The Power of Normalised Word Vectors for Automatically Grading Essays

Robert Williams
School of Information Systems, Curtin University of Technology
Perth, Australia

Bob.Williams@cbs.curtin.edu.au

Abstract

Latent Semantic Analysis, when used for automated essay grading, makes use of document word count vectors for scoring the essays against domain knowledge. Words in the domain knowledge documents and essays are counted, and Singular Value Decomposition is undertaken to reduce the dimensions of the semantic space. Near neighbour vector cosines and other variables are used to calculate an essay score. This paper discusses a technique for computing word count vectors where the words are first normalised using thesaurus concept index numbers. This approach leads to a vector space of 812 dimensions, does not require Singular Value Decomposition, and leads to a reduced computational load. The cosine between the vectors for the student essay and a model answer proves to be a very powerful independent variable when used in regression analysis to score essays. An example of its use in practice is discussed.

Keywords: Automated Essay Grading, Latent Semantic Analysis, Singular Value Decomposition, Normalised Word Vectors, Electronic Thesaurus, Multiple Regression Analysis.

Introduction

Automated Essay Grading (AEG) systems are now appearing in the educational marketplace, and are increasingly being accepted as a way of efficiently grading large numbers of essays (Shermis & Burstein, 2003). There are many theoretical constructs underpinning the various AEG systems (Williams, 2001; Valenti, Neri & Cucchiarelli, 2003). One of the major systems, the Intelligent Essay Assessor (Pearson Knowledge Technologies, 2005), makes use of a mathematical technique known as Latent Semantic Analysis (LSA) (Landauer, Foltz & Laham, 1998). This system is interesting because of the way it derives the knowledge contained in an essay from the words comprising the essay. The MarkIT system (Williams & Dreher, 2005), being developed by the author and colleagues, uses an alternative way of deriving content from an essay, but still based on the words making up the essay. This paper discusses these two alternative word-based content representations, presents new material on the grading algorithm for MarkIT, and compares the performances of the two systems.

In this paper we do not have space to give a detailed coverage of the issues associated with AEG systems. For a comprehensive coverage of AEG systems, their algorithms, and performance details, see Hearst (2000), Williams (2001), and Valenti, Neri and Cucchiarelli (2003).

Latent Semantic Analysis

LSA is a mathematical technique based on vector algebra. It is used to derive a representation of the content of a collection of text documents in a particular domain of knowledge. This content representation is generally termed the semantic space. This space is built from text segments that may consist of the complete documents, or subsets of the documents, such as paragraphs or sentences. Each word in the segment is represented as a row in a matrix, and each segment is represented as a column in the same matrix. The counts of the number of times the words appear in the segments are entered in the corresponding elements in the matrix.

The following example, taken from Landauer, Foltz, and Laham (1998) and used with permission from the authors and Lawrence Erlbaum Associates, the publishers, illustrates the technique. The titles of five documents relating to human computer interaction and four relating to mathematical graph theory are shown below.

c1:	*Human* machine *interface* for ABC *computer* applications
c2:	A *survey* of *user* opinion of *computer system response time*
c3:	The *EPS user interface* management *system*
c4:	*System* and *human system* engineering testing of *EPS*
c5:	Relation of *user* perceived *response time* to error measurement
m1:	The generation of random, binary, ordered *trees*
m2:	The intersection *graph* of paths in *trees*
m3:	*Graph minors* IV: Widths of *trees* and well-quasi-ordering
m4:	*Graph minors*: A *survey*

The matrix below shows the word count for the selected words occurring in at least two of the titles. These words are shown in italics in the document titles.

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

A vector algebra technique, known as Singular value Decomposition (SVD) is then applied to this matrix. SVD breaks the matrix into 3 component matrices that can be matrix multiplied to produce the original matrix. However the dimensions of these 3 matrices are reduced before the remultiplication. The remultiplied matrix is now approximately equivalent to the original matrix in terms of its element values, but now contains values for elements that were previously zero. In other words, the reconstituted matrix now has relationships for words and segments that were not explicitly displayed in the original matrix, but have been induced by the SVD process from the hidden or latent relationships amongst the words and segments. The reconstructed approximation to the original matrix, based upon the first two columns in the three component matrices (not shown), is

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	0.16	0.40	0.38	0.47	0.18	-0.05	-0.12	-0.16	-0.09
interface	0.14	0.37	0.33	0.40	0.16	-0.03	-0.07	-0.10	-0.04
computer	0.15	0.51	0.36	0.41	0.24	0.02	0.06	0.09	0.12
user	0.26	0.84	0.61	0.70	0.39	0.03	0.08	0.12	0.19
system	0.45	1.23	1.05	1.27	0.56	-0.07	-0.15	-0.21	-0.05
response	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
time	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
EPS	0.22	0.55	0.51	0.63	0.24	-0.07	-0.14	-0.20	-0.11
survey	0.10	0.53	0.23	0.21	0.27	0.14	0.31	0.44	0.42
trees	-0.06	0.23	-0.14	-0.27	0.14	0.24	0.55	0.77	0.66
graph	-0.06	0.34	-0.15	-0.30	0.20	0.31	0.69	0.98	0.85
minors	-0.04	0.25	-0.10	-0.21	0.15	0.22	0.50	0.71	0.62

What was originally a sparsely populated matrix of relationships amongst words and segments is now a rich array of associations. This is now the semantic space for this collection of document titles.

"This text segment is best described as having so much of abstract concept one and so much of abstract concept two, and this word has so much of concept one and so much of concept two, and combining those two pieces of information (by vector arithmetic), my best guess is that word X actually appeared 0.6 times in context Y." (Landauer, et al., 1998, p 264)

Essays on a particular topic are graded as follows. The appropriate semantic space is built – this can be done by processing electronic texts on the topic, or from a collection of several hundred human graded essays on the topic. The essay to be graded is then processed using the SVD technique to build a document vector in this space. An essay score is then computed from near neighbour human scored essay vectors in this space, and other variables.

The IEA is a commercial implementation of the LSA approach to AEG. Landauer indicates that this system builds the semantic space as follows:

> "IEA/LSA always starts from a reduced dimensional space based on a large relevant corpus to which it adds text special to the topic and the student essays" (personal email communication, 16 November, 2005).

Evaluation of LSA and Essay Grading

Nichols has evaluated the IEA. He concludes

> "All four of the measures of the relationship between essay scores and expert scores (percent agreement, Spearman rank-order correlation, kappa statistic and Pearson correlation) indicated a stronger relationship between the IEA and experts than between readers and experts. In addition, the results of examining the scoring processes used by the IEA showed that the IEA used processes similar to a human scorer. Furthermore, the IEA scoring processes were more similar to processes used by proficient human scorers than to processes used by non-proficient or intermediate human scorers." (Nichols, 2005, p 21).

Vector Representation of Documents using a Thesaurus to Normalise Document Words

The MarkIT AEG system is a software system that automatically grades essays against an ideal content answer at the same level of accuracy as human graders (MarkIT, 2005; Williams & Dreher, 2005). This section explains how vector algebra techniques are used to represent similarities in content between documents in MarkIT. In order to build this vector representation, a thesaurus is used to “normalise” words in the documents by reducing all words to a thesaurus root word appropriate to the concept the word belongs to. Counts of these concepts are then used for the vector representation. Consider the following start of sentence fragments from successive sentences in 3 separate documents:

Document Number	Document Text
(1)	The little boy… A small male…
(2)	A minor boy… A funny girl…
(3)	The large boy… Some minor day…

Suppose a thesaurus exists with the following root concept numbers and words:

Concept Number	Words
1.	the, a
2.	little, small, minor
3.	boy, male
4.	large
5.	funny
6.	girl
7.	some
8.	day

Three dimensional vector representations of the above document fragments on the first 3 concept numbers (1-3) can be constructed by counting the number of times a word in that concept number appears in the document fragments. These vectors are:

Document Number	Vector on first 3 concepts	Explanation
(1)	[2, 2, 2]	[The, a; little, small; boy, male]
(2)	[2, 0, 1]	[A, a; ; boy]
(3)	[1, 1, 1]	[The; minor; boy]

Figure 1 shows these 3 dimensional vectors pictorially.

Computing the Variable CosTheta

If we assume that document 1 is the model answer, then we can see how close semantically documents 2 and 3 are to the model answer by looking at the closeness of their corresponding vectors. The angle between the vectors varies according to how “close” the vectors are. A small angle indicates that the documents contain similar content, a large angle indicates that they do not have much common content. Angle Theta1 is the angle between the model answer vector and the vector for document 2, and angle Theta2 is the angle between the model answer vector and the vector for document 3.

The cosines of Theta1 and Theta2 can be used as measures of this closeness. If documents 2 and 3 were identical to the model answer, their vectors would be identical to the model answer vector, and would be collinear with it, and have a cosine of 1. If on the other hand, they were completely different, and therefore orthogonal to the model answer vector, their cosines would be 0.

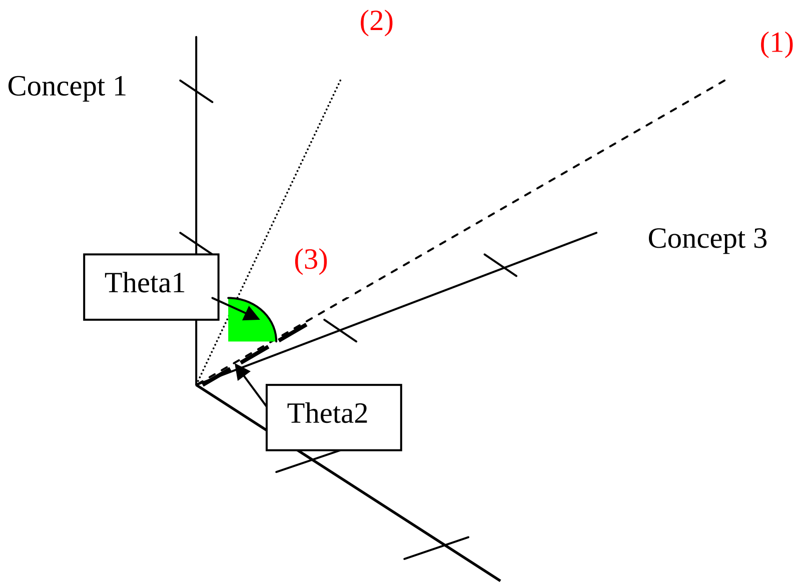

Figure 1. Vector representation (dashed lines) of documents

Generally in practice, a document's cosine is between these upper and lower limits.

The variable CosTheta used in the scoring algorithm is this cosine computed for the document being scored.

In general, these ideas are extended to the 812 concepts in the Macquarie Thesaurus from Macquarie Library Pty Ltd (Macquarie Library, 2005), and all words in the documents. This means that the vectors are constructed in 812 dimensions, and the vector theory carries over to these dimensions in exactly the same way – it is of course hard to visualise the vectors in this hyperspace. (The system developers approached a number of thesaurus publishers with a view to obtaining a research licence to use an electronic thesaurus, and Macquarie Library Pty Ltd was the only company willing to grant one; hence its usage).

Computing the Variable VarRatio

We now discuss another powerful essay grade predictor, VarRatio, which is based on these concept vectors. The number of concepts that are present in the model answer (document 1) above is 3. This can be determined from the number of non-zero counts in the numerical vector representation.

The number of concepts that are present in document 2 above is 2 – the second vector index is 0. To compute the VarRatio for this document 2 we divide the non-zero concept count for document 2 by the non-zero concept count in the model answer i.e. VarRatio = 2/3 = 0.67. The corresponding VarRatio for document 3 is 3/3 = 1.00.

This simple variable provides a remarkably strong predictor of essay scores, and is generally present as one of the components in the scoring algorithm.

Scoring Student Essays by Matching a Model Answer against Student Answers

MarkIT makes use of a multiple regression equation to assign a grade to a student essay. The regression equation is developed from about 100 human graded training essays and an ideal or model answer. The document vectors described above are constructed. Values are then computed for many variables from the relationships between the content and vectors of the model answer and the training essays. Once the training has been performed, and the grading algorithm built, each unmarked essay is processed to obtain the values for the independent variables, and the regression equation is then applied. Generally CosTheta and VarRatio are significant predictors in the scoring equation. An example taken from a trial of the system is now discussed.

In the trial, Year 10 high school students hand wrote essays on paper on the topic of "The School Leaving Age". Three trained human graders then graded these essays against a marking rubric. The essays, 390 in total, were then transcribed to Microsoft Word document format. The essay with the highest average human score was selected as the model answer. It had a score of 48.5 out of a possible 54, or 90%. In one test of the system, 100 essays were used to build the scoring algorithm. The scoring algorithm was built using the first 100 essays in the trial when ordered in ascending order of the identifier. Table 1 shows the results of the multiple regression procedure built upon the output of the MarkIT system for these 100 essays. The multiple R is 0.89 and the prediction equation is

Student Grade = -22.35 + 11.00*CosTheta + 15.70*VarRatio +7.64*Characters Per Word + 0.20 Number of NP Adjectives

Table 1. Multiple Regression Analysis for First 100 Essays

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.89
R Square	0.79
Adjusted R Square	0.78
Standard Error	4.16
Observations	100.00

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4.00	6079.76	1519.94	87.71	0.00
Residual	95.00	1646.21	17.33		
Total	99.00	7725.97			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-22.35	6.67	-3.35	0.00
CosTheta	11.00	3.74	2.94	0.00
VarRatio	15.70	2.86	5.49	0.00
Characters Per Word	7.64	1.74	4.40	0.00
Number of NP Adjectives	0.20	0.08	2.41	0.02

- CosTheta is computed as per the explanation above.
- VarRatio is computed as per the explanation above.
- Characters Per Word is the average number of characters in the words in the essay
- Number of NP Adjectives is the number of Adjectives in Noun Phrases in the essay

Notice that only 4 independent variables are needed for the predictor equation in this example.

Once this scoring algorithm was coded into the scoring program, the remaining 290 essays were graded by it. Figure 2 shows the results.

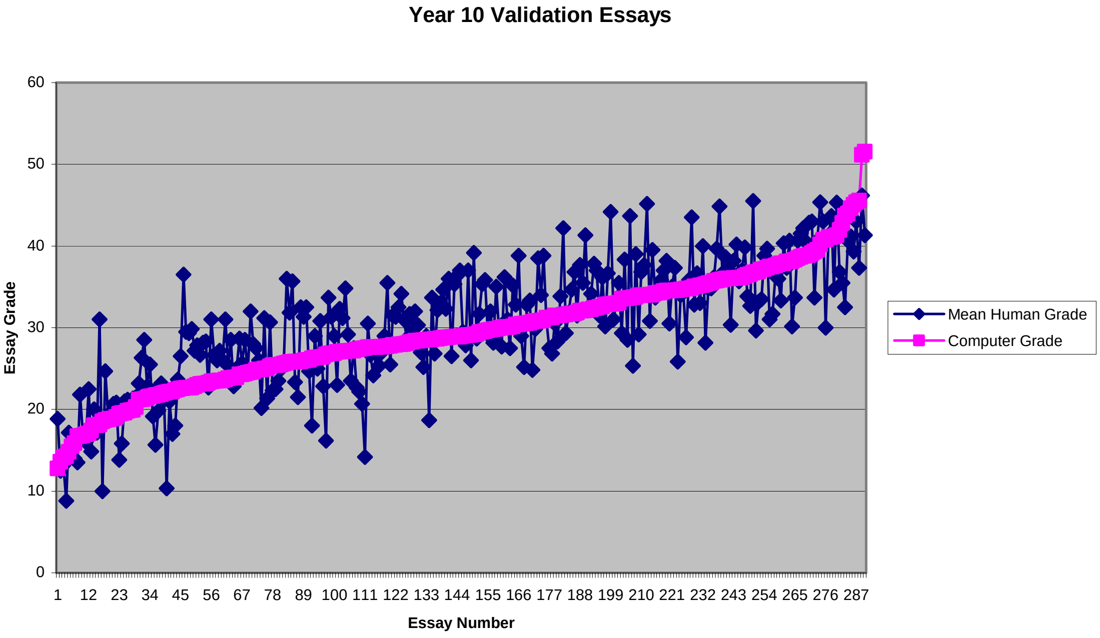

Figure 2. Results of Computer Scoring of Last 290 Essays

The mean score for the human average grade for these 290 essays was 30.34, while the mean grade given by the computer was 29.45, a difference of 0.89. The correlation between the human and computer grades was 0.79. The mean absolute difference between the two was 3.90, representing an average error rate of 7.23% when scored out of 54 (the maximum possible human score).

The correlations between the three humans amongst themselves were 0.81, 0.78 and 0.81.

The benefits of averaging the scores from the human graders are shown by the fact that the correlation between the computer and the mean score of the three humans is higher, at 0.79, than the individual correlations at 0.67, 0.75 and 0.75.

Conclusion

LSA makes use of SVD to reduce the large number of dimensions generated when each word in a document is counted as a separate dimension. Typically the dimensions are reduced to about 300 (Landauer, 2005). The processing involved for the SVD takes a few hours on a common small Linux cluster (Landauer, personal email communication, 16 November 2005).

While the number of dimensions resulting from normalising words against thesaurus index numbers is 812, much less processing is involved – typically the training session to build the scoring algorithm for a prompt using 100 essays takes 5 minutes on a Pentium 3.4GHz machine under Windows XP. Similar accuracy of the resultant scores, when compared to human scores, is maintained. For example, IEA achieved a correlation of 0.81 with human scores for GMAT essays (Landauer, Laham & Foltz, 2003), compared to the 0.79 achieved by MarkIT for the Year 10 High School essays reported above.

The power of the resultant document vectors to represent the essay content is also impressive, as only the cosine of the model and student essay vectors, and three other predictors, are needed for scoring the student essay, in the example discussed. This low number of predictors appears to be unique to MarkIT. Other documented systems appear to require substantially more (Shermis & Burstein, 2003).

Acknowledgement

The author wishes to express his appreciation to Professor Tom Landauer of the University of Colorado, Boulder, for his valuable comments and suggestions for improvements during the writing this paper.

References

Hearst, M. (2000). The debate on automated essay grading. *IEEE Intelligent Systems, 15* (5), 22-37, IEEE CS Press.

Landauer, T., Foltz, P., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*, 259-284.

Landauer, T., Laham, D. & Foltz, P. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor™. In M. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp 87–112). New Jersey, USA: Lawrence Erlbaum Associates.

Macquarie Library. (2005). Retrieved November 29, 2005 from http://www.macquariedictionary.com.au/

MarkIT (2005). Retrieved November 28, 2005 from http://www.essaygrading.com/

Nichols, P. (2005). Evidence for the interpretation and use of scores from an automated essay scorer. *PEM Research Report 05-02*, Retrieved November 28, 2005 from http://www.pearsonedmeasurement.com/downloads/research/RR_05_02.pdf

Pearson Knowledge Technologies (2005). Retrieved November 28, 2005 from http://www.knowledge-technologies.com/

Shermis, M. & Burstein, J. (2003). *Automated essay scoring: A cross-disciplinary perspective*. New Jersey, USA: Lawrence Erlbaum Associates.

Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading, *Journal of Information Technology Education*, 2, 319 – 330. Retrieved January 30, 2006 from http://jite.org/documents/Vol2/v2p319-330-30.pdf

Williams, R. (2001) Automated essay grading: an evaluation of four conceptual models. In M. Kulski & A. Herrmann (Eds.), *New horizons in university teaching and learning: Responding to change*. Perth, Australia: Curtin University of Technology.

Williams, R. & Dreher, H. (2005). Formative assessment visual feedback in computer graded essays. *Journal of Issues in Informing Science and Information Technology*, *2*, 23-32. Retrieved from http://proceedings.informingscience.org/InSITE2005/I03f95Will.pdf

Biography

Robert Williams has over 25 year's experience in the Information Systems industry, as a practitioner, researcher and lecturer. He currently is a lecturer in the School of Information Systems at Curtin University of Technology in Perth, Western Australia. He has extensive experience in systems analysis and design, and programming, on a variety of mainframe, mini and personal computers, and a variety of operating systems and programming languages. Applications he has worked with include mathematical, statistical, bridge and road engineering, financial, corporate resource allocation, business simulation and educational systems. He has published a number of articles on system users' personalities and satisfaction, decision support systems, and automated essay grading systems. In 2001 he led a team of researchers in the School of Information Systems at Curtin University of Technology which conducted what is believed to be the first trial in Australia of an Automated Essay Grading system. Robert holds a Bachelor of Arts degree with double majors in Mathematics and Economics from the University of Western Australia, a Graduate Diploma in Computing from the Western Australian Institute of Technology, and a Master of Information Systems degree from Curtin University of Technology.