Kakkonen, T., Myller, N., Sutinen, E., & Timonen, J. (2008). Comparison of Dimension Reduction Methods for Automated Essay Grading. *Educational Technology & Society*, *11*(3), 275–288.

Comparison of Dimension Reduction Methods for Automated Essay Grading

Tuomo Kakkonen, Niko Myller, Erkki Sutinen and Jari Timonen
Department of Computer Science and Statistics, University of Joensuu, Finland // niko.myller@cs.joensuu.fi //
Tel. +358 13 251 7929 // Fax. +358 13 251 7955

ABSTRACT

Automatic Essay Assessor (AEA) is a system that utilizes information retrieval techniques such as Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), and Latent Dirichlet Allocation (LDA) for automatic essay grading. The system uses learning materials and relatively few teacher-graded essays for calibrating the scoring mechanism before grading. We performed a series of experiments using LSA, PLSA and LDA for document comparisons in AEA. In addition to comparing the methods on a theoretical level, we compared the applicability of LSA, PLSA, and LDA to essay grading with empirical data. The results show that the use of learning materials as training data for the grading model outperforms the k-NN-based grading methods. In addition to this, we found that using LSA yielded slightly more accurate grading than PLSA and LDA. We also found that the division of the learning materials in the training data is crucial. It is better to divide learning materials into sentences than paragraphs.

Keywords

Automatic essay grading, Dimensionality reduction, Latent semantic analysis, Probabilistic latent semantic analysis, Latent Dirichlet allocation

Introduction

In this paper, we compare in the context of automated essay grading three well-known dimensionality reduction methods, namely *Latent Semantic Analysis* (LSA) (Deerwester et al., 1990; Landauer et al., 1998), and related statistical models *Probabilistic LSA* (PLSA) (Hofmann, 2001) and *Latent Dirichlet Allocation* (LDA) (Blei et al., 2003), all of which are commonly used in *information retrieval* (IR). All of these techniques assume that one can model documents as mere collections of words without giving any consideration to word order. Because these methods try to disambiguate words with multiple meanings (i.e. polysemes) and to find groups of words that belong to the same context or that have the same meaning (i.e. synonyms), they are suitable for comparing similarities between two documents at the semantic level rather than simply at the level of keywords. These three methods all take a data-driven approach to the problem of learning the meanings of words. Because the meanings are based on the contexts of word usages, no thesauri or dictionaries are needed in the process.

It is possible to automatically grade essays (i.e. free-text responses in examinations) by comparing them to select learning materials and human-graded essays. We used these learning materials as a comparative basis for determining the amount of relevant content in an essay. In order to fine-tune this method so that it produces grades that correspond to those given by human assessors, it is necessary first to train the system with human-graded essays. *Dimensionality reduction* refers to the process of giving individual words in the model weights that correspond to their significance in the context of the topic as a whole. The purpose of the dimensionality reduction step is to reduce the noise and unimportant details in the data so that the underlying semantic structure can be used to compare the content of essays.

Since the dimensionality reduction problem in automated essay grading is intuitively clear, we can state it in the following way. The more we make use of pure word occurrences, the more we emphasize various details (such as completely irrelevant words in the corpus) as the basis for grading essays. But what is really required is to focus on the most important of those words that embody the conceptual content of the learning corpus. This sort of procedure is in fact similar to the kind of work that is performed by a human assessor. As she marks, a human assessor identifies important concepts but skims over or pays less attention to the peripheral content of whatever essay she may be marking. One could characterize this process of paying less attention to peripheral contents in manual essay grading in IR-terms as dimensionality reduction.

Even though LSA performs well in various IR tasks apart from automatic essay grading, it is characterized by some undesirable characteristics that we will discuss later in this article. In order to resolve the problems that appear in

LSA, new probabilistic models such as PLSA and LDA have been proposed. While PLSA and LDA have been shown to produce better results in IR tasks (Blei et al., 2003; Hofmann, 2001; Yu-Seop et al., 2002), their performance has not yet been tested in automated essay grading. We have therefore implemented them as a part of our essay grading system.

We will begin with a brief review of earlier work on automatic essay grading and on LSA, PLSA and LDA. This is followed by an introduction to the architecture of the essay grading system, *Automatic Essay Assessor* (AEA). Descriptions of LSA, PLSA and LDA in essay grading domain can be found in subsequent sections. We will report on the results of an empirical comparison of the methods in following section. Finally, we will offer our conclusions and indicate the research that we still intend to carry out in the field of automated essay grading and IR.

Previous Work

Text Categorization and Dimensionality Reduction

Automatic essay grading is closely related to automatic text categorization, which has been researched since the 1960s. Many of the methods that are used in other IR tasks have been found to be applicable to text categorization. Thus, for example, *support vector machines* (SVMs) and dimensionality reduction methods such as LSA, have been successfully applied to the problem (see Sebastiani (2002) for a detailed overview of research into text categorization).

LSA, PLSA and LDA have all been demonstrated to be reliable methods in IR. The earliest model LSA (also known as Latent Semantic Indexing (LSI)) has been successfully applied to various IR tasks from information filtering (Foltz and Dumais, 1992) and classification (Zelikovitz and Hirsh, 2001) to image retrieval (Praks et al., 2003). Early applications of PLSA and LDA also include document retrieval and classification (Blei et al., 2003; Hofmann, 2001; Brants, 2005).

Several studies have shown that PLSA outperforms LSA in document modeling and classification tasks (Hofmann, 2001; Yu-Seop et al., 2002). When Hofmann (2001) compared the performance of the two methods in document modeling by measuring the perplexity of the language models and by comparing them to a unigram baseline model, he found that the improvement of the standard LSA model was less than a factor of two whereas PLSA yielded improvements of a factor of more than three.

Similar results were observed on a document retrieval task performed with the four standard medium-size document collections, MED, CRAN, CACM, and CISI. In these cases, the performance of LSA and PLSA was compared against a baseline term frequencies-based vector space model. While LSA improved the precision results by 0.8% and 16.7% on two of the test sets, it delivered lower precision rates with the two other test sets. PLSA improved the precision results 14.7%...58.3% on all the test sets. The difficulties encountered during the selection of the optimal dimension for LSA is a well-reported problem (Landauer et al., 1998; Bingham and Mannila, 2001; Globerson and Tishby, 2003), and the selection is typically performed by *ad hoc* heuristics. Blei et al. (2003) have shown that LDA outperforms PLSA in document classification and collaborative filtering tasks with medium-size collections.

Automatic essay grading

Research to develop computer systems for automatic essay grading has been carried out since the 1960s and several approaches have been proposed. *Project Essay Grade* (PEG) (Page, 1966; Page and Petersen, 1995) uses multiple regression techniques, and Larkey applies text categorization techniques and linear regression methods. *Bayesian essay testing system* (BETSY) (Rudner and Liang, 2002) is based on Bayesian Networks, *Intelligent Essay Assessor* (Foltz et al., 1999) uses LSA for the content analysis of essays, and *E-RATER* (Burstein et al., 1998) is a hybrid system that combines *Natural Language Processing* (NLP) and statistical techniques.

An automatic grading module is central to any essay assessment system. Content-based grading can be performed by means of two methods: (1) by comparing an essay to human-graded essays and assigning the grade based on the grades of the *k* nearest neighboring essays, and (2) by basing the grading on both human-graded essays and course content. Of the three methods discussed in this article, only LSA has been previously used for automatic grading and other educational applications such as an intelligent tutoring system (Wiemer-Hastings et al., 1999) and for assessing student summaries (Wade-Stein and Kintsch, 2003). LSA has proved to be one of the most successful methods for content-based essay grading. Depending on the test set, Landauer et al. (1997) and Foltz et al. (2000) have, for example, reported correlations from 0.64 to 0.84 between grades given by two human assessors and correlations from 0.59 to 0.89 between the LSA-based grading system and human graders. This means that LSA-based systems perform as well as the human graders. Our aim was to determine whether the use of PLSA and LDA in our grading system would improve the accuracy of the grading and help us to avoid the problems that are characteristic of the LSA model.

According to Kaplan, Wolff, Burstein, Li, Rock, and Kaplan, the four quality criteria for an automated essay grading system are *accuracy, defensibility, coachability* and *cost-efficiency*. For a system to be acceptable, it must deliver on all these criteria. An *accurate* system is capable of producing reliable grades measured by the correlation between a human grader and the system. In order to be *defensible*, the grading procedure employed by the system must be traceable and educationally valid. In other words, it should be possible rationally to justify and explain the grading method and the criteria for given grades. *Coachability* refers to the transparency of the grading method. If the system is based on simple, surface-based methods that ignore content, students could theoretically train themselves to circumvent the system and so obtain higher grades than they deserve. It is also self-evident that an automated grading system must be *cost-efficient* because its ultimate purpose is to reduce the total costs of assessment.

Of the four requirements, *accuracy* is most readily accounted for by all the systems represented in Table 1. Results reported as early as the nineteen sixties indicated that automated grading systems can grade essays as accurately as humans (Page, 1966). Since the cost-effectiveness of the systems is dependent on the number of graded essays that are graded, it would not make sense to collect a set of several hundred human-graded essays in order to grade just a few. It is from this point of view that the methods based on LSA are most effective because LSA can apply course materials as a basis for assessment.

The most problematic requirement for the current systems is probably coachability. If a writer is familiar with the grading principles of the system, he can mislead the system to render better grades than are deserved. One possible solution to this problem is to automatically identify ‘suspicious’ essays and give them to a human assessor for checking.

Table 1. Comparison of five essay assessment methods according to the criteria proposed by Kaplan et al. (1998)

Method	Accuracy	Defensibility	Coachability	Cost
PEG	Grades as accurate as human graders measured by the correlation.	Relies heavily on measuring surface features. But the scoring principles are easily traceable.	Because the applied measures are simple, coachability can cause problems.	Relatively large number of pre-graded essays are needed, but it is less costly to compute than, for example, LSA.
TCT	Like PEG.	Uses both surface features and content measures, and is thus more defensible than PEG.	Does not pay attention to the structure.	Like PEG.
BETSY	Like PEG.	Measures only the content.	Measures only the content.	Like PEG.
LSA	Like PEG.	Like BETSY.	Like BETSY.	In addition to pre-scored essays, course material can be used for training.
E-RATER	Like PEG.	One of the design principles. The strongest of the systems in this respect.	Also analyzes the structure and organization.	270 essays are required for training.

Automatic Essay Assessor: AEA

Our particular grading system, *Automatic Essay Assessor* (AEA), is designed to grade essays automatically (Kakkonen and Sutinen, 2004). It bases grading on a scoring model that has been created by comparing a set of manually graded essays to course materials. While this system is not limited to a particular language, it currently supports only Finnish. It is also possible, in addition, to apply several dimensionality reduction methods for modeling and measuring the content similarities between documents. The system currently contains implementations of LSA, PLSA (Kakkonen et al., 2005a), and LDA (Kakkonen et al., 2006), and this is discussed in more detail in the following section below. The structure of AEA is represented in Figure 1.

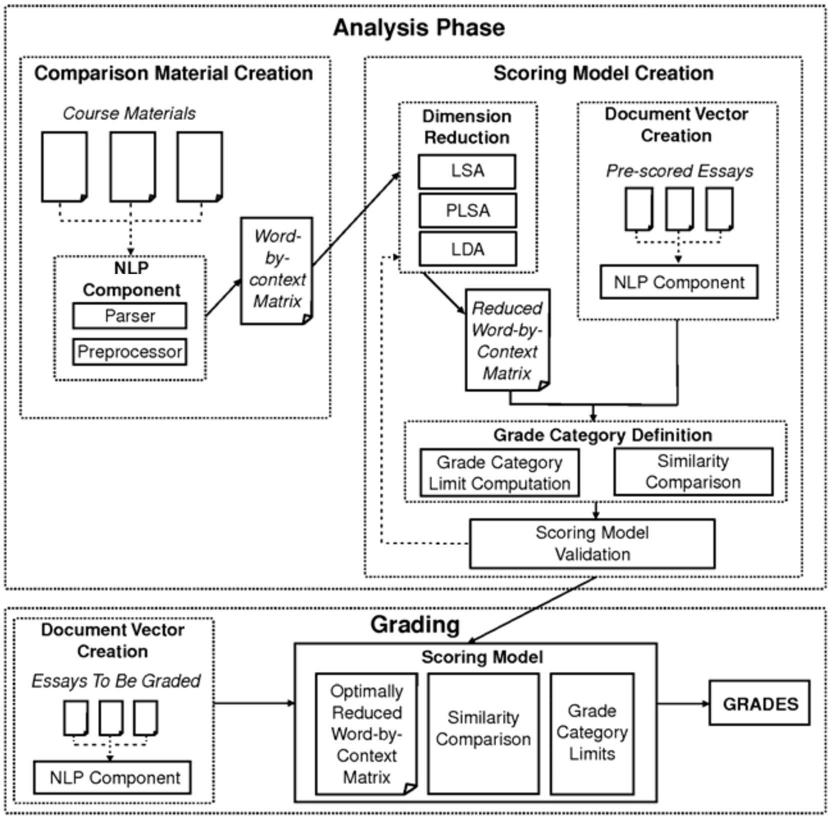

Figure 1. The architecture of AEA

Figure 1 shows how the system logically consists of two main parts: *analysis* and *grading* phases. In the analysis phase, the *scoring model* is created from the course materials (passages from course textbooks, lecture notes and pre-graded essays). The text passages are selected from the course materials on the basis of their relevance to the essay prompt; all the passages that the students are supposed to have read in order to be able to answer the question successfully are used as the bases for the evaluation. In the grading phase, the essays are graded according to the scoring model created in the analysis phase.

There are three main components at the implementation level. These are NLP, the grade definition, and the dimension reduction components. The *NLP component* consists of a syntactic parser and preprocessor. We use the Constraint Grammar Parser for Finnish (Lingsoft Inc., 2007) in order to produce the base forms of words in the input documents. The *preprocessor* performs the standard preprocessing stages that are used in LSA, PLSA and LDA, namely, stopword removal and entropy-based term weighting. If AEA is applied to languages other than Finnish, the parser and stopword lists must be replaced with relevant ones. We used the NLP component for two distinct tasks: firstly, to create a *word-by-context matrix* (WCM) representing the course materials, and, secondly, to build document vectors of the *pre-scored essays* and essays to be graded. *Pre-scored essays* in this context refer to those essays that have been given a grade by the teacher and that are used for calibrating the AEA before the actual grading. Because WCM contains the number of occurrences of each word in each context in the course materials (i.e. document contexts, the paragraph contexts, or the sentence contexts), it is indeed a collection of document vectors from all possible contexts.

Scoring model creation progresses in the following way. Firstly, the WCM (representing the course materials) is processed in the *dimensionality reduction component* with LSA, PLSA, or LDA in order to create a reduced-dimensional representation of the WCM. The reason for this step is to reduce the noise in the WCM. This allows the unimportant details to dissipate and the underlying semantic structure to become more patent. (While issues relating to these dimensionality reduction methods will be further discussed in the next section, a more detailed description of the scoring model creation can be found in Kakkonen et al. (2005a) and Kakkonen et al. (2005b).)

The next step is to compare the document vectors that have been created from the manually graded essays to the reduced WCM so as to determine the similarities between each essay and the course materials. This comparison between an individual student essay and the reduced WCM refers to the process whereby the distances between a document vector of an essay and every document vector in the reduced WCM are calculated and summed to form the similarity value between the essay and the WCM. In order to do this, we apply the cosine between the document vectors as a measure of their semantic similarity. On the basis of these similarities, the predefined grade categories, $g_0, g_1,...,g_{G-1}$ are associated with similarity value limits, $l_0, l_1,..., l_G$, where G is the number of grade categories, and $l_G = \infty$ and normally $l_0 = 0$ or $l_0 = -\infty$ and l_j ($0<j<G$) are, for example, the average of similarity values of training essays in the grade categories g_{j-1} and g_j.

The next step is *the grading phase*. In this phase, the NLP component creates document vectors from each of the essays to be graded. Grades are determined with the scoring model consisting of the reduced WCM and limits for the grade categories. The essay's document vector *d* is compared to WCM in order to define the similarity value as described above. An essay is assigned to a single grade category on the basis of the similarity value, *sim*(*d*), and the limits of grade categories as follows: $l_i < sim(d) \le l_{i+1} \rightarrow d \in g_i$.

The scoring model is then *validated* by measuring the Spearman correlation between the grades given to test essays by the system those given by the human assessors. This phase is essential (especially when applying LSA) because the selection of dimensionality in the method is somewhat arbitrary.

Dimensionality Reduction Methods in AEA

Dimensionality reduction in AEA refers to the process by means of which the individual words that will be used for comparing essays with learning materials are assigned a weight according to their significance. In order to determine the optimal parameters for dimensionality reduction, one needs to train AEA with different parameters and to predict the corresponding grades for essays by using each of these models (Kakkonen et al., 2005b). We shall now review the three dimensionality reduction methods (LSA, PLSA and LDA) that are used by AEA. The results of this comparison of LSA, LDA and PLSA are reported in later section.

Latent Semantic Analysis

Dimensionality reduction in LSA is based on *singular value decomposition* (SVD), a form of factor analysis. In LSA the original WCM is approximated by decreasing the number of singular values in the SVD of WCM. This has been shown to increase the dependencies between contexts and words (Deerwester et al., 1990; Landauer et al., 1998).

While LSA offers a feasible method to compare the similarities between two documents, there are some problems inherent in this method. The number of singular values in the dimensionality reduction is usually selected by means of an *ad hoc* heuristics. While dimensionality reduction can result in a reduced WCM that contains negative values, this is not necessarily a problem because it allows the document vectors to be defined in a larger subspace (by comparison with a subspace that allows only vectors with positive components), and this could be helpful, especially when LSA is used with small training sets. In spite of this, the definition of the WCM becomes problematic because it is unclear what the meaning of a context with a negative number of words is. This would, in addition, prohibit the use of the reduced WCM to define probability distributions. Problems with dimensionality reduction have also been reported in other LSA-based applications (Landauer et al., 1998; Bingham and Mannila, 2001; Globerson and Tishby, 2003).

The selection of dimensionality in LSA is fraught with problems (Bingham and Mannila, 2001; Globerson and Tishby, 2003; Landauer et al., 1998). Kim et al. (2003) concluded on the basis of their experiments that LSA and PLSA dimensionality do not have a specific linkage to semantic knowledge construction. Kim et al. (2003) were unable, in other words, to find the constant dimension that fits their experimental data. This implies that a single dimension is not applicable to all essay collections in AEA. The problems generated by dimensionality reduction led us to develop ADRM, an automatic dimensionality reduction method that searches case-specifically for the dimension that best fits each set of essays (Kakkonen et al., 2005b). But since we have not yet modified ADRM for PLSA and LDA, the experiments reported in these articles were run on the standard LSA that repeats the scoring for all the possible dimensions.

Probabilistic Latent Semantic Analysis

Although we found that the practical performance of LSA is good in many IR tasks as well as in essay grading, it has been discovered that it possesses additional flaws (Quesada, 2003; Hofmann, 2001) apart from the dimensionality selection problem. LSA, for example, does not define a proper probability distribution, and, even more seriously, the reduced matrix can contain negative values. To solve these problems, Hofmann (2001) proposed PLSA, a probabilistic extension to LSA, which we had already utilized PLSA in automated essay grading (Kakkonen et al., 2005a). PLSA is based on a statistical model that is referred as *an aspect model*. An *aspect model* is a latent variable model for co-occurrence data, which associates unobserved class variables $z_k, k \in \{1,2,...,K\}$, with each observation, where *K* is the number of latent classes. While the number of latent classes is an important parameter that needs to be selected in the same way as LSA, there are, as we will point out below, several existing solutions to this problem. In our settings, the *observation* is an occurrence of a word $w_j, j \in \{1,2,...,M\}$, in a particular document/context $d_i, i \in \{1,2,...,N\}$, as in WCM, where *M* is the number of words and *N* is the number of documents in the collection. *Latent classes* can be understood as the *topics* that comprise the text. The probability distributions that associate the latent variables with words and documents describe how closely they are associated with each topic. The generative model for the observation is defined as follows:

1. Obtain a document d_i in which a word occurrence will be observed with probability $P(d_i)$.

2. When the document d_i is known, select the topic z_k of the word with probability $P(z_k \mid d_i)$. This probability distribution is also a measure of the extent to which the document is relevant to each topic.

3. When the topic is known, select a word w_j whose occurrence is observed with probability $P(w_j \mid z_k)$.

In this way the observation pair (d_i, w_j) can be formulated so that the latent class variable can be summed out. Equation (1) shows the probability of observing a pair (d_i, w_j).

$$P(d_i, w_j) = P(d_i)P(w_j \mid d_i)\text{, where } P(w_j \mid d_i) = \sum_{k=1}^{K} P(w_j \mid z_k)P(z_k \mid d_i) \quad (1)$$

When one uses PLSA in essay grading or IR, one's first task is to construct the model. This means approximating the probability mass functions from the training data with machine learning techniques. In our own case we used the comparison materials consisting of the assignment of specific texts as training material. Hofmann (2001) proposes the use of the Expectation Maximization algorithm to identify the parameters for the probability mass function from the training materials.

The *Expectation Maximization (EM)* algorithm can be used to build a model with a maximum likelihood formulation of the learning task (Dempster et al., 1977). In EM, the algorithm alternates between the following two steps: (i) an *expectation (E)* step in which posterior probabilities are computed for the latent variables on the basis of the current estimates of the parameters (see Equation (2)), and (ii) a *maximization (M)* step in which parameters are updated on the basis of the minimization criteria and in dependence on the posterior probabilities computed in the E-step (see Equations (3) and (4)).

$$P(z_k \mid d_i, w_j) = \frac{P(w_j \mid z_k) P(z_k \mid d_i)}{\sum_{i=1}^{K} P(w_j \mid z_l) P(z_l \mid d_i)} \qquad (2)$$

$$P(w_j \mid z_k) = \frac{\sum_{-=1}^{N} n(d_i, w_j) P(z_k \mid d_i, w_j)}{\sum_{m=1}^{M} \sum_{i=1}^{N} n(d_i, w_m) P(z_k \mid d_i, w_m)} \qquad (3)$$

$$P(z_k \mid d_i) = \frac{\sum_{j=1}^{N} n(d_i, w_j) P(z_k \mid d_i, w_j)}{\sum_{m=1}^{M} n(d_i, w_m)} \qquad (4)$$

In the equations, $n(d_i, w_j)$ stands for the count of the word w_j in the document d_i.

The standard EM algorithm can, however, overfit the model to the training data and thus perform poorly with unseen data. Because this algorithm is iterative and converges relatively slowly, it can increase runtime dramatically, especially with large data sets. Because of this, Hofmann (2001) proposed another approach that he calls *Tempered EM* (TEM), which is a derivation of standard EM algorithm. (We refer the interested reader to Hofmann (2001) for further information.)

In PLSA, a query *q*, or, in our case, an essay, can be 'added' or *folded into* the model by using TEM. When folding in a new essay, the only difference is that the probabilities $P(w_j \mid z_k)$ are retained as they are, and only the probabilities $P(z_k \mid q)$ are updated during the M-step.

The similarity between a document or comparison material d_i in the model and an essay *q* folded into the model can be calculated in a similar way to LSA, with the cosine of the angle between the vectors containing the probability distributions $\left(P(z_k \mid q)\right)_{k=1}^{K}$ and $\left(P(z_k \mid d_i)\right)_{k=1}^{K}$ (Hofmann, 2001). Another similarity measure is *the logarithm of the a posteriori probability for the comparison material passage* d_i *given the essay q* formulated by Girolami and Kabán (2003). This is shown in Equation (5).

$$sim(d_i, q) = \sum_{w_j \in d_i} n(d_i, w_j) \log \sum_{k=1}^{K} P(w_j \mid z_k) P(z_k \mid q) \qquad (5)$$

Unlike LSA, PLSA defines proper probability distributions to the documents. PLSA is interpretable with its generative model, latent classes or topics and graphical illustrations in *K*-dimensional space (Hofmann, 2001). In IR, PLSA yielded equal or better results when compared to LSA (Yu-Seop et al., 2002; Hofmann, 2001). Hofmann (2001) demonstrated that the accuracy of PLSA can be increased by increasing the numbers of latent variables. The combination of several similarity scores (such as the cosines of angles between two documents) from models with different number of latent variables also increases accuracy. The selection of the dimension is not therefore as crucial as it is in LSA.

A problem with PLSA is that the algorithm used to compute the model – EM or its variant – is probabilistic and it can converge to a local maximum. Hofmann (2001) believes this is not a severe problem and that the differences between separate runs are small. Baldi et al. (2003) and Si and Jin (2005) are among those who have pointed out that PLSA often overfits the model to the training data. The recent research of Brants (2005) shows that this is not necessarily the case. The generative model of PLSA and the assumptions behind it have also been characterized as flawed by Blei et al. (2003), because there is a need to estimate probability to acquire an unseen document. Blei et al. (2003) solved this problem by proposing another statistical framework, LDA, which we will examine in the next section.

Latent Dirichlet Allocation

The principles of LDA are similar to those of PLSA. The difference is that there is no need in LDA to estimate the probability of obtaining a document, and there is thus no need to perform the difficult estimation process when adding unseen documents to the model. Instead this is achieved by changing the generative model in such a way that it separates the process for each document and uses the word-latent class distribution to determine the document-latent class distribution. LDA assumes the following generative process for each document d_i in a corpus consisting of *N* documents that contains *M* distinct words and *K* distinct *latent variables* or *topics*:

1. Choose the length of the document *L~Poisson*(*x*). Note that we are (for most of the time) dealing with each sequence of words in a single document and not the distinct words in the corpus. We therefore use indexing w_l for words in a single document $d_i = \{w_1,..., w_L\}$ and w_m for distinct words in a corpus.

2. Choose a parameter vector for the topic distribution θ~*Dirichlet*(α), the parameter α is a *K*-vector with components $a_k > 0$ and θ is a *K*-vector so that $\theta_k \geq 0$ and $\sum_{k=1}^{K} \theta_k = 1$ and *P*(θ|α) is the probability density function of the Dirichlet distribution.

3. For each of the *L* words w_l:

a) Choose a topic $z_l \sim Multinomial(q)$. Note that z_l or z_L is used when we discuss the topic of each word in a document, and that all topics in a document are referred as $z_{d_i} = \{z_1,..., z_l\}$.

b) Choose a word w_l from $P(w_l \mid z_l, \beta)$, a multinomial probability conditioned on the topic z_l, where β is a *K*×*M* matrix so that $\beta_{kj} = P(w_m \mid z_k)$ for all 1≤*m*≤*M* and 1≤*k*≤*K*, where M is the number of distinct words in the corpus.

In order to build up the model for LDA, one should compute the posterior distribution of the latent variables for a given document in the way shown in Equation (6).

$$P(\theta, z_{d_i} \mid d_i, \alpha, \beta) = \frac{P(\theta, z_{d_i}, d_i \mid \alpha, \beta)}{P(d_i \mid \alpha, \beta)} \qquad (6)$$

But because Equation (6) is intractable, it needs to be approximated. Blei et al. (2003) introduce an EM-based variational algorithm to approximate the equation and maximize the log likelihood of the model based on the α and β parameters. We will describe the algorithm briefly at this point. Further details can be found in Blei et al. (2003) and Minka and Lafferty (2002).

In the E-step, the density function in Equation (6) needs to be approximated with a tractable model. The idea is to minimize the Kullback-Leibner Divergence between the tractable and intractable model by finding the minimal values for Dirichlet parameter γ and multinomial parameters $(\phi_1,..., \phi_L)$ in the tractable model. In order to find the optimal γ and ϕ for each document d_i, Blei et al. (2003) obtained the updated Equations (7) and (8) for these parameters,

$$\phi_{mk}(d_i) \propto \beta_{km} \exp\{E_P[\log(\theta_k) \mid \gamma(d_i)]\},$$
$$\text{where } E_P[\log(\theta_k) \mid \gamma(d_i)] = \Psi(\gamma(d_i)) - \Psi\left(\sum_{j=1}^{K} \gamma_j(d_i)\right) \qquad (7)$$

$$\gamma_k(d_i) = \alpha_k + \sum_{i=1}^{N} \phi_{nk}(d_i) \qquad (8)$$

Because the γ parameter vector describes the topic distribution for each document, it can be used in a similar way to $P(z_k \mid d_i)$ in the PLSA model. These two equations are computed repeatedly for all *l*, *k* and d_i until the lower bound achieved from Jensen's inequality converges.

In the M-step, α and β parameters need to be estimated once the new values of ϕ and γ have been calculated. Blei et al. (2003) propose the use of the Newton-Raphson optimization technique to find the stationary point of the α function by iterating Equation (9). The conditional multinomial parameters α and β are also updated as in Equations (9) and (10).

$$a_{new} = a_{old} - H(a_{old})^{-1} g(a_{old}) \qquad (9)$$

$$\beta_{km} \propto \sum_{i=1}^{N} \sum_{l=1}^{L_{d_i}} \phi_{lk}(d_i) eq(w_l, w_m), \qquad (10)$$

where *H*(α) and *g*(α) are the Hessian matrix and gradient respectively at point α and $eq(w_l, w_m)$ is 1 if a word w_l from the document d_i is the same word as *m*th distinct word w_m in the corpus otherwise 0. After each cycle in the EM algorithm the convergence of the model building is measured by means of the log-likelihood of the model.

A new document or query can be added to the model by using the same procedure. Blei et al. (2003) propose methods for smoothing the distributions in order to avoid zero probabilities when adding new documents that contain previously unseen words.

The γ vector of a document contains the information about how the document belongs to the different latent classes or topics. Because ϕ contains the same information for each word in the document, the similarity between documents can be compared with the same methods that were used in PLSA – by applying either the cosine of the angle between the documents' γ vectors or the logarithm of the a posteriori probability for the document. Another distance measure that can be used with LDA is the *entropic cosine similarity* that was formulated by Girolami and Kabán (2003). This is shown in Equation (11).

$$sim(d_i, q) = \sum_{w_j \in q} n(q, w_j) \log \sum_{k=1}^{K} P(w_j \mid z_k) P(z_k \mid d_i) \quad (11)$$

PLSA and LDA appear to be very similar methods. Girolami and Kabán (2003) have in fact shown that PLSA may be regarded as a maximum a posteriori/maximum likelihood estimate of the theoretical model in LDA – although it is one that uses different assumptions about the distributions. The variational method used in LDA seems to produce lower perplexity language models than PLSA and to perform better in the text categorization tasks (Blei et al., 2003). In spite of this, the variational method has been claimed to overfit the model into the training data in some cases, and another method, namely Expectation Propagation, has been proposed to compute the LDA model in order to achieve lower perplexity (Minka and Lafferty, 2002). But it is not clear how the perplexity correlates with the results in IR or in essay grading. In the next section, we will compare the performances of LSA, PLSA and LDA in essay-grading contexts in order to analyze their performances on this task.

Empirical Comparison of the Methods

The purpose of using the language modeling methods Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA) in the context of automated essay grading is to reduce noise and to compare the similarities of documents (in the case of AEA, the similarities will be between the essays and the course materials). In this research we compared the performance of these methods in order to analyze their differences in these settings. To validate the suitability of our essay grading process (illustrated in Figure 1), we compared it to the *k-Nearest Neighbors* (k-NN) method that is used in essay grading systems (cf. (Larkey, 1998)).

We compared the essay grading accuracy of our implementations of LSA, PLSA and LDA described in section "Automatic Essay Assessor: AEA" with the essay test sets described in Table 2. In Table 2, the column headed *Field* shows the topic of the course (Education, Communication or Software Engineering). The column headed *Train essays* shows the number of training essays that we used for creating the scoring model before grading, and the column headed *Test essays* gives the number of essays that we used for testing the accuracy of the scoring model. We divided the learning materials from the courses either into paragraphs or sentences as shown in the column headed by *Div. type*. The total count of text passages (the number of columns in the word-by-context-matrix (WCM)) is recorded in the column headed *No. pass,* and the column headed *No. words* indicates the total length of the course materials.

The essay sets in Table 2 were graded by a professor and a teacher or a course assistant. The Software Engineering test sets (5 and 6) were graded by two human graders. The correlation between these graders was .88 on all the essays and the correlation between the two human graders for the training essays and the test essays were .89 and .87, respectively.

We ran the experiment with LSA for all the possible dimensions (i.e. from two to the number of contexts in the WCM). In contrast to the number of dimensions in LSA, which is limited to the number of contexts in the WCM, there is no upper limit for the number of latent variables or topics in the PLSA and LDA models. In order to make a fair comparison, we set the same upper limit for the number of variables and topics in PLSA and LDA that we had in LSA. In addition to this we also conducted experiments with a PLSA model (referred to henceforth as PLSA-C) in which the similarity score was defined as the linear combination of similarity values obtained from PLSA models with predefined numbers of latent variables $K \in \{16,32,48,64,80,96,112,128\},$. When we built up the PLSA models with TEM, we used twenty essays from the training set to test the stopping condition. We utilized the cosine of the angle between the vectors and the logarithm of the a posteriori probability and the entropic cosine as the similarity measures in all the methods where applicable, and selected the highest score.

Table 2. The essay sets used in the experiments

Set No.	Field	Level	Train. essays	Test essays	Grad. scale	Grader	Div. type	No. pass.	No. words
1	Educ.	Under grad.	70	73	0–6	Prof.	Par.	26	2397
2	Educ.	Under grad.	70	73	0–6	Prof.	Sent.	147	2397
3	Comm.	Voca-tional	42	45	0–4	Course teacher	Par.	45	1583
4	Comm.	Voca-tional	42	45	0–4	Course teacher	Sent.	139	1583
5	Soft. Eng.	Grad.	26	27	0–10	Assist.	Par.	27	965
6	Soft. Eng.	Grad.	26	27	0–10	Assist.	Sent.	105	965

We applied similar procedures for the k-NN-based grading methods *KNN-LSA*, *KNN-PLSA*, *KNN-PLSA-C*, and *KNN-LDA*. The models were computed by using the training essays alone. We conducted these experiments with the values of *k* between 1 and 10 and determined the grade for an essay as a similarity-score-weighted average of the grades of the *k* neighboring essays. We reported the correlation for each k-NN method with the value of *k* that resulted in the most accurate grading.

Table 3. The results of the comparison between the grading methods accuracy in Spearman Correlations

Set No.	LSA	KNN -LSA	PLSA	KNN-PLSA	PLSA -C	KNN-PLSA -C	LDA	KNN -LDA
1	.78	.53	.75	.28	.74	.40	.61	.44
2	.80	*)	.78	*)	.71	*)	.64	*)
3	.54	.45	.51	.34	.36	-.02	.25	.44
4	.57	*)	.55	*)	.43	*)	.42	*)
5	.88	.81	.88	.88	.88	.64	.82	.88
6	.90	*)	.95	*)	.90	*)	.83	*)

*) = Same as the previous one because the course materials were not used.

Table 3 shows the results of the experiment as measured by means of the Spearman correlation between the grades given by a human assessor and the system. The achieved correlations are comparable to the ones found in the literature. While the correlations for essay sets 5 and 6 vary from .64 to .95, most of the correlations vary between .81 and .95. A comparison with the correlation between the two human graders for the essay sets 5 and 6 (.87) indicates that AEA is able to grade essays as accurately as the human graders (the correlation between the grades by AEA and the course teacher was .90).

The results clearly show that k-NN-based methods are outperformed by the grading method of AEA that uses course materials in the grading process. All the methods yielded the lowest grading accuracy for the Communications test set. We suspect that this was caused by the open-endedness of the essay prompt and by the fact that several students used real-life examples as the basis of their answers. It is hypothesized that because of this, the comparison with course content or other students' essays did not render meaningful results.

An important finding is that the system performs better when the course materials used for comparison are divided into sentences and not paragraphs. This is because the more separate passages of the course materials there are, the better is the distinction between the grade categories. This division also increases the sparseness of the training data that might have an effect on the distinction between grade categories.

Even though LSA seems to outperform the other two methods (PLSA and LDA), the differences, especially between LSA and PLSA, are small. PLSA also outperforms LDA in accuracy. Although PLSA and LDA have been shown to perform better than LSA in IR tasks (Hofmann, 2001; Blei et al., 2003), this was not the case in our experiment. One possible explanation for this difference might be the size of the collections that we used to train the model. LSA performs better with a small document collection. But there might be other explanations as well. While Hofmann (2001) and Blei et al. (2003) trained systems with collections of 1,000-3,000 documents in earlier studies, we used fewer than 150 documents. Since it is more likely from a practical point of view that relatively small collections of essays are graded, the results obtained from our experiments are valid in automatic essay grading context. It will be necessary to test these assumptions in future experiments if the size of document collections does indeed exert an effect on the performance of LSA, PLSA and LDA.

Although PLSA-C performs worse than PLSA, it is a useful method for removing the dimensionality selection phase in PLSA. This is caused by the fact that the differences between PLSA and PLSA-C are relatively small.

Conclusion and Future Work

In this paper we have described an automated grading system that is based on comparisons between course materials and teacher-graded essays. We have also reported a set of experiments that used the system for comparing three dimensionality reduction methods, LSA, PLSA and LSA. We have also showed that the use of both course materials and human-graded essays makes grading more accurate than the k-NN-based grading method based on human-graded essays alone. The highest correlation between the grades assigned by the system and by a human grader (0.95) was achieved on the Software Engineering test set using PLSA for document comparisons.

The overall results show that our content-based grading model is acceptable in terms of costs (a low number of pre-graded essays is sufficient for training) and coachability (AEA is not based on surface measures). While our results show that all the dimensionality reduction methods that we have considered can provide an acceptable level of accuracy, it was rather surprising to observe that LSA slightly outperforms PLSA and LDA in the essay-grading domain. Because the probabilistic models of these two methods allow for other developments, they are worthy of further investigation. We observed, for instance, during the experiments with PLSA-C, that the combination of dimensions that lead to better results might be dependent on the features of the used document collection (such as the number of passages in comparison materials or the number of essays in training data). It is possible that the combination of dimensions can be optimized for each essay set by using such collection-specific features. This would obviate the need for using computationally demanding dimensionality selection methods such as those that are used for LSA.

The most important distinction between the three dimensionality reduction methods in the context of automatic essay grading is the way in which they support the defensibility of the grading system. PLSA and LDA offer a better means than LSA of giving students feedback about the essay by finding topics that are not covered by the essay writer because the topic distributions can be directly used as estimates of how well a specific topic has been covered. An essay-grading system that uses methods such as these can support a writer by offering keywords from each of the underlying topics that the essay has not addressed.

Defensibility is admittedly the weakest point of AEA (and other automatic grading systems). The next steps in research into AEA include increasing its defensibility by incorporating automatic feedback generation, writing style and plagiarism-detection modules into the system.

Acknowledgements

This work was supported by the Academy of Finland under grant no. 111350, “Automated Assessment Technologies for Free Text and Programming Assignments”. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Academy of Finland.

References

Baldi, P., Frasconi, P., & Smyth, P. (2003). *Modeling the Internet and the web: probabilistic methods and algorithms*, Chichester, England: Wiley.

Bingham, E., & Mannila, H. (2001). Random projection in dimensionality reduction: Applications to image and text data. *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY: ACM Press, 245–250.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3* (5), 993–1022.

Brants, T. (2005). Test data likelihood for PLSA models. *Information Retrieval, 8* (2), 181–196.

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris, M. D. (1998). Automated scoring using a hybrid feature identification technique. *Proceedings of the Annual Meeting of the Association of Computational Linguistics*, Morristown, NJ: Association for Computational Linguistics, 206–210.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science, 41* (6), 391–407.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, 39* (B), 1–38.

Foltz, P. W., & Dumais, S. T. (1992). Personalized information delivery: an analysis of information filtering methods. *Communications of the ACM, 35* (12), 51–60.

Foltz, P. W., Gilliam, S., & Kendall, S. (2000). Supporting content-based feedback in online writing evaluation with LSA. *Interactive Learning Environments, 8* (2), 111–129.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The intelligent essay assessor: applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 1 (2), retrieved June 2, 2008, from http://imej.wfu.edu/articles/1999/2/04/.

Girolami, M., & Kabán, A. (2003). On an equivalence between PLSI and LDA. *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA: ACM Press, 433–434.

Globerson, A., & Tishby, N. (2003). Sufficient dimensionality reduction. *Journal of Machine Learning Research, 3*, 1307–1331.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning, 42* (1-2), 177–196.

Kakkonen, T., Myller, N., & Sutinen, E. (2006). Applying latent Dirichlet allocation to automatic essay grading. In Salakoski, T., Ginter, F., Pyysalo, S., Pahikkala, T. (Eds.), *Proceedings of the 5th International Conference on Natural Language Processing*, Berlin/Heidelberg: Springer, 110–120.

Kakkonen, T., Myller, N., Timonen, J., & Sutinen, E. (2005a). Automatic essay grading with probabilistic latent semantic analysis. *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, New Brunswick: The Association for Computational Linguistics, 29–36.

Kakkonen, T., & Sutinen, E. (2004). Automatic assessment of the content of essays based on course materials. *Proceedings of the International Conference on Information Technology: Research and Education*, New York, NY: IEEE Press, 126–130.

Kakkonen, T., Timonen, J., & Sutinen, E. (2005b). Applying validation methods for noise reduction in LSA-based essay grading. *WSEAS Transactions on Information Science and Applications, 9* (2), 1334–1342.

Kaplan, R. M., Wolff, S., Burstein, J., Li, C., Rock, D., & Kaplan, B. (1998). *Scoring essays automatically using surface features*, Technical Report 94-21P, New Jersey, USA: Educational Testing Service.

Kim, Y.-S., Chang, J.-H., & Zhang, B.-T. (2003). An empirical study on dimensionality optimization in text mining for linguistic knowledge acquisition. *Lecture Notes in Artificial Intelligence, 2637*, 111–116.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25* (2&3), 259–284.

Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. (1997). How well can passage meaning be derived without using word order? A comparison of latent semantic analysis and humans. In Shafto, M. G., Langley, P. (Eds.), *Proceedings of the 19th Annual Conference of the Cognitive Science Society*, Stanford, USA: Cognitive Science Society, 412–417.

Larkey, L. (1998). Automatic essay grading using text categorization techniques. *Proceedings of the 21st Annual International Conference on Research and Development in Information Retrieval*, New York: ACM Press, 90–95.

Lingsoft Inc. (2007). *Lingsoft Oy*, Retrieved June 28, 2008 from http://www.lingsoft.fi/.

Minka, T., & Lafferty, J. (2002). Expectation-propagation for the generative aspect model. In Breese, J., Koller, D. (Eds.), *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, San Francisco, CA, USA: Morgan Kaufmann, 352–359.

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappa, 47* (1), 238–243.

Page, E. B., & Petersen, N. S. (1995). The computer moves into essay grading. *Phi Delta Kappa, 76* (7), 561–565.

Praks, P., Dvorský, J., & Snášel, V. (2003). Latent semantic indexing for image retrieval systems. *Proceedings of the SIAM Conference on Applied Linear Algebra. International Linear Algebra Society (ILAS)*, Philadelphia, PA: SIAM, Retrieved June 28, 2008 from http://www.siam.org/meetings/la03/proceedings/Dvorsky.pdf.

Quesada, J. (2003). *Latent problem solving analysis (LPSA): a computational theory of representation in complex, dynamic problem solving tasks*, PhD thesis, Psychology, University of Granada.

Rudner, L. M., & Liang, T. (2002). Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning and Assessment, 1* (2), 3–21.

Sebastiani, G. (2002). Machine learning in automated text categorization. *ACM Computing Surveys, 34* (1), 1–47.

Si, L., & Jin, R. (2005). Adjusting mixture weights of gaussian mixture model via regularized probabilistic latent semantic analysis. *Lecture Notes on Computer Science, 3518*, 622–631.

Wade-Stein, D., & Kintsch, E. (2003). *Summary street: Interactive computer support for writing*, Technical report, Boulder, CO: Institute for Cognitive Science, University of Colorado.

Wiemer-Hastings, P. M., Wiemer-Hastings, K., & Graesser, A. C. (1999). Approximate natural language understanding for an intelligent tutor. *Proceedings of the Twelfth International Florida Artificial Intelligence Research Society Conference*, Menlo Park, California, USA: AAAI Press, 192–196.

Yu-Seop, K., Jeong-Ho, C., & Byoung-Tak, Z. (2002). A comparative evaluation of data-driven models in translation selection of machine translation. *Proceedings of the 19th International Conference on Computational Linguistics*, Morristown, NJ, USA: Association for Computational Linguistics, 1–7.

Zelikovitz, S., & Hirsh, H. (2001). Using LSI for text classification in the presence of background text. *Proceedings of the 10th International Conference on Information and Knowledge Management*, New York: ACM Press, 113–118.