaluated must then be read by the human raters in the light of this rubric and scores assigned. This group of raters must be trained to correctly apply the scoring rubric in a uniform manner, and their performance must be monitored to ensure some level of inter-rater agreement. Human assessment requires redundant scoring of items, perhaps at multiple levels, to both moderate bias and provide comparability between rater groups. If the measured inter-rater reliability is not at the desired level, additional rater training is often prescribed or perhaps the rubric is adjusted and the raters re-trained. In either case items are generally rescored yet again.

2.2.2 Language Dependencies

Aside from the issues of volume, human raters are also limited in the type of items they can read and process. A human rater will typically only be able to evaluate response items within a single language, so generalizing the evaluation of essay responses across multiple languages presents an obvious difficulty. Even within a single language, however, multiple raters have different understanding and vocabulary dependent on several factors (regional idioms, fluency, use of technical jargon, age, etc.). While training of the raters attempts to mitigate these inconsistencies, they still remain at some level. These issues manifest themselves in the form of poor inter-rater agreement or inconsistency in rating reliability over large groups of essays or long periods of time.

2.2.3 Consistency Issues

Key to large-scale generalizable evaluation of open response items is consistency. This consistency must exist at the individual rater level, across teams of multiple raters, and over time as responses are collected, perhaps over a span of years.

Consistent objective analysis by human raters is difficult to achieve for both individual and multiple raters. In many cases, an individual rater will not agree with their own scores when repeating their assessment of the same essay some time later after they have forgotten the first score. The correlation of repeat ratings of the same items has been shown to be only about 0.70 [3]. One factor that certainly contributes to this is rater fatigue. Faced with a large collection of items to be assessed, it is no surprise that the time and attention given to the evaluation of item number 1 will differ from that given to item number 300. The other potential is that after examining the same sort of item repeatedly the evaluation will devolve from a holistic analysis to a simple check-off of certain prominent attributes.

Multiple raters used to increase the throughput of open item scoring present another set of problems impacting consistency. It has been noted that "with human raters, it is seldom clear that any two are grading an essay on exactly the same criteria, even when trained" [4]. Individual raters agree with each other weakly, generally correlating about 0.50–0.60 of the time. The average grade given to an essay by a group of graders is more dependable. For example, if a panel of four experts scores an essay, each expert will correlate about 0.80 with the average grade of the four experts [1]. Many tests are high stakes and need humans to grade in the capacity of objective testing, requiring better than a 0.65 correlation with a reliability of 0.78 [3]. Achieving this goal requires more raters, more training, and consequently more expense.

Consistency Over Time

Many assessments are gathered over a long period of time, perhaps even years. In such cases rater consistency becomes a critical issue as raters change and even those that do not may be subject to greater drifting in their scoring. The quality and consistency of different groups of human raters is difficult to measure and requires repeated rating of the same items to establish a basis for comparison. This results in some amount of redundant work if a quality comparison is to be performed. Human judgments do not meet the need of group descriptions or group trends due to the fact that raters usually change from year to year or for each time a test is given and assessed. Human raters tend to adapt their grading or rating to the essays in hand or at the time. Therefore, it is hard to draw conclusions on trends over time. How can current tests indicate how a group of students are doing compared to last year or another state or other demographic groups [3]? This is important for improvement in education of any kind.

2.3 Automated Scoring Is Needed

The use of human raters for essay evaluation is limited in both consistent repeatability and scalability. The only way to scale up to handle large volumes of responses while providing reliable and generalizable ratings is through automated systems. There has been much work and research on using computers to grade essays and perform evaluation with the goal of producing "pedagogically adequate, psychometrically sound, and socially acceptable machine assessment and tutorial feedback for expository essays" [5]. Several such systems have been developed, demonstrating that computerized evaluation of open response items and essays is feasible [5–8].

Just as with human raters, the performance of automated scoring systems must meet standards for reliability and quality. Research has shown that computerized essay scoring systems correlate to expert human ratings as well as or more highly than the human experts do with each other [3, 8]. Of course automated systems have no issue with consistency and are not subject to fatigue. When properly developed they can eliminate the issue of rater bias [5]. The use of computerized systems also provides capabilities that are not possible using human raters, such as allowing an evaluation item to be compared with every other evaluation item. Essay items being evaluated can then be described in different aspects of the full collection of items as groups exhibiting certain similarities, differences, trends, etc. All of this can be performed at large volumes, with shorter turnaround times, and at lower cost.

2.3.1 Creating Automated Methods Requires Learning Systems

The objective of all this is not to have computers replace human judgments or grading of essay items completely. Rather, the computer is used to replicate human judgments by learning the process of grading, reaching an acceptable agreement with humans, and then is deployed for large-scale assessment [3]. Due to the need to handle varying prompts across many different subject areas and multiple languages, an automated evaluation system must be easily adaptable to different applications and different contexts. One such technology that supports this adaptable learning approach is LSA.

3 LSA as an Unsupervised Learning System

LSA has evolved into a theory of learning, a computational model of human thought, and a powerful text analytics tool. It takes its name from the fact that it presumes the existence of an underlying or "latent" structure relating the meanings (semantic value) of words within a body of text [9]. This section discusses the historical development of LSA, briefly describes the mathematics that make up its foundations, and explores the application of LSA as a learning system to analyze language and meaning.

3.1 Brief History of LSA

LSA has its historical roots in work performed during the 1980s at Bell Communications Research (Bellcore) aimed at what is known as the Vocabulary Problem. In computerized data retrieval systems of the time, the ability to access a relevant object or item was dependent upon lexical matching which required referencing the correct word tied to the information of interest. It was noted that users of such a retrieval system often used the "wrong" words to indicate the object of their request, and were therefore unable to retrieve the items they desired. A research study was conducted collecting data from human language usage in five different domains. For each of the five experiments, subjects were asked to provide a word to identify a given object or item. One of the key results from the study was the finding that people use the same term to describe a given object less than 20 % of the time. This fact leads to the Vocabulary Problem [10].

The Vocabulary Problem has come to be known in the information retrieval (IR) field as the problem of synonymy. It refers to the diversity in word usage by people referencing the same object [11]. Further complicating matters, there is also the case where people use the same word to reference different objects, a condition known as polysemy. Both synonymy and polysemy complicate the task of information retrieval. The key to the performance of an information retrieval system is addressing these issues [11–13].

Following up their earlier work, the group at Bellcore proceeded to develop an automated technology to overcome the lexical keyword matching problems that were being encountered in attempts to access their document databases. This approach, LSA, provided new leverage for indexing and retrieval and was applied to a number of internal projects to improve success of retrieval. LSA showed remarkable improvement in information retrieval because objects in a database were retrieved for a given query based on semantics and were no longer restricted to lexical keyword matching [9].

Additional work at Bellcore showed that LSA improved recall (the finding of relevant documents for a query), solving the synonym problem, but was less successful improving precision (the number of relevant documents in the top k spots retrieved for a query), which would address the polysemy problem [14]. It is important to note that all of this early Bellcore work was performed on relatively small data sets (a couple thousand documents or less) due in part to the limited computational power available at the time. The issue of polysemy has been addressed more recently as it has come to be recognized that a great many more documents are needed to support the learning of the linguistic constructs necessary for addressing the problem, and the computational abilities of modern processors have grown to allow these larger sets to be processed (to be elaborated on later in this chapter). In this early work LSA showed potential for being a viable option, better than lexical term matching, for use as an effective retrieval method. The LSA approach was able to overcome the Vocabulary Problem, was widely applicable, fully automated (not requiring human intervention), and gave a compact representation of the original textual data with noise removed [9].

Using the performance measures of precision and recall, initial experiments with LSA showed 20 % better performance than lexical matching techniques and a 30 % improvement over standard vector methods. In the early 1990s much progress was made applying LSA in the field of information retrieval. One of the first adjustments explored was the application of different term weighting schemes. Significantly improved retrieval performance was noted when using the log-entropy weighting scheme. This yielded a 40 % improvement in performance over the straight term frequency weighting that had been used up to that point [12]. Over time, it was realized that performance improved as increasing numbers of documents were added to the data collections being processed. Growth in the size of the available corpora for analysis and the continued improvement in results led to further development of LSA as a theory of meaning and its recognition as a learning system.

3.2 Mathematical Background

Before the discussion of how LSA can be used as an unsupervised learning system, the mathematical basis for the LSA needs to be presented. After all, it is this mathematical representation of the original textual database and its subsequent processing that give LSA its power. A brief overview of the concepts is presented here, but for a detailed description one should refer to Martin and Berry 2007 or Martin and Berry 2010 [11, 15].

3.2.1 Parsing: Turning Words into Numbers

The first step in the process is the conversion of strings of text into a suitable numeric representation. This is a fairly straightforward process. Given a textual body of information, a matrix (conceived of as a simple table) is constructed where the rows are unique token types and the columns are contexts in which the token types are used. Token types are typically formed from the single word items in the text. Contexts, also referred to as documents or passages, can be phrases, sentences, paragraphs, or multiple paragraphs, but are usually selected to be single paragraphs [16]. A number of policy rules must be established for how the text stream will be broken down into tokens in the parsing process. Some of these rules are fairly standard, such as using a single word per token, treating whitespace characters as boundary markers for tokens, or ignoring character case distinctions. Other rules are decided based on the application and the nature of the text being processed. These might include rules on how to handle punctuation characters, numeric values, or even designating certain words to be ignored altogether. In general, these policy rules are meant simply to establish a uniform parsing scheme, not to influence the interpretation of meaning in the text.

Once the parsing process has been completed, every cell in this matrix will have been assigned a value indicating the number of times each token type appears in each individual context. The result is a large sparse matrix, often referred to as the term-by-document matrix.

To this sparse matrix a weighting function is applied. The use of a weighting function is common practice in vector space models for information retrieval. The weighting function serves to somewhat normalize the importance of token types (terms) within contexts and across the collection of contexts (documents). The value of each element is adjusted by a global weight, the overall importance of the term in the collection, and local weight, the importance of a term within a document. Local weighting schemes are generally intended to dampen the effect of terms that occur frequently within a single document. Local weighting functions that might be applied include the use of straight unadjusted term frequency, a binary weighting indicating simply the absence or presence of a term, and finally the natural log of term frequency $+1$.

Similar to the local weight, the global weighting function is intended to dampen the effect of terms appearing frequently across the documents in a collection. Common global weighting functions include:

Normal

$$\sqrt{\frac{1}{tf_{ij}{}^2}},$$

where tf_{ij} is the number of times term i appears in document j,

GfIdf

$$\frac{gf_i}{df_i},$$

where gf_i is the number of times term i appears in the collection (global frequency of term i) and df_i is the number of documents in which term i appears (document frequency of term i),

Idf

$$\log_2\left(\frac{n}{df_i}\right)+1,$$

where n is the total number of documents in the collection, and

Entropy[1]

$$1+\sum\frac{p_{ij}\log_2\left(p_{ij}\right)}{\log_2(n)},\quad \text{where } p_{ij}=\frac{tf_{ij}}{gf_i}\,.$$

[1]There was an error in the entropy formula that was published in Dumais [12]. The formula as it appears here has been corrected. This correction also appeared in Martin and Berry [15].

Entropy is the most sophisticated global weighting scheme, taking into account the distribution of terms over documents in the collection. In tests of various combinations of these local and global weighting schemes, log-entropy yielded the best retrieval performance, performing 40 % better than just straight term frequency alone [12].

These weighting transformations are akin to classical conditioning in psychology because the transformations depend on co-occurrence of words in context weighted first by local co-occurrence frequency (local weighting) and then inversely by the extent to which its occurrences are spread evenly over the collection (global weighting). Using log as the local weighting factor approximates the standard empirical growth function of simple learning and using entropy as the global factor indicates the degree to which seeing the word indicates its context [17].

3.2.2 Singular Value Decomposition

The data represented in the term by document matrix at this point describes the content as a co-occurrence of words (terms) and contexts (documents). A document derives its meaning from the terms it contains. Each term contributes something to its meaning. At the same time, two documents can be similar in meaning and not contain the same terms. Similarly, terms that appear together in one document do not necessarily have similar meaning in different document contexts. This leads to the definition of the Compositionality Constraint, that the meaning of a document is the sum of the meaning of its terms [16]. Recognizing that this matrix reflects a coefficient matrix for a system of simultaneous linear equations with each equation representing a passage, the basis of LSA is the computation of a solution to this system in order to infer the meaning for each term based on the contexts in which they do and do not appear. Singular Value Decomposition (SVD) is used to solve this system of equations, yielding as the output a set of vectors in a high-dimensional "semantic space" with each passage and word represented by a vector. A document, or context, vector can then be considered as the vector sum of all the term vectors corresponding to the terms it contains. Similarly, a term vector represents a term in all its different senses [18].

The SVD produces a factorization of the original term by document matrix A in three parts:

$$A = U\Sigma V^{\mathrm{T}}$$

The rows of matrix U correspond to the term vectors, and the rows of matrix V correspond to the document vectors. The nonzero diagonal elements of Σ, the scaling factors, are known as the singular values.

From a full SVD, it is possible to reconstruct the original matrix A from the three matrices U, Σ, and V; however, the truncated SVD that is computed for LSA will equate to the best k-rank approximation of A. This dimensional reduction is desirable because it has the effect of removing noise from the original representation

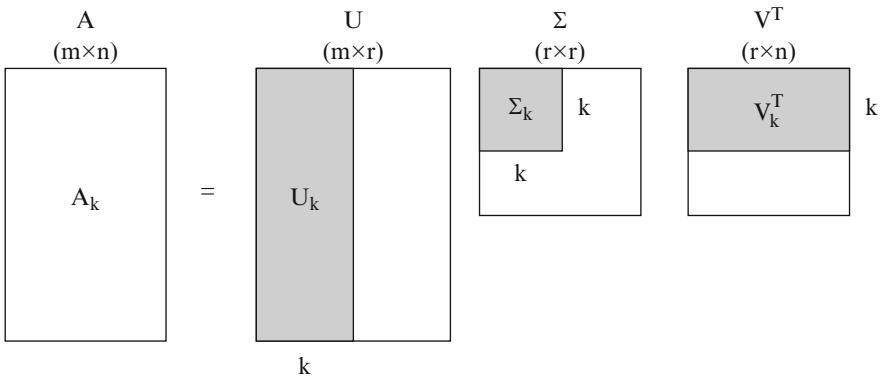

Fig. 1 Pictorial representation of the truncated SVD

of *A* (countering the dissimilarity of related documents that use synonymous terms and separating those that contain polysemic terms).

The truncated SVD (see Fig. 1) is defined as

$$A_k = U_k \sum_k V_k^{\mathrm{T}} .$$

Use of the truncated SVD is a hallmark of LSA. The truncated SVD is used to produce a set of dimensions based on the k extremal singular values of the term-by-document matrix A. This results in the most significant dimensions of the space being used to define the truncated vector space used for the LSA representation. The importance of this has become more apparent over time as the study of LSA has continued. The truncated vector space can be considered as a multi-dimensional hyperspace where each item is represented by a vector projecting into this space. This concept can be roughly pictured with a simple 3-dimensional representation where the vectors point out into the 3D space, see Fig. 2.

This illustration is extremely simplified to facilitate visualization. In practice, k is typically selected to be anywhere from 300 to 500 dimensions. Empirical testing has shown dimensionalities in this range to be most effective for synonym recognition applications [17, 20]. Within this hyper-spatial representation, information items are left clustered together based on the latent semantic relationships between them. The result of this clustering is that items which are similar in meaning are clustered close to each other in the space and dissimilar items are distant from each other. In many ways, this is how a human brain organizes the information an individual accumulates over a lifetime [19].

Calculation of the SVD is mathematically nontrivial and computationally intensive. Early research in LSA was certainly restricted by the available computing power of the era. What would today be considered very small data sets (2000 documents or less), were used in much of this initial testing. Advances in available computing power have made the processing of much larger bodies of content

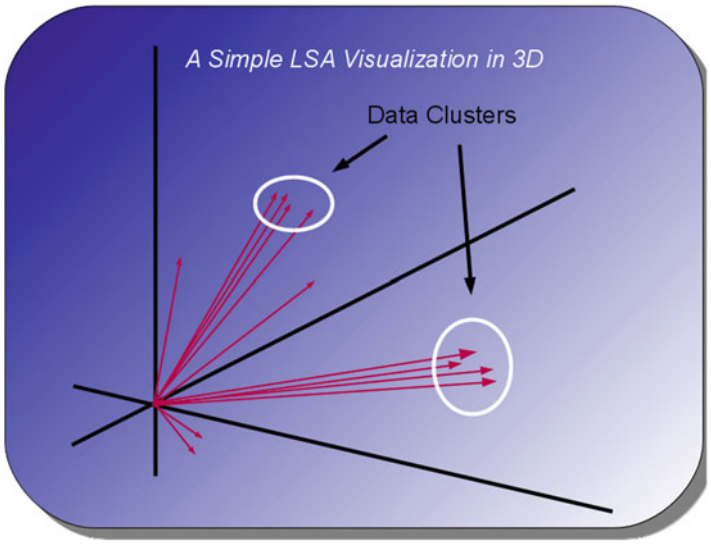

Fig. 2 A simplified visualization of a LSA semantic space [19]

feasible, as have advances in the algorithms and software for performing this work. Our own efforts have contributed to the processing of larger data sets through the optimization of resource usage [21] and implementation of large-scale parallelized processing software [22]. For our current research, we use the LSA_Toolkit™, a commercially available library for both efficiently constructing large LSA spaces and performing many additional analysis functions [23].

3.2.3 Query and Analysis Processing

Once the truncated SVD has been computed for a text collection, two forms of analysis can be performed. Items within the space can be compared for semantic similarity, and new document items can be constructed and projected into the space mapped according to their semantic content.

Comparison of items within the space can be used to analyze the content of a document collection, looking for items that are closely related or perhaps clustered together by their similar semantic mapping. By evaluating term proximities, it is possible to identify synonyms and examine term usage across the collection. An item within the LSA space, whether a document or an individual term, is represented by a k-dimensional vector. Vectors for two such items may be compared by computing a distance measure or a similarity measure that quantifies their semantic proximity or separation. The similarity measure that is typically used for LSA is the cosine similarity which has been shown to be a reliable measure of semantic relatedness within the LSA space [24].

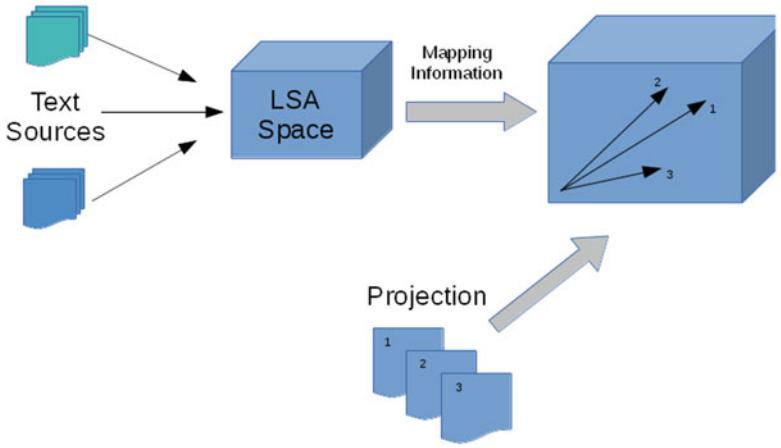

Fig. 3 Conceptual view of items projected into a LSA space

Cosine similarity is defined as:

$$\frac{u \cdot v}{\|u\| \ \|v\|},$$

that is, the dot product of vectors u and v divided by the product of their vector lengths. Other options for similarity measures include Euclidean distance and the dot product itself.

Projection of New Content: Subject Items for Evaluation

Projection of items (see Fig. 3) exploits the term definitions provided by the space to map new document items that were not included in the original content within the semantic context of the space. This can be used to probe the space as in the case of an information retrieval query, or for simply establishing semantic relationships between items of interest within the context of meaning represented by the LSA space. In mathematical terms, a new document projection is computed as the weighted sum of the term vectors corresponding to the terms in the item being projected scaled by the inverse of the singular values.

Parsing

Projection begins by parsing the text of the item to be evaluated. This is performed in the same manner as the parsing of the content used to form the base space. Policy for handling punctuation, casing, numeric values, etc. are all followed to break the evaluation item down into a set of terms. Unlike the formation of the base space, however, new terms that are not present in the base space cannot be added to the

vocabulary and cannot be used in the projection mapping. The terms that do exist in the vocabulary of the base space are identified so that their term vectors may be used in the computation of the projection. The occurrence of each mappable term in the evaluation item is counted, and a term frequency vector (TFV) is formed representing its content.

Weighting

Just as when the original LSA space was computed, a weighting is applied to the evaluation item TFV. This weighting scheme must correspond to the one used in the construction of the LSA space. If the typical log entropy weighting method is used, for each term in the evaluation item compute the local weighting for the TFV: $\ln\left(tf_{ij}+1\right)$. This value is then multiplied by the global entropy value for that term within the base space. The global entropy value was computed for each term in the original sparse matrix when it was weighted in preparation for formation of the LSA space.

Composition

Finally, to compose the projection vector, z, the weighted term frequency values q for the terms in the evaluation item are multiplied by their respective term vectors U_k from the base space. The sum of these vector products is then the new projection vector for the evaluation item being considered.

The projection vector computation is

$$z = \frac{q^{\mathrm{T}} U_k}{\Sigma_k}\,.$$

Once computed, the projection of an item can be compared to items in the space or to other projection items using the same distance measurement functions described previously.

3.3 LSA Learns Language

The pioneer research presenting LSA as an unsupervised learning system was presented in the 1997 paper by Landauer and Dumais describing LSA as "a solution to Plato's problem" [17]. Plato's problem is the name of a persistent problem in cognitive science. It is basically the question of how humans develop the knowledge they have based on the relatively limited amount of information they have received. Peoples' knowledge exceeds simply the facts and experiences that they have been exposed to over time. In their paper, Landauer and Dumais put forward LSA as a model of human learning as an answer to this question, marking a change from viewing LSA as retrieval method to that of a powerful learning system.

Landauer and Dumais proposed that underlying domain knowledge is a vast number of interrelations. When properly exploited, these interrelations can increase learning by the process of inference. This mechanism, under the right choice of dimensionality, can represent the meaning of words and contexts and their similarities in such a way that the acquisition of knowledge is enhanced, bridging the gap between "information in local contiguity and what people know after large amounts of experience" [17]. More succinctly, people learn things that they were not directly taught. LSA is a mathematical model of the meaning represented in human language and the acquisition of knowledge. While it lacks certain capabilities present in our human cognitive model such as word order, syntax, morphology, etc., the representation produced by the LSA process is highly similar to that of humans. It is based on concept and semantics not keywords or syntactic constructs [17, 25].

The LSA representation of words and contexts as vectors in a vector space where the meaning of each is represented as a point is a model of how people acquiring meaning from experience. The assumption is that people learn by associating words and contexts that are experienced together over time. This provides a dynamic framework for predicting new experiences. Human cognition takes in all these experiences, word and context co-occurrences, and fits them into a semantic map that represents how each word, object, and context is related to each other. The mathematical model of LSA does much the same thing. It constructs a semantic space by first digesting a large body of textual information, and then it combines all these links to form a common semantic space. This second step is what constitutes LSA's understanding of meaning. LSA will bring those words that never co-occur together, like synonyms, together and separate those contexts that share common terms apart if they do not mean the same thing [26].

3.3.1 Unsupervised Learning

Using LSA a machine can learn in much the same way that a human does in the fact that it can be fed much of the same empirical associational data that literate humans use in learning. While the input is electronic text and a human has multiple learning sources, LSA does remarkably well as a learning system. Even without prior specific, or any, knowledge, LSA can produce a linguistic and cognitively effective representation of word meanings [18]. LSA is different from traditional natural language processing (NLP) methods because it requires no prior human intervention: no dictionaries, no grammars, no syntactic parsers, no morphologies, and no knowledge base. LSA simply takes raw text as input, separated into passages, and parsed into a collection of distinguishable words [20]. Traditional NLP uses cues from the structure of a language to infer meaning. It is syntax based. The rules for such systems must be developed with expert human input, are language specific, perhaps domain specific (limited to a certain form of usage for a particular language), and are subject to aging. NLP is not an unsupervised learning system; it may not even be considered a learning system in many cases.

3.3.2 The LSA Model of Learning

As previously mentioned, LSA is based on what has come to be known as the Compositionality Constraint: The meaning of a document is a sum of the meaning of its words, and the meaning of a word is defined by all the contexts in which it appears (and does not appear). By using the SVD to solve a set of simultaneous linear equations representing the content of a collection, a hyper-dimensional semantic space is obtained where each term, and each document, is mapped by a vector. This representational technique exploits mutual constraints on the occurrence of many words in many contexts, and the resulting representation allows for similarities to be observed between the words and contexts based on their position in the mapping space. Word association is extremely important to human cognition, and that association is what LSA models [18]. The resulting system can be viewed as both a way to automatically learn the meaning of words and contexts as well as a computational model for the process of human learning itself [27].

It is important to note that since LSA only uses the supplied set of input text, it does not always capture all human knowledge. Some of this shortcoming is due to the insufficiency of the training corpora to represent the language experience of a typical person [27]. LSA only induces the meaning of words and contexts through the analysis of text alone, not from perceptual information, instinct, feelings, etc. However, LSA can access knowledge about those processes through the written word and produce a close enough approximation to peoples' knowledge to represent those concepts. Landauer describes LSA's capabilities "to be analogous to a well-read nun's knowledge of sex, a level of knowledge often deemed a sufficient basis for advising the young" [20].

Dimensionality

An important factor in the formation of the LSA space is the choice of dimensionality. LSA does not use the complete SVD, but rather a truncated SVD that will equate to a k-rank approximation of the term-by-document matrix A. This dimensional reduction is desirable because it has the effect of removing noise from the original representation of A. Selection of k, the number of dimensions, is crucial. Including too few dimensions will leave out important underlying concepts in the data, but using too many dimensions will introduce excessive differentiation between the items making it difficult to find the underlying important semantic relationships [8, 14, 16]. It is theorized that reducing the number of dimensions to be smaller than the initial number of contexts or words produces the same approximate relations as occur in human cognition. Ideally, finding the right number of LSA dimensions produces a representation that is analogous to the same dimensionality as the source that generates the semantic space of a human [25]. The high dimensionality of LSA's semantic space is key and mimics the structure of the brain along with the statistical structure of experience [26].

Orthogonal Axes

One of the products of the SVD computation is a set of orthogonal axes that are the mapping dimensions of the semantic space. Terms and documents derive their meaning from the mapping on these axes, but they do not define the axes [8, 16]. Dimensional axes in the semantic space are an abstract feature. They are the foundation or framework for the vector space, not the words or any nameable characteristic, and they should not be interpreted as such [18, 50]. Attempts to force a labeling of these axes in humanly recognizable concepts tend to be contrived. There is no reason to require that these axes be more than what they are, namely a coordinate system for locating items within the semantic space. Identifying concepts and features represented in the LSA space is best achieved by identifying semantic clusters within the space and other analysis techniques for investigating the relationships it contains.

Meaning

Success in inferring meaning from this system of linear equations is critically dependent on the availability of a sufficient amount and adequate quality (garbage in = garbage out) of input. In a LSA learning system, the meaning of a word is learned not just by the number of times it has been seen in a context, but also by the tens of thousands of times it has been observed absent from a context. "Greater amounts of text help define the space by providing more contexts in which words can co-occur with other words" [28]. LSA must be trained on enough passages before the meaning of any word can be distinguishable from other words. LSA cannot learn word meanings used in contexts on which it was not trained. Small or domain specific corpora are not sufficient to train the LSA learning system [18].

The SVD uses all the linear relations in the given dimensionality to determine the word vector that best predicts the meaning of all of the contexts (pieces of text) in which the word occurs. Additionally, an LSA learning system will not recall a passage verbatim, but it will convey a meaning in the semantic space [18]. "This expresses the belief that a representation that captures much of how words are used in natural context captures much of what we mean by meaning" [17].

LSA mimics thinkable meaning. Word order is not considered, but it is not the most important thing in learning. Inductive association from the word usage in contexts is sufficient for learning the meaning of words and performing the way humans do. It is true that LSA will not represent the meaning of passages that depend strongly on word order syntax or perceptual knowledge not captured in the textual corpus from which LSA learns. The comprehension and representation of meaning is modeled by LSA, not how the passages are constructed or produced [18].

3.3.3 Evidence of the Model

The semantic representation employed by LSA has been shown to induce meaning of words and passages and their similarities in a way that simulates human like similarity judgments and behavior. LSA is not necessarily the way the brain actually acquires knowledge, but is a model, though not a complete one. The pioneering study described in detail in Landauer and Dumais [17] demonstrated this through an experiment using an LSA simulation of an older grade-school child which learned on average 10 or more new words after reading a modest amount of text (approximately 50 paragraphs). Notably, the learned words were not necessarily included in the text that was read. This excessive learning rate observed in the experiment was analogous to the vocabulary acquisition of a school-age child suggesting that the LSA model was an answer to "Plato's problem." Evidence that LSA is mimicking how a human learns language and acquires knowledge comes from this and other simulation of human performance. The performance of LSA on multiple choice and domain knowledge tests as well as essay grading applications show this is true [18].

Text Comprehension and Coherence

One of the first research works to indicate that LSA could be used for much more than information retrieval was in text comprehension research. The semantic structure revealed in the vector representation of LSA allowed for semantic similarity between textual information to be measured quantitatively. In the study of text comprehension, it is essential to understand what components affect the ability of a reader to comprehend a passage. Common practice for gauging reading comprehension is to have a student read a given text and then write a summary for that text. The summarization is examined to determine how much information the student has learned from the text [28].

One method for studying text comprehension is to develop a human coded cognitive model of student's representation of the text read. This can be done using propositions, which are sets of semantic components drawn from clauses in the text, to represent the semantic information in a passage. A linking of the propositions within the textual information is constructed to represent the semantic structure of both the original text and the summary. Then, a semantic comparison is performed between the two propositional analyses [29]. LSA provides a fully automated way to implement this cognitive modeling process based on propositions without the need for human intervention. The cosine similarity between two texts represented in a LSA semantic space provides an estimated semantic similarity.

First, LSA was used to look at a subject's summary to determine how much information he/she had learned by examining what textual documents had the most influence on the subject's recall. Sentences from the subject's summary were compared to the sentences in the original text to determine which ones had the most influence on the subject's understanding. The results were promising, performing within 7 % agreement of human raters. A second experiment involved characterizing

the content quality of each summary as a whole. The results indicated that LSA performs as well as human raters at judging the quality of the summary. Finally, a third experiment was performed to measure text coherence in the summaries. LSA was used to examine the semantic similarity or overlap between sections of a summary. The LSA predictions of coherence were generated by calculating the semantic overlap between adjoining sentences in a text and then computing a mean for all the cosines between the sentences. This single number or score was used to represent the mean coherence of the text and was compared to scoring based on straight word overlap between sentences. The results indicated that LSA did capture coherence based on semantics not shared words between adjoining sentences [28]. As demonstrated in these experiments and in extensions of them, LSA provides an accurate model of the coherence exhibited in a text that is similar to the propositional modeling that can be done by humans. The representation for a given text by LSA corresponds to the subject's semantic interpretation of the same text, whether written or read. It was also shown that the LSA coherence measurement was, in many cases, an accurate predictor of a reader's comprehension of the text [25].

Synonym Test

In one early study assessing how LSA mimics humans in synonym recognition, an LSA simulated student was given a Educational Testing Services (ETS) multiple-choice synonym test for students where "English is a Foreign Language" (TOEFL). The overall test score achieved by the LSA-simulated student was the same as the average score for students from non-English speaking countries successfully seeking entrance into a college in the USA. From the performance on this test, the LSA model could be considered to be similar to the behavior of a group of moderately proficient English readers, with respect to judging the meaning of similarity between two words [17].

Semantic Priming

Semantic priming is an experiment in which a person reads short paragraphs word by word, and then at selected time points he/she is presented with a lexical decision. This lexical decision involves either choosing a word that is related to one or another sense of an ambiguous word in the paragraph that was just presented, or picking words related to words that did not appear in the paragraph but which are related inferentially to the text the subject should have comprehended up to the stopping point.

In this experiment a Latent Semantic Analysis simulation LSA simulation of semantic priming was performed. In the cases of the second lexical decision, the average cosine between the vector for the ambiguous word presented in the paragraphs and the vectors of the related words presented were significantly higher than the average vector cosine with the unrelated word vectors. For the first type of lexical decision, it was observed that the cosines of the related words to the homograph in the paragraphs and passage vector representation were significantly

higher than the cosines of the unrelated, different sense, word vectors, and the passage vector. The LSA representation of the passage provided the correct meaning to select the "homograph's contextually appropriate associate" [17]. The similarity relations examined in the study support the LSA simulation of the construction-integration theory in cognition [29].

Learning from Text

Learning is a constructive process. Just as when building a house one cannot start with the roof, learning advanced knowledge requires a foundation of primary learning. Much of what a person learns, they learn from reading. In order to learn effectively from a particular piece of text, that text cannot be too difficult or too easy to comprehend based on the background knowledge a person already possesses. New information being processed must integrate with prior knowledge, both in terms of understanding the text and in remembering it for use later in other situations. This information must be somewhat relevant to prior knowledge for this learning to occur; otherwise, a person has no relevant contextual basis with which to integrate the new information.

This study was performed to test whether LSA could be used to judge prior knowledge and then to determine if learning had occurred by analyzing a series of essays written by humans. This involved presenting a group of students with varied amounts of prior knowledge texts of different difficulty level on the same subject. Students provided answers to open-ended prompts on the subject before reading a selected text and again after reading the text. LSA was used to compare the pre-reading responses to the text in question to determine the amount of prior knowledge possessed by the student. The post-reading responses were then compared to the text to measure the amount of learning that had occurred. The findings showed that a student's prior knowledge needs to be at an appropriate level to the instructional text (moderate cosine between pretest essay and text) for effective learning to occur. Students who displayed either little prior knowledge or a high level of prior knowledge compared to the text showed little change in their response evaluation between the pre- and post-reading prompts. The highest degree of learning was exhibited when the student demonstrated some but not complete prior knowledge of the subject before reading the text. The results indicated the capability of LSA to both characterize prior knowledge and assess learning, and support the idea that learning is optimized in the context of proper foundational knowledge [30].

Assessing Knowledge

Another early study examined properties surrounding the ability of LSA to assess knowledge in the work by Wolfe et al. 1998. The semantic comparison (cosine similarity), as estimated by LSA, between certain domain specific instructional information and a student's essay on the topic was shown to be a reliable measure of a student's knowledge of the subject and a predictor of how much a student can learn

from a text. Additionally it was demonstrated that all terms in a student's essay are equally important in assessment of the student's knowledge, not just the technical or seemingly pertinent words. Another notable finding suggests that some aspect of a student's knowledge can be assessed with a relatively small amount of written text from a student [24].

Essay Grading

Very early experiments using LSA in reference to essay grading started by using LSA to model subject-matter knowledge. Given only an introductory psychology textbook as input, a Latent Semantic Analysis simulated student was trained on the text and then tested using a multiple choice test that had been provided with the textbook. The simulated student exhibited performance that was better than chance, receiving a passing score on the test and showing promise in the modeling of subject-matter knowledge. Building on this early work, the use of LSA was explored for grading expository essays. Students wrote essays on a particular subject, which were then scored by both human graders as well as a prototype LSA scoring system. Several different scoring methods were applied. In the resulting analysis, all scoring methods, using multiple different instructional texts, correlated as well with the human scores as the raters' scores did with each other. Additionally, LSA showed better correlation with individual expert raters than the humans did with each other [20].

3.4 LSA Applications

While LSA is not a complete or perfect theory of meaning, it does simulate many aspects of human understanding of words and meanings of text, as shown in these early experiments. It demonstrates a major component of language learning and use. Therefore, it can be used in many applications to replace the workload of humans. Over the past two decades, since the inception of LSA and thinking of it as theory of meaning, LSA has been employed in many research projects and applications.

Many research projects and applications have used LSA technology either as a method for identifying similarities of words and passages in a large body of text based on their meaning or as new model of knowledge representation. LSA has been widely used in many information retrieval situations with much success. It has also been applied to cross-language information retrieval, indexing content in multiple languages simultaneously [13]. Data mining applications of LSA include CareerMap, a tool that matches military personnel to certain occupational tasks based on their training data, and performance assessment tools for team communication [31, 32].

The automated assessment of essays using LSA has been a major area of research and has been shown successful in the Intelligent Essay Assessor (IEA), which

has been used in numerous applications to both score essays as well as provide instant feedback and analysis on essays of various sorts [5–8, 33]. IEA has also demonstrated ability to score creativity writing as well as expository essays [8].

LSA has also been incorporated as a key component of several educational systems. It has been shown that the exercise of writing a summary of informational text leads to improved comprehension and learning. Summary Street, an educational technology tool, helps students by interactively guiding the student through the process improving their written summarization, providing individualized feedback to prompt revision and enhance learning. Summary Street utilizes LSA to analyze the semantic content of the student's summary examining topical coverage and identifying redundant, relevant, and irrelevant information [34]. Another software literacy tool that has incorporated the capabilities and skills of IEA and Summary Street is WriteToLearn (http://www.writetolearn.net/). This technology tool is used in the classroom to help students with writing and reading [35]. AutoTutor is an example interactive intelligent system that tutors a student in certain topic areas. Using LSA to analyze the student's text responses to questions, LSA constructs a cognitive model of the student's response, determines the student's knowledge level, and then interacts with a student to help the student gain more understanding of the topic [36]. Other variants of this technology are being researched and tested [37].

Another area where LSA is used as cognitive model is in the analysis of team discourse. Transcriptions of spoken discourse can be examined to determine if a team is effectively accomplishing a certain mission [38]. Tools have been developed to analyze the content of team communication to characterize the topics and quality of information and give measures of situational awareness [39]. These tools were built upon KnowledgePost, a system using LSA to analyze online discussion groups or online course activity to generate summaries of the discussion and assess the thinking and contribution of individual participants [40–42]. LSA has been used in many additional educational tools which are discussed in The Handbook of LSA [43].

In the area of general text analytics recent applications of LSA include the development of the Word Maturity metric, a technique to model vocabulary acquisition which estimates the age at which different word meanings are learned providing insight to guide vocabulary development [44]. Cohesion and coherence metrics for text based on LSA analysis have been developed in the form of the Coh-Metrix system for assessing the difficulty of a written text [45]. LSA has also been used in Operations Management Research to uncover intellectual insights in unstructured data [46].

Of particular interest in reference to this chapter, a recent research project used LSA as a text mining tool to analyze open-ended responses to a questionnaire. In this application, military service men and women responded to open-ended prompts covering health issues. LSA was successfully used to evaluate the participants' responses to identify important health concerns that were perhaps not identified by the structured part of the questionnaire. Analysis of the open responses was used to identify trends, needs, and concerns [47].

4 Methodology

Using LSA as an unsupervised learning system has proven to be especially beneficial for evaluation of essays and open response items. Its automated nature overcomes the scaling issues attached to the use of human raters while providing consistent, repeatable, non-biased evaluation. The fact that LSA captures meanings, not keywords, makes a concept-based approach to assessment possible. An LSA system is easy to train, though it does require large volumes of general training text as well as some representation of domain expertise to provide an adequate foundation for semantic mapping. LSA provides many of the same capabilities as a human rater without the limitations on throughput and at less expense. The methodology used in the application of LSA toward the task of text analysis and more specifically to the evaluation of essay and open-ended response items is significantly different than that for information retrieval.

4.1 Objective

The objective in analyzing open-ended responses obviously cannot be retrieval oriented. Rather, it is an analytic task that is comparative in nature. For a given response the assessment task is to determine what sort of concepts it expresses, if any desired concepts are missing, and how the information it contains compares to other responses for the same prompt. Since specific content is not being retrieved from the collection, the LSA space is used as a semantic mapping system for concepts and the comparison of projected items. The content of the LSA space forms context or basis for understanding the meaning expressed in the response items as those items are projected into the space. It becomes an interpretive tool for text analysis. Evaluation methods using this semantic mapping system must exploit the concept identification it affords in ways that can yield useful output, whether in the form of comparative scores, feedback recommendations, or other indicators.

4.2 The Base Interpretive Space

As already described, LSA forms a model of meaning by producing a mapping system for representing the semantic associations of both individual words and multi-word pieces of text. In text analysis applications, this mapping system forms a semantic background, what we will call a base interpretive space (BIS), for processing the items to be analyzed. Because the meaning contained in the BIS is learned from the initial content used to build the space, the nature of the meanings represented in the space are completely dependent on the content provided as input

in the space construction process. It is important that the BIS be well formed in order to represent the range of meanings that will be required by the application.

The mapping system provided by the BIS can not only be used for evaluation analysis, but is also useful for other applications where the data set of interest is small or narrowly focused, not providing enough material to establish a basis of meaning. The use of a BIS in these situations provides a contextual background that augments the meaning represented in the data set. While a set of 10,000 documents might seem like a large volume of data to a person, it is an insufficient amount from which to learn meaning. For any LSA application using a BIS, there are several important considerations when selecting content that will be used for its construction. These include the overall size of the training corpus, the presence of relative and distributed content within the corpus, and its overall term coverage.

4.2.1 Corpus Size

Initial research using LSA was attempted on relatively small bodies of content (2000 documents or less) with varied but often promising results. As the availability of computational power and electronic text both increased, the processing of larger corpora became feasible and improvements in the performance of LSA were observed. Poor results reported from some past research studies using LSA were based on the processing of ridiculously small corpora (sometimes as few as ten short paragraphs). Since LSA must learn the meanings of words and documents by forming associations, a sizeable amount of content is required to provide sufficient learning to mimic human understanding of language constructs. Providing too little training data gives insufficient meaning for developing the underlying mapping system. It has been suggested that a minimum of 100,000 paragraph sized passages is necessary to represent the language experience of an elementary student [16]. Our own work has observed that the association of domain specific content is significantly improved as even randomly selected background material is added to the content used to build a space [48].

Currently, the best policy is to use as much content as can be obtained and make the training corpora as large as possible. This of course must be tempered with knowledge of the available computing resources. It is still quite possible to build a corpus so large that it simply cannot be processed. It is also currently unknown if there is a point at which adding content makes no difference or in fact degrades the space. More research in this area is needed.

4.2.2 Relevant and Distributed Content

In addition to a large volume of input text, LSA also needs good quality content that provides concept information not just in the domain of interest, but also enough general content to represent language usage as a whole. LSA forms its representations of meaning from the analysis of text alone, and to build this notion of meaning it needs to have many representative textual associations both in the

present and in prior knowledge of a potential user of the system [16, 20]. The size and content of the corpus that is used to construct the BIS influences whether the LSA representation is similar to an amateur, novice, or expert level of knowledge. If LSA is trained on content including highly technical texts of a particular domain, then it would behave similar to an expert in the field because the LSA representation is much more elaborate [25]. The desire is for the LSA representation to map similar items in close proximity, yet with enough separation to be distinguishable. This sensitivity is especially true in evaluation applications as the evaluation items are typically expected to be about a common body of subject matter [48]. Perhaps one of the failings of past attempts using LSA for information retrieval is the common practice of constructing a space using solely the retrieval targets for content. The BIS must include not just items in the domain but background meanings representative of the wide range of expression.

4.2.3 Term Coverage

Term coverage is an important aspect in the construction of a background space. Terms that are in an evaluation item can only be mapped if they are represented in the BIS. This, in turn, only occurs if the terms are present in the original content used in its construction. Unmapped terms cannot be assessed for meaning. It is essential to provide content that gives adequate term coverage for possible terms used in the expected evaluation items.

Because the meaning of a document is construed as the sum of the meaning of its terms, when some of those terms cannot be assigned a meaning the interpretation of the entire item is questionable. This is no different than if a human reader encountered a document with terms that were simply unknown to him or her. Depending on the number or significance of the unknown terms the entire document may be misunderstood or simply unintelligible. A human reader would at this point give some indication that they did not know what was going on. Unless this is explicitly monitored and reported, an automated system would simply return wrong answers. This very issue has led to flawed reports of poor LSA results due to missing terms for the items being examined [18].

Monitoring the term "hit-rate" as items are evaluated must be performed to flag individual items where the scoring may be questionable and to indicate whether the BIS is adequately formed for the domain of the items being considered. Too many missing terms would indicate that the base space is insufficient for evaluating the content of interest [48].

In general, obtaining sufficient term coverage is not difficult as most of the expected terms in a response set will be typical of the overall language and already in the vocabulary of a reasonably sized corpus. Depending on the type of items being evaluated, consideration must be made for the formality of the language anticipated. Working with transcribed conversational data will differ significantly from a proper essay or an open response short answer item. Adding domain specific content and

specifically including content generated by the subject group in question are both effective strategies for obtaining appropriate term coverage.

4.3 Evaluation Algorithms

Various methods for evaluating essays and open responses have been explored. Several potential scoring methods have been proposed and tested: holistic scoring, target based scoring, essay to all other essay ranking, and other comparisons to an instructional text, either based on selected components of the text or as a whole. In several different tests all of these scoring methods were demonstrated to produce scores that correlated with human rater scoring of the items as well as the raters' scores did with each other. Additionally, LSA correlated better with individual expert raters than the humans did with each other [20]. Out of these methods, three primary evaluation approaches have emerged: target based scoring, near neighbor scoring, and additive analysis.

4.3.1 Target Based Scoring

Of the three primary evaluation methods, target based scoring is possibly the most intuitive approach. Target based scoring is essentially a measurement process comparing how the meaning articulated in the subject response compares to an expected meaning. This expected meaning is the target or "gold standard" to which the responses must map in close proximity [2]. This measurement can be produced quite easily in an LSA system simply by projecting both the target and subject response in the BIS and then computing the cosine similarity, or other desired distance measure, between the projection vectors. The nature of this method limits its output to the production of a single score for the subject essay.

The main challenge in using target based scoring is the establishment of the target. In some cases it is appropriate for a subject matter expert (SME) to provide a target by producing what might be considered an ideal response against which all of the subjects may be compared [8]. The disadvantage to this approach is that it emphasizes the influence of the single SME on the evaluation results. The target might be technically correct but unrealistic for the subject audience. Also, it may not be possible, or appropriate, to define a single target as the "correct" response. Other methods for constructing a target mitigate some of these issues, such as using the average, or centroid vector, of multiple high scoring responses or multiple SME constructed responses. Another method for constructing a target is to use a related instructional text as a target, in part or in whole [30]. Application of this method is limited to cases where such a related text exists.

A secondary challenge for this scoring method is the question of how to interpret the similarity measure as a score. Depending on the application, a simple pass/fail score may be easy to generate, but producing scores at a finer granularity

might not be possible. The magnitude of the distance measure between the target and the subject responses will vary with the application, so relating the distance measurement to a score value requires adjustment for each new scenario.

4.3.2 Near Neighbor Scoring

Near neighbor scoring (referred to as holistic scoring in the IEA related publications) is a method for leveraging multiple pre-scored responses to guide the evaluation of subject response items. A collection of pre-scored items are projected in the BIS and used to provide a scoring reference for new subject items. To generate a score, a new subject item is projected in the BIS and the *k*-nearest pre-scored items are identified using a distance measure. The scores related to these items are then weighted by their distance from the subject item and then averaged to compute a new score for the subject. This scoring method has the advantage of allowing subjects with differing content to receive similar scores, reflecting the idea that a good response might be expressed in multiple ways or possibly focus on different concepts of equal merit [5, 8]. Another benefit of this method is the ability to produce scores on a continuous range.

Limiting factors to this scoring approach are the availability and quality of pre-scored data. A sufficient number of pre-scored items is required before any evaluation can be performed using this method. The scores must be reliable and broadly distributed. If the pre-scored content does not adequately represent the full range of possible scores, the scoring system will not be able to assign scores across the entire range.

4.3.3 Additive Analysis

In our own research we have developed an evaluation technique referred to as additive analysis. This is a component based technique that can be used to produce a holistic score, a component score, or for other non-scoring purposes. Leveraging the idea that a desired subject response should exhibit certain key concepts, it is possible to detect the presence of those concepts by adding content to a subject response to observe how the augmented response projection moves in the BIS. If the content was already present in the subject response, the projection should not move significantly, but if the content is absent from the response then the projection will move to the degree that the new content alters the meaning. The magnitude of the change in the projection can be used to weight scoring assignments or as shown in the case study below to identify appropriate feedback items.

This approach provides advantages over the target based or near neighbor approaches. Desired concepts can be expressed without tying them together as a single target. Analysis concepts can simply be enumerated rather than requiring them to be composed collectively. This method also allows for the detection of individual concepts in the subject response without requiring that the response item

be decomposed in any fashion. Identification of the presence of key concepts can be used to guide the selection of appropriate feedback items.

4.4 Feedback Selection

Scoring is one thing, but providing feedback is another. Scoring is used primarily to indicate the measure of knowledge on a specified topic that has been expressed in a response. Feedback goes beyond simply measurement and provides teaching information for the subject. There are several different purposes for feedback:

- Improvement—supplies the user with concepts missing from their response
- Correction—identifies components of the response that are in error and detract from the score
- Reinforcement—identifies components of the response that are correct
- Recommendation—provides related information that may be of interest to the user based on their response

Feedback may be selected based on comparison to a target in the case of target based scoring, or may be selected based on the additive analysis method alone. Ideally, the selection should be driven by detection of the concepts expressed in the subject response being evaluated.

Definition of the feedback items is highly dependent on the application domain and presents a challenge in the development of the feedback system. Some applications require the construction of custom feedback items, while other domains have a readily available pool of potential feedback from which appropriate items may be selected. Construction of useful feedback often requires input from SMEs that can author the specific items desired.

5 Case Study NICHD Project

As an example of a working application that uses LSA as an unsupervised learning system for evaluating open-ended questionnaire responses, we discuss the *Practiced Driver* application. This system was developed with colleagues at Parallel Consulting, LLC (PC) as a functional prototype of an interactive Web-based driver's education system for beginning and novice drivers as part of an NIH-funded research program.

5.1 Background: Driver Training

The purpose of the Practiced Driver system is to assist novice drivers by facilitating the rehearsal of higher-order driving skills and the development of tacit knowledge

for handling the demanding situations involved in operating a motor vehicle. Tacit knowledge is a different sort of knowledge than can be explicitly articulated and taught through direct instruction, such as traffic rules, but which can be gained through experiential learning. The Practiced Driver system provides this as a two part training process where novice drivers first engage in a Web-based exercise presenting them with a driving scenario and then prompting them to provide open-ended responses to questions about how they would handle the situation described in the scenario. During the exercise, the participants are immediately presented with tailored constructive feedback based on their responses suggesting areas of concern and providing them with strategies to consider in similar situations. The second part of the process involves suggesting behind the wheel practice recommendations tied to their responses to the scenario prompts and the suggested areas of concern. This is serviced though a client application that can be accessed using a smart phone or other mobile device.

5.1.1 Open-Ended Responses to Scenario Prompts

For the initial version of the Practiced Driver system, two driving scenarios were developed by PC that presented the user with realistic driving situations based on typical circumstances encountered by newly licensed drivers. The scenarios incorporated several hazards that are known to contribute to accidents involving novice drivers. After being presented with the scenario, users are prompted with a series of open-ended questions about how they would handle the scenario and their prior experience with any of the aspects of the scenario.

For each scenario, representative responses were collected from a large pool of novice and experienced drivers in a controlled data collection effort conducted by PC. After filtering responses for completeness and other inclusion criteria, a set of 471 responses for each scenario was available for use. Portions of this response set were used for training and testing the prototype system.

5.1.2 Provide Feedback Suggestions for Improvement

Unlike other applications focused on essay scoring, in this project the interest was almost completely based in the semantic content of the questionnaire responses. Scoring based on construction or readability was not needed in this application, so concerns about grammar and formation of the response items were minimal. The goal of the system was to automatically analyze each scenario response and give the user feedback consisting of tailored, relevant, and safe driving strategies. It was important to avoid making suggestions in areas where a participant already exhibited adequate knowledge, so the system had to recognize those aspects that were already present in the response.

Feedback items selected for the driving scenarios were also tied to suggested practice driving skills to be pushed to the mobile client application which would recommend targeted practice skills. These suggestions were designed to exercise aspects of driver knowledge that were indicated as needing improvement based on the scenario responses of the participant.

5.2 Construction of the Background Space

For this project it was desired to have a background space that first represented general English language usage as well as adequate representation of driving specific language and specific language usage for the teenage demographic that were the primary audience for the driving scenario questions. We experimented with a series of background data sets, noting several improvements in the grouping of related items as content was added to the BIS. In our work on this project, we also explored several cursory metrics for analyzing the characteristics of a LSA semantic space [48].

The final content set used to construct the BIS for the Practiced Driver prototype consisted of a total 112,627 documents including 185,730 unique terms. The content areas represented in the total set include 100,000 general language documents randomly selected from the RTRC corpus [49] to provide a basic foundation of language usage. To this were added 6629 driving subject documents collected from an array of websites. This content consisted of state licensure instruction manuals, driver training material, and articles about driving related topics. An additional 5056 teen-authored articles were collected from high school newspapers published on the Web in order to round out the subject audience vocabulary. After initial training and testing, the 942 scenario response items that had been collected in the initial data gathering effort were also included in the content set.

5.3 Establish Target and Feedback Items

As with all scoring and evaluation systems, the first challenge was to determine how responses would be rated. Since the LSA approach involves quantitative measurements of semantic similarity, one way to achieve this objective is to define a target against which response items could be evaluated. It was decided early in the project effort to use the responses gathered from the data collection effort to guide the selection of both response targets and feedback items. These response items were initially processed and projected into the BIS both to observe the semantic groupings of the response items and to test the performance of the BIS.

5.3.1 Human Input: The SME

To form the initial basis for validating the performance of the system, two human raters reviewed all of the responses to identify the most common strategies used in both good (safe) and poor (unsafe) responses. This information was used to establish a scoring rubric, and the responses were then hand scored on a 4-point scale based the recognition of hazards and the driving strategies employed, with 4 representing the best or "most safe" responses. Using the projection vectors for the response items in each score group, a representative centroid vector was calculated for each of the four groups and the distribution of the response vectors was compared to the score group centroids to verify that the groups were semantically separated in the space.

The safest responses (highest scoring) were selected and then reviewed by SME to identify the very best responses based on safety factors, eliminating any that exhibited strategies that could be dangerous for novice drivers. Professional driving instructors also reviewed the initial scenario response essays and selected the best responses for each scenario prompt.

Using the top rated selected safe responses, a target response was generated by calculating a centroid from the selected group of items. This had the advantage of producing a target that was informed by the SME while also being based on language usage from the participant responses in the semantic representation of the target. Additionally, using a composite target of this sort was an attempt to capture the possibility of there being multiple good responses that each contained different concepts. This target served to describe an ideal "best response" that became the standard against which all responses would be measured for sufficiency. Since score outputs were not being produced by the system, mappings for the lower score points did not need to be created.

5.3.2 Human Selected Feedback Items

While not reporting a score, the Practiced Driver system did need to return constructive feedback to the user. This feedback consisted of individual suggestion items that could be reported to the user based on the concepts they expressed in their response. Feedback snippets were developed by selecting individual sentences from the safest (highest scoring) responses covering all of the driving hazards included in the scenarios. The snippets were chosen to represent concepts or strategies that might be used to mitigate the hazards contained in the scenarios. These snippets were reviewed by SMEs to verify that the suggestions were both safe and practical for novice drivers. The snippets were edited to refine the wording so as to improve their readability for novice drivers. The final feedback database consisted of 26 snippets for scenario one and 28 snippets for scenario two.

5.4 Feedback Selection Method

In this application we selected appropriate feedback items by using an additive analysis process in combination with a target based approach to determine which snippets provided the most improvement based on the relative proximity of a response to the target.

Each user response was projected into the BIS compared to the target response mapping and then augmented with an individual feedback snippet by simply concatenating the text and then projecting the augmented response into the BIS. This projection of the augmented response was again compared to the target response mapping and the change recorded. After performing this process iteratively for each candidate feedback item, the feedback item pairing that resulted in the most improved augmented response closest to the target was selected as a feedback item to be returned. This process was repeated with the new augmented response to find a possible second appropriate feedback item, and again for the selection of a possible third feedback item (see Fig. 4). A minimum threshold of improvement was set to eliminate the production of feedback items that offered no significant contribution. This threshold was established through empirical testing.

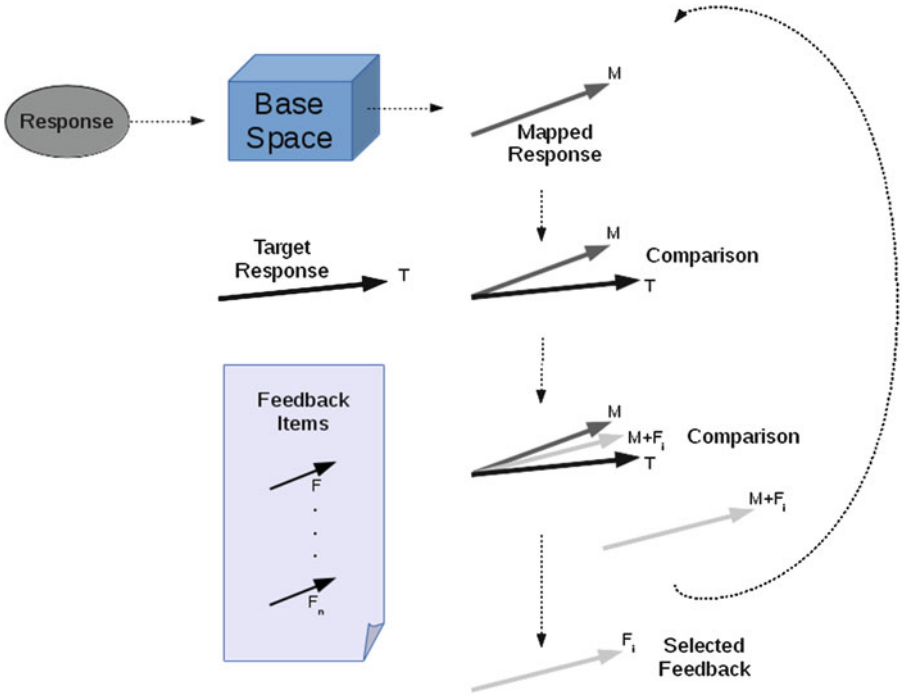

Fig. 4 Flowchart of the feedback selection process

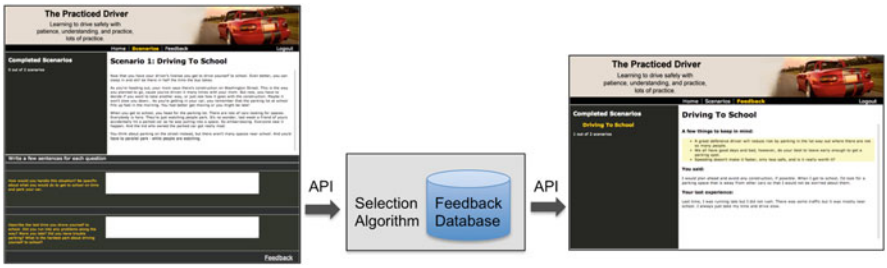

Fig. 5 The Practiced Driver prototype

5.5 Results

The feedback generation process was first validated using the original scenario responses as input. For all teen driver participants, this process produced at least two feedback items with a positive effect (i.e., moving the response closer to the target). The feedback produced from validation testing was reviewed by SMEs to verify that it was appropriate. Testing demonstrated that the feedback items generated were both specific to the concepts not addressed in the user's original response and not redundant with each other.

The prototype system (see Fig. 5) was fielded for demonstration and further testing purposes. The Practiced Driver system was able to interactively provide response feedback in real time for multiple users. Reviews from driving academies and individual driving instructors were all positive. The adaptive nature of the feedback generation, the fact that users will not receive a suggested strategy if they have already included in their response, was seen as a significant quality and desirable capability by the reviewers.

6 Conclusion

While essays and open-ended questionnaires are useful analysis tools, scoring and evaluating them are often problematic due to the need for human input and effort. Human raters limit the scalability of the evaluation task which prevents the use of open response items on a wider scale. Automated systems for evaluating open response items are necessary, but must be able to perform comparably to human raters in order for such systems to be accepted. LSA provides a unique unsupervised learning approach to address the need for automated evaluation. The feasibility of applying LSA to this purpose has been demonstrated by the case study and the several projects described in this chapter.

The types of assessment to be performed vary widely depending on the application domain or the need of the users. Our own current research projects using

LSA in assessment cover the range of measuring diverse and creative thinking, identifying chains of reasoning exhibited in essay responses, and monitoring situational awareness. The development of evaluation methods is ongoing and there remains much work to be done both in the refinement of evaluation algorithms and in exploiting and enhancing the unsupervised learning capabilities of LSA. Many current systems are based on measuring proximity to a defined target or targets in the semantic space, but there are areas of application where nearness to a semantic target is not the true measure of response quality or potentially not even of interest. Development of additional evaluation methods, such as additive analysis, to exploit the capabilities of LSA for analyzing open response type items will allow the extension of LSA into broader areas of application.

Acknowledgments Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute Of Child Health & Human Development of the National Institutes of Health under Award Number R41HD074300. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Computations reported in this research were performed using the LSA_Toolkit™ produced by Small Bear Technologies, Inc. (http://SmallBearTechnologies.com).

References

1. Page, E.B.: The imminence of... grading essays by computer. Phi Delta Kappan **47**(5), 238–243 (1966)
2. Foltz, P.W., Laham, D., Landauer, T.K.: The intelligent essay accessor: applications to educational technology. Interactive Multimedia Electron J Comput Enhanced Learn **1**(2) (1999)
3. Page, E., Petersen, N.S.: The computer moves into essay grading: updating the ancient test. Phi Delta Kappan **76**(7), 561–565 (1995)
4. Page, E.B.: Computer grading of student prose, using modern concepts and software. J. Exp. Educ. **62**(2), 127–142 (1994)
5. Landauer, T.K., Laham, D., Foltz, P.: Automatic essay assessment. Assess. Educ. Principles Policy Pract. **10**(3), 295–308 (2003). doi:10.1080/0969594032000148154
6. Foltz, P.W., Streeter, L.A., Lochbaum, K.E., Landauer, T.K.: Implementation and applications of the intelligent essay accessor. In: Shermis, M., Burstein, J. (eds.) Handbook of Automated Essay Evaluation, pp. 68–88. Routledge, New York (2013)
7. Hearst, M.A.: The debate on automated essay grading. IEEE Intell. Syst. Appl. **15**(5), 22–37 (2000). doi:10.1109/5254.889104
8. Landauer, T.K., Laham, D., Foltz, P.W.: Automated scoring and annotation of essays with the intelligent essay accessor. In: Shermis, M.D., Burstein, J. (eds.) Automated Essay Scoring: A Cross-Disciplinary Perspective, pp. 87–112. Lawrence Erlbaum Associates, Mahwah (2003)
9. Dumais, S. T., Furnas, G., Landauer, T. K., Deerwester, S., Harshman, R.: Using latent semantic analysis to improve access to textual information. In: SIGCHI Conference on Human Factors in Computing Systems, pp. 281–285. ACM (1988)
10. Furnas, G.W., Landauer, T.K., Gomez, L.M., Dumain, S.T.: The vocabulary problem in human-system communication: an analysis and a solution. Commun. ACM **30**(11), 964–971 (1987)
11. Martin, D.I., Berry, M.W.: Latent semantic indexing. In: Bates, M.J., Maack, M.N. (eds.) Encyclopedia of Library and Information Sciences (ELIS), vol. 3, pp. 2195–3204. Taylor & Francis, Oxford (2010)

12. Dumais, S.T.: Improving the retrieval of information from external sources. Behav. Res. Methods Instrum. Comput. **23**(2), 229–236 (1991)
13. Dumias, S.T.: LSA and information retrieval: getting back to basics. In: Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W. (eds.) Handbook of Latent Semantic Analysis. Lawrence Erlbaum Associates, Mahwah (2007)
14. Deerwester, S., Dumais, S.T., Furnas, G., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. **41**(6), 391–407 (1990)
15. Martin, D.I., Berry, M.W.: Mathematical foundations behind latent semantic analysis. In: Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W. (eds.) Handbook of Latent Semantic Analysis. Lawrence Erlbaum Associates, Mahwah (2007)
16. Landauer, T.K.: LSA as a theory of meaning. In: Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W. (eds.) Handbook of Latent Semantic Analysis. Lawrence Erlbaum Associates, Mahwah (2007)
17. Landauer, T.K., Dumais, S.T.: A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychol. Rev. **104**(2), 211–240 (1997)
18. Landauer, T.K.: On the computational basis of learning and cognition: arguments from LSA. In: Ross, N. (ed.) Psychology of Learning and Motivation, vol. 41, pp. 43–84. Elsevier, Amsterdam (2002)
19. Martin, J.C.: How does LSA work. http://smallbeartechnologies.com/lsa-technology/how-does-lsa-work/ (2012). Retrieved 19 Mar 2015
20. Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. Discourse Process. **25**(2–3), 259–284 (1998)
21. Martin, D.I., Martin, J.C., Berry, M.W., Browne, M.: Out-of-core SVD performance for document indexing. Appl. Numer. Math. **57**(11–12), 1230–1239 (2007). doi:10.1016/j.apnum.2007.01.002
22. Berry, M.W., Martin, D.: Principal component analysis for information retrieval. In: Kontoghiorghes, E. (ed.) Statistics: A Series of Textbooks and Monographs: Handbook of Parallel Computing and Statistics, pp. 399–413. Chapman & Hall/CRC, Boca Raton (2005)
23. Martin, J.C.: The LSA_Toolkit™.http://smallbeartechnologies.com/products-and-services/lsa-toolkit/ (2015). Retrieved 19 Mar 2015
24. Rehder, B., Schreiner, M.E., Wolfe, M.B.W., Laham, D., Landauer, T.K., Kintsch, W.: Using latent semantic analysis to assess knowledge: some technical considerations. Discourse Process. **25**(2–3), 337–354 (1998)
25. Foltz, P.W., Kintsch, W., Landauer, T.K.: The measurement of textual coherence with latent semantic analysis. Discourse Process. **28**(2–3), 285–307 (1998). doi:10.1080/01638539809545029
26. Landauer, T.K.: Learning and representing verbal meaning: the latent semantic analysis theory. Curr. Dir. Psychol. Sci. **7**(5), 161–164 (1998)
27. Landauer, T.K., Laham, D., Foltz, P.W.: Learning human-like knowledge by singular value decomposition: a progress report. In: Jordan, M.I., Kearns, M.J., Solla, S.A. (eds.) Advances in Neural Information Processing Systems, pp. 45–51. MIT Press, Cambridge (1998)
28. Foltz, P.W.: Latent semantic analysis for text-based research. Behav. Res. Methods Instrum. Comput. **28**(2), 197–202 (1996)
29. Kintsch, W.: The use of knowledge in discourse processing: a construction-integration model. Psychol. Rev. **95**(2), 163–182 (1988)
30. Wolfe, M.B.W., Schreiner, M.E., Rehder, B., Laham, D., Foltz, P.W., Kintsch, W., Landauer, T.K.: Learning from text: matching readers and texts by latent semantic analysis. Discourse Process. **28**(2–3), 309–336 (1998)
31. Foltz, P.W., Oberbreckling, R.J., Laham, R.D.: Analyzing job and occupational content using latent semantic analysis. In: Wilson, M.A., Bennett, W., Gibson, S.G., Alliger, G.M. (eds.) The Handbook of Work Analysis Methods, Systems, Applications and Science of Work Measurement in Organizations. Taylor & Francis Group, New York (2012)

32. Laham, D., Bennett, W., Landauer, T.K.: An LSA-based software tool for matching jobs, people, and instruction. Interact. Learn. Environ. **8**(3), 171–185 (2000). doi:10.1076/1049-4820(200012)8:3;1-D;FT171
33. Foltz, P.W., Gilliam, S., Kendall, S.: Supporting content-based feedback in online writing evaluation with LSA. Interact. Learn. Environ. **8**(2), 111–127 (2000)
34. Kintsch, E., Caccamise, D., Franzke, M., Johnson, N., Dooley, S.: Summary street: computer-guided summary writing. In: Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W. (eds.) Handbook of Latent Semantic Analysis. Lawrence Erlbaum Associates, Mahwah (2007)
35. Landauer, T.K., Lochbaum, K.E., Dooley, S.: A new formative assessment technology for reading and writing. Theory Pract. **48**(1), 44–52 (2009). doi:10.1080/00405840802577593
36. Graesser, A., Penumatsa, P., Ventura, M., Cai, Z., Hu, X.: Using LSA in autotutor: learning through mixed-initiative dialogue in natural language. In: Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W. (eds.) Handbook of Latent Semantic Analysis. Lawrence Erlbaum Associates, Mahwah (2007)
37. D'mello, S., Graesser, A.: AutoTutor and affective autotutor: learning by talking with cognitively and emotionally intelligent computers that talk back. ACM Trans. Interact. Intell. Syst. **2**(4) (2012). http://doi.org/10.1145/2395123.2395128
38. Foltz, P.: Automated content processing of spoken and written discourse: text coherence, essays, and team analyses. Inform. Des. J. **13**(1), 5–13 (2005)
39. Foltz, P., Lavoie, N., Oberbreckling, R., Rosenstein, M.: Automated performance assessment of teams in virtual environments. In: Schmorrow, D., Cohn, J., Nicholson, D. (eds.) The PSI Handbook of Virtual Environments for Training and Education: Developments for the Military and Beyond. Praeger Security International, Westport (2008)
40. Boyce, L., Lavoie, N., Streeter, L., Lochbaum, K., Psotka, J.: Technology as a tool for leadership development: effectiveness of automated web-based systems in facilitating tacit knowledge acquisition. Mil. Psychol. **20**(4), 271–288 (2008). doi:10.1080/08995600802345220
41. Lavoie, N., Streeter, L., Lochbaum, K., Wroblewski, D., Boyce, L.A., Krupnick, C., Psotka, J.: Automating expertise in collaborative learning environments. J. Asynchronous Learn. Netw. **14**(4), 97–119 (2010)
42. Streeter, L., Lochbaum, K., Lavoie, N., Psotka, J.: Automated tools for collaborative learning environments. In: Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W. (eds.) Handbook of Latent Semantic Analysis. Lawrence Erlbaum Associates, Mahwah (2007)
43. Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W. (eds.): Handbook of Latent Semantic Analysis. Lawrence Erlbaum Associates, Mahwah (2007)
44. Biemiller, A., Rosenstein, M., Sparks, R., Landauer, T.K., Foltz, P.W.: Models of vocabulary acquisition: direct tests and text-derived simulations of vocabulary growth. Sci. Stud. Read. **18**(2), 130–154 (2014). doi:10.1080/10888438.2013.821992
45. McNamara, D.S., Graesser, A.C., McCarthy, P.M., Cai, Z.: Automated Evaluation of Text and Discourse with Coh-Metrix. Cambridge University Press, Cambridge (2014)
46. Kulkarni, S.S., Apte, U., Evangelopoulos, N.: The use of latent semantic analysis in operations management research. Decis. Sci. **45**(5), 971–994 (2014). doi:10.1111/deci.12095
47. Leleu, T.D., Jacobson, I.G., Leardmann, C.A., Smith, B., Foltz, P.W., Amoroso, P.J., Smith, T.C.: Application of latent semantic analysis for open-ended responses in a large, epidemiologic study. BMC Med. Res. Methodol. **11**, 136 (2011). doi:10.1186/1471-2288-11-136
48. Martin, J.C., Martin, D.I., Lavoie, N., Parker, J.: Quantitative Metrics Assessing the Quality of a Large Hyper-dimensional Space for Latent Semantic Analysis (To Appear)
49. Lewis, D., Yang, Y., Rose, T., Li, F.: RCV1: a new benchmark collection for text categorization research. J. Mach. Learn. Res. **5**, 361–397 (2004)
50. Martin, D.I., Berry, M.W.: Text mining. In: Higham, N. (ed.) Princeton Companion to Applied Mathematics, pp. 887–891. Princeton University Press, Princeton, NJ (2015)