ISSN 1744-1986

Technical Report N° 2008/ 14

EMMA – a Computer Assisted Assessment System based on Latent Semantic Analysis

Debra Trusso Haley
Pete Thomas
Marian Petre
Anne De Roeck

27th August, 2008

Department of Computing
Faculty of Mathematics, Computing and Technology
The Open University

Walton Hall, Milton Keynes, MK7 6AA
United Kingdom

http://computing.open.ac.uk

EMMA - a Computer Assisted Assessment System based on Latent Semantic Analysis

Debra Trusso Haley, Pete Thomas, Marian Petre, Anne De Roeck
The Centre for Research in Computing
The Open University
Walton Hall, Milton Keynes MK7 6AA UK
D.T.Haley, P.G.Thomas, M. Petre, A.DeRoeck at open.ac.uk

Abstract

We present EMMA (ExaM Marking Assistant), a Latent Semantic Analysis (LSA) based Computer Assisted Assessment System (CAA) we have developed as part of ELeGI (www.elegi.org) – "A semantic Grid for human learning is the vision behind the European ELeGI Integrated Project for the implementation of future learning scenarios based on ubiquitous, collaborative, experiential-based and contextualized learning through the design, implementation and validation of the Learning Grid". Assessment is an important component of learning and can have a strong impact on student progress. EMMA can provide both formative and summative assessment that is unbiased and repeatable as well as the almost instant feedback that is most useful for student learning. Our work has demonstrated that, even though the theory of LSA is over 15 years old, many of the details that make LSA a practical assessment technique are not known by the research community beyond the original LSA developers. In this paper, we summarise what we have learned about LSA, give an overview of how EMMA works, describe the types of questions that EMMA can assess, and evaluate its results as compared to human markers, and outline a plan for further research.

1 Introduction

Reliable, repeatable, and rapid assessment is crucial for education (Berglund, 1999; Daniels, Berglund, Pears, & Fincher, 2004). Unfortunately, frequent assessment can be an onerous task for educators, thus prompting the development of various Computer Assisted Assessment Systems (CAA) to mark essays or short answers. For example, see (Burstein, Chodorow, & Leacock, 2003) for a CAA that grades essays and (Wiemer-Hastings, Graesser, & Harter, 1998) for a tutoring system that evaluates short answers.

We are interested in providing tools, such as formative online tests, that improve the learning of programming and computing in general. We have developed a tool (P. G. Thomas, Waugh, & Smith, 2005) that is part of an online system to mark diagrams produced by students in a database course. We are developing EMMA, a Latent Semantic Analysis-based CAA (D. Haley, P. Thomas, A. De Roeck, & M. Petre, 2007) to mark short answers about html and other areas in computer science. We chose LSA as the technology underlying our CAA because it had been used successfully in the past to mark general knowledge essays (Landauer, Foltz, & Laham, 1998) and a pilot study (P. Thomas, Haley, De Roeck, & Petre, 2004) showed it had promise in our area of short answers in the domain of computer science.

Our experience with LSA has highlighted a significant challenge – the developer must choose many options that are intrinsic to the success of any LSA-based marking system. A review of the literature (Haley, Thomas, De Roeck, & Petre, 2005) revealed that although many researchers have reported work with LSA, it is difficult to get a full picture of these systems. Some of the missing information includes type of training corpus and examples of questions being marked as well as the fundamental LSA options such as weighting function and number of dimensions in the reduced matrix. (See Section 2 for a description of what these terms mean.)

Evaluation of a CAA is a crucial topic because automatic marking systems will not be used if people do not have faith in their accuracy. Not only is there no agreed-upon *level* of acceptable accuracy, there is no agreed-upon *method* by which to measure the accuracy of these CAA systems.

This paper provides details about EMMA using a framework we developed. We believe that the state of knowledge about CAA would be improved if researchers were able to share each others' experience in a meaningful way. It is difficult to compare research efforts and existing systems because there is no uniform

procedure for reporting CAA results. Our framework attempts to fill that gap by providing a coherent, compact, and comprehensive outline for reporting on and evaluating automatic assessment tools.

Before introducing our framework, we give an overview of how LSA and EMMA work, an understanding of which is necessary to appreciate the need for the framework. After explaining the framework, we describe the types of questions that EMMA can assess and the type of training data we used. Next, we evaluate its results as compared to human markers, and conclude with a plan for further research

1.1. Contribution

A major contribution of this paper is a description of our CAA. We summarise LSA for readers new to the field. We try to take away some of the mystery surrounding LSA by using our framework to comprehensively describe EMMA. We give examples of the types of questions EMMA can assess as well as the text used to train EMMA. We describe a study involving five human markers and use the results to evaluate the marks given by EMMA.

1.2. Organization of the paper

Section 2 introduces the reader to LSA and EMMA. Section 3 describes our framework for thoroughly describing a CAA. Section 4 provides information about the kinds of questions that EMMA assessed in addition to the types and amount of text we used to train it. Those readers most interested in evaluation might want to skip to Section 5, where we explain the method we used to evaluate the marks given by EMMA. and discuss the results of a study to determine inter-rater reliability and compare the human to human inter-rater reliability with LSA to human inter-rater reliability. We provide a plan for further research in Section 6 followed by concluding remarks in Section 7.

2 About Latent Semantic Analysis

2.1. Background

Researchers at Bellcore invented LSA, which is a statistical-based method for inferring meaning from a text. A seminal paper (Landauer et al., 1998) gives a more formal definition: “Latent Semantic Analysis is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text”. It was first used as an information retrieval technique (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990) in the late 1980s. By 1997, Landauer and Dumais (1997) asserted that LSA could serve as a model for the human acquisition of knowledge. They developed their theory after creating a mathematical information retrieval tool and observing unexpected results from its use. They claimed that LSA solves Plato’s problem, that is, how do people learn so much when presented with so little? The answer, oversimplified but essentially accurate, is the inductive process: LSA “induces global knowledge indirectly from local co-occurrence data in a large body of representative text” (1997).

2.2. How it works

To use LSA, researchers amass a suitable corpus of text. They create a term-by-document matrix where the columns represent documents and the rows represent terms (Deerwester, et al., 1990). A term is a subdivision of a document; it can be a word, phrase, or some other unit. A document can be a sentence, a paragraph, a textbook, or some other unit. In other words, documents contain terms. The elements of the matrix are weighted word counts of how many times each term appears in each document. More formally, each element, a_{ij} in an i x j matrix is the weighted count of term *i* in document *j*.

LSA decomposes the matrix into three matrices using Singular Value Decomposition (SVD), a well-known technique (Miller, 2003) that is the general case of factor analysis. Deerwester et. al., (1990) describe the process as follows.

Let t = the number of terms, or rows
d = the number of documents, or columns
X = a t by d matrix

Then, after applying SVD, X = TSD, where

m = the number of dimensions, m <= min(t,d)
T = a t by m matrix
S = an m by m diagonal matrix, i.e., only diagonal entries have non-zero values
D = an m by d matrix

LSA reduces S, the diagonal matrix created by SVD, to an appropriate number of dimensions k, where k << m, resulting in S'. The product of TS'D is the least-squares best fit to X, the original matrix (Deerwester, et al., 1990).

The literature often describes LSA as analyzing co-occurring terms. Landauer and Dumais (1997) argue it does more and explain that the new matrix captures the "latent transitivity relations" among the terms. Terms not appearing in an original document are represented in the new matrix as if they actually were in the original document (Landauer & Dumais, 1997). LSA's ability to induce transitive meanings is considered especially important given that fewer than 20% of paired individuals will use the same term to refer to the same common concept (Furnas, Gomez, Landauer, & Dumais, 1982).

LSA exploits what can be named the transitive property of semantic relationships: If A→B and B→C, then A→C (where → stands for is semantically related to). However, the similarity to the transitive property of equality is not perfect. Two words widely separated in the transitivity chain can have a weaker relationship than closer words. For example, LSA might find that copy → duplicate → double → twin → sibling. Copy and duplicate are much closer semantically than copy and sibling.

Finding the correct number of dimensions is critical; if it is too small, the structure of the data is not captured. Conversely, if it is too large, sampling error and unimportant details remain, e.g., grammatical variants (Deerwester, et al., 1990; Miller, 2003; Wade-Stein & Kintsch, 2003). Empirical work shows the correct number of dimensions to be about 300 (Kintsch & Bowles, 2002; Landauer & Dumais, 1997; Wade-Stein & Kintsch, 2003).

Creating the matrices from a huge corpus of training data using SVD and reducing the number of dimensions, often referred to as training the system, requires a lot a computing power; it can take hours or days to complete the processing (Miller, 2003). Fortunately, once the training is complete, it takes just seconds for LSA to evaluate a text sample (Miller, 2003). The need for lots of memory and lots of computing power is a clear indication of the value of the grid for a production, as opposed to a research, LSA CAA.

2.3. EMMA: our LSA-based marking system

EMMA (ExaM Marking Assistant) is an LSA-based marking system we are developing to mark short answers to questions in the domain of computer science. To mark a student answer, EMMA chooses the five answers in the training data (using the matrix, D, modified by S') that are closest (using the cosine similarity measure) to the answer being marked. EMMA assigns the weighted average of these human-assigned marks to the answer being marked.

EMMA requires the use of a database containing several types of information: basic course information (e.g. number of questions, question text), general training data in the domain being tested (e.g. course textbook) and previously marked answers. Any LSA-based system requires a server with a huge amount of RAM and a fast processor – we have been developing EMMA using a dual processor AMD Opteron with 16G of RAM. With this powerful computer, it takes between 2 and 20 minutes to perform the LSA matrix manipulations, depending on the number of dimensions.

3 The Framework

3.1. Background and usefulness

Our two-part framework for comparing CAA systems is based on a research taxonomy (Haley et al., 2005) we developed to compare LSA based educational applications. It was the result of an in-depth, systematic review of the literature concerning LSA research in the domain of educational applications. The taxonomy was designed to present and summarise the key points from a representative sample of the literature.

The taxonomy highlighted the fact that others were having difficulty matching the results reported by the original LSA researchers (Landauer & Dumais, 1997). We found a lot of ambiguity in various critical implementation details (e.g. weighting function used) as well as unreported details. We speculated that the

conflicting or unavailable information explains at least some of the inability to match the success of the original researchers.

The framework can be of value to both producers and consumers of CAA. Producers are researchers and developers who design and build assessment systems. They can benefit from the framework because it provides a relatively compact yet complete description of relevant information about the system. If producers of CAAs use the framework, they can contribute to the improvement of CAA state-of-the-art by adding to a collection of comparable data.

Consumers are organisations, such as universities, that wish to use a CAA system. CAA consumers are, or should be, particularly interested in two areas. The most important area is the accuracy of the results. But what does accuracy mean and how do we measure it? We believe that a CAA system is *good enough* if its marks correlate to human markers as well as human markers correlate with each other.

3.2. Details of the framework

The first part of the framework, which is for describing a CAA, can be visualised as the top half of the jigsaw puzzle in Figure 1. The bottom half of Figure 1 shows the second part of the framework - the evaluation of the CAA. We contend that all the pieces of this puzzle must be present if a reviewer wants to see the whole picture.

The important categories of information for specifying a CAA are the items assessed, the training data, and the algorithm-specific technical details. The general type of question (e.g., essay, multiple choice) is crucial for indicating the power of a system. The granularity of the marking scale provides important information about the accuracy – it is easier to mark a 3 point question than one worth 100 points. The number of items assessed provides some idea of the generalise-ability and validity of the results. Both the number of unique questions and the number of examples of each question contribute to the understanding of the value of the results. The second category comprises the technical details of the algorithm used. Haley, et al. (2005) discuss why these options are of interest to producers of an LSA-based CAA. The central piece of Figure 1 shows LSA-specific options, but these could be changed if the CAA is based on a different method. The corpus used to train the CAA is another crucial category. Both the type and amount of text help to indicate the amount of human effort needed to gather this essential element of CAAs. Some systems (LSA for one (D. Haley et al., 2007)) need two types of training data – general text about the topic being marked and specific previously marked answers for calibration. Information about both these types of training data should be included.

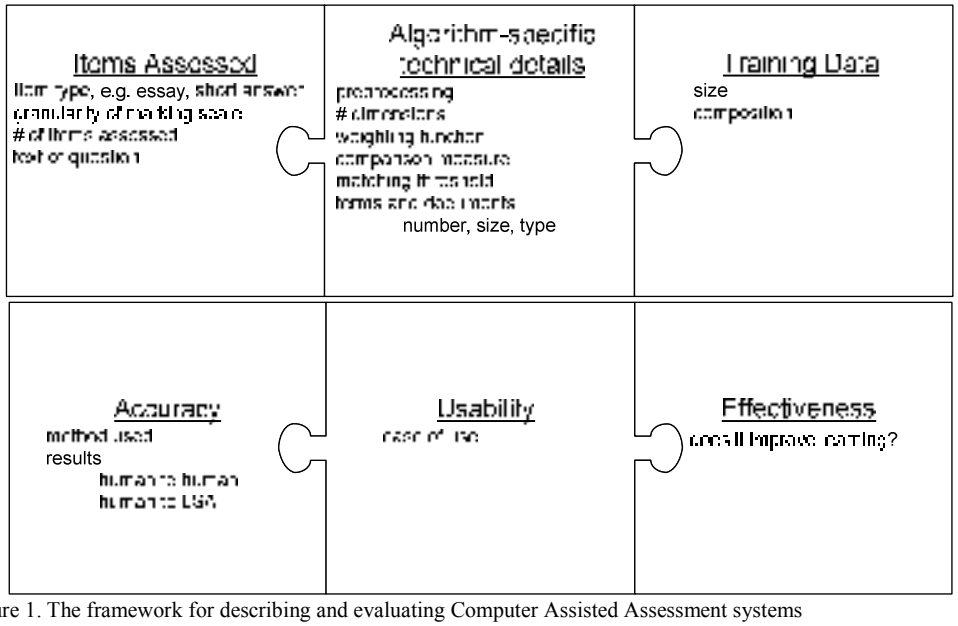

Figure 1. The framework for describing and evaluating Computer Assisted Assessment systems

Anyone interested in developing or using a CAA will be interested in its evaluation. The accuracy of the marks is of primary importance. A CAA exhibiting poor agreement with human markers is of little value. Our previous work (Haley et al., 2005) showed that different researchers report their results using different

methods. Ideally, all researchers would use the same method for easily comparable results. If researchers fail to reach a consensus, they should at least clearly specify how they determined the accuracy of their results. The other two pieces of the evaluation picture are usability and effectiveness. These pieces are of interest to consumers wanting to choose among deployed systems.

3.3. Using the framework for an LSA-based CAA

Table 1 is an example of how the framework could be used to compare different research results. It starts with an overview and proceeds with the pieces in the puzzles of Figures 1 and 2. The particular study described attempted to quantify the optimum amount of training data needed (D. Haley et al., 2007). The experience of creating and using it served to crystallize our thinking about the important elements of reporting on a CAA.

		OverView						Items Assessed				Algorithm-specific Technical Details											
																	Terms			Documents			
System Name	Refer-ence	Who / Where	What / Why	Stage of Development/ Type of work	Innovation	Major Result / Key points	Human Effort	Type of Item	Granularity of Marking Scale	# of Items assessed	text of question	preprocessing	# dimensions	weighting function	comparison measure	matching threshold	Number	Size	Type	Number	Size	Type
EMMA	HTD 07	Thomas, Haley, De Roeck, Petre; The Open University	assess computer science short answers for summative assessment	research prototype	demonstrated the use of a research taxonomy for CAA, marked questions about html	amount of training data that works best: 50 marked answers for question A	gather training data, gather marked answers	short answers about html	4 points	50	Correct the following fragments of HTML. For each case, write the correct HTML and write one or two sentences about the problem with the original HTML. HTML: <I>It is <B>very</I> </B> important to read this text carefully. The desired appearance: It is very important to read this text carefully.	stemming, stop words	90	log / entropy	cosine	none	12k	1 word	text	45k	1 paragraph	text
						amount of training data that works best: 80 marked answers for B			4 points	50	Correct the following fragments of HTML. For each case, write the correct HTML and write one or two sentences about the problem with the original HTML. HTML: Things to do: Pack suitcase, Book taxi. The desired appearance: Things to do: Pack suitcase, Book taxi.		500	log / entropy	cosine	none						

Table 1 Part 1. Filling in the framework

	Training Data		Evaluation				
			Accuracy			Effectiveness	Usability
Refer-ence	Size	Composition	method used	Human to LSA	Human to Human	Does it improve learning?	How easy is it to use?
HTD07	1) 45k paragraphs 2) 50	1) course texts 2) human marked answers	compared LSA marks with 5 human markers and calculated average	average % identical 53 off by 1 34 off by 2 12 off by 3 1 off by 4 1	54 32 11 1 1	not relevant - a research prototype	not relevant - a research prototype
HTD07	1) 45k paragraphs 2) 80	1) course texts 2) human marked answers	compared LSA marks with 5 human markers and calculated average	average % identical 43 off by 1 45 off by 2 6 off by 3 3 off by 4 3	61 28 9 1 1	not relevant - a research prototype	not relevant - a research prototype

Table 1 Part 2. Filling in the framework

Our previous work (Haley et al., 2005) highlighted the insights revealed by the framework. The major result of using the framework is the conclusion that researchers need to know all of the details to fully evaluate and compare reported results. Research results cannot be reproduced and validated if researchers do not provide more detailed data regarding their LSA implementations.

4 Questions to be Marked and Training Data

4.1. Questions to be marked

Table 2 shows the text of the 18 questions that EMMA has marked. There are several types of questions; however, they are all in the domain of computer science. The questions were from the first 2 homework assignments from an introductory computer literacy course. Some of the questions (e.g. 13, 14, 16) require quite concise, short, straight-forward answers while others (e.g. 4, 20) require longer, more open-ended answers. Some (e.g. 1 and 2) are multi-part and worth many more points than others that are worth just 2 points. Five questions (8-12) are about html. Thus, there is a fair variety of question types, although the main point is that they are all short answer, rather than multiple choice or true/false type questions.

4.2. Training data

We used two types of training data for EMMA – general and specific. The general training data was 45,000 paragraphs from course text books for both this course and other computer science courses. The specific data was a set of 1,000 answers for each of the 18 questions previously marked by human markers. We performed tests to determine what number of previously marked answers provided the best agreement with human markers and found different amounts worked best for different questions (D. Haley et al., 2007). We plan to investigate whether the quality of the training data is more important than the quantity.

Text of Questions

Q 1	Name two elements of the course materials that will be distributed to you via the M150 course website?
	What is the role of the Study Calendar?
	What is the cut-off date for TMA 02?
	Find the learning outcomes for M150 which are listed in both the Course Companion and the Course Guide – these tell you what you should be able to do after studying the course. Write down the learning outcome that you feel you are most interested in achieving, and one or two sentences to describe why you have chosen that learning outcome.
	What does eTMA stand for?
	What is the name of the document you should read in order to prepare yourself for submitting an eTMA?
	Who should you contact with queries about course software?
Q 2	Find the UK AltaVista site. What is its URI?
	What is the name of the large aquarium in Hull?
	Which query led you to the answer?
	What is the URI of the site?
	What is the minimum number of intervening web pages you have to visit between the main site and the page that contains the information on the ballan wrasse?
	List the URI of each intervening web page.
	How big can a ballan wrasse grow?
	Does the ballan wrasse page tell you anything about the age a ballan wrasse can reach?
	What age can a ballan wrasse reach?
	What is the URI of the web page where you found the information?
	What is the role of the Study Calendar?
	What is the cut-off date for TMA 02?
	Which search engine, and which query got you to the page that contained your answer?
Q 3	Explain, with examples, the difference between an analogue and a discrete quantity.
Q 4	Give an example of a computer standard, explaining its purpose. Why is there a general need for standards in computing?
8-12	For each case; write the correct HTML and write one or two sentences about the problem with the original HTML. (The first line is the original HTML. The second line is the desired appearance.)
Q 8	<B>Always look left and right before crossing the road.
	Always look left and right before crossing the road.
Q 9	<B>Important!<B>Do <B> not place metal items in the microwave.
	Important! Do **not** place metal items in the microwave.
Q 10	<I>It is <B>very</I> </B> important to read this text carefully.
	It is ***very*** important to read this text carefully.
Q 11	Things to do:
	Pack suitcase, </BR>
	Book taxi.
	Things to do:
	Pack suitcase,
	Book taxi.
Q 12	More information can be found here.
	More information can be found here.
13-16	Victoria uses her computer to write up a report. When the report is complete, she saves it to the hard disk on her computer. Later she revises her report and saves the final version with the same document name.
Q 13	Considering the contents of the report as data, at what point does the data become persistent?
Q 14	What happens to the first saved version of the document?
Q 15	Suggest an improvement in Victoria's work practice, giving a reason for your answer.
Q 16	Give two examples of persistent storage media other than the hard disk.
Q 17	Victoria then wishes to email a copy of her report, which includes data on identifiable individuals, to John, a work colleague at her company's Birmingham office. Write two sentences to explain the circumstances under which, within UK law, she may send the unedited report to John.
Q 18	Explain briefly the property of internet email that allows the contents of the report to be sent as an attachment rather than as text in the body of the email message.
Q 19	John's email address is John@Birmingham.office.xy.uk Which parts of the address are: the user name, the name of the domain, the top-level domain?
20-21	Victoria then prepares her report for publication on a website, so that people can read her report using a browser.
Q 20	In no more than 100 words, explain what she has to take into account when making her report public.
Q 21	Which of the following should she publish on the website with her report and why? Company address, personal telephone number, email address

Table 2. Text of questions

5 Evaluation of EMMA

Our framework shows three areas of importance for evaluating a CAA – accuracy, usability, and effectiveness. We will discuss accuracy in this section. Usability (how easy the system is to use) and effectiveness (how well the system improves learning) have not been addressed.

In order for an automatic marking tool to be accepted, it must produce marks at an acceptable level of accuracy. Automatic marks must correlate to human marks as well as human marks correlate with other human marks. If humans agreed with each other all the time, EMMA would have to show perfect agreement with human markers. It is widely accepted, however, that humans do not agree with each other all the time, thus, EMMA needs to be *good enough*, not perfect. EMMA is good enough if it matches or exceeds the agreement of humans.

We evaluated EMMA by a two-step process. First, we conducted a study to determine how well humans agree with each other. Then, we compared the average human agreement with the average LSA to human agreement. The following subsections give the details.

5.1. Human to Human agreement

We conducted a large scale study to quantify how well human marks agree with each other. We recruited five expert markers and asked them to mark the same random set of 60 student answers to 18 questions. Two of the five markers had not completed the last few questions at the time of the analysis so only 13 questions are included in this evaluation. The marks for the first 10 answers of each question were discarded to guard against problems relating to markers judging the first answers differently than later answers. Table 3 shows the results for question 8. Each of the 5 markers was paired with the other 4 for a total of 10 comparisons (markers 1 and 2, 1 and 3, and so on). None of these pairs agreed perfectly with each other. Table 4 summarises the results of Table 3. For the marks that agreed perfectly, the worst rater pair agreed 74%. For the best rater pair, 94% of the marks agreed. The average of all 10 rater pairs was 83%.

A similar analysis was done for all of the questions. The tables are not included here; the results are summarised in later graphs.

Question 8					
Comparison of Human Marks					
	% of marks off by:				
Marker Pair	0	1 point	2 points	3 points	4 points
1-2	74	22	4	0	0
1-3	80	16	4	0	0
1-4	80	14	4	2	0
1-5	76	20	4	0	0
2-3	94	6	0	0	0
2-4	82	12	4	2	0
2-5	94	6	0	0	0
3-4	84	10	4	2	0
3-5	92	8	0	0	0
4-5	78	16	6	0	0

Table 3. Comparison of human marks

Question 8			
Summary of Comparison of Human Marks			
	Stats for 10 marker pairs		
	minimum	maximum	average
% of marks			
off by 0	74	94	83.4
off by 1	6	22	13
off by 2	0	6	3
off by 3	0	2	0.6
off by 4	0	0	0

Table 4. Summary Comparison of human marks

5.2. LSA to Human agreement

The marks given by the human markers were compared to the marks given by EMMA for the same 50 answers to the 13 questions. Tables 5 and 6 give the figures for comparing LSA and humans for question 8.

Question 8					
Comparison of Human and LSA Marks					
	% of marks off by:				
Marker Pair	0	1 point	2 points	3 points	4 points
LSA - 1	70	16	6	2	6
LSA - 2	82	8	2	0	8
LSA - 3	80	10	2	0	8
LSA - 4	70	14	6	0	10
LSA - 5	86	2	4	0	8

Table 5. Comparison of human and LSA marks

Question 8			
Summary of Comparison of Human and LSA Marks			
	Stats for 5 markers and LSA		
	minimum	maximum	average
% of marks			
off by 0	70	86	77.6
off by 1	2	16	10
off by 2	2	6	4
off by 3	0	2	0.4
off by 4	6	10	8

Table 6. Summary of Comparison of Human and LSA marks

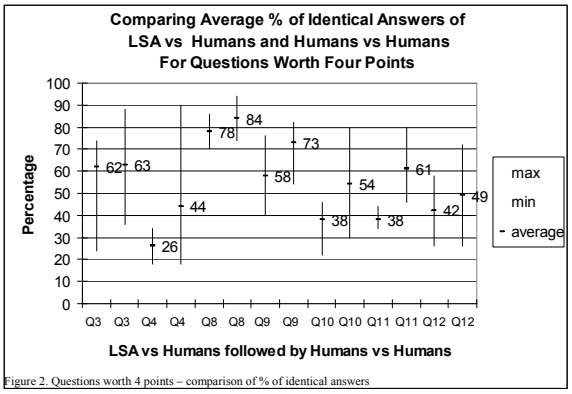

Figure 2. Questions worth 4 points – comparison of % of identical answers

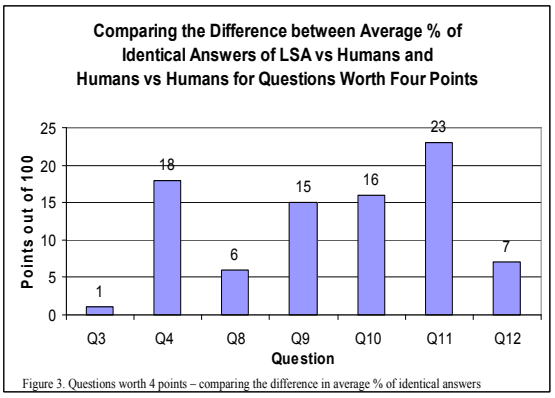

Figure 3. Questions worth 4 points – comparing the difference in average % of identical answers

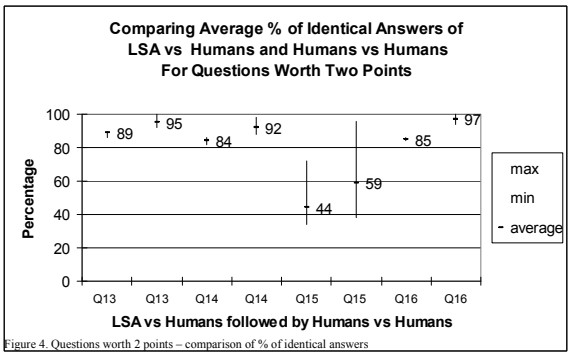

Figure 4. Questions worth 2 points – comparison of % of identical answers

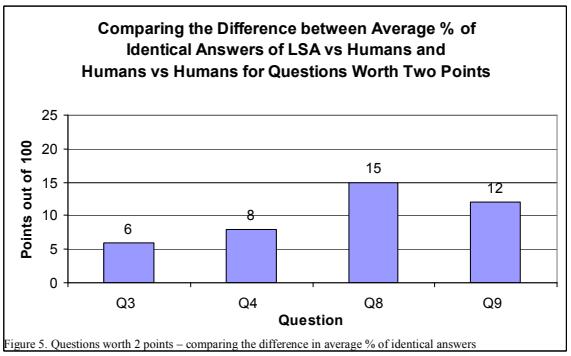

Figure 5. Questions worth 2 points – comparing the difference in average % of identical answers

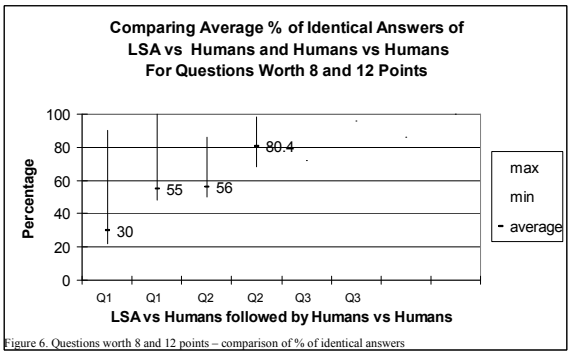

Figure 6. Questions worth 8 and 12 points – comparison of % of identical answers

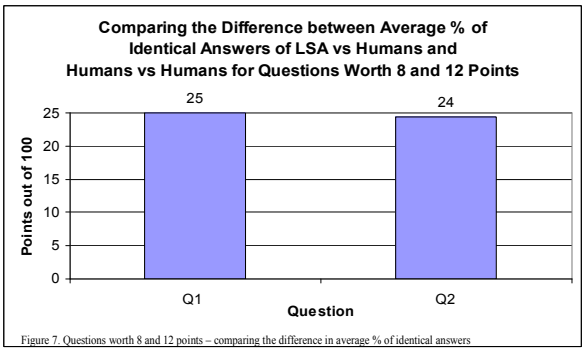

Figure 7. Questions worth 8 and 12 points – comparing the difference in average % of identical answers

5.3. Putting it all together

The previous subsections gave information for question 8. They showed how the 10 rater-pairs agreed with each other and how LSA agreed with each of the 5 raters. The tables for the other questions are not listed here but are summarised in Figures 2 - 7. The questions are grouped according to the total number of points available for the question. Questions 3, 4, 8, 9, 10, 11, and 12 are all worth 4 points and are shown in Figures 2 and 3. Questions 13, 14, 15, and 16 are all worth 6 points; they are shown in Figures 4 and 5. Question 1 is worth 8 points and question 12 is worth 12 points. They are shown in Figures 6 and 7.

Figures 2, 4, and 6 are meant to show the range of agreement for LSA to humans and human to humans. The graphs are ordered by question, and within a question, first appears the data for the average LSA to human agreement followed by the average of the human to human agreement. For each vertical line on the graphs, the labelled point in the middle is the average % of identical answers. The top of the line indicates the best agreement; the bottom of the line indicates the worst agreement. So for question 3, LSA agreed with humans, on average, on 62% of the marks given. Humans agreed with other humans, on average, on 63% of the marks given. In general, one can see that human agreement with other humans was better than LSA

agreement with humans for every question. Another general observation is that the level of agreement differed for every question.

Figures 3, 5, and 7 allow one to see how well they agreed. The bars show the number of points that differed between human to human and LSA to human. For question 3, EMMA performed very well. One can see on Figure 3 that LSA to human agreement was only 1 point below human to human agreement for question 3. EMMA did not do well for other questions. For example, the worst result was for question 1, where the LSA to human marks were 25 points lower than they were for human to human. By comparing the 3 figures, one can see that, in general, EMMA did best for questions worth fewer points and did worse for questions worth more points.

6 A roadmap for future research

Our results show that EMMA, in its current state of implementation, is not as good as human markers and is therefore unacceptable for use as a CAA. However, our research is by no means complete. We plan several further experiments to improve the results. We believe that increasing the amount of our general training data will make the largest improvement and this will be our first future effort. The subsections below discuss different improvements we plan to implement.

6.1. The corpus

LSA results depend on both corpus size and corpus content.

6.1.1. Corpus size

Existing LSA research stresses the need for a large corpus. For example, Summary Street, an LSA-based instructional software system, uses a corpus of 11 million words (Wade-Stein & Kintsch, 2003). We have used a small corpus of just 50,000 documents. We need to identify and acquire a larger corpus for future studies.

6.1.2. Corpus content

An earlier paper (Wiemer-Hastings, Wiemer-Hastings, & Graesser, 1999) reports that size is not the only important characteristic of the corpus. Not surprisingly, the composition of the corpus affects the results of essay grading by LSA. They claim that two types of training data are necessary: general documents in the form of textbooks and other domain-specific text and specific documents comprising previously human-marked essays. They found the best composition to be about 40% general documents and 60% specific documents.

An ideal corpus would provide specific documents that give a spread of marks across the mark range and a variety of answers for each mark. While we believe that we have such a specific corpus, we need a much larger general corpus as mentioned in the previous subsection.

6.2. Weighting function

Local weighting is the most basic form of term weighting. Local weighting is defined as tf_{ij} (the number of times term i is found in document j) dampened by the log function: local weighting = $1 + \log (tf_{ij})$. This dampening reflects the fact that a term that appears in a document *x* times more frequently than another term is not *x* times more important.

Most LSA systems use a combination of local and global weighting. Global weighting is defined as 1 – the entropy or noise. Global weighting attempts to quantify the assumption that a term appearing in many documents is less important than a term appearing in fewer documents.

6.2.1. Log-entropy

One of the original investigators (Dumais, 1991) recommended using log-entropy weighting, which is local weighting times global weighting. The log-entropy term weight for term i in doc j =

$$\log(1+tf_{ij})*\left[1-\frac{\sum \frac{tf_{ij}}{gf_i}*\log\frac{tf_{ij}}{gf_i}}{\log(numdocs)}\right]$$

where

tf_{ij} – term frequency – the frequency of term i in document j

gf_i – global frequency – the total number of times term i occurs in the whole collection

6.2.2. tfidf

More recently, a researcher (Sebastiani, 2002) claims the most common weighting is tfidf, or term frequency inverse document frequency.

$$tfidf(t_k,d_j) = \#(t_k,d_j)*\log\frac{|Tr|}{\#Tr(t_k)}$$

where #(tk, dj) denotes the number of times tk occurs in dj
#Tr(tk) denotes the document frequency of term tk, that is, the number of documents in Tr in which tk occurs.

6.2.3. Our plans for term weighting

Dumais recommended the use of log-entropy weighting based on her results from the field of information retrieval. Sebastiani was reporting on text categorization, of which essay assessment can be seen as a sub-part. We think the choice of weighting function should be based on a comprehensive analysis based on assessment. To our knowledge, no such study has been reported. We plan to fill this gap by using tfidf to repeat the experiments we have already carried out using log-entropy.

6.3. Corpus pre-processing

Removing stop words and stemming are two types of pre-processing we have used. Stop words (e.g. a, the, my) are considered non-meaningful. Stemming involves conflating word forms to a common string, e.g., write, writing, writes, written, writer would be represented in the corpus as *writ*.

We plan one more form of pre-processing that has not yet been studied, to our knowledge. We want to use compound nouns as LSA terms. Currently, only single nouns are used. We conjecture that, in the domain of computer science, such terms as floppy disk and hard drive are ubiquitous and could make a real difference in our results.

6.4. Dimension reduction

Choosing the appropriate dimension, k, for reducing the matrices in LSA is a well known open issue. The current consensus is that k should be about 300. No theory yet exists to suggest the appropriate value for k. Currently, researchers determine k by empirically testing various values of k and selecting the best one. The only heuristic says that k << min(terms, documents).

Our previous work (D. T. Haley, P. Thomas, A. De Roeck, & M. Petre, 2007) showed that different dimensions worked best for different questions. We want to repeat this work with a much larger corpus to ascertain whether the same results hold.

7 Conclusion

This paper reports on the status of the work on EMMA, a Latent Semantic Analysis (LSA) based Computer Assisted Assessment (CAA) system developed with partial support from ELeGI. Although ELeGI is almost over, our research is not. We believe that we can improve our results by following the roadmap laid out in Section 6. Although we did not achieve our goal of creating a *good enough* CAA in the course of the ELeGI project, we believe we are well on the way. We have learned much about LSA, identified gaps in the knowledge, recommended a framework to the LSA community for uniform, comprehensive reporting of research results, built a prototype CAA system, and established a method of evaluating our system.

Acknowledgements

The work reported in this study was partially supported by the European Community under the Innovation Society Technologies (IST) programme of the 6th Framework Programme for RTD - project ELeGI, contract IST-002205. This document does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of data appearing therein.

References

Berglund, A. (1999). ***Changing Study Habits - a Study of the Effects of Non-traditional Assessment Methods. Work-in-Progress Report.*** **Paper presented at the 6th Improving Student Learning Symposium, Brighton, UK.**

Burstein, J., Chodorow, M., & Leacock, C. (2003). Criterion Online Essay Evaluation: An Application for Automated Evaluation of Student Essays. In ***Proc. of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence*****. Acapulco, Mexico.**

Daniels, M., Berglund, A., Pears, A., & Fincher, S. (2004). ***Five Myths of Assessment.*** **Paper presented at the 6th Australasian Computing Education Conference (ACE2004), Dunedin, New Zealand.**

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. ***Journal of the American Society for Information Science, 41*****(6), 391-407.**

Dumais, S. T. (1991). Improving the retrieval of information from external sources. ***Behavioral Research Methods, Instruments & Computers, 23*****(2), 229-236.**

Furnas, G. W., Gomez, L. M., Landauer, T. K., & Dumais, S. T. (1982). Statistical semantics: How can a computer use what people name things to guess what things people mean when they name things? In ***Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*** **(pp. 251-253): ACM.**

Haley, D., Thomas, P., De Roeck, A., & Petre, M. (2007, 31 January 2007). ***Measuring Improvement in Latent Semantic Analysis-Based Marking Systems: Using a Computer to Mark Questions about HTML.*** **Paper presented at the Proceedings of the Ninth Australasian Computing Education Conference (ACE2007), Ballarat, Victoria, Australia.**

Haley, D. T., Thomas, P., De Roeck, A., & Petre, M. (2005, 21-23 September 2005). ***A Research Taxonomy for Latent Semantic Analysis-Based Educational Applications.*** **Paper presented at the International Conference on Recent Advances in Natural Language Processing'05, Borovets, Bulgaria.**

Haley, D. T., Thomas, P., De Roeck, A., & Petre, M. (2007). ***Tuning an LSA-based Assessment System for Short Answers in the Domain of Computer Science: The Elusive Optimum Dimension.*** **Paper presented at the 1st European Workshop on Latent Semantic Analysis in Technology Enhanced Learning, Heerlen, The Netherlands.**

Kintsch, W., & Bowles, A. R. (2002). Metaphor comprehension: What makes a metaphor difficult to understand? ***Metaphor and Symbol, 17*****, 249-262.**

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. ***Psychological Review, 104*****(2), 211-240.**

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to Latent Semantic Analysis. ***Discourse Processes, 25*****, 259-284.**

Miller, T. (2003). Essay assessment with Latent Semantic Analysis. ***Journal of Educational Computing Research, 28*****.**

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. ***ACM Computing Surveys, 34*****(1), 1-47.**

Thomas, P., Haley, D., De Roeck, A., & Petre, M. (2004). E-Assessment using Latent Semantic Analysis in the Computer Science Domain: A Pilot Study. In ***Proc. of the eLearning for Computational Linguistics and Computational Linguistics for eLearning Workshop at COLING 2004.*** **(pp. 38-44). Geneva.**

Thomas, P. G., Waugh, K., & Smith, N. (2005). ***Experiments in the Automatic Marking of ER-Diagrams.*** **Paper presented at the Proceedings of the 10th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education, Monte de Caparica, Portugal.**

Wade-Stein, D., & Kintsch, E. (2003). ***Summary Street: Interactive computer support for writing*****. University of Colorado, USA.**

Wiemer-Hastings, P., Graesser, A., & Harter, D. (1998). The foundations and architecture of Autotutor. In ***Proceedings of the 4th International Conference on Intelligent Tutoring Systems*** **(pp. 334-343). San Antonio, Texas.**

Wiemer-Hastings, P., Wiemer-Hastings, K., & Graesser, A. C. (1999). Improving an intelligent tutor's comprehension of students with Latent Semantic Analysis. In S. P. Lajoie & M. Vivet (Eds.), ***Artificial Intelligence in Education*****. Amsterdam: IOS Press.**