Measuring Improvement in Latent Semantic Analysis-Based Marking Systems: Using a Computer to Mark Questions about HTML

Debra T. Haley, Pete Thomas, Anne De Roeck, Marian Petre

Centre for Research in Computing, Department of Computing
The Open University
Walton Hall, Milton Keynes MK7 6AA UK

[D.Haley, P.G.Thomas, A.DeRoeck, M.Petre] [at] open [dot] ac [dot] uk

Abstract

This paper proposes two unconventional metrics as an important tool for assessment research: the Manhattan (L1) and the Euclidean (L2) distance measures. We used them to evaluate the results of a Latent Semantic Analysis (LSA) system to assess short answers to two questions about HTML in an introductory computer science class. This is the only study, as far as we know, that addresses the question of how well an LSA-based system can evaluate answers in the very specific and technical language of HTML. We found that, although there are several ways to measure automatic assessment results in the literature, they were not useful for our purposes. We want to compare the marks given by LSA to marks awarded by a human tutor. We demonstrate how L1 and L2 quantify the results of varying the amount of training data necessary to enable LSA to mark the answers to two HTML questions. Although this paper describes the use of the metrics in one particular case, it has more general applicability. Much fine-tuning of an LSA marking system is required for good results. A researcher needs an easy way to evaluate the results of various modifications to the system. The Manhattan and the Euclidean distance measures provide this functionality.

Keywords: Latent Semantic Analysis, assessment, metrics, measures

1 Introduction

Marking essays by computer has been researched since the 1960s (Page, 1967). Educational institutions hope to save time, and therefore, money by using computerised marking systems. In addition to the possible cost savings, the computer offers some advantages over the human. Human markers may mark differently as they become fatigued in addition to being affected by the order of marking. For example, if a brilliant answer is the first read by a marker, it could cause the marker to be harsher for the remainder of the answers. Even the most scrupulous people might show bias based on personal feelings towards a student. While they may successfully avoid awarding better marks to their favourite students, they may mark unfavoured students more highly than they deserve in an attempt to be unbiased. Automatic markers can be an improvement over human markers because their results are reliable and repeatable. They do not get tired, they do not show bias based on personal feelings towards students, their results will be the same without regard to the order in which the answers are presented, and they are able to return results much faster than humans.

This paper reports on research taking place to improve an automatic, Latent Semantic Analysis (LSA) based marking system. We are using LSA to examine the feasibility of marking short answers about HTML. The HTML questions we are marking are two-part. Part one asks the learner to correct a given fragment of incorrect HTML. Part two asks for a few sentences explaining the error(s) in the original HTML. To our knowledge, this has never been done before. Although LSA researchers have marked essays and short answers, we could find nothing in the literature describing efforts to mark short answers in the very specific and technical language of HTML.

LSA is a theory and method invented in the late 1980s (Deerwester et al., 1990) for information retrieval (IR) and was first known as Latent Semantic Indexing (Furnas et al., 1988). The IR community still refers to LSI whereas the researchers using the technique for other applications refer to it as LSA (Landauer and Dumais, 1997, Foltz et al., 1998, Miller, 2003). Automatic marking of essays and short answers is one application of LSA with great appeal for educational institutions.

Researchers using LSA-based grading systems have done a great deal of work to replicate the results of the early researchers (Landauer et al., 1998, Landauer and Dumais, 1997) with varying degrees of success (Nakov et al., 2003, Foltz et al., 1999). There are two factors that contribute to the fact that many researchers have not been able to match the results (Haley et al., 2005). One factor is that developers must make many choices that affect the results of their assessment systems. The second factor is that there is no standard way of reporting the choices made or the results of these choices (Haley et al., 2005). Thus, critical features that improve LSA-based marking systems remain unpublished and unknown to the research community. Adding to the problem, researchers have difficulty comparing various systems and modifications to the basic LSA algorithm (Wiemer-Hastings, 2000, Kanejiya et al., 2003, Foltz et al., 1999, Lemaire and Dessus, 2001).

One of the under-researched areas in LSA-based marking systems is the amount of training data needed for good results. Training data comprises both general and specific documents (Wiemer-Hastings et al., 1999). General training data are in the form of course textbooks; specific training data are human-marked answers. This paper addresses the question of the optimum amount of specific training data as well as a related question: How do you evaluate the results of using various amounts of training data? These questions are explored in the two major strands of the paper: a description of an experiment to ascertain the required amount of training data for an LSA-based automatic marking system and suitable metrics for judging the results. The experiment used EMMA (ExaM Marking Assistant), an automatic, Latent Semantic Analysis based system we are developing to mark short answers to exam questions in the domain of Computer Science.

Section 2 addresses the first strand by summarising how LSA works and explaining why it is important to know the optimum amount of training data. Section 3 deals with the second strand; it discusses existing ways to measure the success of LSA-based assessment systems and motivates the need for new metrics.

Section 4 discusses several standard statistical tests that could be used to measure the success of automatic assessment systems and argues that none of them is suitable for our purposes. Section 5 describes the experiment to determine an appropriate amount of training data.

Sub-section 6.1 provides one of the main contributions of the paper – a suggestion of two possible metrics for comparing automatic assessment systems. These metrics are the Manhattan distance (L1) and the Euclidean distance (L2). Sub-section 6.2 presents another contribution of the paper. It uses the two metrics to draw some conclusions about the amount of training data needed for an automatic marking system. Section 7 summarises the conclusions and lists further work.

2 Using EMMA to mark exams

This study used EMMA, our LSA-based automatic marking system, to determine the effectiveness of automatically assessing answers to questions about HTML and the optimum amount of training data. We have exactly 1,000 tutor-marked answers for two questions and wished to evaluate EMMA by comparing the tutor marks with marks assigned by our system.

A very brief summary of LSA follows. A more thorough explanation can be found in (Landauer et al., 1998). LSA transforms training data into a term-frequency matrix M, where m_{ij} is the weighted number of times term i appears in document j. It then decomposes M by Singular Value Decomposition into three matrices such that M = TSD. The M matrix can be very large. For instance, we are using over 45,000 documents which yield over 11,000 terms. LSA reduces S to k dimensions (where k is of the order of 300) resulting in the matrix S'. The matrices T, S', and D can be multiplied, resulting in M', the least squares best fit of M, where M' = TS'D and M ≈ M'.

To mark a student answer, EMMA chooses the five answers in the training data that are closest (using the cosine similarity measure) to the answer being marked. EMMA assigns the weighted average of these tutor-assigned marks to the answer being marked. Although our current implementation uses five answers, we plan to experiment with varying the number of answers in the future and also with using a threshold to limit the number of similar answers to those above a certain cosine value.

The amount and type of training data are two of many areas that affect the results of an LSA-based marking system (Haley et al., 2005). The training data for EMMA comprises both general and specific documents. The general documents are the course textbooks, which were divided into over 45,000 documents, each a paragraph long. The specific training documents came from the 1,000 tutor-marked answers. We split these tutor-marked answers into two sets, a set to be marked of 333 answers and a training set. To ensure good coverage for training purposes, we allocated the remaining 667 answers to be available for the training set. The experiment used varying numbers of the 667 training data answers to find an adequate amount.

The number of required tutor-marked answers is an important criterion for evaluating the feasibility of using an LSA-based marking system. One reason to use an automatic marker is to avoid the necessity of human marking. It would be hard to justify on the basis of cost savings the use of an automatic marker that requires more effort to create the training data than it would to mark the answers manually. (This analysis ignores the other benefits of automatic markers given in the Introduction.) It takes less human effort and is thus more practical if *few* answers are required for *good results*. The rest of this paper attempts to quantify the meanings of these two terms.

3 The inadequacy of existing success measures for LSA marking systems

Many choices need to be made when implementing an LSA-based marking system. The LSA literature leaves many parameters unspecified including number of dimensions in the reduced matrix, amount and type of training data, types of pre-processing, and weighting functions (Haley et al., 2005). The choice of these parameters is an intrinsic aspect of building an LSA marking system. Therefore, researchers need an adequate way to measure and compare the results of the various selections, as we shall now explore.

3.1 A simple metric

It is tempting to determine the percentage of marks where the tutor and Emma gave identical marks and use this number as the success measure. However, this simple metric gives an incomplete picture of the results. Consider the hypothetical case illustrated in Table 1. It is debatable which is the better marking system – A or B. Although Marking System A has a higher percentage of identical answers than does Marking System B, Marking System B has 100% of its marks disagreeing with the human by at most one point while Marking System A has only 80% of its marks disagreeing by at most one point and 20% that differ by three or more points.

Marking System A		Marking System B	
% of Questions	Point Difference between Human and Computer Scores	% of Questions	Point Difference between Human and Computer Scores
80	0	75	0
0	1	25	1
0	2	0	2
5	3	0	3
15	4	0	4

Table 1. Hypothetical results for two marking systems that show that the simple metric of the percentage of identical scores for a four-point question hides important details

3.2 Widely used metrics

The literature offers two techniques to evaluate marking systems – precision and recall, and correlation.

3.2.1 Precision and recall

The first technique is the use of precision and recall; these measures are used widely in LSI and LSA research (Manning and Schütze, 1999, Dumais, 1991, Graesser et al., 2000, Nakov et al., 2003). Precision looks at how relevant the collection of retrieved documents is; it is the ratio of correctly retrieved, i.e. relevant, documents to all retrieved documents. Recall is a measurement of completeness. It is the ratio of correctly retrieved documents to all relevant documents i.e., those that were retrieved plus those that the retrieval system failed to retrieve (Foltz, 1990). As recall goes up, precision tends to go down; in the trivial case, a system achieves 100% recall if all the documents are retrieved, which would give the lowest precision. Information retrieval (IR) researchers plot values of precision for various levels of recall to provide a good picture of the effectiveness of their techniques (Dumais, 2003).

It is important to have a good metric to measure success when tuning a marking system. Dumais, in a widely cited study (Dumais, 1991), used precision and recall to justify the use of log-entropy weighting in the term-frequency matrix. The decision of a weighting function is a critical choice to be made by LSA researchers. Nakov et al. used precision and recall figures to argue that the choice of a weighting function is the most crucial of all tuning techniques (Nakov et al., 2003). Many researchers continue to justify the use of the log-entropy weighting factor (Foltz et al., 1998) by relying on the early work of Dumais (Dumais, 1991). Although log-entropy may well be the best weighting function, it should be justified for LSA-based assessment systems on research done *with* LSA-based assessment systems instead of IR systems. Researchers need to remember that Dumais is primarily interested in information retrieval rather than essay assessment. Although precision and recall are useful for evaluating IR techniques, we believe that using them to measure automatic marking systems is irrelevant. Recall is not important – it makes no difference how many documents are returned because the marking system looks at only a pre-determined number that are the closest matches to the document being marked. Precision, on the other hand, is very important – the documents judged by the marking system to be relevant must actually *be* relevant. Precision, however, is a binary measure. It assumes that the documents are relevant or not. EMMA uses the cosine similarity measure to rank the documents in terms of their similarity to the answer to be marked. It then awards a mark by calculating the weighted average (using the cosine measure) of the five most similar answers. This feature of LSA provides a finer-grained measure than precision and recall, which are better suited to information retrieval.

3.2.2 Correlation

The second technique to evaluate marking systems is statistical correlation, which is used by many researchers (Foltz et al., 2000, Wiemer-Hastings et al., 1999). The most widely known correlation measures are Pearson's r, Spearman's rho, and Kendall's tau_b (Dancey and Reidy, 2002). Correlation statistics indicate how well one variable can be used to predict another variable. If human-assigned marks and computer-assigned marks agree, the correlation would be perfect. Even if the human marks were always twice the computer marks, the correlation would once again be perfect. In this case, a good correlation would not mean a good marking system. Another problem with the correlation statistic is that it would be low in the case where computer marks are off by plus-or-minus one point. In this situation, the computer mark could not be used to predict the human mark even though we argue that the overall results of the marking system could be very good if all the contradictory marks differ by only one point in either direction. For these reasons, the standard correlation statistics may not be useful for evaluating automatic marking systems.

Section 4 looks at a traditional statistical test and explains why it fails to help us evaluate our automatic marking system.

4 Problems with the traditional t-test

Having considered and rejected the metrics described in Section 3, we turn to the field of statistics for possible success metrics.

The traditional t-test for comparing two groups is a candidate for evaluating automatic marking systems. It comes in parametric and non-parametric versions. The parametric tests are more powerful than the non-parametric versions but the data must meet three assumptions to use them: normally distributed populations, approximately equal variations of the populations, and no extreme scores.

The t-test compares the means of two groups. For marking systems, one group is the human-assigned scores and the other group is the computer-assigned scores. When all participants take place in both conditions (short answer marked by tutor and short answer marked by

EMMA), the study design is known as within-participants (also called repeated measures or related design) and the appropriate parametric statistical test for comparing the groups is the t-test (Dancey and Reidy, 2002). The output of the t-test includes the mean scores for each group, the difference between them, and the standard deviations. With these values, one can compute the effect size, which is the difference of the means divided by the mean of the standard deviations. Confidence intervals around the effect sizes are an additional tool for evaluating results (Aberson, 2002). Therefore, if the data meet the three assumptions for using parametric tests, employing effect sizes with confidence intervals is a good way to evaluate automatic marking systems.

It is highly unfortunate for us that we cannot use the t-test and effect sizes with confidence intervals because our data are not normally distributed. The marks given by tutors are highly negatively skewed because the marks tend to cluster towards the high end of the marking scale. If the marks are not normally distributed, the effect sizes and confidence intervals will be incorrect (Thompson, 2002) and the t-test is not applicable. We can, however, use the Wilcoxon signed ranks test, which is the non-parametric version of the t-test. This test statistic is calculated by ranking the differences between the two scores. But the scores with zero difference are ignored because "they do not give us any information" (Dancey and Reidy, 2002).

There are two problems with the Wilcoxon test. The first problem is the elimination of those cases where the difference is zero. Dancey and Reidy (2002) claim that these cases do not give us any information, which may be true when trying to establish that there *is* a difference between two groups. However, when evaluating marking systems, we want to establish that there is *no* difference between two groups or that the difference is very small. If, for example, a marking system produces marks that agree with the human 95% of the time, that figure *is* informative, contradicting one of the assumptions of the Wilcoxon test. We need a test statistic that takes into account the number of cases where the difference between two marks is zero.

The second problem with the Wilcoxon test statistic is that it shows *whether* two groups are different but not by how much. To solve that problem, we can look at the mean difference given by the descriptive statistics – no difference or very small differences would allow us to conclude that there is no significant difference between two groups. However, as mentioned in Section 2, tuning an LSA-based marking system is critical. How should we compare the results of tuning the system? We cannot use mean differences by themselves; we must consider the standard deviations. This requirement leads us back to effect sizes, but as mentioned in Section 3, our results will be invalid because our data are not normally distributed.

For the reasons given above, we cannot use correlation statistics, t-tests, or effect sizes with confidence intervals. Section 6 presents two possible alternative metrics and provides an example of using them to evaluate test results. Before presenting the alternatives, however, we digress to describe the experiments carried out to tune EMMA for HTML questions.

5 Evaluating EMMA for assessing questions on HTML

As discussed in Section 2, we are developing EMMA to mark answers to exams given in an introductory computer science course. In this study, we investigated how much training data are needed for good results marking questions on the use of HTML. (Note that we are treating the HTML as text. We are not running the HTML through a browser to evaluate the results.) Table 2 gives the text of the two questions. Each question was worth a maximum of four points. The questions were preceded with the instructions: 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.

Question ID	HTML	The desired appearance
A	<I>It is <B>very</I> </B> important to read this text carefully.	It is ***very*** important to read this text carefully.
B	Things to do: Pack suitcase,<BR></BR> Book taxi.	Things to do: Pack suitcase, Book taxi.

Table 2. Exam questions on HTML

The training data for EMMA comprises the course textbook and tutor-marked questions. The purpose of this study was to determine the optimum number of marked answers needed for best results. We have 1,000 marked answers for each of the two questions shown in Table 2. We used EMMA to mark each of the questions in the following way. We submitted 333 of the previously marked answers to be marked by EMMA; we used none of these 333 answers for training data. We discarded 40 answers for Question A and 41 answers for Question B due to marker error. (The human marker erroneously gave an aggregate score for the entire exam rather than providing a separate mark for each question.) We used the remainder of the 1,000 marked answers (minus those with input errors) to train the system. We ran sixteen experiments (for each question) using different amounts of training data: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600 and 627 (for Question A) and 626 (for Question B) marked answers. Figures 1 and 2 show the results of marking each of the two questions.

The figures contain a lot of information, which makes them difficult to understand and interpret. How can one determine the best amount of training data by looking at these charts? (This difficulty in interpreting the data is the main motivation for finding an alternative success metric.) The y-axis shows the percentage of marks. The x-axis shows the amount of training data. The data points show the percentage of marks where the human and the computer agree, or differ by from zero to four points. The

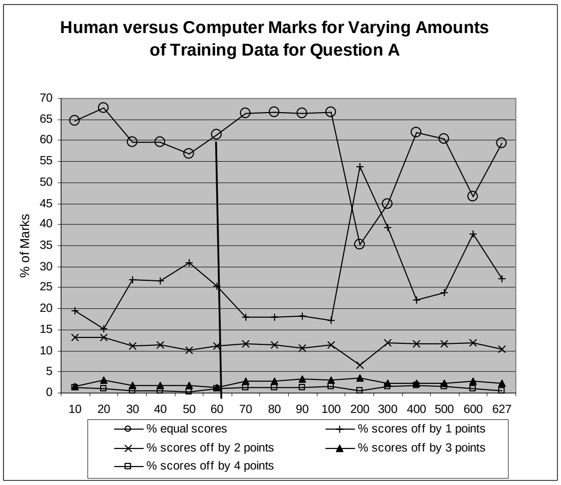

Figure 1. Comparison of tutor and computer marks for various amounts of training data for Question A

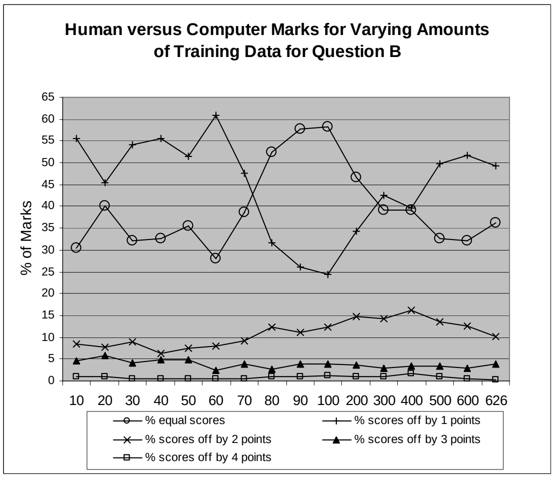

Figure 2. Comparison of tutor and computer marks for various amounts of training data for Question B

first set of data points (marked by an open *0*) indicates the cases where there was zero difference between the tutor mark and the computer mark, i.e., they are identical. The second set of data points (marked by a +) is where there was a difference of plus or minus one point. The fifth set (marked with an open square) indicates those questions with the worst results: either the tutor awarded four points and the computer awarded zero points, or the tutor awarded zero points and the computer awarded four points. The legend below the graph shows the correspondence between each set of data points and the amount by which the human and computer scores differ.

The viewer can see all of the results for a particular amount of training data by looking at a vertical slice of the graph. For example, the vertical line in the graph shows that when 60 training examples were used, EMMA matched the human about 61% of the time, differed by one point about 25% of the time, differed by two points about 11% of the time, differed by three points about 1% of the time, and differed by 4 points about 1% of the time.

Looking at a vertical slice of the graph shows the performance of EMMA for a particular amount of training data. Looking at a horizontal set of data points gives another point of view. The set shows how much the performance varies over different amounts of training data. For example, the set indicated with an *x* shows that the marks that differed by plus or minus two points ranged from about 7% for 200 training data to about 13% for 10 and 20 training data.

Manhattan distance measure (1-norm, or L1):

$$M(\mathbf{X},\mathbf{Y}) = \sum_{i=1}^{n} \left| x_i - y_i \right|$$

Euclidean distance measure (2-norm, or L2):

$$M(\mathbf{X},\mathbf{Y}) = \sqrt{\sum_{i=1}^{n} \left(x_i - y_i \right)^2}$$

Where $\mathbf{X} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{Y} = (y_1, y_2, \ldots, y_n)$ are two *n*-dimensional vectors.

Table 3: Two metrics from vector arithmetic

We tried several different graphical ways to display the data – figures 1 and 2 show the clearest way we found. Even so, it is difficult to evaluate the overall effectiveness of varying the amount of training data by analysing these figures. The next section explains the metrics we used to answer the primary question of this paper: What is the appropriate amount of training data for computer assessment of short answers using an LSA marking system?

6 Two unconventional success metrics

We planned to evaluate the results of these 32 experiments by comparing the marks given by EMMA with the marks given by the tutor using a success metric from a standard statistical test. The inability to locate an appropriate metric, as described in Sections 3 and 4, combined with the difficulty in interpreting Figures 1 and 2 led us to look at unconventional metrics. We took two useful metrics from the field of vector space theory. Section 6 explains these metrics and shows how they can be used to evaluate the results of marking the two questions with varying amounts of training data.

6.1 Metrics from vector space theory

The Manhattan Distance measure (L1) and the Euclidean Distance measure (L2) are two metrics used to calculate the distance between two vectors (Gerald and Wheatley, 1970). Their application to marking exams is as follows. One vector is the list of scores given to a question by a human marker; the other vector is the list of scores assigned by EMMA. If the vectors are identical, the distance between the vectors is zero and the automatic marker would agree perfectly with the human.

The two measures are calculated using the well-known formulas shown in Table 3. These formulas compute the

Question A							
# of Marked Answers	% Equal Scores	Tutor and Computer differ by ±1	Tutor and Computer differ by ± 2	Tutor and Computer differ by ± 3	Tutor and Computer differ by ± 4	Manhattan Distance L1	Euclidean Distance L2
10	64.7	19.5	13.2	1.5	1.2	104.8	10.2
20	67.7	15.3	13.2	3.0	0.9	109.3	10.5
30	59.6	26.9	11.1	1.8	0.6	97.0	9.8
40	59.6	26.6	11.4	1.8	0.6	97.9	9.9
50	56.9	30.8	10.2	1.8	0.3	92.5	9.6
60	61.4	25.4	11.1	1.2	0.9	94.9	9.7
70	66.5	18.0	11.7	2.7	1.2	108.1	10.4
80	66.8	18.0	11.4	2.7	1.2	106.9	10.3
90	66.5	18.3	10.8	3.3	1.2	110.2	10.5
100	66.8	17.4	11.4	3.0	1.5	113.8	10.7
200	35.3	53.9	6.6	3.6	0.6	122.2	11.1
300	44.9	39.2	12.0	2.4	1.5	132.6	11.5
400	62.0	22.2	11.7	2.4	1.8	119.2	10.9
500	60.5	24.0	11.7	2.4	1.5	116.2	10.8
600	46.7	37.7	12.0	2.7	0.9	124.3	11.1
627	59.3	27.2	10.5	2.4	0.6	100.3	10.0

Table 4. Number of answers used for training data over 16 Values ranked in order of amount of training data

Question B							
# of Marked Answers	% Equal Scores	Tutor and Computer differ by ±1	Tutor and Computer differ by ± 2	Tutor and Computer differ by ± 3	Tutor and Computer differ by ± 4	Manhattan Distance L1	Euclidean Distance L2
10	30.5	55.7	8.4	4.5	0.9	144.0	12.0
20	40.1	45.5	7.8	5.7	0.9	142.2	11.9
30	32.0	54.2	9.0	4.2	0.6	137.4	11.7
40	32.6	55.7	6.3	4.8	0.6	133.5	11.6
50	35.6	51.5	7.5	4.8	0.6	134.1	11.6
60	28.1	60.8	8.1	2.4	0.6	124.3	11.1
70	38.6	47.6	9.3	3.9	0.6	129.3	11.4
80	52.4	31.7	12.3	2.7	0.9	119.5	10.9
90	57.8	26.2	11.1	3.9	0.9	120.5	11.0
100	58.1	24.4	12.3	3.9	1.2	128.3	11.3
200	46.7	34.4	14.7	3.6	0.9	139.8	11.8
300	39.2	42.5	14.4	3.0	0.9	141.3	11.9
400	39.2	39.5	16.2	3.3	1.8	162.6	12.8
500	32.6	49.7	13.5	3.3	0.9	147.6	12.1
600	32.0	51.8	12.6	3.0	0.6	138.6	11.8
626	35.7	50.5	8.7	4.2	0.9	137.5	11.7

Table 5. Number of answers used for training data over 16 Values ranked in order of amount of training data

distance between the vectors in slightly different ways. The L1 computes the sum of the differences between each point in the two vectors; the L2 computes the square root of the sum of the squares of the differences. The next sub-section shows how each of these two metrics evaluates the 32 experiments discussed in the previous section: the effects of 16 different amounts of training data for two different questions.

6.2 Example of L1 and L2 using questions on HTML

Tables 4 and 5 show the results of the experiments described in Section 5. The final two columns show the values for L1 and L2, as calculated by the formulas in Table 3. The results are sorted in increasing order of amount of training data. By careful inspection, one can see that the smallest L2 distance corresponds to the best result. A perfect automatic marking system, i.e., one that agrees with the tutor 100% of the time would have a Euclidean distance of zero.

Figure 3 displays a graphical representation of the Euclidean distance (L2) measures from Tables 4 and 5. Recall that a lower distance is a better result than a higher distance. The graph shows that for Question A, the L2 metric ranges from about 9.5 to 11.5. For Question B, L2 ranges from about 11 to 12.8.

A few conclusions can be drawn from studying Figure 3. The most obvious, given that lower L2 values indicate a better result, is that EMMA works better for Question A than it does for Question B for all amounts of training data tested. Another result is that the graphs look slightly different for each question. Question A shows that the marking system works best at around 50 and gets progressively worse until about 300 where it then gets better but not as good as for 50. Question B shows a smoother curve from 10 to 80 where the results get better before getting worse from 80 to 400, and then like Question A, get better again but not as well as for 80.

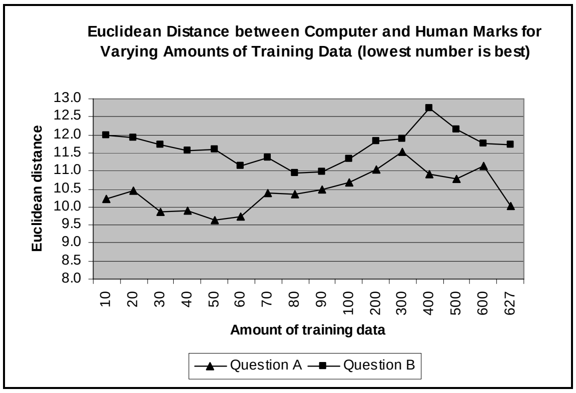

Figure 3. Euclidean distance measure for Question A and Question B

A surprising result was that the largest amount of training data did not produce the best results for either of the questions.

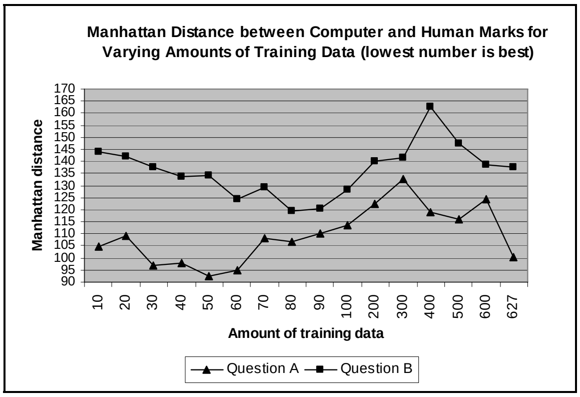

Figure 4. Manhattan Distance Measure for Question A and Question B

Figure 4 is similar to Figure 3. It displays a graphical representation of the Manhattan distance (L1) measures from Tables 4 and 5. As for L2, a smaller distance for L1 is a better result than a larger distance. The graph shows that for Question A, the L1 metric ranges from about 92 to 133. For Question B, L1 ranges from about 120 to 163. The shape of these two graphs is quite similar to the shapes in Figure 3. Thus, the L1 metric provides similar insights to the L2 metric.

The L1 and L2 metrics provided a figure that could be used to assess the results of the different experiments quickly. This evaluation agreed with a careful hand analysis of all of the figures given in Tables 4 and 5.

7 Conclusions and further work

7.1 Conclusions

This study resulted in three significant insights to the field of automatic assessment of short answers.

Contrary to expectations, more training data are not better for these questions. In fact, the better results were from around 50 to 90 for both of the questions. This is good news for those wishing to use an LSA-based marking system because it is easier, quicker, and cheaper to supply fewer human marked answers.

We had different results for the two questions. EMMA worked best for Question A at 50 marked answers used for training data; for Question B, the most effective amount of training data was 80. Therefore, fine-tuning an automatic assessment system requires analysis to be done for each question separately.

The Manhattan and Euclidean distance measures seem to be promising tools for automatic marking researchers to evaluate their systems. These two metrics take into consideration the values of agreement between human and computer over the whole range of possibilities. That is, they evaluate the results where human and computer marks are identical, where they are off by plus or minus one point, plus or minus two points, and so on until the worst result which is where the human and computer differ by the maximum point value of the question. The simple metric that uses the value where the human and computer marks are identical can lead to ambiguity, as demonstrated in section 3.1. L1 and L2 give a richer picture of the effectiveness of an automatic marking system than the simple metric and are no more difficult to analyse than the simple metric.

7.2 Further work

We plan further work to, on the one hand, better understand our results and on the other hand better understand how to tune our LSA-based automatic marking system:

- We plan to meticulously examine training data for these questions to determine why the results were variable.
- We want to experiment with a new hypothesis based on the results reported in this study: the quality of the training data is more important than the quantity.
- We will use L1 and L2 to compare results of varying other LSA parameters such as the number of dimensions, weighting method, stop words, stemming, and spell checking.

8 Acknowledgement

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.

9 References

Aberson, C., (2002). Interpreting Null Results: Improving Presentation and Conclusions with Confidence Intervals. *Journal of Articles in Support of the Null Hypothesis* **1**(3):36-42.

Dancey, C. P. and Reidy, J., (2002). *Statistics Without Maths for Psychology: Using SPSS for Windows,* Essex, England, Prentice Hall.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and 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.

Dumais, S. T., (2003). Data-driven approaches to information access. *Cognitive Science* **27**:491-524.

Foltz, P. W., (1990). Using latent semantic indexing for information filtering. *Proc. Conference on Office Information Systems.* (Ed, Allen, R. B.), Cambridge, MA. 40-47.

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

Foltz, P. W., Kintsch, W. and Landauer, T. K., (1998). The Measurement of Textural Coherence with Latent Semantic Analysis. *Discourse Process* **25**(2&3):285-307.

Foltz, P. W., Laham, D. and Landauer, T. K., (1999). Automated Essay Scoring: Applications to Educational Technology. *Proc. ED-MEDIA '99.* Seattle.

Furnas, G. W., Deerwester, S., Dumais, S. T., Landauer, T. K., Harshman, R. A., Streeter, L. A. and Lochbaum, K. E., (1988). Information retrieval using a singular value decomposition model of latent semantic structure. *Proc. of 11th annual int'l ACM SIGIR conference on Research and development in information retrieval.* ACM. 465-480.

Gerald and Wheatley, (1970). *Applied Numerical Analysis*, Addison-Wesley.

Graesser, A. C., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D. and The Tutoring Research Group, (2000). Using latent semantic analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Environments. [Special Issue, J. Psotka, guest editor]* **8**(2):129-147.

Haley, D. T., Thomas, P., De Roeck, A. and Petre, M., (2005). A Research Taxonomy for Latent Semantic Analysis-Based Educational Applications. *Proc. International Conference on Recent Advances in Natural Language Processing'05.* (Eds, Angelova, G., Bontcheva, K., Mitkov, R., Nicolov, N. and Nikolov, N.), Borovets, Bulgaria. 575-579.

Kanejiya, D., Kumar, A. and Prasad, S., (2003). Automatic Evaluation of Students' Answers using Syntactically Enhanced LSA. *Proc. HLT-NAACL 2003 Workshop: Building Educational Applications Using Natural Language Processing,.* 53-60.

Landauer, T. K. and 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. and Laham, D., (1998). An introduction to Latent Semantic Analysis. *Discourse Processes* **25**:259-284.

Lemaire, B. and Dessus, P., (2001). A system to assess the semantic content of student essays. *Journal of Educational Computing Research* **24**(3):305-320.

Manning, C. D. and Schütze, H., (1999). *Foundations of Statistical Natural Language Processing,* Cambridge, Massachusetts, MIT Press.

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

Nakov, P., Valchanova, E. and Angelova, G., (2003). Towards Deeper Understanding of the LSA Performance. *Proc. of Recent Advances in Natural Language Processing.* Borovetz, Bulgaria. 311-318.

Page, E., (1967). Statistical and linguistic strategies in the computer grading of essays. *Proc. 1967 International Conference On Computational Linguistics.* 1-13.

Thompson, B., (2002). What Future Quantitative Social Science Research Could Look Like: Confidence Intervals for Effect Sizes. *Educational Researcher* **31**(3):25-32.

Wiemer-Hastings, P., (2000). Adding syntactic information to LSA. *Proc. 22nd Annual Conference of the Cognitive Science Society.* 989-993.

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