Proceedings of the 5th WSEAS Int. Conf. on SIMULATION, MODELING AND OPTIMIZATION, Corfu, Greece, August 17-19, 2005 (pp184-189)

Noise Reduction in LSA-based Essay Assessment

TUOMO KAKKONEN, ERKKI SUTINEN, AND JARI TIMONEN
Department of Computer Science
University of Joensuu
P.O. Box 111, FI-80101 Joensuu
FINLAND
http://www.cs.joensuu.fi

Abstract: - With the Latent Semantic Analysis (LSA), it is possible to automatically grade essays, i.e., free-text responses to examinations, by comparing them to a corpus of available learning materials. In order to get grades that correspond to those given by human assessors, it is crucial to train the system with essays that have already been graded. Noise reduction refers to a process in which individual words used for comparing essays with learning materials are given weight according to their significance. To find out the optimal parameters for noise reduction, the system is trained with different parameters, and the corresponding grades for essays are predicted by each of these models. Three standard validation methods, holdout, bootstrap, and k-fold cross-validation, were applied for noise reduction. In an experiment that consisted of 283 essays from three examinations, each of a different subject, the holdout validation method turned out to give the best predictions, and hence, reduce most of the noise.

Key-Words: - Automated Essay Assessment, Latent Semantic Analysis

1 Introduction

An *essay* is a free-text response to a question in a written examination. Assessing essays is a challenging and time-demanding task for a teacher, especially in mass courses [5]. *AEA* (Automatic Essay Assessor) is a system based on the Latent Semantic Analysis (LSA), which automatically grades essays written in the agglutinative Finnish language [7]. Because of its design, it is not, however, limited only to one language.

Basically, as an LSA-based method, AEA determines the grade of an essay based on its similarity to the learning material (textbook passages, lecture notes etc.) that we call *corpus*. The similarity is computed using a *word-by-context matrix,* which essentially contains the occurrence information of each word in the corpus. The corpus is represented as matrix columns, or document vectors, each of which represents a certain sentence, paragraph, or passage of the corpus. An essay to be graded is represented as a *query vector* whose similarity to the corpus decides the grade.

The problem with AEA, as with any LSA-based method, is intuitively clear. The more we make use of word occurrences, the more we emphasize details, such as completely irrelevant words in the corpus, as the basis for the grade of a particular essay. However, we need to focus on the most important words reflecting the conceptual contents of the corpus. This interpretation is also in accordance with a human assessor's work: s/he needs to identify the important concepts but neglect or pay less attention to peripheral contents of the essay s/he is marking. Peripheral contents could also be characterized as *noise*.

The *noise reduction* in LSA is based on the singular value decomposition (SVD), a form of factor analysis. SVD reduces the *dimensionality* of the original word-by-context matrix and increases the dependence between contexts and words [12]. An approximation matrix with reduced *k*-dimensional representation of the original word-by-context matrix is acquired with the operation. In the reduced-dimensional vector space, documents are not represented as sets of independent words, but as "continuous values on each of the *k* orthogonal indexing dimensions" [2]. The aim of the noise reduction step is to trim down noise or unimportant details in the data and to allow the underlying semantic structure to become evident. Awkwardness of the reduction of dimensions is a well-reported problem in applying LSA [1] [4] [12]. Hence, we use also the term *dimension reduction* for noise reduction.

A potential technique for automatic noise reduction is a *model validation method*, commonly applied to information retrieval and data mining. A validation method divides the training data, which in this case is a set of essays graded by a human marker, into two sets: a *training set* and a *test set*. The model that predicts the grade is built with the training set. The test set of essays is used to evaluate the model by comparing the predicted grades to marks given by human assessors. For dimension reduction, a validation method can be applied with several dimension candidates; the one giving the most accurate model will be chosen.

We equipped the AEA system with three different validation methods: *holdout*, *k-fold cross-validation* and *bootstrap*. We studied the following questions:

1. How accurate models can we develop by noise reduction with validation methods?

2. Which of the validation methods is the most reliable when comparing the predictions to essays outside the test set?

In our experiments we used essays and corpora from three different courses.

2 Automatic Essay Assessor AEA

AEA is a Java application for assessing essays written in Finnish. The system consists of three main components: a natural language parser, a method for comparing the similarity between texts, and a method for determining the grades. The system applies LSA in order to measure the content similarity between the essays and the course materials [6] [7].

As Finnish is a morphologically complex language, and words are formed by adding suffixes into the base forms, base forms have to be used instead of inflectional forms when applying LSA to Finnish, especially if a relatively small corpus is utilized. A syntactic parser, the Constraint Grammar Parser for Finnish (FINCG) based on a framework originally proposed by Karlsson [8], is applied in order to get the base forms of each word in the texts [13]. Morphological analysis, which in the case of an agglutinative language such as Finnish is a complicated task, is based on the Two-level model of Koskenniemi [10]. Disambiguation is done by means of Constraint Grammar (CG). In CG constraints are applied to the possible part-of-speech tags generated by the morphological analyzer to eliminate the tags that are inconsistent with the context.

The assessment procedure consists of two phases. In Phase 1, the basis of assessment, the *reference material,* is created from the corpus (Fig. 1). First, a word-by-context matrix giving the number of occurrences of each lemmatized word in each corpus document is constructed. The words occurring in the stopword list (articles, prepositions and alike) are not included in the matrix. Next, entropy-based term weighting is applied to the matrix. The *LSA representation,* a reduced dimensional version of the original word-by-context matrix is finally built by SVD.

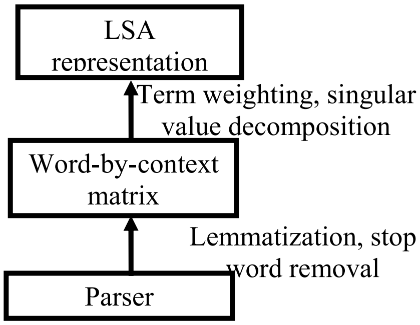
LSA representation
Term weighting, singular value decomposition
Word-by-context matrix
Lemmatization, stop word removal
Parser

Fig 1. Creating the reference material.

In Phase 2, AEA uses the essays in the training set for determining threshold similarity values for each grade category by comparing essays to the reference material. In order to compare the similarity of an essay to the course content a query vector representing its content is created. The query vector is created by calculating the number of occurrences of words in the essay and applying entropy-weighting as when creating the reference material. The comparison of the query vector to each column of the LSA representation results in word-by-context matrix *similarity values* between the essay and each of the documents in the reference material.

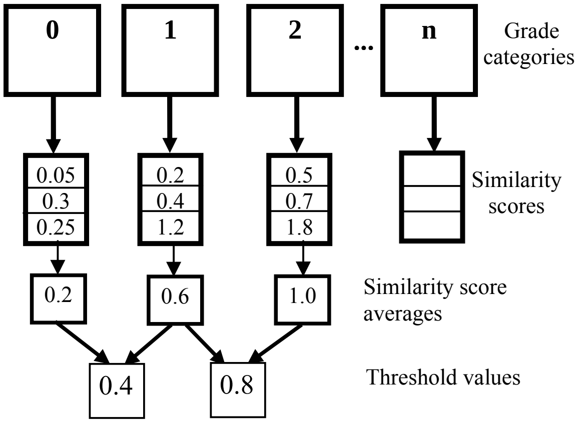

Fig 2. An example of determining threshold values for individual grades.

Taking the sum over the similarity values between an essay and all the documents gives the *similarity score* of the essay to the reference material (Fig. 2). The similarity scores are averaged for all essays that have the same grade given by a human marker. For example, the essays that have got the grade 0 in Fig. 2, have similarity scores 0.05, 0.3, and 0.25, with the average of 0.2. This gives similarity scores to average essays for each grade. After that the threshold values for each grade category are evaluated by splitting the similarity score of an average essay into different categories. The number of categories depends on the applied grade scale (e.g. for a scale of 0…6, the number of different categories is 7).

According to experiments [7], the AEA system is able to achieve the Spearman correlation of up to 0.82 between the grades by the system and a human grader. The correlation was derived from 86 essays and textbook sentences used for training. After training the assessment system, 57 essays were graded using all the possible LSA dimensions.

3 Data Validation Methods Applied

The original AEA method cannot be used in real-life automatic essay assessment. This is because the dimensionality that produces the most accurate grades has to be obtained automatically, not by human-guided experiments.

The selection of the appropriate model arises often in many statistical problems [15]. The dimensions in LSA can also be interpreted as models, which represent different meanings of contents in a corpus or an essay. When applying LSA to essay assessment, we try to detect relevant content from an essay that is indirectly, or latently, present in the corpus material. One of the LSA dimensionalities seems to be the optimal one and models the information, relevant to grading, better than other dimensionalities. What information is relevant in each case depends not only on the essay assignment and reference material, but also on the teacher who has graded the essays. Thus, the goal of the system is to model both the relevant content of the course as well as the teacher's grading.

The correlation between grades given by two human assessors varies typically from 0.50 to 0.75 [5] [11] [14]. In automatic assessment, training the system with essays graded by a particular human marker allows us to take the diverse assessment styles into consideration.

The validation methods that were applied in this study automatically perform the noise reduction in AEA aiming at the discovery of the optimal number of LSA dimensions. This results in the highest correlation between the grades given by the system and a human assessor. The process of noise (or dimension) reduction is called *system training*, and it is carried out by a validation method as follows.

The *graded essays* are divided into two separate sets. One set is used in the *training* phase and the other in the *testing* phase. The training phase takes place as explained in Section 2 for AEA, and its result is a model for a given LSA dimension. In the testing phase, the model predicts the grades to the essays in the test set. Training and testing is repeated for all possible LSA dimensions. Because all the essays have a grade given by a human grader, it is possible to calculate the correlation between human and system given grades. As a result of the repeated training and testing phases, we obtain correlations for all the dimensions. Finally, the LSA dimension that results in the highest correlation on average is selected.

Next, we will introduce the validation methods that were applied for the noise reduction process. The validation method determines the selection of the training and test sets. These sets are always mutually exclusive.

3.1 Holdout

In the holdout validation method, the most commonly used technique is to isolate 1/3 of the data into the test set and use the remaining 2/3 of the data for training. In *random subsampling* (or *repeated holdout*) the division is performed k times and the final estimate is calculated by averaging the estimates from different runs [9].

In our implementation, the holdout method divides the essays used in system training into two sets. 2/3 of the total amount of essays is used for training and the remaining 1/3 is used for testing. All the other phases of the grading process remain as reported in Section 2.

3.2 *k*-fold Cross-Validation

In k-fold cross-validation the whole set D of graded essays is randomly divided into k subsets (folds) $D_1, D_2,\ldots, D_k$ [9]. After the division each subset $D_1, D_2,\ldots, D_k$ is alternately used for testing. The remaining subsets are united into one and used for training. This means that in k-fold cross-validation the training and testing is repeated k times. Final estimation, in our case correlation, is calculated by averaging the estimates from each of the training and testing runs.

In cases where the original dataset contains instances from different classes it is possible to use *stratified* k-fold cross-validation. In our studies classes of instances mean different grade categories of essays. Stratification means that the division into k folds takes the classes of instances into account. When selecting training and testing sets randomly, without using stratification, it is not possible to guarantee that each class is properly represented in both sets [16]. After stratification each cross-validation fold contain approximately the same number of class-instances as the original dataset. In our case, after stratification, each cross-validation fold contains the same number of essays from different grade categories as the original essay set.

Complete cross-validation (or *leave-one-out cross-validation*) can be represented as n-fold cross-validation where n is the total number of all instances in the original dataset. Complete cross-validation uses the largest amount of data in each training and testing case, which presumably increases the accuracy of the estimate drawn [16]. Leave-one-out cross-validation does not use random splitting of the whole dataset because all of the instances are used alternately for testing and the remaining for training. A problem with the method is its time consumption. The computational costs may become unpractical, if complete cross-validation is applied to large datasets.

In our experiments, stratified and non-stratified versions of 3- and 10-fold cross-validation were applied.

3.3 .632 Bootstrap

The bootstrap validation method was originally introduced by Efron and Tibshirani [3]. It is a method for statistical conclusions based on experimental data. In our system the experimental data consists of human-graded essays used in system training. In cross-validation methods the same instance cannot occur more than once in the training set. As opposed to cross-validation methods bootstrap uses *replacement*. In replacement, the same instance from the original dataset can be selected more than once into training set.

.632 bootstrap is a variant of the bootstrap method. The decimal fraction .632 describes the probability of an item being picked into the training set. Given a dataset of size *n*, the bootstrap training set is created by sampling uniformly *n* instances using replacement from the original dataset. After the creation of the training set the instances that were not selected into training set are used as a test set.

The probability of an instance being picked each time into training set is 1/*n*, and so a 1-1/*n* probability of *not* being picked. The number of picking opportunities is *n* so the probability of an instance not being chosen after *n* samples is $(1 - 1/n)^n \approx e^{-1} \approx .368$ [9]. This means that the test set contains approximately 36.8 % and training set approximately 63.2 % of the total number of instances in the original dataset.

In our studies the creation of the samples is repeated 10, 20, 50 or 100 times consecutively. In each run the grade limits are defined by using the essays in the training set, and the accuracy of the system grades is tested with essays in the test set. The final estimation is an average of each bootstrap repetition.

4 Experimental Results

For our experiments we collected essays and corpora from three courses. Table 1 shows the courses from which the materials were collected, the subjects of the course materials, the total number of essays, and the status of the human graders. In all the experiments, the essays were divided into two parts. The first part was used for training and testing with all the validation methods, and the second part was used for evaluating the accuracy of the derived model with other essays (which had not been used either for training or testing).

Table 1. Summary of the essay material collected for the experiments 1..3.

Experiment	Subject	Level	No. Essays	Grader
1	Education	Undergraduate	143	Professor
2	Communi-cation	Vocational	87	Course teacher
3	Software engineering	Graduate	53	Assistant

Some of the results are represented in Tables 2..4. In tables, the column *Training of the system* indicates the applied validation method and its rank compared to the accuracy of other validation methods. The essays were divided into test and training sets. The column *Grading accuracy* shows the results gained after the system was trained with the first part of the essays and tested with second part of the essays. The column *Correlation* shows the Spearman correlation between the grades given by the human assessor and those given by the system, when the dimension found in the training phase was applied by the system for scoring the essays. *Correlation match* shows the relation C_2/C_1, where C_1 is the Spearman correlation between the most accurate predicted grades and those given by a human marker, and C_2 is the Spearman correlation that is created by the dimension found in the training phase. The most accurate prediction is obtained by using a model based on all the essays. For example, the correlation match of 100 % means that the method found the optimal dimension in the training phase. Tables 2..4 are vertically divided into two sections, separated by a thick border. The upper section shows the results gained when the corpus material was divided into paragraphs, and the lower section shows the ones gained when the corpus material was divided into sentences. In Tables 2..4 we report the ranks of the five methods that gave the most accurate results.

Essays in Experiment 1 were graded on a scale from zero to six. The length of the essays varied from 18 to 445 words. The size of the corpus that gave an answer to the assignment was 2397 words. The corpus was divided alternately into paragraphs and sentences, and all the validation methods were tested for each of the divisions. We used 70 essays in order to train the system. 73 essays were used for testing the accuracy of the LSA dimension found during the training phase. For results see Table 2.

Table 2. Results from Experiment 1. The corpus was divided into paragraphs and sentences.

Training of the system		Grading accuracy	
Rank	Method	Correlation	Correlation match (%)
1.	Bootstrap 10	0.76	96.68
2.	Holdout 10	0.76	96.68
3.	Bootstrap 20	0.76	96.61
4.	Bootstrap 50	0.76	96.61
5.	Bootstrap 100	0.76	96.61
1.	3-fold cross-validation	0.80	100.00
2.	Stratified 3-fold cross-validation	0.80	100.00
3.	Holdout 10	0.80	100.00
4.	Holdout	0.78	97.07
5.	Stratified 10-fold cross-validation	0.76	94.44

In Table 2, ten times repeated bootstrap and holdout gave the highest correlation match between the optimal correlation and the correlation when the dimension found during the training phase was used. When the corpus material that gave the answer to the essay assignment was divided into paragraphs, non-stratified and stratified 3-fold cross-validations and ten times repeated holdout resulted in the optimal dimension during the training phase.

In Experiment 2, two textbooks were used in the course. The textbook on communication was 125 pages long and the other dealing with psychology 51 pages long. The essay assignment consisted of two parts: writing definitions of two terms and explaining the use of the terms. Because one part of the assignment required some own explanation of the term use by the student, we added one example answer to the corpus material. The total size of the corpus was 1583 words. We divided the corpus into paragraphs and sentences and, in turn, we tested all the validation methods by using both paragraphs and sentences one by one. 42 essays were used for training the system. 45 essays were used for grading and calculating the accuracy of the dimension found during the training phase. Table 3 shows the results.

Table 3. Results from Experiment 2. The corpus was divided into paragraphs and sentences.

Training of the system		Grading accuracy	
Rank	Method	Correlation	Correlation match (%)
1.	Holdout 10	0.54	100.00
2.	3-fold cross-validation	0.53	98.09
3.	Stratified 3-fold cross-validation	0.53	98.09
4.	Bootstrap 10	0.53	98.01
5.	Holdout 20	0.53	98.01
1.	Stratified 3-fold cross-validation	0.57	100.00
2.	Stratified 10-fold cross-validation	0.57	100.00
3.	Bootstrap 50	0.57	100.00
4.	Holdout 10	0.57	100.00
5.	3-fold cross-validation	0.50	87.08

As shown in Table 3, ten times consecutively repeated holdout resulted in the optimal dimension during the training phase when the corpus material was divided into paragraphs. Stratified 3- and 10-fold cross-validations, fifty times repeated bootstrap and ten times repeated holdout resulted in the correlation match of 100 %, meaning that the method found the optimal dimension in the training phase.

In Experiment 3, the essays were graded by an assistant on a scale from zero to ten. Corpus material for the LSA was constructed from the course handout with teacher's comments included. In addition, the transparencies presented to the students at the lectures were included in the corpus. From the text material we selected the parts that gave the correct answer to the essay assignment and used that as a corpus for LSA. In total 26 essays were used for training the system and 27 for grading and testing the accuracy of each applied method. Table 4 shows the results of the experiment.

Table 4. Results from Experiment 3. The corpus material was divided into paragraphs and sentences.

Training of the system		Grading accuracy	
Rank	Method	Correlation	Correlation match (%)
1.	3-fold cross-validation	0.88	100.00
2.	Holdout 20	0.88	100.00
3.	Bootstrap 20	0.87	98.41
4.	Holdout 10	0.87	98.40
5.	10-fold cross-validation	0.87	98.23
1.	Holdout 10	0.90	99.05
2.	Holdout 20	0.90	99.05
3.	3-fold cross-validation	0.88	97.04
4.	10-fold cross-validation	0.83	91.37
5.	Holdout	0.81	89.49

Table 5 shows the top ten ranks of the applied methods and their variants. It summarizes the results from Tables 2..4. *Average of correlation* matches was calculated as the average of the correlation matches for each three sets of essays described above.

Table 5. Correlation matches from the experiments represented in Tables 2..4 averaged by each applied method.

Rank	Method	Average of correlation matches (%)
1.	Holdout 10	99.02
2.	Bootstrap 50	97.02
3.	Stratified 3-fold cross-validation	96.28
4.	Stratified 10-fold cross-validation	95.96
5.	3-fold cross-validation	95.95
6.	10-fold cross-validation	93.01
7.	Holdout 20	92.38
8.	Bootstrap 20	89.29
9.	Bootstrap 10	88.94
10.	Holdout	84.96

In Table 5, ten times repeated holdout results highest correlation matches on average. Fifty times repeated bootstrap gives an average of 97.02 % for the correlation matches represented in Tables 2..4. Also other methods and their variants are ranked according to their correlation match averages.

5 Conclusions

We have presented the usage of three validation methods, holdout, bootstrap and *k*-fold cross-validation, for automating the noise, or dimensionality, reduction in LSA-based essay assessment system for Finnish. We have also presented results of experiments with three test sets collected from university and vocational school courses concerning diverse topics. Results of our experiments indicate that ten times repeated holdout resulted in the most accurate grades compared to the grades given by a human. On average the method resulted the accuracy of 99 % between two correlations. First correlation was calculated by using the dimension found in the training phase between human and system grades. The second correlation was calculated between human and system grades by using the dimension that produced the most accurate grades.

Research results indicate that noise reduction can be automated and the sufficient accuracy of the system grades can be achieved. We see the automation of dimensionality selection as a crucial step towards developing a truly ready-for-production assessment system. Our long-term goal is to develop a semi-automated evaluation environment, which offers support for both students and teachers.

References:

[1] Bingham, E., Mannila, H. 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*. San Francisco, California, USA, 245-250, 2001.

[2] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, Vol. 41, No. 6, 1990, pp. 391-407.

[3] Efron, B., Tibshirani R., J. *An Introduction to the Bootstrap*. Chapman and Hall, New York, New York, USA, 1993.

[4] Globerson, A., Tishby, N.. Sufficient Dimensionality Reduction. *Journal of Machine Learning Research*. Vol. 3, No. 1, 2003, pp. 1307-1331.

[5] Hopkins, K. D., Stanley, J. C., Hopkins, B. R. *Educational and Psychological Measurement, Seventh Edition.* Prentice-Hall, Englewood Cliffs, USA, 1990.

[6] Kakkonen, T., Myller, N., Sutinen, E. Semi-Automatic Evaluation Features In Computer-Assisted Essay Assessment. *Proceedings of Computers and Advanced Technology in Education (CATE)*, Kauai, Hawaii, USA, 456-461, 2004.

[7] Kakkonen, T., Sutinen, E. Automatic Assessment of the Content of Essays Based on Course Materials. *Proceedings of International Conference on Information Technology: Research and Education (ITRE)*, London Metropolitan University, London, 126-130, 2004.

[8] Karlsson, F. Constraint Grammar as a Framework for Parsing Running Text. *Proceedings of the 13th Conference on Computational Linguistics - Volume 3*. Helsinki, Finland, 1990.

[9] Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference on Artificial Intelligence (IJCAI)*, 1137-1143, 1995.

[10] Koskenniemi, K. A General Computational Model for Word-Form Recognition and Production. *Proceedings of the 22nd conference on Association for Computational Linguistics*. Stanford, California, USA, 1984.

[11] Landauer, T. K., Laham, D., Rehder, B., Schreiner, M. E. How Well Can Passage Meaning be Derived without Using Word Order? A Comparison of Latent Semantic Analysis and Humans. *Proceedings of the 19th annual meeting of the Cognitive Science Society*. Mawhwah, NJ: Erlbaum, 412-417, 1997.

[12] Landauer, T. K., Foltz, P. W., Laham, D. Introduction to Latent Semantic Analysis. *Discourse Processes,* Vol. 25, No. 2&3, 1998, pp. 259-284.

[13] Lingsoft Inc. http://www.lingsoft.fi (Accessed 20.3.2005).

[14] Page, E. B., Petersen, N. S. The Computer Moves into Essay Grading, *Phi Delta Kappan,* Vol. 76, No. 7, 1995, pp. 561-565.

[15] Picard, R. R., Cook, R.D. Cross-validation of regression models. *Journal American Statistical Association*, Vol. 79, No. 387, 1984, pp. 575-583.

[16] Witten, I. H., Frank, E. *Data Mining : Practical Machine Learning Tools and Techniques with Java Implementations*. Academic Press San Diego, California, USA, 2000.