Measuring MT Adequacy Using Latent Semantic Analysis

Florence Reeder
MITRE Corporation
7515 Colshire Dr.
McLean VA 22102
freeder@mitre.org

Abstract

Translation adequacy is defined as the amount of semantic content from the source language document that is conveyed in the target language document. As such, it is more difficult to measure than intelligibility since semantic content must be measured in two documents and then compared. Latent Semantic Analysis is a content measurement technique used in language learner evaluation that exhibits characteristics attractive for re-use in machine translation evaluation (MTE). This experiment, which is a series of applications of the LSA algorithm in various configurations, demonstrates its usefulness as an MTE metric for adequacy. In addition, this experiment lays the groundwork for using LSA as a method to measure the accuracy of a translation without reliance on reference translations.

1 Introduction

Translation adequacy can be defined as the amount of semantic content from the source language document that is conveyed in the target language document (White & Reeder, 2002). As such, it is difficult to measure since semantic content must be measured in two documents and then compared. Latent Semantic Analysis (Furnas et al., 1988; Deerwester et al., 1990; Landauer, et al., 1998a; Foltz et al., 1998; Foltz et al., 2000) is a content measurement technique used in language learner evaluation that exhibits characteristics attractive for re-use in machine translation evaluation (MTE). In particular, LSA measures semantic content, demonstrates independence over individual word choice, tolerance of syntactic errors, ability to train for domains and applicability to language testing problems. For instance, it has been used for assessing the Test of English as a Foreign Language (TOEFL) essays (Burstein et al., 1998a; Burstein & Chodorow, 1999; Landauer et al., 1998a). This experiment, which is a series of applications of the LSA algorithm in various configurations, demonstrates its usefulness as an MTE metric for adequacy. In addition, this experiment lays the groundwork for using LSA as a method to measure the accuracy of a translation without reliance on reference translations.

2 Latent Semantic Analysis

Latent Semantic Analysis (LSA) was developed from the information retrieval technique, Latent Semantic Indexing (LSI) (Furnas et al., 1988; Deerwester et al., 1990). LSA has been successfully applied to the problem of automated essay grading (Foltz, 1996; Foltz et al., 1996; Foltz et al., 1998; Foltz et al., 1999; Foltz et al., 2000; Laham, 1997; Landauer et al., 1998a, 1998b; Rehder, et al., 1998; Landauer & Psotka, 2000; Burstein et al., 1998a; Burstein & Chodorow, 1999); intelligent tutoring systems (Foltz et al., 2000; Wiemer-Hastings, et al., 1999; Graesser et al., 2000a) and other educational problems (Landauer & Psotka, 2000). The basic concept behind LSA is that words are known by the company they keep, or co-occurrence, and also by the functions they serve, or usage patterns. The developers also assume word dependence unlike other information retrieval techniques (Deerwester et al., 1990). LSA attempts to measure this dependence while reducing the noise of surface representation.

The first step in applying LSA is to train a knowledge space to reflect the domain of interest. The corpus can be generalized or specific (Landauer, et al., 1998a; Foltz et al., 1998) with the use of the knowledge space changing slightly for each type. The corpora used are a collection of docu-

Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, pages 176-184,
Cambridge, August 2006. ©2006 The Association for Machine Translation in the Americas

ments representing the body of knowledge to be tested on: such as an introductory psychology text (in Foltz et al., 2000), all the reading a high school student should have been exposed to (in Landauer et al., 1998a) or in this case, a collection of news text for a given year. Training occurs in the following sequence of steps:

1. Given text corpus $\Delta = \{ \delta_1 .. \delta_m \}$ with vocabulary $\{w_1 \ldots w_n\}$, an N x M matrix is constructed where each cell, A_{ij} is the occurrence count for w_i in δ_j
2. Each entry is then scaled according to an information theoretic weighting scheme. For global weighting, an entropy measurement is used where entropy is defined as:

   $$1 - \Sigma_{\Delta} ([(A_{ij} / \Sigma_m A_i) * \log (A_{ij} / \Sigma_m A_i)] / \log m) \qquad \textbf{(2)}$$

   with m equaling the number of documents in Δ. (Dumais, 1991)
3. The resulting matrix is then decomposed using singular value decomposition (Furnas et al., 1988), yielding $A = USV^T$ where:
   a. U = Eigenvectors AA^T
   b. V = Eigenvectors A^TA
   c. S = Nonnegative square root of eigenvalues AA^T

   In this context, AA^T is the word similarity matrix.
4. After singular value decomposition (SVD), the first K columns are selected.

Because of the sparseness of the word-document matrix, it is desirable to have an array with reduced dimensions as a "best approximation" of AA^T (Landauer & Littman, 1996). By reducing the dimensions and collapsing the representational space, documents which are semantically related become similar even if the individual words are not greatly overlapping. Part of successful LSA application lies in picking the correct K, which is typically between 100-500 and usually around 300 (Landauer et al., 1998a).

Once the semantic space has been calculated, LSA is ready to be utilized. Typically LSA is used to compare two "documents"[1]. For each document, a vector is created where each entry in the vector represents the word count for a given word in that document. The vector is scaled by entropy and converted into the LSA knowledge space. Two such vectors (X, Y) can then be compared, frequently with a cosine comparison where:

$$\cos \Theta = (X \cdot Y) / ||X|| * ||Y|| \qquad \textbf{(3)}$$

A score closer to 1 or –1 means the documents are very similar. They are deemed less similar as the cosine approaches zero (Figure 1). This is the second area where flexibility in design exists and a system can have thresholds tailored to best meet needs.

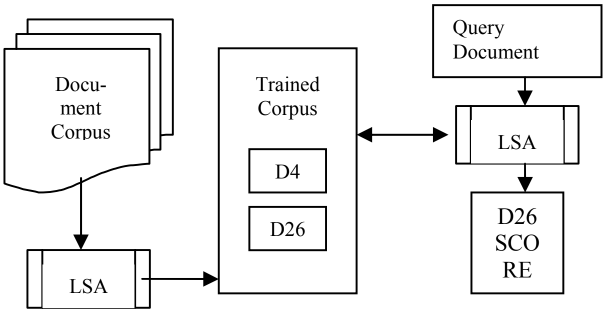


Figure 1: Use of LSA as Query

3 Determining Essay Fidelity with Latent Semantic Analysis

LSA has been developed as a method for measuring the adequacy of student essays covering the semantic content in a domain. In educational contexts this has been used in two different key ways: successfully taking and passing tests designed for English students and as a grading agent for student-authored essays on a course topic. Landauer et al (1998a) used an LSA-trained system to take vocabulary tests. In that experiment, researchers trained on the TASA[2] corpus (Soto, 1998) which represents the amount of textual knowledge a student is exposed to through high school. They then obtained the vocabulary portions of the Test of English as a Foreign Language (TOEFL) examination. For each question an LSA vector was computed and another set of vectors for each response option. The option whose vector was closest as determined by the cosine metric was selected as the correct answer. The LSA system passed the vocabulary test with a grade well-above chance, although on the low end of the passing scale

[1] Where a document can be as small as a word or phrase or as large as a multi-paragraph essay, although paragraph is preferred as the minimum level of granularity (Foltz, et al., 1998).

[2] Touchstone Applied Science Associates, Inc.

(roughly a score of 60% correct). This method of using LSA demonstrates a vocabulary flexibility necessary for adequacy judgments in MTE since it picked semantically close words without having to select exact string matches.

In the grading context, LSA can be used in its retrieval form (Wiemer-Hastings et al., 1999). Given a new essay to be graded, the closest essay is retrieved from the pre-graded pool. The retrieved document's grade is then assigned to the query document. In the second application, LSA was used to grade through five methods (Foltz et al., 2000; Foltz et al., 1999; Landauer et al., 1998a; Foltz, 1996; Landauer et al., 1998b). The difference between the various methods was in the comparison basis: to other pre-graded essays, to reference essays, to reference texts, to reference propositions and in relation to all other submitted essays.

In the first method, a student-authored essay set was graded by judges. This set was used to train the LSA space. A candidate document was then compared to the pre-graded essays in the LSA space. The document score was assigned as the average of the scores for the N-closest documents in the vector space. All documents within the space were in response to the same essay question.

The second method used a teacher authored essay as a reference text. The LSA space was trained using course materials. The reference text was then transformed to a vector within the LSA space. After calculating the vector representation for the candidate text, the cosine between the two vectors was measured. The degree to which the candidate essay matches the reference was used to compute the score for the text.

The third method also trained on the text read by the students. For each sentence in the student essay, a vector was computed within the LSA space. The cosine between the sentence vector and the trained space was measured and used to compute a score. The scores for the sentences were accumulated into a final score. The fourth method was a variation on this where only the document vectors marked as important by the instructor were considered in the scoring.

In the fifth method, the essays were measured in relation to each other. The measurements were then clustered and the category scores are assigned according to the category breaks in the clustering. All five methods worked reasonably well, as consistently as human scorers. Although some were marginally better than others, Foltz (1998) does not identify either the best candidate method or the reason behind the selection. Due to software limitations, we used only the first two methods in this experiment.

4 Experimental Setup

Due to the unavailability of the LSA essay grading software, we used Latent Semantic Indexing (LSI++) implementation developed at University of Tennesssee, Knoxville by Michael Berry (Giles et al., 2003). In addition, a number of data formatting and result calculation scripts were written to account for the fact that this is an information retrieval platform rather than an essay scoring one.

4.1 Test Data

The data for this experiment was selected because it had been used in previous MT evaluations (White & O'Connell, 1994), representing a body of previously scored data. The corpus is known as the DARPA-94 corpus as it had been designed for the DARPA MT evaluation held in 1994. Consisting of three language pairs, French-English, Spanish-English and Japanese-English, it contains one hundred documents per language pair taken from news texts, selected for a diversity of topics. A given document consists of a headline and the accompanying news story. The texts are roughly 400 words apiece. For each source language document, one reference translation and one expert, human judged translation accompany the machine translation outputs. The scores used in this experiment are the adequacy scores. Adequacy was judged by showing scorers segments of reference text along side a translation output sentence. Adequacy measured the degree to which the MT output reflected the semantic content of the reference segment. Judgments were therefore performed monolingually. The scoring was set up so as to account for human factors, although no document was seen by more than one scorer.

4.2 Experiment Execution

The data was divided into two components, an experiment set which contains the documents selected for testing each method and a test set which contains the documents reserved for later experimentation. The experiment set contains 75% of the documents, selected by removing every fourth document from the document set. Within the experiment set, the data was further divided into the training set which is used to train the LSA data space from which results are selected and the query set which contains the MT output to be scored. The training set consists of the reference translations and some combination of expert and MT translations, depending on the method used. Table 1 describes each method with the training set and query set for each run. Each method and run was performed for both document and paragraph level granularity.

Table 1: Experiment Configurations

Method	**Training Set**	**Query Set**

Method 1 –
{K=300,I=100}
{K=200,I=100}
{K=200,I=50}
	Reference translations
	Expert translations
		All MT system outputs

Method 1 – {K=200, I=100}	Reference translations only	All MT system outputs
Method 1 – {K=200,I=100}	Expert translations only	All MT system outputs
Method 2 – Run 1	Reference translations Expert translations For each document, N-2 system outputs	For each document, rest of system outputs
Method 2 – Run 2	Reference translations Expert translations First N-2 systems' outputs	Last two systems' outputs
Method 2 – Run 3	Reference translations Expert translations Last N-2 systems' outputs	First two systems' outputs

Two scores are calculated for method one. The first score is the cosine score returned by the LSI++ algorithm, representing the strength of closeness between the MT output (query document) and the reference or expert translation (returned document). The second score is calculated as the product of:

- the returned document's score: a 1.0 for reference translations, the human rated score for expert translations;
- a penalty for the query document not matching the reference document. That is, the reference document returned is not a translation of the source document corresponding to the MT output. For document level matching, the penalty is 0.5. For paragraph level matching, it is 0.75, due to paragraph alignment issues in the data.
- the returned cosine score.

For method two, the calculated score is the average of the scores belonging to the top two documents returned. Again, these are conditioned by the cosine scores returned by the LSI software as well as penalties for document mismatches.

4.3 Expected Results

LSA has been used successfully to grade the semantic content of a given essay through the assignment to a grade category which matches the grade category assigned by human raters, usually on a one to six scale. In the essay grading case, individual essay scores as calculated by LSA were shown to correlate with the human graded essay scores. To test this we vary run parameters: from 300 to 200 **K** factors, from 200 to 50 run iterations, from document to paragraph granularity, and the scoring method used. We expect that the lower **K** factor and larger granularity to score better. We expect that in the MT output case, LSA can be used to grade MT output similarly.

5 Results and Analysis

The results show that LSA can be used to grade MT output, but not at the desired level of granularity in its current form. Additionally, the individual document correlation is not nearly as good as was seen with educational applications. On the other hand, using correlation score calculations based on averaging the scores for a given system, as is common in MTE, shows strong correlations between the LSI scores and the human judged adequacy scores. These results, however, are strong enough to investigate the use of a cross-language

LSI implementation for evaluating MT output without the benefit of a reference translation.

5.1 Method 1: Trained on Reference and Expert Translations

In training on the reference and expert translations, the best indicator of MT adequacy is the cosine score for the retrieved document, based on SVD **K** parameters of 200, 50. In calculating the per system correlation between adequacy score (QADE) and computed LSA score (SCORE), a correlation of $R^2 = 0.91$ is obtained. For system adequacy compared to returned document cosine score (QSCORE), a correlation of $R^2 = 0.95$ is measured. Per system scores are calculated by averaging the individual document scores for a given system. The individual document scores do not correlate strongly (Pearson correlation = 0.34). This is also true when applying a category-assignment measure as opposed to a sliding scale score. Given the claims of LSA proponents, this is a disappointing result, but may be accounted for by the fact that there are only individual scores for each of the documents. The individual document granularity issues in MTE have been demonstrated (e.g., Reeder & White, 2003). Without a second human judgment score, expected agreement cannot be baselined for a kappa result. Removing the expert translation yields additional gains in correlation. The penalties for incorrect documents returned and for expert translations returned rather than reference translations tended to drag down the correlations.

Table 2: Results for Run 1

QTAG	QADE	MATC	DSCOR	QSCOR	SCOR
FR-C	0.669	1.507	0.96	0.946	0.908
FR-GP	0.694	1.446	0.964	0.944	0.911
FR-MS	0.721	1.513	0.963	0.948	0.913
FR-SY	0.786	1.48	0.967	0.956	0.925
FR-XS	0.346	1.507	0.954	0.9	0.86
JA-L	0.303	1.493	0.934	0.868	0.801
JA-P	0.24	1.453	0.926	0.855	0.792
JA-PN	0.37	1.467	0.924	0.884	0.79
JA-SY	0.315	1.473	0.926	0.842	0.771
SP-GP	0.78	1.48	0.961	0.974	0.936
SP-L	0.652	1.52	0.966	0.957	0.914
SP-PA	0.811	1.507	0.967	0.975	0.932
SP-P	0.522	1.527	0.964	0.932	0.888
SP-SY	0.774	1.568	0.966	0.973	0.928

5.1.1 Article Level

The article level scores for the first run are shown (Table 2). QTAG indicates the language and system. QADE is the adequacy score average for the system. MATCH was calculated by assigning a two if the returned document was the reference translation, a one if the expert translation was returned or zero if neither was returned. The MATCH scores were then averaged across all documents. DSCORE is the average expert translation adequacy score. QSCORE is the cosine score and SCORE is the final calculated score. These last two scores are depicted graphically (Figure 2; Figure 3). The run scores are shown as well (Table 3). While the correlation is good when averaging across documents ($R^2 = 0.91$ for SCORE versus adequacy), individual document correlations are weak ($R^2 = 0.33$ for SCORE versus adequacy). Additionally, the various runs show that lower dimensions are optimal for this application. This is most likely due to the small data set and relatively short document size.

Table 3: Summary of Method One Run Scores

RUN	DESCRIPTION	SCORE	COS-SCORE
1	Base run, K= 300, 200	0.89	0.92
2	K-factors = 200, 100	0.90	0.95
3	K-factors = 200, 50	0.91	0.95
4	Remove expert translations	0.94	0.95
5	Remove reference translations	0.86	0.95

5.1.2 Paragraph Level Experiment

At the paragraph level, the scores drop. This is not surprising, given that there are alignment issues with the MT output. That is, translators and translation systems will author paragraph breaks at different points in a text, therefore, there are situations where the numbers of paragraphs do not match and the paragraph contents do not match. In addition, by reducing the size of the query, the possibility of a query with few or no content words increases. This situation was seen in processing at the paragraph level. Both run 1 and run 2 in this method have one outlying instance which have significantly dropped the correlation score, that is the FR-C system. The results for the runs are reported in tabular form (Table 4). Further analysis is needed to determine if this is an artifact of the fact that FR-C, the Candide system, is a statistical MT system. The other systems were rule-based systems and therefore the relationship between the metric type and the MT system type needs to yet be explored.

Figure 2: Query Adequacy versus Score for Run 1

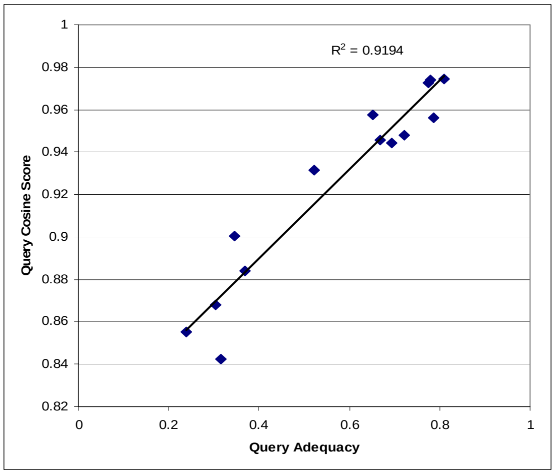


Figure 3: Query Adequacy vs. Cosine Score for Run 1

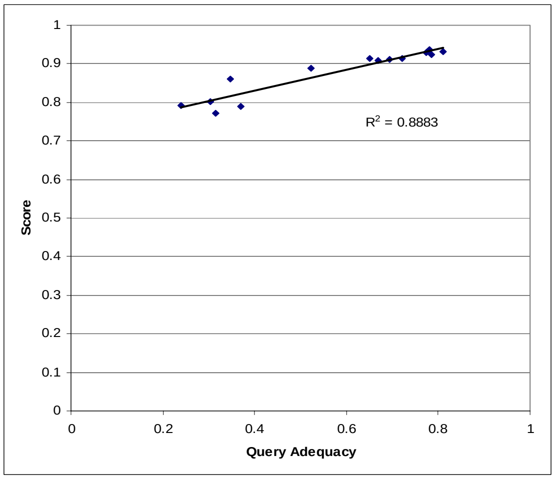


Table 4: Summary of Runs for Method 2

RUN	DESCRIPTION	SCORE	COS-SCORE
1	Base run, K= 300, 200	0.79	0.96
2	K-factors = 200, 100	0.73	0.93
3	K-factors = 200, 50	0.89	0.88
4	Remove expert translations	0.94	0.95
5	Remove reference translations	0.88	0.93

5.2 Method 2: Trained on Reference, Expert and other Systems

When trained on reference, expert and other systems, the results are predictably more uneven. The major reason for this is the small sample size and therefore the lack of exemplars of the same essay. Unlike in the use for Educational Testing Service ETS) where thousands of essays are written on the same topic, in this instance, there are a limited number of MT outputs from which to choose. If the training systems are like the test system in terms of capabilities and lexical coverage, then the scores tend to be more accurate. If the systems are more capable, assigned scores tend to be higher than warranted. Additionally, from a statistical perspective, the sampling process tends to drop the number of test documents which leads to the correlation granularity issues previously discussed.

Table 5: Summary Scores per System for Average Score Method

QNUM	QTAG	QADE	SCORE
15	FR-C	0.69	0.84
15	FR-GP	0.76	0.79
15	FR-MS	0.69	0.82
15	FR-SY	0.82	0.80
15	FR-XS	0.41	0.70
18	JA-L	0.28	0.45
19	JA-P	0.22	0.50
19	JA-PN	0.47	0.55
19	JA-SY	0.36	0.29
14	SP-GP	0.74	0.82
15	SP-L	0.65	0.80
15	SP-P	0.53	0.76
15	SP-PA	0.78	0.78
14	SP-SY	0.78	0.81

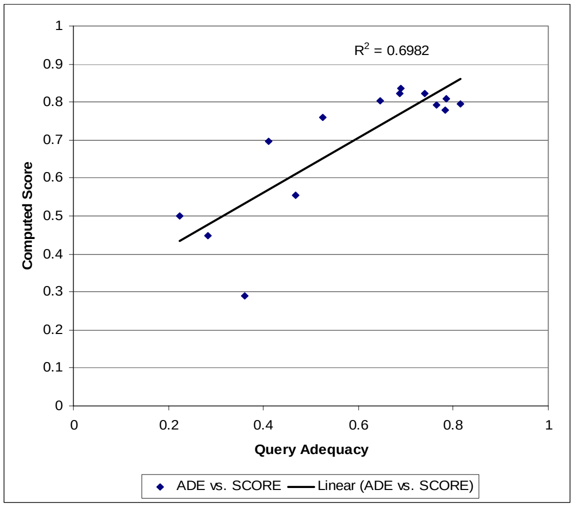
5.2.1 Article Level Experiment

The human judged adequacy scores are compared with the returned system scores (Table 5). System scores are computed by averaging the prejudged adequacy scores of the documents that are in the trained space. As can be seen (Figure 4), the correlation between adequacy and the LSA-derived score is not as strong (R^2 = 0.70) as with method one. On the other hand, it is comparable to results reported with other LSA and essay grading applications (e.g., Burstein et al., 1998a). When varying the pool by holding out entire system outputs as opposed to samples of system outputs, the correlations improve to $R^2 = 0.82$ and 0.98 for two different combinations (Figure 5; Figure 6, respectively). This is due to the fact that in all cases, the best MT system remained in the training set and the fact that the systems held out reflected the systems being trained on more accurately in terms of abilities as judged in DARPA-94.

5.2.2 Paragraph Level Experiment

Due to the increased sample size, the correlation scores improve at the paragraph level (Figure 7) for the general case. Outliers still exist, however, primarily from the Japanese-English translation which had many cases of paragraphs with a returned translation of "X". Because a significant number of Japanese translation systems returned "X" as their paragraph output, zero scores had to be assigned. This tended to reduce the LSA scores more than predicted.

Figure 4: Query Adequacy versus Computed Score



6 Conclusion and Future Work

The LSA results indicate that it can be used further. Three major thrusts will be of interest in the future. The first is in parameter tuning for finer grained adequacy judgments. A multi-judge adequacy assessment would facilitate determining if the metric can be used as a human judge substitute, in the way that it is used in educational applications. Additionally, the weighting scheme for incorrect selections needs tuning. The second thrust is in the application of cross-language LSI to provide reference-less MT evaluation. The last experiment indicates the possibility of this, within a given domain. Of note, however, is the fact that LSI is designed for segmented languages, therefore difficulties may arise with unsegmented languages such as Chinese. Finally, the recent advances in MT and the availability of corpora and these MT systems means that the technique should be tested with newer data.

Figure 5: Scores with Systems First 2 Systems Held Out

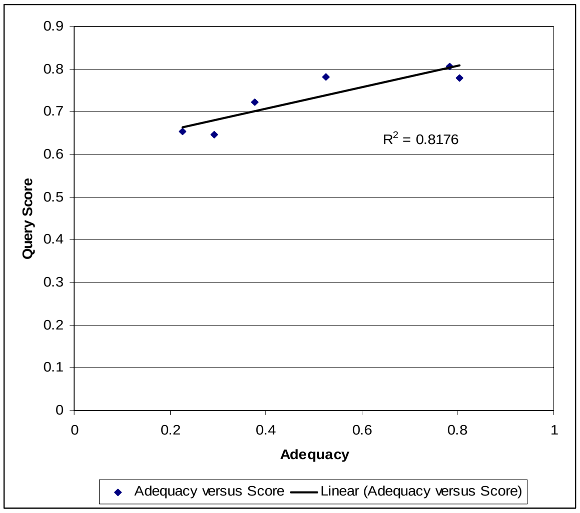


Figure 6: Scores with last two systems held out

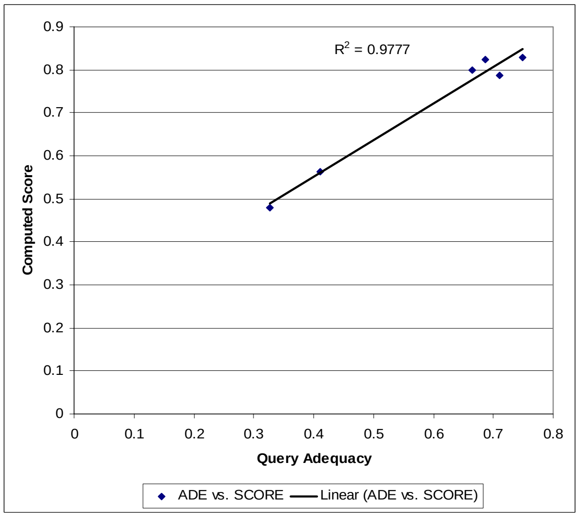
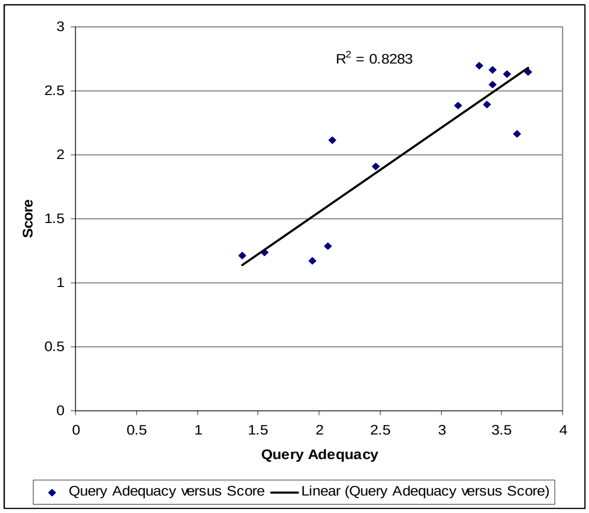


Figure 7: Summary Scores for Method 2, paragraph level



References

Burstein, J. & Chodorow, M. 1999. Automated Essay Scoring for Non-native English Speakers. In M. Olsen, ed., *Computer-Mediated Language Assessment and Evaluation in Natural Language Processing*, Proceedings of a Symposium by ACL/IALL. University of Maryland, 68-75.

Burstein, J., Braden-Harder, L., Chodorow, M., Hua, S., Kaplan, B., Kukich, K., Lu, C., Nolan, J., Rock, D. & Wolff, S. 1998a. Computer Analysis of Essay Content for Automated Score Prediction: A Prototype Automated Scoring System for GMAT Analytical Writing Assessment Essays. Educational Testing Services Technical Report, RR-98-15.

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

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

Foltz, P. 1996. Latent Semantic Analysis for text-based research. Behavior Research Methods, Instruments and Computers. 28(2),197-202.

Foltz, P., Britt, M. & Perfetti, C. 1996. Reasoning from multiple texts: An automatic analysis of readers' situation models. In G. W. Cottrell ed., Proceedings of the 18th Annual Cognitive Science Conference. Lawrence Erlbaum, NJ, 110-115.

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

Foltz, P., Kintsch, W. & Landuaer, T. 1998. The Measurement of Textual Coherence with Latent Semantic Analysis. Discourse Processes 25(2 & 3), 285-307.

Foltz, P., Laham, D. & Landauer, T. 1999. The Intelligent Essay Assessor: Applications to Educational Technology. Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1(2).

Furnas, G., Deerwester, S., Dumais, S., Landauer, T., Harshman, R., Streeter, L. & Lochbaum, L. 1988. Information Retrieval using a singular value decomposition model of latent semantic structure. In SIGIR-88. Grenoble, France

Giles, J., Wo, L & Berry, M. 2003. GTP (General Text Parser) Software for Text Mining. In Bozdogan, H. ed., Statistical Data Mining and Knowledge Discovery. CRC Press, Boca Raton, pp. 455-471

Graesser, A., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Person, N. & Tutoring Research Group. 2000a. Using Latent Semantic Analysis to Evaluate the Contributions of Students in AutoTutor. Interactive Learning Environment

Laham, D. 1997. Automated holistic scoring of the quality of content in directed student essays using Latent Semantic Analysis. Unpublished Master's Thesis. University of Colorado, Boulder.

Landauer, T. & Dumais, S. 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. & Psotka, J. 2000. Simulating text understanding for educational applications with Latent Semantic Analysis: Introduction to LSA. Interactive Learning Environments, 8(2), 73-86.

Landauer, T., Foltz, P. & Laham, D. 1998a. An Introduction to Latent Semantic Analysis. Discourse Processes, 25: 259-284.

Landauer, T., Laham, D. & Foltz, P. 1998b. Computer-based grading of the conceptual content of essays. Unpublished manuscript.

Landauer, T., Laham, D. & Foltz, P. 1998c. Learning human-like knowledge by Singular Value Decomposition: A progress report. In M. I. Jordan, M. J. Kearns & S. A. Solla (Eds.), Advances in Neural Information Processing Systems 10, pp. 45-51. Cambridge: MIT Press.

Landauer, T., Laham, D., Rehder, B. & Schreiner, H. 1997. How well can passage meaning be derived without using word order? A Comparison of Latent

Semantic Analysis and Humans. Proceedings of the 19th Conference of Cognitive Science Society.

Reeder, F. & White, J. 2003. Granularity in MT Evaluation. In Proceedings of MT Evaluation Workshop, Machine Translation Summit IX.

Rehder, B., Schreiner, M., Wolfe, M., Laham, D., Landauer, T. & Kintsch, W. 1998. Using Latent Semantic Analysis to assess knowledge: Some technical considerations. Discourse Processes, 25:337-354

Soto, R. 1998. Learning and Performing by Exploration: Label Quality Measured by Latent Semantic Analysis. Submitted manuscript

White, J., & O'Connell, T. 1994. The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. Proceedings of the 1994 Conference, Association for Machine Translation in the Americas.

Wiemer-Hastings, P., Wiemer-Hastings, K. & Graesser, A. 1999. Improving an intelligent tutor's comprehension of students with Latent Semantic Analysis. In S. Lajoie & M. Vivet (eds.), Artificial Intelligence in Education (AI-ED '99). Amsterdam: IOS Press