On the Automatic Scoring of Handwritten Essays

Sargur Srihari, Rohini Srihari, Pavithra Babu, Harish Srinivasan
Center of Excellence for Document Analysis and Recognition (CEDAR)
University at Buffalo, State University of New York Amherst, New York 14228, U.S.A.
srihari@cedar.buffalo.edu

Abstract

Automating the task of scoring short handwritten student essays is considered. The goal is to assign scores which are comparable to those of human scorers by coupling two AI technologies: optical handwriting recognition and automated essay scoring. The test-bed is that of essays written by children in reading comprehension tests. The process involves several image-level operations: removal of pre-printed matter, segmentation of handwritten text lines and extraction of words. Recognition constraints are provided by the reading passage, the question and the answer rubric. Scoring is based on using a vector space model and machine learning of parameters from a set of human-scored samples. System performance is comparable to that of scoring based on perfect manual transcription.

1 INTRODUCTION

Handwritten essays are widely used for student performance evaluation in schools and colleges. Since this approach to evaluation is efficient and reliable it is likely to remain a key component of learning. Assessing large numbers of handwritten essays is a relatively time-consuming and monotonous task. In statewide examinations on reading comprehension in the U.S. there is an intense need to speed up and enhance the process of rating handwritten essays while maintaining cost effectiveness. The assessment can also be used as a source of timely, relatively inexpensive and responsible feedback about writing.

Writing done by hand is the primary means of testing students on state assessments of reading comprehension. Consider as an example the New York State English Language Assessment (ELA) administered statewide in grades 4 and 8. In the reading part of the test the student is asked to read a passage such as that given in Fig 1 and answer several questions in writing.

An example of a reading comprehension question based on the passage of Fig. 1 is the following:"How was Martha Washington's role as First Lady different from that of Eleanor Roosevelt? Use information from American First Ladies in your answer." The completed answer sheets of three different students to the question are given in Fig. 2. The answers are

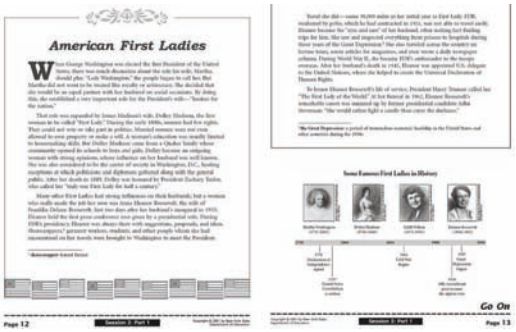

Figure 1: From the New York English Language Arts assessment for Grade 8, 2001 – two of three pages of the story "American First Ladies" are shown.

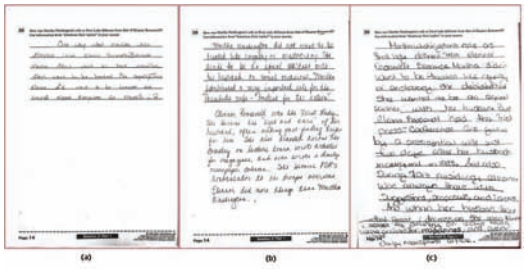

Figure 2: Sample answer sheets of three students (a-c) based on the reading comprehension passage of Fig. 1. The human assigned scores for these essays, on a scale of 0-6, were 2, 4 and 4 respectively.

scored by human assessors on a seven-point scale of 0-6. A rubric for the scoring is given in Table 1. This is referred to as a *holistic* rubric– which is in contrast to an analytic rubric that captures several writing traits.

There is significant practical and pedagogical value in computer-assisted evaluation of such tests. The task of scoring and reporting the results of these assessments in a timely manner is difficult and relatively expensive. There is also an intense need to test later in the year for the purpose of capturing the most student growth and at the same time meet the requirement to report student scores before summer break. The biggest challenge is that of reading and scoring the handwritten portions of large-scale assessments.

6	5	4	3	2	1
Understanding of text	Understanding roles of first ladies	Logical and Accurate	Partial Understanding	Readable	Brief
understanding of similarities and differences among the roles	Organized	Only literal understanding of article	Drawing conclusions about roles of first ladies	Not Logical	Repetitive
Characteristics of first ladies	Not thoroughly eleaborate	Organized	Sketchy	Limited Understanding	Understood only sections
Complete, Accurate and Insightful		Too generalized	Weak		
Focused, Fluent and Engaging		Facts without synchronization			

Table 1: Holistic Rubric Chart for "How was Marth Washington's role as First Lady different from that of Eleanor Roosevelt?"

The assessment problem is a well-constrained problem in artificial intelligence (AI) whose solution will push forward existing technologies of handwriting recognition and automatic essay scoring. The task is a first step in solving an inverse of a grand challenge of AI– that of a computer program to read a chapter of a freshman physics textbook and answer the questions at the end [Reddy, 2003].

2 COMPONENT AI SUBSYSTEMS

The major component AI systems for solving the task are optical handwriting recognition (OHR) and automatic essay scoring (AES). Both subsystems involve a learning phase.

2.1 Handwriting Recognition

OHR is concerned with transforming an image of handwritten text into its textual form; a survey of OHR is [Plamondon and Srihari, 2000]. While computers have become indispensable tools for two of three R's, viz., arithmetic and writing, their use in the third R of reading is still emerging. OHR involves several processing steps such as form (or rule line) removal, line/word segmentation and recognition of individual words. Word recognition relies on a lexicon of words-which could be derived from the passage, question and rubric available in statewide tests.

Recognition of characters and words is performed in a two step process of feature extraction followed by classification. Features can be at the character level (called analytic recognition) or at the word level (holistic recognition). Word recognition, which is the task of assigning a word image to a member of a list of words (or lexicon), can be performed well for correctly segmented words with small lexicons. The process is error prone for mis-segmented text and large lexicons.

When the lexicon is limited, a majority of the words are correctly recognized although there are substitution errors and missed words. These errors can be reduced by better word segmentation and by using linguistic context in the form of transitional probabilities between words, between parts-of-speech tags, noun phrases, etc. It is possible that certain words, when considered in isolation, are illegible. Local context can resolve such ambiguity. The reading comprehension passage and the rubric provide a rich source of contextual information that can be exploited to get high recognition rates. However, the task itself is one of correct interpretation rather than that of recognizing every illegible word. It includes character recognition (OCR), word recognition, part-of-speech tagging, etc.

Prior to OHR, several image processing steps need to be performed on answer sheets, e.g., detecting and eliminating extraneous information such printed instructions, questions, ruled lines and margin lines. Within the handwritten text the ordering of the lines has to be determined and within each line the words need to be segmented. These operations are similar to those for analyzing unconstrained handwritten pages for forensic, or questioned document, analysis [Srihari *et al.*, 2003]).

Handwriting Interpretation is where the goal is not so much one of recognizing every character and word perfectly but to perform some overall task using recognition results. It involves using contextual information when there is uncertainty in the specific components. Such approaches have found success when the domain is limited and contextual information is available, e.g., in the postal domain the destination ZIP+4 code can be assigned even when individual components are poorly written [Srihari and Keubert, 1997].

2.2 Automatic Essay Scoring (AES)

Automatic scoring of computer readable essays has been a topic of research for over four decades. A limitation of all past work is that the essays or examinations have to be in computer readable form. A survey of previous AES methods has been made by Palmer, et. al (2002). Project Essay Grade (PEG) (Page, 1961) uses linguistic features from which a multiple regression equation is developed. In the Production Automated Essay Grading System a grammar checker, a program to identify words and sentences, software dictionary, a part-of-speech tagger, and a parser are used to gather data. E-rater (Burstein, 2003) uses a combination of statisti-

cal and NLP techniques to extract linguistic features. Larkey (1998) implemented an AES approach based on text categorization techniques (TCT). One approach to AES is based on an information retrieval technique known as latent semantic indexing. Its application to AES, known as latent semantic analysis (LSA), uncovers lexical semantic links between an essay and a gold standard. Landauer, et. al. (1998) developed the Intelligent Essay Assessor using LSA. A matrix for the essay is built, and then transformed by the algebraic method of singular value decomposition (SVD) to approximately reproduce the matrix using reduced dimensional matrices built for the topic domain. Using SVD new relationships between words and documents are uncovered, and existing relationships are modified to represent their significance. Using LSA the similarity between two essays can be measured despite differences in individual lexical items. Where a gold standard exists, LSA is a robust approach. It correlates as closely with human raters as human raters correlate with each other [Landauer *et al.*, 2003].

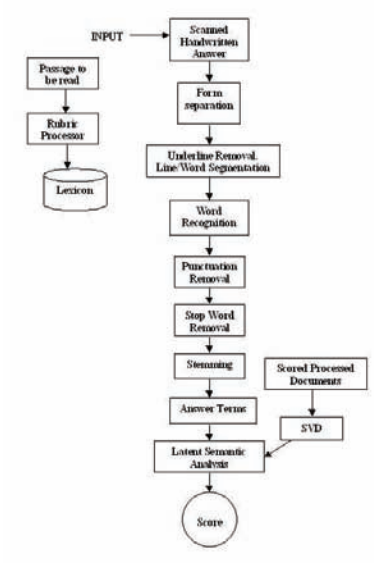

Figure 3: Answer Processor Architecture.

3 SYSTEM INTEGRATION

The overall task is that of handwriting interpretation for essay scoring. A first level of integration is to sequentially couple the OHR and AES systems by regarding OHR simply as a transcription system. Both the OHR and AES components involve machine learning at several levels. In the case of OHR, lexicons are acquired from three sources: reading passage, question and rubric. Learning of handwriting styles in the formation of letters and words is a classic pattern recognition problem. In the case of AES the learning process acquires a method associating content to score by learning from a set of human scored essays. A system to analyze and score the scanned answer sheet(s)is shown in Fig. 3.

3.1 OHR

After performing image pre-processing steps, e.g., foreground/background extraction, eliminating non-informative material (rule lines and printed markings), determining the presence of handwriting, etc., the main tasks are:

(1)Word segmentation into lines and words in the presence of ambiguity. To determine whether a gap is a true gap or not by learning from the current document.

(2) Word recognition: When vocabularies are large contextual information needs to be exploited to dynamically limit word choices. Contextual information is available in the form of the passage to be read and the answer rubric.

After words are recognized the resulting word sequences are written to text files. These text files are then pre-processed for AES which include the following steps: (**a**). Removing punctuations and special characters, (**b**). Converting upper case to lower case for generalization, (**c**). Stop word removal - removing common words such as *a* and *the* which occur very often and are not of significant importance, (**d**). Stemming - morphological variants of words have similar semantic interpretations and therefore a stemming algorithm is used to reduce the word to its stem or root form. The algorithm [Porter, 1980]] uses a technique called suffix stripping where an explicit suffix list is provided along with a condition on which the suffix should be removed or replaced to form the stem of the word, which would be common among all variations. For example the word *reading* after suffix stripping is reduced to *read*.

3.2 AES

In the LSA approach, a good approximation of the computer score to a human score heavily depends on the optimal reduced dimensionality. This optimal dimension is related to the features that determine the term meaning from which we can derive the hidden correlations between terms and answer documents. Reducing the dimensions is done by omitting inconsequential relations and retaining only significant ones. A factor analysis method such as Singular Value Decomposition (SVD) helps reduce the dimensionality to a desired approximation.

The first step in LSA is to construct a t x n term-by-document matrix M whose entries are frequencies. SVD or two-mode factor analysis decomposes this rectangular matrix into three matrices [Baeza-Yates and Ribeiro-Neto, 1999]. The SVD for a rectangular matrix M can be defined as

$$M = TSD', \tag{1}$$

where prime ($'$) indicates matrix transposition, M is the rectangular term by document matrix with t rows and n columns, T is the t x m matrix, which describes rows in the matrix M as left singular vectors of derived orthogonal factor values, D is the n x m matrix, which describes columns in the matrix M as right singular vectors of derived orthogonal factor values, S is the m x m diagonal matrix of singular values such that when T, S and D' are matrix multiplied $M$is reconstructed, and m is the rank of M = min(t , n).

To reduce dimensionality to a value k from the matrix S we have to delete $m - k$ rows and columns starting from those which contain the smallest singular value to form the matrix

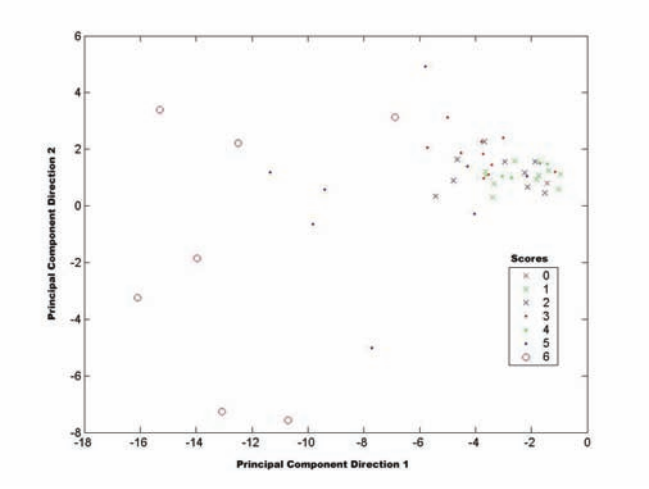

Figure 4: Projected locations of 50 Answer Documents in two dimensional plane

S_1. The corresponding columns in T and rows in $D^{'}$ are also deleted to form matrices T_1 and $D_1^{'}$ respectively. The matrix M_1 is an approximation of matrix M with reduced dimensions as follows

$$M_1 = T_1 S_1 D_1^{'}. \quad (2)$$

Standard algorithms are available to perform SVD. To illustrate, from the document-term matrix constructed from 50 essays from the *American First Ladies* example shown in Fig 1 and Fig 2, the first two principal components are plotted in Fig.4. The principal components are the two most significant dimensions of the term by document matrix shown in Table 2 after applying SVD. This is a representation of the documents in semantic space. The similarity of two documents in such a semantic space is measured as the cosine of the angle made by these documents at the origin.

The testing set consists of a set of scored essays not used in the training and validation phases. The term-document matrix constructed in the training phase and the value of k determined from the validation phase are used to determine the scores of the test set.

4 PERFORMANCE EVALUATION

The corpus for experimentation consisted of 96 handwritten answer essays for the "American First Ladies" task shown in Fig. 1. Of these essays 73 were by students and 23 were by teachers. Each of the 96 answer essays were manually assigned a score (the "gold standard") by education researchers. The essays were scored manually using the holistic grading rubric shown in Table 1. The essays were divided into 50 training samples (each of which also served as validation samples in the leave one out cross validation method employed) and 46 testing samples. The training set had a human-score distribution on the seven-point scale as follows: 1,9,9,11,5,8,7. The testing set had human-score distributions of 0,8,9,10,5,8,6. The answer sheets were scanned as gray scale images at a resolution of 300 pixels per inch.

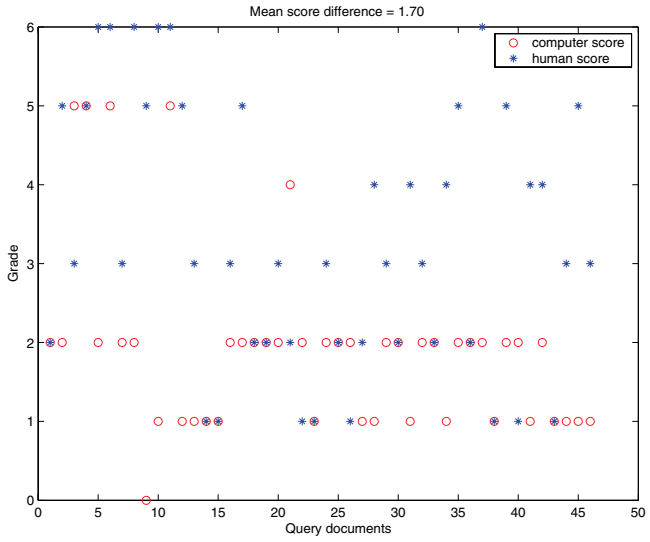

Figure 5: Comparison of human scores and Manual Transcription - Latent Semantic Analysis (MT-LSA) scores on 46 student responses to the *American First Ladies* question: MT-LSA scores (open circles) are within 1.70 of human scores (stars).

Two different sets of 96 transcribed essays were created, the first by manual transcription (MT) and the second by the OHR system. The lexicon for the OHR system consisted of unique words from the passage to be read, which had a size of 274. Separate training and validation phases were conducted for the MT and OHR essays. For the MT essays, the document-term matrix M had $t = 521$ and $m = 50$ and the optimal value of k was determined to be 8. For the OHR essays, the corresponding values were $t = 164$, $m = 50$ and $k = 5$. The smaller number of terms in the OHR case is explained by the fact that several words were not recognized.

Comparisons of the human-assigned scores (the gold-standard) with (i) automatically assigned scores based on MT is shown in Fig. 5 and (ii) automatically assigned scores based on OHR is shown in 6. Using MT the human-machine mean difference was 1.70 (Fig. 6). Using OHR the human-machine difference was 1.65 (Fig. 6). Thus a 0.05 difference is observed between MT and OHR using LSA scoring. These preliminary results demonstrate the potential of the method for holistic scoring and robustness with OHR errors.

5 SUMMARY AND DISCUSSION

Automatically evaluating handwritten essays involves the integration of optical handwriting recognition and automatic essay scoring methodologies. Handwriting recognition is assisted by constraints provided by the reading passage, question and rubric. Scoring based on latent semantic analysis (LSA)is robust with respect to recognition inadequacies. Results on a small testing set show that with manually transcribed (MT) essays, LSA scoring has on an average less than a two-point difference from human scoring. With the same test set, OHR-LSA scoring has a minor difference from MT-

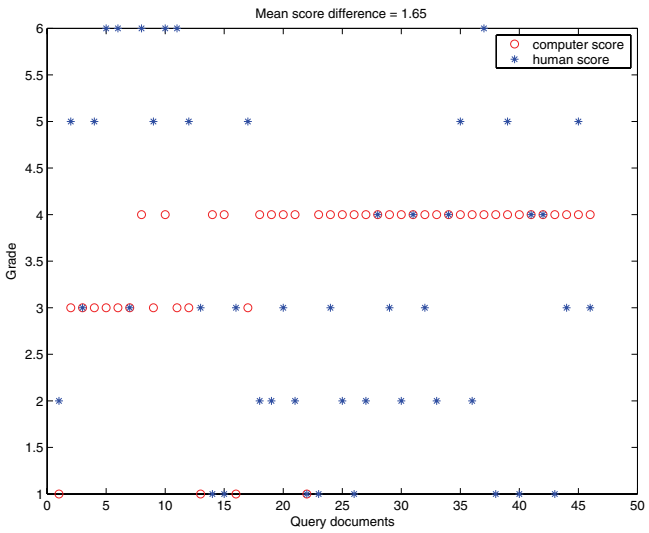

Figure 6: Comparison of human scores and OHR-LSA scores on 46 student responses to the *American First Ladies* question: OHR-LSA scores (open circles) are within 1.65 of human scores (stars).

LSA scoring.

The results point out that despite errors in word recognition the overall scoring performance is good enough to have practical value. This points out that when the evaluation of an OHR system is based not so much on word recognition rates but in terms of the overall application in which it is used, the performance can be quite acceptable. The same phenomenon has been observed in other OHR applications such as postal address reading where the goal is not so much as to read every word correctly but achieve a correct sortation.

The LSA approach has the advantage that as a "bag of words" or holistic technique it is robust with respect to word recognition errors. However it ignores linguistic structures. The analytic approach to scoring is based on idea development, organization, cohesion, style, grammar, or usage conventions. The result of analytic scoring will be more useful to teachers and education researchers.

Essay scoring based on language features such as general vocabulary, passage related vocabulary, percentage of difficult words, percentage of passive sentences, rhetorical features and usage of conjunctions, pronouns, punctuations etc for connectedness could play a significant role in improving the performance of this system. This approach is employed in the automated Japanese Essay Scoring System: Jess [Ishioka and Kameda, 2004] where the final weighted score is calculated by penalizing a perfect score based on features recognized in the essay.

References

[Baeza-Yates and Ribeiro-Neto, 1999] R. Baeza-Yates and B Ribeiro-Neto. *Modern information retrieval*. New York: Addison-Wesley, 1999.

[Ishioka and Kameda, 2004] T. Ishioka and M. Kameda. Automated japanese essay scoring system: Jess. 2004.

[Landauer *et al.*, 2003] T. Landauer, D. Laham, and P. Foltz. Automated scoring and annotation of essays with the intelligent essay assessor. *Automated Essay Scoring*, 2003.

[Plamondon and Srihari, 2000] R. Plamondon and S. N. Srihari. On-line and off-line handwriting recognition: A comprehensive survey. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(1):63–84, 2000.

[Porter, 1980] M.F Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[Reddy, 2003] R Reddy. Three open problems in artificial intelligence. *Journal of the ACM*, 50(1):1–4, 2003.

[Srihari and Keubert, 1997] S.N. Srihari and E. J. Keubert. Integration of handwritten address interpretation technology into the united states postal service remote computer reader system. *Proc. Int. Conf. Document Analysis and Recognition, Ulm, Germany*, pages 892–896, 1997.

[Srihari *et al.*, 2003] S. N. Srihari, B. Zhang, C. Tomai, S. Lee, Z. Shi, and Y. C. Shin. A system for handwriting matching and recognition. *Proc. Symp. Document Image Understanding Technology, Greenbelt, MD*, pages 67–75, 2003.