Automatic Essay Grading with Probabilistic Latent Semantic Analysis

Tuomo Kakkonen, Niko Myller, Jari Timonen, and Erkki Sutinen
Department of Computer Science, University of Joensuu
P.O. Box 111, FI-80101 Joensuu, FINLAND
firstname.lastname@cs.joensuu.fi

Abstract

Probabilistic Latent Semantic Analysis (PLSA) is an information retrieval technique proposed to improve the problems found in Latent Semantic Analysis (LSA). We have applied both LSA and PLSA in our system for grading essays written in Finnish, called Automatic Essay Assessor (AEA). We report the results comparing PLSA and LSA with three essay sets from various subjects. The methods were found to be almost equal in the accuracy measured by Spearman correlation between the grades given by the system and a human. Furthermore, we propose methods for improving the usage of PLSA in essay grading.

1 Introduction

The main motivations behind developing automated essay assessment systems are to decrease the time in which students get feedback for their writings, and to reduce the costs of grading. The assumption in most of the systems is that the grades given by the human assessors describe the true quality of an essay. Thus, the aim of the systems is to "simulate" the grading process of a human grader and a system is usable only if it is able to perform the grading as accurately as human raters. An automated assessment system is not affected by errors caused by lack of consistency, fatigue or bias, thus it can help achieving better accuracy and objectivity of assessment (Page and Petersen, 1995).

There has been research on automatic essay grading since the 1960s. The earliest systems, such as PEG (Page and Petersen, 1995), based their grading on the surface information from the essay. For example, the number of words and commas were counted in order to determine the quality of the essays (Page, 1966). Although these kinds of systems performed considerably well, they also received heavy criticism (Page and Petersen, 1995). Some researchers consider the use of natural language as a feature for human intelligence (Hearst et al., 2000) and writing as a method to express the intelligence. Based on that assumption, taking the surface information into account and ignoring the meanings of the content is insufficient. Recent systems and studies, such as e-rater (Burstein, 2003) and approaches based on LSA (Landauer et al., 1998), have focused on developing the methods which determine the quality of the essays with more analytic measures such as syntactic and semantic structure of the essays. At the same time in the 1990s, the progress of natural language processing and information retrieval techniques have given the opportunity to take also the meanings into account.

LSA has produced promising results in content analysis of essays (Landauer et al., 1997; Foltz et al., 1999b). Intelligent Essay Assessor (Foltz et al., 1999b) and Select-a-Kibitzer (Wiemer-Hastings and Graesser, 2000) apply LSA for assessing essays written in English. In Apex (Lemaire and Dessus, 2001), LSA is applied to essays written in French. In addition to the essay assessment, LSA is applied to other educational applications. An intelligent tutoring system for providing help for students (Wiemer-

Proceedings of the 2nd Workshop on Building Educational Applications Using NLP,
pages 29–36, Ann Arbor, June 2005.

Hastings et al., 1999) and Summary Street (Steinhart, 2000), which is a system for assessing summaries, are some examples of other applications of LSA. To our knowledge, there is no system utilizing PLSA (Hofmann, 2001) for automated essay assessment or related tasks.

We have developed an essay grading system, *Automatic Essay Assessor* (AEA), to be used to analyze essay answers written in Finnish, although the system is designed in a way that it is not limited to only one language. It applies both course materials, such as passages from lecture notes and course textbooks covering the assignment-specific knowledge, and essays graded by humans to build the model for assessment. In this study, we employ both LSA and PLSA methods to determine the similarities between the essays and the comparison materials in order to determine the grades. We compare the accuracy of these methods by using the Spearman correlation between computer and human assigned grades.

The paper is organized as follows. Section 2 explains the architecture of AEA and the used grading methods. The experiment and results are discussed in Section 3. Conclusions and future work based on the experiment are presented in Section 4.

2 AEA System

We have developed a system for automated assessment of essays (Kakkonen et al., 2004; Kakkonen and Sutinen, 2004). In this section, we explain the basic architecture of the system and describe the methods used to analyze essays.

2.1 Architecture of AEA

There are two approaches commonly used in the essay grading systems to determine the grade for the essay:

1. The essay to be graded is compared to the human-graded essays and the grade is based on the most similar essays' grades; or

2. The essay to be graded is compared to the essay topic related materials (e.g. textbook or model essays) and the grade is given based on the similarity to these materials.

In our system, AEA (Kakkonen and Sutinen, 2004), we have combined these two approaches. The relevant parts of the learning materials, such as chapters of a textbook, are used to train the system with assignment-specific knowledge. The approaches based on the comparison between the essays to be graded and the textbook have been introduced in (Landauer et al., 1997; Foltz et al., 1999a; Lemaire and Dessus, 2001; Hearst et al., 2000), but have been usually found less accurate than the methods based on comparison to prescored essays. Our method attempts to overcome this by combining the use of course content and prescored essays. The essays to be graded are not directly compared to the prescored essays with for instance k-nearest neighbors method, but prescored essays are used to determine the similarity threshold values for grade categories as discussed below. Prescored essays can also be used to determine the optimal dimension for the reduced matrix in LSA as discussed in Kakkonen et al. (2005).

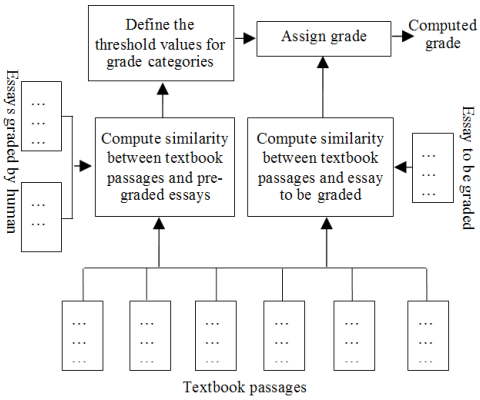

Figure 1: The grading process of AEA.

Figure 1 illustrates the grading process of our system. The texts to be analyzed are added into *word-by-context matrix* (WCM), representing the number of occurrences of each unique word in each of the contexts (e.g. documents, paragraphs or sentences). In WCM M, cell M_{ij} contains the count of the word i occurrences in the context j. As the first step in analyzing the essays and course materials, the lemma of each word form occurring in the texts must be found. We have so far applied AEA only to essays written in Finnish. Finnish is morphologically more complex than English, and word forms are formed by adding suffixes into base forms. Because of that,

base forms have to be used instead of inflectional forms when building the WCM, especially if a relatively small corpus is utilized. Furthermore, several words can become synonyms when suffixes are added to them, thus making the word sense disambiguation necessary. Hence, instead of just stripping suffixes, we apply a more sophisticated method, a morphological parser and disambiguator, namely Constraint Grammar parser for Finnish (FINCG) to produce the lemmas for each word (Lingsoft, 2005). In addition, the most commonly occurring words (stopwords) are not included in the matrix, and only the words that appear in at least two contexts are added into the WCM (Landauer et al., 1998). We also apply entropy-based term weighting in order to give higher values to words that are more important for the content and lower values to words with less importance.

First, the comparison materials based on the relevant textbook passages or other course materials are modified into machine readable form with the method described in the previous paragraph. The vector for each context in the comparison materials is marked with Y_i. This WCM is used to create the model with LSA, PLSA or another information retrieval method. To compare the similarity of an essay to the course materials, a query vector X_j of the same form as the vectors in the WCM is constructed. The query vector X_j representing an essay is added or *folded in* into the model build with WCM with the method specific way discussed later. This folded-in query $\tilde{X}_j$ is then compared to the model of each text passage $\tilde{Y}_i$ in the comparison material by using a similarity measure to determine the similarity value. We have used the cosine of the angle between ($\tilde{X}_j$, $\tilde{Y}_i$), to measure the similarity of two documents. The *similarity score* for an essay is calculated as the sum of the similarities between the essay and each of the textbook passages.

The document vectors of manually graded essays are compared to the textbook passages, in order to determine the similarity scores between the essays and the course materials. Based on these measures, threshold values for the grade categories are defined as follows: the grade categories, $g_1, g_2, \ldots, g_C$, are associated with similarity value limits, $l_1, l_2, \ldots, l_{C+1}$, where C is the number of grades, and $l_{C+1} = \infty$ and normally $l_1 = 0$ or $-\infty$. Other category limits $l_i, 2 \leq i \leq C$, are defined as weighted averages of the similarity scores for essays belonging to grade categories g_i and g_{i-1}. Other kinds of formulas to define the grade category limits can be also used.

The grade for each essay to be graded is then determined by calculating the similarity score between the essay and the textbook passages and comparing the similarity score to the threshold values defined in the previous phase. The similarity score S_i of an essay d_i is matched to the grade categories according to their limits in order to determine the correct grade category as follows: `For each` i, $1 \leq i \leq C$, `if` $l_i < S_i \leq l_{i+1}$ `then` $d_i \in g_i$ `and break`.

2.2 Latent Semantic Analysis

Latent Semantic Analysis (LSA) (Landauer et al., 1998) is a corpus-based method used in information retrieval with vector space models. It provides a means of comparing the semantic similarity between the source and target texts. LSA has been successfully applied to automate giving grades and feedback on free-text responses in several systems as discussed in Section 1. The basic assumption behind LSA is that there is a close relationship between the meaning of a text and the words in that text. The power of LSA lies in the fact that it is able to map the essays with similar wordings closer to each other in the vector space. The LSA method is able to strengthen the similarity between two texts even when they do not contain common words. We describe briefly the technical details of the method.

The essence of LSA is dimension reduction based on the singular value decomposition (SVD), an algebraic technique. SVD is a form of factor analysis, which reduces the dimensionality of the original WCM and thereby increases the dependency between contexts and words (Landauer et al., 1998). SVD is defined as $X = T_0 S_0 {D_0}^T$, where X is the preprocessed WCM and T_0 and D_0 are orthonormal matrices representing the words and the contexts. S_0 is a diagonal matrix with singular values. In the dimension reduction, the k highest singular values in S_0 are selected and the rest are ignored. With this operation, an approximation matrix $\tilde{X}$ of the original matrix X is acquired. The aim of the dimension reduction is to reduce "noise" or unimportant details and to allow the underlying semantic structure to be-

come evident (Deerwester et al., 1990).

In information retrieval and essay grading, the queries or essays have to be folded in into the model in order to calculate the similarities between the documents in the model and the query. In LSA, the folding in can be achieved with a simple matrix multiplication: $\tilde{X}_q = X_q^T T_0 S_0^{-1}$, where X_q is the term vector constructed from the query document with preprocessing, and T_0 and S_0 are the matrices from the SVD of the model after dimension reduction. The resulting vector $\tilde{X}_q$ is in the same format as the documents in the model.

The features that make LSA suitable for automated grading of essays can be summarized as follows. First, the method focuses on the content of the essay, not on the surface features or keyword-based content analysis. The second advantage is that LSA-based scoring can be performed with relatively low amount of human graded essays. Other methods, such as PEG and e-rater typically need several hundred essays to be able to form an assignment-specific model (Shermis et al., 2001; Burstein and Marcu, 2000) whereas LSA-based IEA system has sometimes been calibrated with as few as 20 essays, though it typically needs more essays (Hearst et al., 2000).

Although LSA has been successfully applied in information retrieval and related fields, it has also received criticism (Hofmann, 2001; Blei et al., 2003). The objective function determining the optimal decomposition in LSA is the Frobenius norm. This corresponds to an implicit additive Gaussian noise assumption on the counts and may be inadequate. This seems to be acceptable with small document collections but with large document collections it might have a negative effect. LSA does not define a properly normalized probability distribution and, even worse, the approximation matrix may contain negative entries meaning that a document contains negative number of certain words after the dimension reduction. Hence, it is impossible to treat LSA as a generative language model and moreover, the use of different similarity measures is limited. Furthermore, there is no obvious interpretation of the directions in the latent semantic space. This might have an effect if also feedback is given. Choosing the number of dimensions in LSA is typically based on an ad hoc heuristics. However, there is research done aiming to resolve the problem of dimension selection in LSA, especially in the essay grading domain (Kakkonen et al., 2005).

2.3 Probabilistic Latent Semantic Analysis

Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 2001) is based on a statistical model which has been called the *aspect model*. The aspect model is a latent variable model for co-occurrence data, which associates unobserved class variables z_k, $k \in \{1, 2, \ldots, K\}$ with each observation. In our settings, the observation is an occurrence of a word w_j, $j \in \{1, 2, \ldots, M\}$, in a particular context d_i, $i \in \{1, 2, \ldots, N\}$. The probabilities related to this model are defined as follows:

- $P(d_i)$ denotes the probability that a word occurrence will be observed in a particular context d_i;
- $P(w_j|z_k)$ denotes the class-conditional probability of a specific word conditioned on the unobserved class variable z_k; and
- $P(z_k|d_i)$ denotes a context specific probability distribution over the latent variable space.

When using PLSA in essay grading or information retrieval, the first goal is to build up the model. In other words, approximate the probability mass functions with machine learning from the training data, in our case the comparison material consisting of assignment specific texts.

Expectation Maximization (EM) algorithm can be used in the model building with maximum likelihood formulation of the learning task (Dempster et al., 1977). In EM, the algorithm alternates between two steps: (i) an *expectation (E)* step where posterior probabilities are computed for the latent variables, based on the current estimates of the parameters, (ii) a *maximization (M)* step, where parameters are updated based on the loglikelihood which depends on the posterior probabilities computed in the E-step. The standard E-step is defined in equation (1).

$$P(z_k|d_i, w_j) = \frac{P(w_j|z_k)P(z_k|d_i)}{\sum_{l=1}^{K} P(w_j|z_l)P(z_l|d_i)} \qquad (1)$$

The M-step is formulated in equations (2) and (3) as derived by Hofmann (2001). These two steps

are alternated until a termination condition is met, in this case, when the maximum likelihood function has converged.

$$P(w_j|z_k) = \frac{\sum_{i=1}^{N} n(d_i, w_j) P(z_k|d_i, w_j)}{\sum_{m=1}^{M} \sum_{i=1}^{N} n(d_i, w_m) P(z_k|d_i, w_m)} \quad (2)$$

$$P(z_k|d_i) = \frac{\sum_{j=1}^{M} n(d_i, w_j) P(z_k|d_i, w_j)}{\sum_{m=1}^{M} n(d_i, w_m)} \quad (3)$$

Although standard EM algorithm can lead to good results, it may also overfit the model to the training data and perform poorly with unseen data. Furthermore, the algorithm is iterative and converges slowly, which can increase the runtime seriously. Hence, Hofmann (2001) proposes another approach called *Tempered EM* (TEM), which is a derivation of standard EM algorithm. In TEM, the M-step is the same as in EM, but a dampening parameter is introduced into the E-step as shown in equation (4). The parameter β will dampen the posterior probabilities closer to uniform distribution, when $\beta < 1$ and form the standard E-step when $\beta = 1$.

$$P(z_k|d_i, w_j) = \frac{(P(w_j|z_k)P(z_k|d_i))^{\beta}}{\left(\sum_{l=1}^{K} P(w_j|z_l)P(z_l|d_i)\right)^{\beta}} \quad (4)$$

Hofmann (2001) defines the TEM algorithm as follows:

1. Set $\beta := 1$ and perform the standard EM with early stopping.
2. Set $\beta := \eta\beta$ (with $\eta < 1$).
3. Repeat the E- and M-steps until the performance on hold-out data deteriorates, otherwise go to step 2.
4. Stop the iteration when decreasing β does not improve performance on hold-out data.

Early stopping means that the optimization is not done until the model converges, but the iteration is stopped already once the performance on hold-out data degenerates. Hofmann (2001) proposes to use the *perplexity* to measure the generalization performance of the model and the stopping condition for the early stopping. The perplexity is defined as the log-averaged inverse probability on unseen data calculated as in equation (5).

$$\mathcal{P} = \exp\left(-\frac{\sum_{i,j} n'(d_i, w_j) \log P(w_j|d_i)}{\sum_{i,j} n'(d_i, w_j)}\right), \quad (5)$$

where $n'(d_i, w_j)$ is the count on hold-out or training data.

In PLSA, the folding in is done by using TEM as well. The only difference when folding in a new document or query q outside the model is that just the probabilities $P(z_k|q)$ are updated during the M-step and the $P(w_j|z_k)$ are kept as they are. The similarities between a document d_i in the model and a query q folded in to the model can be calculated with the cosine of the angle between the vectors containing the probability distributions $(P(z_k|q))_{k=1}^{K}$ and $(P(z_k|d_i))_{k=1}^{K}$ (Hofmann, 2001).

PLSA, unlike LSA, defines proper probability distributions to the documents and has its basis in Statistics. It belongs to a framework called Latent Dirichlet Allocations (Girolami and Kabán, 2003; Blei et al., 2003), which gives a better grounding for this method. For instance, several probabilistic similarity measures can be used. PLSA is interpretable with its generative model, latent classes and illustrations in N-dimensional space (Hofmann, 2001). The latent classes or topics can be used to determine which part of the comparison materials the student has answered and which ones not.

In empirical research conducted by Hofmann (2001), PLSA yielded equal or better results compared to LSA in the contexts of information retrieval. It was also shown that the accuracy of PLSA can increase when the number of latent variables is increased. Furthermore, the combination of several similarity scores (e.g. cosines of angles between two documents) from models with different number of latent variables also increases the overall accuracy. Therefore, the selection of the dimension is not as crucial as in LSA. The problem with PLSA is that the algorithm used to computate the model, EM or its variant, is probabilistic and can converge to a local maximum. However, according to Hofmann (2001), this is not a problem since the differences between separate runs are small. Flaws in the generative model and the overfitting problem

Set No.	Field	Training essays	Test essays	Grading scale	Course materials	Comp. mat. division type	No. Passages	No. Words
1	Education	70	73	0–6	Textbook	Paragraphs	26	2397
2	Education	70	73	0–6	Textbook	Sentences	147	2397
3	Communications	42	45	0–4	Textbook	Paragraphs	45	1583
4	Communications	42	45	0–4	Textbook	Sentences	139	1583
5	Soft. Eng.	26	27	0–10	*)	Paragraphs	27	965
6	Soft. Eng.	26	27	0–10	*)	Sentences	105	965

Table 1: The essay sets used in the experiment. *) Comparison materials were constructed from the course handout with teacher's comments included and transparencies represented to the students.

have been discussed in Blei et al. (2003).

3 Experiment

3.1 Procedure and Materials

To analyze the performance of LSA and PLSA in the essay assessment, we performed an experiment using three essay sets collected from courses on education, marketing and software engineering. The information about the essay collections is shown in Table 1. Comparison materials were taken either from the course book or other course materials and selected by the lecturer of the course. Furthermore, the comparison materials used in each of these sets were divided with two methods, either into paragraphs or sentences. Thus, we run the experiment in total with six different configurations of materials.

We used our implementations of LSA and PLSA methods as described in Section 2. With LSA, all the possible dimensions (i.e. from two to the number of passages in the comparison materials) were searched in order to find the dimension achieving the highest accuracy of scoring, measured as the correlation between the grades given by the system and the human assessor. There is no upper limit for the number of latent variables in PLSA models as there is for the dimensions in LSA. Thus, we applied the same range for the best dimension search to be fair in the comparison. Furthermore, a linear combination of similarity values from PLSA models (PLSA-C) with predefined numbers of latent variables $K \in \{16, 32, 48, 64, 80, 96, 112, 128\}$ was used just to analyze the proposed potential of the method as discussed in Section 2.3 and in (Hofmann, 2001). When building up all the PLSA models with TEM, we used 20 essays from the training set of the essay collections to determine the early stopping condition with perplexity of the model on unseen data as proposed by Hofmann (2001).

3.2 Results and Discussion

The results of the experiment for all the three methods, LSA, PLSA and PLSA-C are shown in Table 2. It contains the most accurate dimension (column *dim.*) measured by machine-human correlation in grading, the percentage of the same (*same*) and adjacent grades (*adj.*) compared to the human grader and the Spearman correlation (*cor.*) between the grades given by the human assessor and the system.

The results indicate that LSA outperforms both methods using PLSA. This is opposite to the results obtained by Hofmann (2001) in information retrieval. We believe this is due to the size of the document collection used to build up the model. In the experiments of Hofmann (2001), it was much larger, 1000 to 3000 documents, while in our case the number of documents was between 25 and 150. However, the differences are quite small when using the comparison materials divided into sentences. Although all methods seem to be more accurate when the comparison materials are divided into sentences, PLSA based methods seem to gain more than LSA.

In most cases, PLSA with the most accurate dimension and PLSA-C perform almost equally. This is also in contrast with the findings of Hofmann (2001) because in his experiments PLSA-C performed better than PLSA. This is probably also due to the small document sets used. Nevertheless, this means that finding the most accurate dimension is unnecessary, but it is enough to com-

Set No.	LSA dim.	LSA same	LSA adj.	LSA cor.	PLSA dim.	PLSA same	PLSA adj.	PLSA cor.	PLSA-C same	PLSA-C adj.	PLSA-C cor.
1	14	39.7	43.9	0.78	9	31.5	32.9	0.66	34.2	35.6	0.70
2	124	35.6	49.3	0.80	83	37.0	37.0	0.76	35.6	41.1	0.73
3	8	31.1	28.9	0.54	38	24.4	35.6	0.41	17.7	24.4	0.12
4	5	24.4	42.3	0.57	92	35.6	31.1	0.59	22.2	35.6	0.47
5	6	29.6	48.2	0.88	16	18.5	18.5	0.78	11.1	40.1	0.68
6	6	44.4	37.1	0.90	55	33.3	44.4	0.88	14.8	40.7	0.79

Table 2: The results of the grading process with different methods.

bine several dimensions' similarity values. In our case, it seems that linear combination of the similarity values is not the best option because the similarity values between essays and comparison materials decrease when the number of latent variables increases. A topic for a further study would be to analyze techniques to combine the similarity values in PLSA-C to obtain higher accuracy in essay grading. Furthermore, it seems that the best combination of dimensions in PLSA-C depends on the features of the document collection (e.g. number of passages in comparison materials or number of essays) used. Another topic of further research is how the combination of dimensions can be optimized for each essay set by using the collection specific features without the validation procedure proposed in Kakkonen et al. (2005).

Currently, we have not implemented a version of LSA that combines scores from several models but we will analyze the possibilities for that in future research. Nevertheless, LSA representations for different dimensions form a nested sequence because of the number of singular values taken to approximate the original matrix. This will make the model combination less effective with LSA. This is not true for statistical models, such as PLSA, because they can capture a larger variety of the possible decompositions and thus several models can actually complement each other (Hofmann, 2001).

4 Future Work and Conclusion

We have implemented a system to assess essays written in Finnish. In this paper, we report a new extension to the system for analyzing the essays with PLSA method. We have compared LSA and PLSA as methods for essay grading. When our results are compared to the correlations between human and system grades reported in literature, we have achieved promising results with all methods. LSA was slightly better when compared to PLSA-based methods. As future research, we are going to analyze if there are better methods to combine the similarity scores from several models in the context of essay grading to increase the accuracy (Hofmann, 2001). Another interesting topic is to combine LSA and PLSA to compliment each other.

We used the cosine of the angle between the probability vectors as a measure of similarity in LSA and PLSA. Other methods are proposed to determine the similarities between probability distributions produced by PLSA (Girolami and Kabán, 2003; Blei et al., 2003). The effects of using these techniques will be compared in the future experiments.

If the PLSA models with different numbers of latent variables are not highly dependent on each other, this would allow us to analyze the reliability of the grades given by the system. This is not possible with LSA based methods as they are normally highly dependent on each other. However, this will need further work to examine all the potentials.

Our future aim is to develop a semi-automatic essay assessment system (Kakkonen et al., 2004). For determining the grades or giving feedback to the student, the system needs a method for comparing similarities between the texts. LSA and PLSA offer a feasible solution for the purpose. In order to achieve even more accurate grading, we can use some of the results and techniques developed for LSA and develop them further for both methods. We are currently working with an extension to our LSA model that uses standard validation methods for reducing automatically the irrelevant content informa-

tion in LSA-based essay grading (Kakkonen et al., 2005). In addition, we plan to continue the work with PLSA, since it, being a probabilistic model, introduces new possibilities, for instance, in similarity comparison and feedback giving.

References

D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent Dirichlet Allocation. *J. of Machine Learning Research*, 3:993–1022.

J. Burstein and D. Marcu. 2000. Benefits of modularity in an automated scoring system. In *Proc. of the Workshop on Using Toolsets and Architectures to Build NLP Systems, 18th Int'l Conference on Computational Linguistics*, Luxembourg.

J. Burstein. 2003. The e-rater scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis and J. Burstein, editors, *Automated essay scoring: A cross-disciplinary perspective*. Lawrence Erlbaum Associates, Hillsdale, NJ.

S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. 1990. Indexing By Latent Semantic Analysis. *J. of the American Society for Information Science*, 41:391–407.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *J. of the Royal Statistical Society*, 39:1–38.

P. W. Foltz, D. Laham, and T. K. Landauer. 1999a. Automated Essay Scoring: Applications to Educational Technology. In *Proc. of Wolrd Conf. Educational Multimedia, Hypermedia & Telecommunications*, Seattle, USA.

P. W. Foltz, D. Laham, and T. K. Landauer. 1999b. The Intelligent Essay Assessor: Applications to Educational Technology. *Interactive Multimedia Electronic J. of Computer-Enhanced Learning*, 1. `http://imej.wfu.edu/articles/1999/2/04/index.asp` (Accessed 3.4.2005).

M. Girolami and A. Kabán. 2003. On an Equivalence between PLSI and LDA. In *Proc. of the 26th Annual Int'l ACM SIGIR Conf. on Research and Development in Informaion Retrieval*, pages 433–434, Toronto, Canada. ACM Press.

M. Hearst, K. Kukich, M. Light, L. Hirschman, J. Burger, E. Breck, L. Ferro, T. K. Landauer, D. Laham, P. W. Foltz, and R. Calfee. 2000. The Debate on Automated Essay Grading. *IEEE Intelligent Systems*, 15:22–37.

T. Hofmann. 2001. Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning*, 42:177–196.

T. Kakkonen and E. Sutinen. 2004. Automatic Assessment of the Content of Essays Based on Course Materials. In *Proc. of the Int'l Conf. on Information Technology: Research and Education*, pages 126–130, London, UK.

T. Kakkonen, N. Myller, and E. Sutinen. 2004. Semi-Automatic Evaluation Features in Computer-Assisted Essay Assessment. In *Proc. of the 7th IASTED Int'l Conf. on Computers and Advanced Technology in Education*, pages 456–461, Kauai, Hawaii, USA.

T. Kakkonen, N. Myller, E. Sutinen, and J. Timonen. 2005. Comparison of Dimension Reduction Methods for Automated Essay Grading. Submitted.

T. K. Landauer, D. Laham, B. Rehder, and M. E. Schreiner. 1997. How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. In *Proc. of the 19th Annual Meeting of the Cognitive Science Society*, Mawhwah, NJ. Erlbaum.

T. K. Landauer, P. W. Foltz, and D. Laham. 1998. Introduction to latent semantic analysis. *Discourse Processes*, 25:259–284.

B. Lemaire and P. Dessus. 2001. A System to Assess the Semantic Content of Student Essays. *J. of Educational Computing Research*, 24:305–320.

Lingsoft. 2005. `http://www.lingsoft.fi/` (Accessed 3.4.2005).

E. B. Page and N. S. Petersen. 1995. The computer moves into essay grading. *Phi Delta Kappan*, 76:561–565.

E. B. Page. 1966. The imminence of grading essays by computer. *Phi Delta Kappan*, 47:238–243.

M. D. Shermis, H. R. Mzumara, J. Olson, and S. Harrington. 2001. On-line Grading of Student Essays: PEG goes on the World Wide Web. *Assessment & Evaluation in Higher Education*, 26:247.

D. Steinhart. 2000. *Summary Street: an LSA Based Intelligent Tutoring System for Writing and Revising Summaries*. Ph.D. thesis, University of Colorado, Boulder, Colorado.

P. Wiemer-Hastings and A. Graesser. 2000. Select-a-Kibitzer: A computer tool that gives meaningful feedback on student compositions. *Interactive Learning Environments*, 8:149–169.

P. Wiemer-Hastings, K. Wiemer-Hastings, and A. Graesser. 1999. Approximate natural language understanding for an intelligent tutor. In *Proc. of the 12th Int'l Artificial Intelligence Research Symposium*, pages 172–176, Menlo Park, CA, USA.