Available online at www.sciencedirect.com

ScienceDirect

Artificial Intelligence 172 (2008) 300–324

Artificial Intelligence

www.elsevier.com/locate/artint

Automatic scoring of short handwritten essays in reading comprehension tests

Sargur Srihari *, Jim Collins, Rohini Srihari, Harish Srinivasan, Shravya Shetty, Janina Brutt-Griffler

Center of Excellence for Document Analysis and Recognition (CEDAR) University at Buffalo, State University of New York, Amherst, NY 14228, USA

Received 14 September 2006; received in revised form 27 June 2007; accepted 29 June 2007

Available online 24 July 2007

Abstract

Reading comprehension is largely tested in schools using handwritten responses. The paper describes computational methods of scoring such responses using handwriting recognition and automatic essay scoring technologies. The goal is to assign to each handwritten response a score which is comparable to that of a human scorer even though machine handwriting recognition methods have high transcription error rates. The approaches are based on coupling methods of document image analysis and recognition together with those of automated essay scoring. Document image-level operations include: removal of pre-printed matter, segmentation of handwritten text lines and extraction of words. Handwriting recognition is based on a fusion of analytic and holistic methods together with contextual processing based on trigrams. The lexicons to recognize handwritten words are derived from the reading passage, the testing prompt, answer rubric and student responses. Recognition methods utilize children's handwriting styles. Heuristics derived from reading comprehension research are employed to obtain additional scoring features. Results with two methods of essay scoring—both of which are based on learning from a human-scored set—are described. The first is based on latent semantic analysis (LSA), which requires a reasonable level of handwriting recognition performance. The second uses an artificial neural network (ANN) which is based on features extracted from the handwriting image. LSA requires the use of a large lexicon for recognizing the entire response whereas ANN only requires a small lexicon to populate its features thereby making it practical with current word recognition technology. A test-bed of essays written in response to prompts in statewide reading comprehension tests and scored by humans is used to train and evaluate the methods. End-to-end performance results are not far from automatic scoring based on perfect manual transcription, thereby demonstrating that handwritten essay scoring has practical potential.

Keywords: Automatic essay scoring; Contextual handwriting recognition; Reading comprehension; Latent semantic analysis; Artificial neural networks

1. Introduction

Reading comprehension is an important component of learning in schools. Tasks that require students to write about texts are ubiquitous at all levels of schooling and assessment, and low-performing writers have difficulty with

* Corresponding author.
E-mail address: srihari@cedar.buffalo.edu (S. Srihari).

0004-3702/$ – see front matter
doi:10.1016/j.artint.2007.06.005

such tasks. For example, a recent New York State assessment of fourth grade English language arts asked students to write after reading an essay and a poem about whales, and the prompt clearly specified that students should use information from the texts they had read in their responses. The test prompt and two responses were as follows.

Test Prompt:
Do you think that fishing boats should be allowed in waters where whales swim? Why or why not? Use details from BOTH the article and the poem to support your answer. In your answer, be sure to
– state your opinion,
– explain your reasons for this opinion,
– support your opinion using information from BOTH the article and the poem.
Low Scoring Response:
"They should not be a loud where whale are. Because whale need to swim or they will die".

High Scoring Response:
"I think fishing boats should not be allowed where whales are because the people might hurt the whale or get it in the fishing net and the whale might eat the fish in the fishing net and the people might throw a spear at it. They might even go and kill the whale for no reason what so ever. They might even hurt the whale with the boat and it might get killed that way. That is why I think that fishing boats and not allowed where whales are".

Whereas the second writer presents a relatively full, logically connected, and error free response, the first writer uses information minimally, far from the extent necessary to form a skilled argument.

While electronically written responses are becoming the standard for college level entrance testing, handwritten responses are the principal means in state-wide testing in schools. This is due to issues such as how early to introduce key-boarding skills, academic integrity with closely spaced test stations, network down-time during testing, etc. Since the approach of using handwritten essays in reading comprehension evaluation is efficient and reliable it is likely to remain a key component of learning.

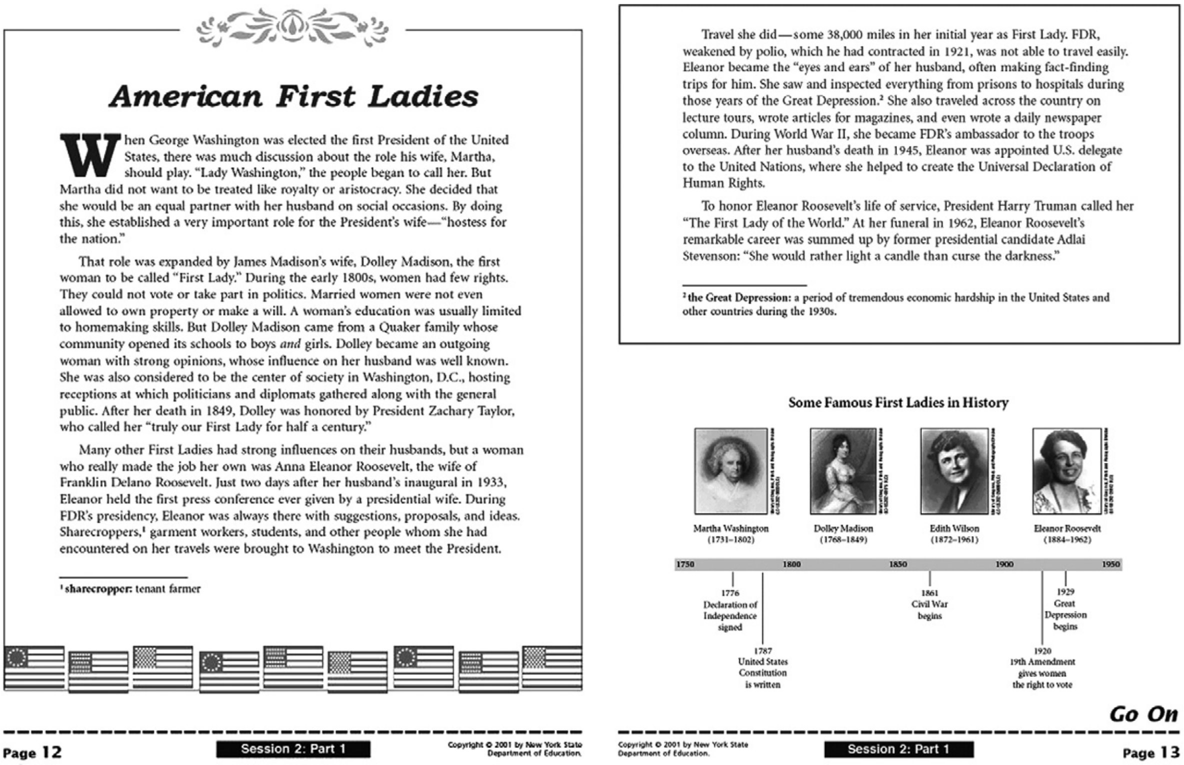
Writing done by hand is the primary means of testing students on state assessments. Consider as an example the New York State English Language Assessment (ELA) administered statewide in grades 5 and 8. In the reading part of the test the student is asked to read a passage such as that given in Fig. 1, which is a grade 8 example, and respond to several prompts in writing. An example prompt is: "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 prompt are given in Fig. 2. The responses are 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.

Assessing large numbers of handwritten responses is a relatively time-consuming and monotonous task. At the same time there is an intense need to speed up and enhance the process of rating handwritten responses, while maintaining cost effectiveness. The assessment can also be used as a source of timely, relatively inexpensive and responsible feedback about writing. The paper describes a first attempt at designing a system for reading, scoring and analyzing handwritten essays from large scale assessments to provide assessment results and feedback. Success in designing such a system will not only allow timely feedback to students but also can provide feedback to education researchers and educators.

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.

From the research viewpoint an automated solution will allow studying patterns among handwritten essays that may be otherwise laborious or impossible. For instance metrics can be obtained for identifying difficulties struggling students are having, for measuring repetition of sections from the original passage, for identifying language constructs specific to the population, etc.

The assessment problem is a well-defined problem whose solution will push forward existing technologies of handwriting recognition and automatic essay scoring. A grand challenge of artificial intelligence (AI) is that of a



American First Ladies

When George Washington was elected the first President of the United States, there was much discussion about the role his wife, Martha, should play. "Lady Washington," the people began to call her. But Martha did not want to be treated like royalty or aristocracy. She decided that she would be an equal partner with her husband on social occasions. By doing this, she established a very important role for the President's wife—"hostess for the nation."

That role was expanded by James Madison's wife, Dolley Madison, the first woman to be called "First Lady." During the early 1800s, women had few rights. They could not vote or take part in politics. Married women were not even allowed to own property or make a will. A woman's education was usually limited to homemaking skills. But Dolley Madison came from a Quaker family whose community opened its schools to boys *and* girls. Dolley became an outgoing woman with strong opinions, whose influence on her husband was well known. She was also considered to be the center of society in Washington, D.C., hosting receptions at which politicians and diplomats gathered along with the general public. After her death in 1849, Dolley was honored by President Zachary Taylor, who called her "truly our First Lady for half a century."

Many other First Ladies had strong influences on their husbands, but a woman who really made the job her own was Anna Eleanor Roosevelt, the wife of Franklin Delano Roosevelt. Just two days after her husband's inaugural in 1933, Eleanor held the first press conference ever given by a presidential wife. During FDR's presidency, Eleanor was always there with suggestions, proposals, and ideas. Sharecroppers,[1] garment workers, students, and other people whom she had encountered on her travels were brought to Washington to meet the President.

[1] **sharecropper:** tenant farmer

Travel she did—some 38,000 miles in her initial year as First Lady. FDR, weakened by polio, which he had contracted in 1921, was not able to travel easily. Eleanor became the "eyes and ears" of her husband, often making fact-finding trips for him. She saw and inspected everything from prisons to hospitals during those years of the Great Depression.[2] She also traveled across the country on lecture tours, wrote articles for magazines, and even wrote a daily newspaper column. During World War II, she became FDR's ambassador to the troops overseas. After her husband's death in 1945, Eleanor was appointed U.S. delegate to the United Nations, where she helped to create the Universal Declaration of Human Rights.

To honor Eleanor Roosevelt's life of service, President Harry Truman called her "The First Lady of the World." At her funeral in 1962, Eleanor Roosevelt's remarkable career was summed up by former presidential candidate Adlai Stevenson: "She would rather light a candle than curse the darkness."

[2] **the Great Depression:** a period of tremendous economic hardship in the United States and other countries during the 1930s.

Some Famous First Ladies in History

Martha Washington (1731–1802)
Dolley Madison (1768–1849)
Edith Wilson (1872–1961)
Eleanor Roosevelt (1884–1962)

1750 1800 1850 1900 1950

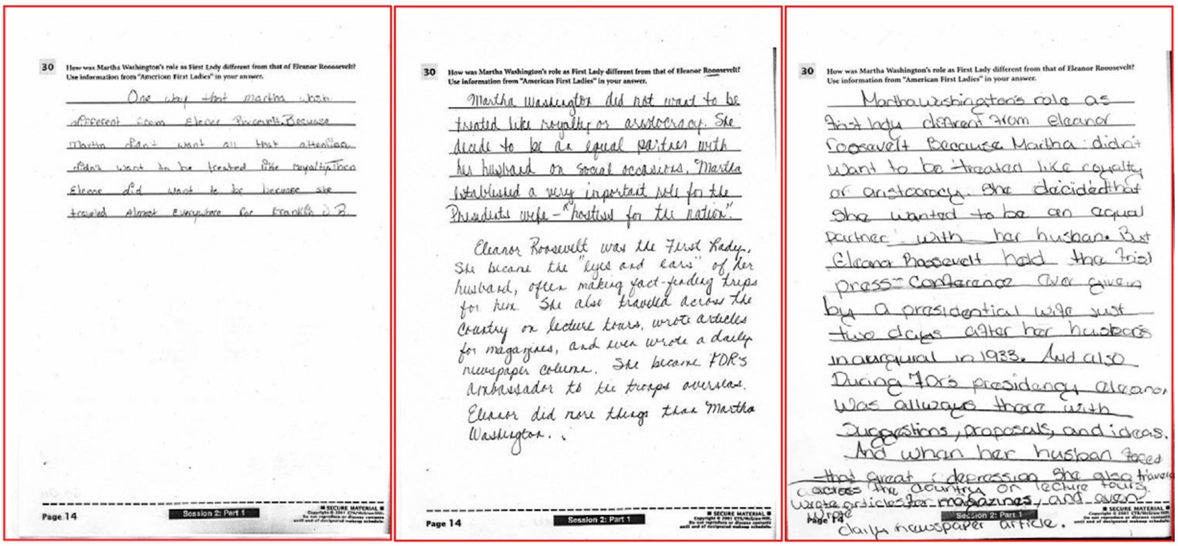
1776 Declaration of Independence signed
1787 United States Constitution is written
1861 Civil War begins
1920 19th Amendment gives women the right to vote
1929 Great Depression begins

Go On

Page 12 Session 2: Part 1 Copyright © 2001 by New York State Department of Education.

Copyright © 2001 by New York State Department of Education. Session 2: Part 1 Page 13

Fig. 1. Text passage to be read. From the New York English Language Arts assessment for Grade 8, 2001—two of three pages of the story "American First Ladies" are shown.



30 How was Martha Washington's role as First Lady different from that of Eleanor Roosevelt? Use information from "American First Ladies" in your answer.

(a)

(b)

(c)

Fig. 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.

computer program to read a chapter in a freshman physics textbook and answer prompts at the end of the chapter [27]. Our challenge is go the other way to evaluate student responses which are handwritten. Much of AI research has progressed in the quest for solutions for specific problems, and this problem promises to be an exciting one both in

Table 1
Holistic rubric chart for the prompt "How was Martha Washington's role as First Lady different from that of Eleanor Roosevelt?"

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 elaborate	Organized	Sketchy	Limited understanding	Understood only sections
Complete, accurate and insightful		Too generalized	Weak		
Focused, fluent and engaging		Facts without synchronization			

terms of the task and its use. Solving the problem also promises to reduce costs and raise efficiency of large-scale assessments.

This is an interdisciplinary project involving three distinct knowledge areas: optical handwriting recognition (OHR), automatic essay scoring (AES) and reading comprehension studies. OHR may first be viewed as largely an engineering enterprise concerning data input. However the challenges posed in deciphering handwriting, particularly that of children, makes it a truly difficult AI task. AES is a topic involving computational linguistics which has practical solutions, however dealing with very noisy textual input calls for new methods. Reading comprehension studies are conducted by education researchers. As any successful AI project demonstrates working with domain experts is key to developing methods and heuristics.

The rest of this paper is organized as follows. Section 2 is a brief review of three areas of research: handwriting recognition, automatic essay scoring and reading comprehension. Choices made in designing different parts of a scoring system are described in Sections 3–4, with the former describing the document analysis and handwriting recognition aspects and the latter describing two different scoring methods (latent semantic analysis and feature-based classification); the methods are illustrated using the Grade 8 prompt described earlier. Section 5 describes the design and evaluation aspects of the methods on a testbed consisting of 300 handwritten responses to the Grade 8 prompt as well as 205 responses to a Grade 5 prompt. Section 6 is a discussion of the results.

2. Previous work

This project involves integrating knowledge and methods from three areas which are reviewed here.

2.1. Handwriting recognition

The first step is that of computer reading of handwritten material in the scanned image of an answer sheet or booklet page. 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.

Handwriting recognition is concerned with transforming an image of handwritten text into its textual form. A survey of both on-line (also called dynamic) and off-line (or static or optical) handwriting recognition is [24]. While the former is now widely used in tablet PCs and PDAs, the latter has been successful only in constrained domains such as postal addresses. To distinguish the two types of handwriting recognition the off-line case is also referred to as optical handwriting recognition (OHR). The higher complexity of OHR stems from the lack of temporal information and the complexity of document analysis.

Word segmentation: The extraction of word images from a page image requires several pre-processing steps to be performed: detecting and eliminating extraneous information such as pre-printed matter, removing ruled lines and margin lines, etc. Separating lines of text and separating individual words is a challenging task that has been addressed in the context of historical manuscripts [21]. Within the handwritten text the ordering of the lines has to be determined and within each line the words need to be segmented. A system for reading unconstrained handwritten pages known as PENMAN was developed [31] which has since been developed into the CEDAR-FOX system for the analysis of handwritten documents for forensic analysis [32]. These systems have tools for gray-scale to binary thresholding, rule line removal, and line/word segmentation. There are interactive user interfaces available for the analysis of the documents by researchers. The CEDAR-FOX system was used in the research described in Section 3.

Word recognition: Once a word image has been isolated it can be subjected to the tasks of character recognition—if the word can be reliably segmented into characters—or segmentation-free word recognition. Recognition of characters and words is performed in a two-step process of feature extraction followed by classification [3]. Word spotting is the task of directly finding key words in a document image [35]. Here the features of handwritten keywords are matched against candidate word images. It does not require segmentation into characters but requires a set of handwritten word prototypes. Features can be either the raw image pixels or shape descriptors. Features can be at the character level (called analytic recognition) or at the word level (holistic recognition). Handwritten word recognition typically involves using a lexicon of possible words. The task becomes one of ranking the lexicon—which can be performed reasonably well for correctly segmented words with small lexicons. The process is error prone for mis-segmented text, large lexicons and words with spelling errors. Word recognition rates for correctly segmented words range from 70% to 95% for lexicon sizes of a few hundred to about 20.

Linguistic constraints: Exploiting statistical dependencies between words was first explored in the OCR domain [13] and then extended to on-line handwriting [29]. Statistical dependencies between word tags corresponding to parts of speech (POS) rather than to words themselves have also been explored.

Handwriting interpretation: Handwriting interpretation is a goal-oriented task where the goal is not so much one of recognizing every character and word perfectly but to perform the overall task in an accurate manner. It involves using basic handwriting recognition tools together with contextual information to solve specific tasks even when there is significant uncertainty in the specific components. For instance, in the domain of postal addresses a system was developed for determining the destination irrespective of whether the individual components were correctly written [20,33]. The strategy was to recognize the most easily recognizable parts of the address first, which in this case consists of the ZIP code and street number. These two "islands" are used to narrow down the lexicon of choices of the street name, which simplifies the task of recognizing the street name. The ZIP code is constrained by city and state names. The mutual constraints lead to a correct interpretation despite spelling errors, mistakes and illegibility. Today, over 90% of all handwritten addresses in the United States are interpreted by OHR. This triangulation is useful for recognition of essay words when constraints imposed by certain words can be used to disambiguate illegible words.

Children's handwriting: Adapting the methods of OHR to children's handwriting is an unexplored frontier. This is attributable to the fact that OHR for general (or adult) handwriting is itself a difficult task. Children's handwriting on the one hand may be easier to recognize due to better formed character shapes. However the layout of words spatially may be poorer whereby the words and lines of text are merged creating significant recognition ambiguity. Also, linguistic constraints on recognition will not work well when there are spelling mistakes and poorly formed sentence constructs.

2.2. *Automatic essay scoring (AES)*

Automated essay scoring has been a topic of research for over four decades. A limitation of all past work is that the essays have to be in computer readable form. A survey of AES methods for electronically represented essays has been made by Palmer et al. [23]. Project Essay Grade (PEG) [22] 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 [4] uses a combination of statistical and NLP techniques to extract linguistic features. Larkey (1998) implemented an AES approach based on text categorization techniques (TCT).

A powerful approach to AES is based on a technique developed in the information retrieval community 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 [18]. 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. A preliminary version of an LSA-based approach to handwritten essay scoring is given in [34].

The Intelligent Essay Assessor (IEA) is the most well-developed and widely used LSA based machine-scoring method [17]. It incorporates other statistical variables that allow it to return more diverse and differentiated scores and tutorial feedback, e.g., coherence, word-choice, plagiarism. IEA does not use a gold-standard. It uses a variant of the nearest-neighbor algorithm. It compares the to be scored essay with 100–200 essays that have been previously scored by expert human readers. It finds a small set thereof (typically around 10) that are the most similar in semantic content to the new essay and applies an algorithm to predict from these what score the same judges would have given the new essay. Thus it is directly mimicking whatever it is humans do. It correlates as closely with human raters as human raters correlate with each other [17].

Hybrid systems, which combine word vector similarity metrics with structure-based linguistic features, are used by ETS to score essays on GMAT, TOEFL, and GRE exams. One such system is the widely-used E-rater [4] which uses a much simpler form of LSA called content vector analysis (CVA). E-rater is trained on about 300 hand-graded essays for each question or prompt. A predictive statistical model is developed from the set of scored essays by beginning with a set of 50–70 features and using stepwise linear regression to select features necessary to assign a score from 1–6. E-rater uses part of speech (POS) tagging and shallow parsing to identify subjunctive auxiliary verbs, more complex clause types (e.g. complement, infinitive, and subordinate clauses), and discourse cue words (e.g., because, in summary, for example). Discourse cue words are used both as individual features and as a way to divide essays into labeled discourse segments. The score is a combination of an overall score and the scores for each discourse segment. Additional features include the presence of words like possibly and perhaps, presence of sentence-initial infinitive phrases, and discourse-deictic words like this and these. To evaluate content and word choice, E-rater uses vectors of word frequencies. The training set is collapsed into six categories (one for each scoring value), and each discourse segment is scored by comparing it with the six categories. The mean of the argument scores is adjusted for the number of arguments (to penalize shorter essays).

Analysis of essays based on linguistic features is of value not only for scoring but also for providing student feedback. Some features are: general vocabulary, passage related vocabulary, percentage of difficult words, percentage of passive sentences, rhetorical features and usage of conjunctions, pronouns, punctuations for connectedness, etc. This approach is employed in the automated Japanese Essay Scoring System: Jess [14] where the final weighted score is calculated by penalizing a perfect score based on features recognized in the essay. C-rater offers automated analysis of conceptual information in short-answer, free responses [19].

Most of the features of advanced technologies such as E-rater and C-rater depend very strongly on accurately representing word sequence. Thus these advanced scoring technologies cannot be expected to work well with the output of handwriting recognition. Simpler methods such as baseline CVA and elaborations/modifications of LSA or feature-based neural networks might be the only feasible ones.

2.3. Reading comprehension

In any AI project domain knowledge is key to success. Studies and findings of reading comprehension are pertinent to: give the context in which automated tools may be useful, obtain heuristics for automatic scoring, meaningfully structure system inputs and outputs, and evaluate system performance.

Reading and writing involve processes of construction, integration, and connection [8,16,28]. Literacy is achieved by using tools to construct, integrate, and connect meanings which are appropriate within cultural settings. Studies of reading comprehension have used guides to writing about reading called thinksheets. They are effective tools for guiding the writing processes of struggling students [6,7,10,26]. Thinksheets and accompanying teacher feedback have had a positive impact on three measures of writing about reading: internal and external connectedness and conventions. Internal connectedness refers to coherence and cohesion in an essay, external connectedness refers to the representation of ideas from the reading in writing, and conventions refer to spelling and mechanics of writing.

Human analysis and prototypical automated analysis of samples from 5th grade students on measures of reading and writing achievement shows improved student performance over time. Essays written at the mid-point of the school year demonstrate higher mean scores for overall quality, internal and external connectedness and conventions. The second key finding pertains to internal connectedness. At the end of the academic year, substantial growth in internal connectedness is seen: the ability to signal without ambiguity relations among clauses, such as logical, causal, coordinate and subordinate. The mean score on internal connectedness rose from $M = 1.70$ to $M = 2.44$ to $M = 3.56$. While substantially increasing the scores on internal connectedness, students were also able to maintain higher quality scores, reached slightly higher values on automated analysis for essay length, conventions, and decreasing scores on overlapping strings of words with the literary selection being written about, when compared to those scores achieved at the mid-point of the year. Their performance on the end-of-the-year assessment suggests a new developmental stage, one characterized by increased scores on internal connectedness, suggesting students' greater control of language and rhetoric and less reliance on direct borrowing or copying from their reading, as compared to the mid-point of the year.

Another relevant work is on measuring coherence. The Coh-metrix [11] system uses a set of 200 features and computes a series of scores that are meant to indicate how coherent a given text would be for a reader. This is accomplished by measuring internal cohesiveness along 50 axes and using LSA to model the mental representation of a reader at particular level (K-12 and college) to which the text will be compared. The assumption here is that high-knowledge readers benefit from gaps in a text's internal cohesion by necessitating inference via previous textual cues and/or the reader's world knowledge, and Coh-metrix can be used to determine how appropriate a text will be for a reader at a particular level. The Coh-metrix system has been designed to reading passages and not student responses.

3. Analysis and recognition of handwritten responses

The process of analyzing handwritten responses begins with the answer sheets being scanned. The standard practice for handwriting recognition is to scan them as gray scale images at a resolution of 300 pixels per inch. Several preliminary image processing steps are first needed. They include: extracting the foreground from the background, eliminating non-informative material such as rule lines and other printed markings, determining the presence of handwritten words, their reading sequence, etc. Several of the operational modules useful for processing scanned essays, viz., rule–line removal, text–line segmentation, word segmentation, word recognition, word spotting and contextual word recognition are briefly described next.

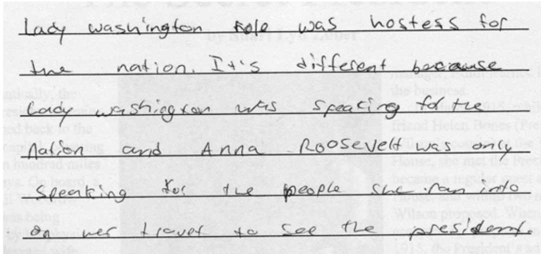
3.1. Rule–line removal

Guide lines are provided in the answer sheets so that the lines of handwriting are straight. However these come in the way of automated recognition and have to be removed from the image. The need for rule–line removal can be eliminated if the lines are printed in an ink that is invisible to the scanner. Since they are indeed present in today's answer sheets they must be dealt with. Line removal algorithms attempt to remove such lines without unduly breaking up the handwriting. One such process detects straight lines using the Hough transform [9]. The process of line removal is illustrated in Fig. 3.

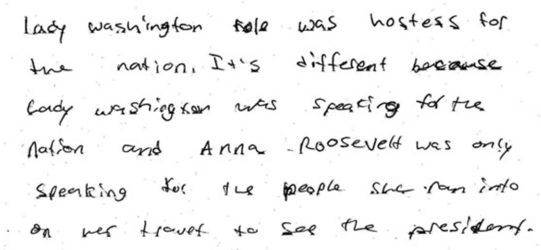
The method basically looks for pixel counts that are high along each direction, by using a polar coordinate representation of the image coordinates, and eliminates such pixels in the image. While it works well for printed straight lines, any guidelines introduced by the writer may cause difficulties due to imperfections.

3.2. Text segmentation

The task of extracting word images is divided into the segmentation of lines and words of handwritten text—although such a division of tasks is not always possible. Line segmentation is a difficult task when the lines overlap. The result of line segmentation when there is no overlap of lines is shown in Fig. 4(a). The algorithm for line segmentation is based on computing the horizontal projection profile for vertical strips of the document image. The valleys in the projection file indicate the presence of line gaps. When components are ambiguous a probability of whether it belongs to the line above or below is employed. When the lines are overlapping a method of cutting through the writing is employed [1].



(a) Original image



(b) Image after guidelines are removed

Fig. 3. Rule–line removal: (a) scanned gray-scale image, and (b) rule–lines detected using the Hough transform and removed.

Word segmentation

Gaps between words are used to segment words of text. Several features are used to perform the classification of whether a gap is a true word gap or not. Examples of such features are: convex hull distance between components, widths and heights of components, sizes of current and neighboring gaps, etc. To determine whether a gap is a true gap or not features are taken into account from the current document rather than solely rely on a learning set. The result of word segmentation performed by using an artificial neural network [15] is shown in Fig. 4(b).

3.3. *Word recognition*

Word recognition is the task of ranking a lexicon of word choices based on the input word image. The shapes of the same word can vary significantly. The method of word recognition adopted was one that combined the results of two processes, one an analytic recognizer and the other a holistic recognizer. The former is based on character shapes and the latter uses global shapes of words. The results of the two methods can be combined using a weighting scheme.

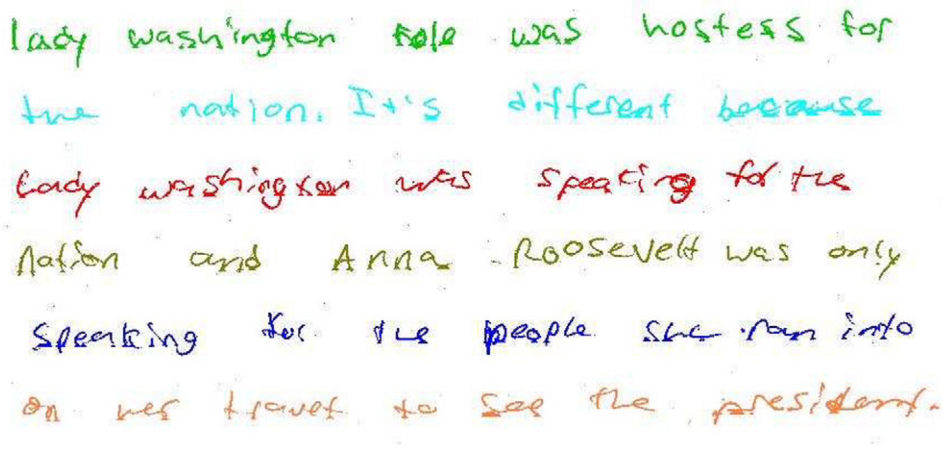
Analytic recognition

This method relies on segmenting words into characters and recognizes them with the aid of a lexicon. Examples of the handwritten word "Eleanor" in student essays in Fig. 5 show different letter formations and inter-letter spacings. For each lexicon entry the word image is divided into the corresponding number of characters and each potential character is then sent to a character recognizer. Word recognition relies on a lexicon of words—which could be derived from several sources, e.g., text passage, prompt, rubric, sample responses, etc. The result of recognition is shown in Fig. 6.

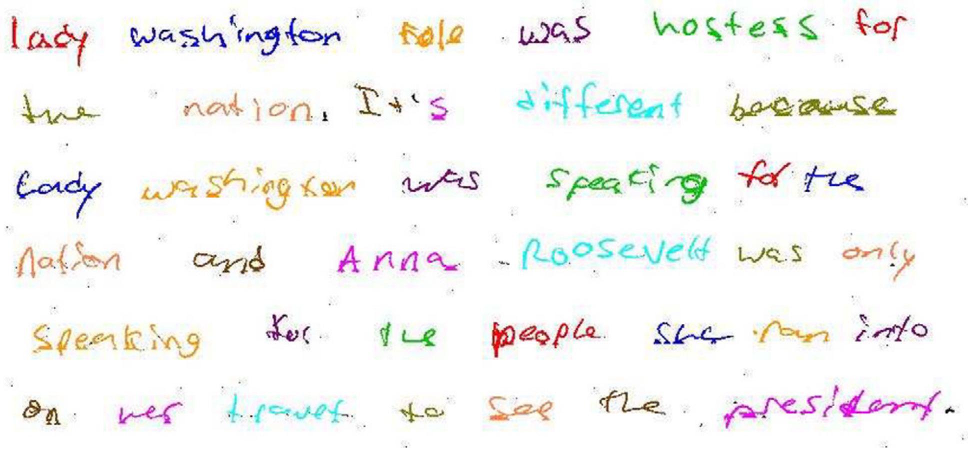
There are four sources for compiling a lexicon: the text passage, the prompt, the answer rubric, and samples of student writing. As an illustration the lexicon compiled from the reading passage of the Grade 8 prompt is shown in Fig. 7. The lexicon has to be increased to accommodate a larger student vocabulary. However performance of a word recognizer decreases with increasing lexicon size.

Holistic recognition and word spotting

The second method uses known prototypes of handwritten words and matches them against segmented words in the scanned image. This is a flexible template matching method known as word spotting. Given a scanned corpus of handwritten responses for system training, prototype templates can be obtained for the words used. These templates can be matched against each word in the query document in a process called word spotting. The number of templates



(a) Lines of text



(b) Words of text

Fig. 4. Segmentation of text lines and words: (a) extracted lines of text, and (b) extracted words.

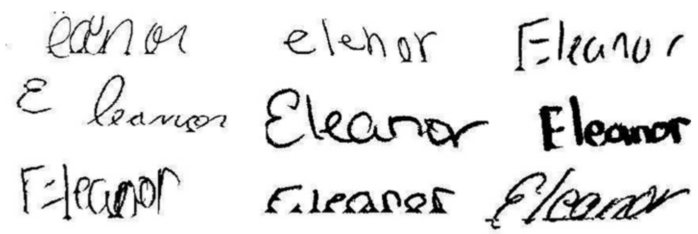


Fig. 5. Samples of handwritten “Eleanor” in handwritten responses.

available depends on the training set available. Examples of templates derived from 150 student responses are shown in Fig. 8—which has only one template for the word “remarkable”, five templates for “doing” and ten for “equal”.

The word spotting algorithm itself is based on extracting a set of 1024 binary features representing the shape of the word image. The shapes are compared using a correlation similarity measure [35]. Each prototype word image is matched against each word in the essay and the results are ranked according to similarity. An example of matching a prototype against all words in an essay are shown in Fig. 9.

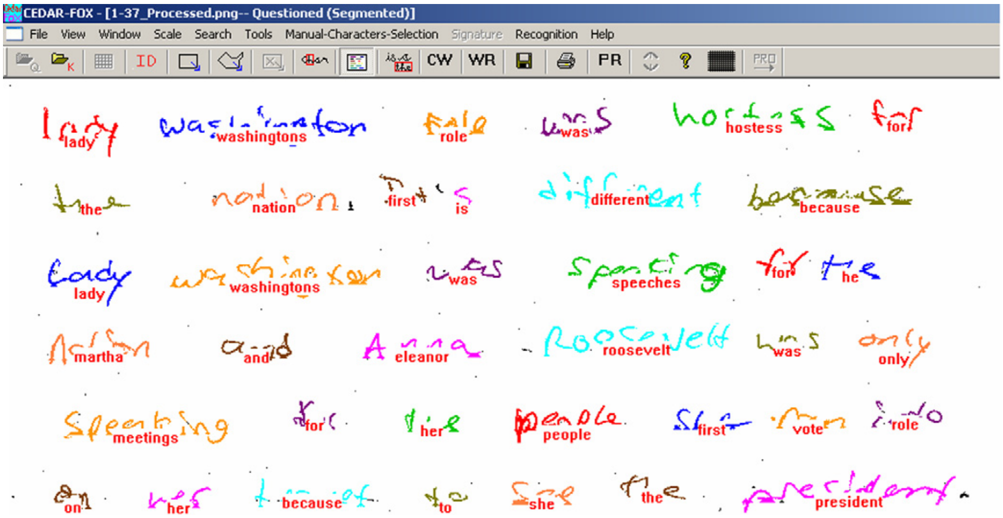


Fig. 6. Word recognition results: words with highest confidence are superimposed on corresponding word images.

1800s	an	center	did	family	held	initial	martha	partner	role	than	us
1849	and	century	diplomats	fdr	helped	inspected	meet	people	roosevelt	that	usually
1921	anna	column	discussion	fdrs	her	its	miles	play	roosevelts	the	very
1933	appointed	community	doing	few	him	james	much	polio	royalty	their	vote
1945	aristocracy	conference	dolley	first	his	job	nation	politicians	saw	there	want
1962	articles	considered	during	for	homemaking	just	nations	politics	schools	they	war
38000	as	contracted	early	former	honor	known	newspaper	presidency	service	this	was
a	at	could	ears	franklin	honored	ladies	not	president	sharecroppers	those	washington
able	be	country	easily	from	hospitals	lady	occasions	presidential	she	to	weakened
about	became	create	education	funeral	hostess	lecture	of	presidents	should	tours	well
across	began	Curse	eleanor	garment	hosting	life	often	press	skills	travel	were
adlai	boys	daily	elected	gathered	human	light	on	prisons	social	traveled	when
after	brought	darkness	encountered	general	husband	like	opened	property	society	travels	where
allowed	but	days	equal	george	husbands	limited	opinions	proposals	some	treated	which
along	by	dc	established	girls	ideas	made	or	public	states	trips	who
also	call	death	even	given	ii	madison	other	quaker	stevenson	troops	whom
always	called	decided	ever	great	important	madisons	our	rather	strong	truly	whose
ambassador	came	declaration	everything	had	in	magazines	outgoing	really	students	truman	wife
american	candidate	delano	expanded	half	inaugural	make	overseas	receptions	suggestions	two	will
	candle	delegate	eyes	harry	influence	making	own	remarkable	summed	united	with
	career	depression	factfinding	he	influences	many	part	rights	take	universal	woman
						married			taylor	up	womans
											women
											workers
											world
											would
											wrote
											year
											years
											zachary

Fig. 7. Lexicon of 276 unique words from the “American First Ladies” reading passage.

Combining analytic and holistic recognition results

The final output of word recognition is obtained by combining the individual distance values returned by the character-based word recognition and word spotting methods for each word in the lexicon. An optimal weighting scheme for the two distance values is found using a simple neuron classifier with two inputs corresponding to the two distance values and a single output specifying the probability that the input corresponds to the distances of a non-matching word. The two distance values were noted for several random choices for both matching and non-matching words from a few sample essays and the optimal weights for the two distances were learnt using a gradient descent algorithm. The final automatic word recognition results are obtained by sorting the sum of the two weighted distance values in an ascending order.

Word or phrase spotting without using analytic recognition becomes an important tool when many of the words cannot be recognized. The presence of a few key phrases can be used in an image feature based approach to scoring.

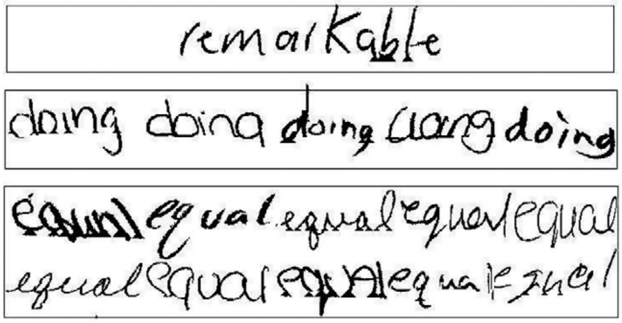


Fig. 8. Word templates: templates for three words are shown.

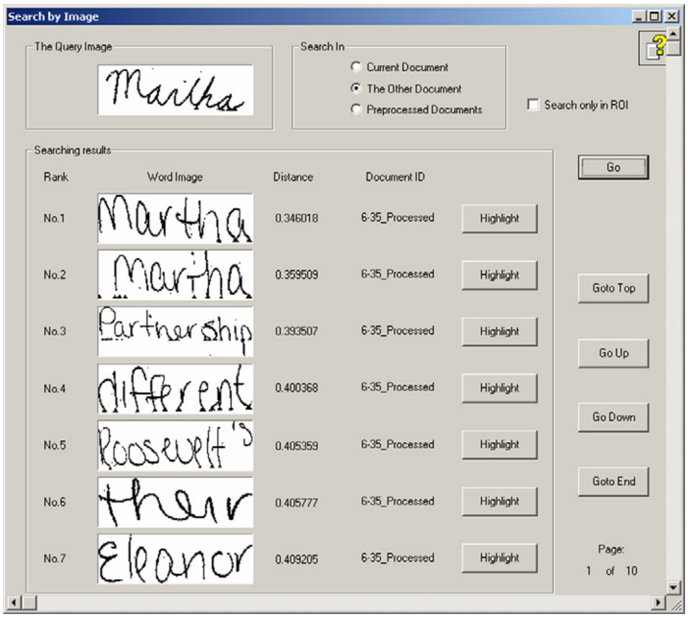


Fig. 9. Word spotting: the query phrase on top returns the images below as the top choices together with their locations.

Transcript mapping

Since word-spotting requires a way of obtaining prototype word images, a method of using a transcription of sample answer essays was used [12]. As an example the following is the transcript for the handwritten essay shown in Fig. 3.

```
Lady Washington role was hostess for the nation.
Its different because Lady Washington was speaking for the nation
and Anna Roosevelt was only speaking for the people she ran into on wer travet
to see the president.
```

It is automatically mapped to the image resulting in the transcript-mapped image shown in Fig. 10(a). Since transcript mapping only gets about 85% of the words right a method correcting the results is needed. An interactive tool for performing the correction task is shown in Fig. 10(b). Here a cursor is automatically positioned under each word image and the user types in the correct truth for the word.

(a) Transcript mapped image (b) Tool for entering word truth

Fig. 10. The transcript of a handwritten essay is used to associate word truth with each word image: (a) transcript mapped image, and (b) interactive tool for correcting transcript map.

3.4. Contextual post-processing

A trigram model based approach is used to correct errors in word recognition. The set of unique words obtained from several sample essays forms the lexicon used for automatic word recognition. The transition probability from a pair of words to any other word in the lexicon can be calculated from these sample essays. The automatic word recognizer outputs a distance to the words in the lexicon which indicates the confidence level with which the handwritten word in the answer document is similar to the one in the lexicon. The word recognizer returns the word with the least distance as the recognized word, but this may result in several words being wrongly recognized. So, the lexicon words with the top m scores for each of the segmented word images in the handwritten answer document are considered for error correction. This is based on the assumption that the actual word is almost always recognized in the top m choices. With this data a Viterbi trellis with T states can be built, where T corresponds to the number of segmented handwritten words in a passage. Thus, using the trigram approach, if there are T number of words in the answer document, then a word at position t depends on the words at position $t-1$ and $t-2$. Given that automatic word recognition on these handwritten answers is done based on the lexicon it is not possible to recognize words from the students own vocabulary that are not in the lexicon. But we can take advantage of possible overlap in strings of words to make sure that at least the sequence of words in common between a student's answer and the sample essays are recognized correctly.

Trigram language model with interpolated Kneser–Ney smoothing

A trigram language model is used to create a transition matrix from the sample essays. This matrix keeps a track of the transition probabilities from any two previous words to the next in the actual passage. For example: If *Eleanor Roosevelt's role* is the combination then

$$P(\mathit{role}|\mathit{Eleanor\ Roosevelt's}) = P(\mathit{Eleanor\ Roosevelt's\ role})/P(\mathit{Eleanor\ Roosevelt's}).$$

The various trigram, bigram and unigram counts from the sample essays are insufficient to give an accurate measure of the transition probabilities due to data sparseness. Smoothing increases low probability values, like zero probabilities, and decreases high probabilities, thereby increasing the accuracy of the model. There are various smoothing techniques [5] among which interpolated Kneser–Ney smoothing has been found to perform well in speech recognition using a perplexity metric. The interpolated Kneser–Ney smoothing interpolates the trigram model with bigram and unigram models and a fixed discount is subtracted from each nonzero count. This technique also ensures that the unigram and bigram counts of a word are not just proportional to the number of occurrences of the word/bigram, instead it depends on the number of different contexts that the word/bigram follows. In this model the probability $p(w_i|w_{i-1},\dots,w_{i-n+1})$, which is the frequency with which word w_i occurs given that the previous n words were $w_{i-1},\dots,w_{i-n+1}$, is given by

$$p_{KN}\big(w_i|w_{i-n+1}^{i-1}\big) = \frac{\max\{[c(w_{i-n+1}^{i}) - D], 0\}}{\sum_{w_i} c(w_{i-n+1}^{i})} + \frac{D}{\sum_{w_i} c(w_{i-n+1}^{i})} N_{1+}\big(w_{i-n+1}^{i-1}\bullet\big) p_{KN}\big(w_i|w_{i-n+2}^{i-1}\big) \tag{1}$$

where n for a trigram model is 3, c is a count of the number of times the n-gram w_{i-n+1}^{i} occurs, D is the absolute discount value set at $n_1/n_1 + 2n_2$ where n_1 and n_2 are the total number of n-grams with exactly one and two counts respectively, and

$$N_{1+}\big(w_{i-n+1}^{i-1}\bullet\big) = \big|\big\{w_i : c\big(w_{i-n+1}^{i-1}w_i\big) > 0\big\}\big| \tag{2}$$

where the notation N_{1+} is meant to evoke the number of words that have one or more counts, and $\bullet$ to evoke a free variable that is summed over.

Viterbi decoding

A second order Hidden Markov Model is used to infer the words of the passage. The segmented handwritten words represent the observed states and the actual words from the lexicon represent the hidden states. The top m results from the output of the word recognition for each segmented handwritten word are the possible hidden states at each time step. The state probabilities at each time step are obtained from the distance values returned by the automatic word recognition. The transition probabilities for a word given the words at the previous two time steps are obtained using the smoothing technique described previously.

Second order Viterbi decoding is used to infer the words of the passage. It evaluates the probabilities of the partial observation sequence ending at time t and the transition between every pair of consecutive states $t-1$ and t. It also stores the previous state at time $t-2$ with the best probability in a vector and outputs the most likely sequence of words. The partial probabilities at each state for a particular word pair is given by

$$\alpha_{jk}(t) = P(s_{t-1} = j, s_t = k, O_{1\ldots t}) = \max_i\big(\alpha_{i,j}(t-1)T_{ijk}\big)A_k(y_t), \tag{3}$$

which is the joint probability of having j as the hidden variable at step $t-1$, k at step t and observing $O_{1\ldots t}$ from time steps $1\ldots t$. Here s refers to the hidden state i.e., the recognized character, T to the transition probability, A to the emission probability and t is the current time step. This can be expressed in terms of α at time $t-1$. T_{ijk} is the probability of transitioning from characters ij to k (similar to a trigram model) and $A_k(y_t)$ is the emission probability of hidden state k emitting the observed y_t at time step t. The state corresponding to the highest probability is the most likely word in the sequence.

Word segmentation errors may lead to erroneous sequences which is handled by adding the null string to the list of m possible states at every time step. This is done to handle cases where a small part of a word has been segmented as a different word, the most likely sequence will probably identify this segment as a null string and the correct word should be identifiable from the remaining part of the segmented word. The transition probabilities are of the form

$$P(s_t|s_{t-1} = \phi, s_{t-2}) = P(s_t|s_{t-2}).$$

The resulting output sequence of words are used for the scoring process.

Example of trigram processing

An example of an essay before and after post-processing is given here. The sample essay after lexicon-based word recognition, word-spotting and their combination, is:

```
lady washingtons role was hostess for the nation
first to different because lady washingtons was speeches for martha
and taylor roosevelt was only meetings for did people
first vote polio on her because to see the president
```

The final result after contextual processing using the trigram model with $m = 15$ is:

```
lady washingtons role was hostess for the nation
but is different because george washingtons was different for the nation
and eleanor roosevelt was only everything for the people
first ladies late on her travel to see the president.
```

While the semantics of the text is garbled in both cases, the number of word errors is fewer after trigram processing as shown in the underlined parts. While there are still several errors in the recognized text, their effect on scoring is the true measure of performance.

4. Automatic scoring methods

When the OHR method produces a reasonable number of correctly recognized words any text-based approach can be used. The LSA approach was chosen since it does not depend as heavily on word sequences as other methods and therefore can be expected to be more robust with word recognition errors. The implementation also uses the nearest-neighbor approach described in [17]. A second approach to scoring was also chosen so that different types of features, semantic as well as image-based, could be used. The features are then used to assign a score using an artificial neural network.

Both approaches need the availability of a training corpus. The training corpus consists of human-scored answer documents. In the training phase the system parameters are learnt from a set of human-scored samples. In the testing phase these parameters are used in scoring.

4.1. Latent semantic analysis approach

The LSA method implemented here is just the core of the approach rather than the full-fledge LSA The flow of processes in the LSA implementation is shown in Fig. 11.

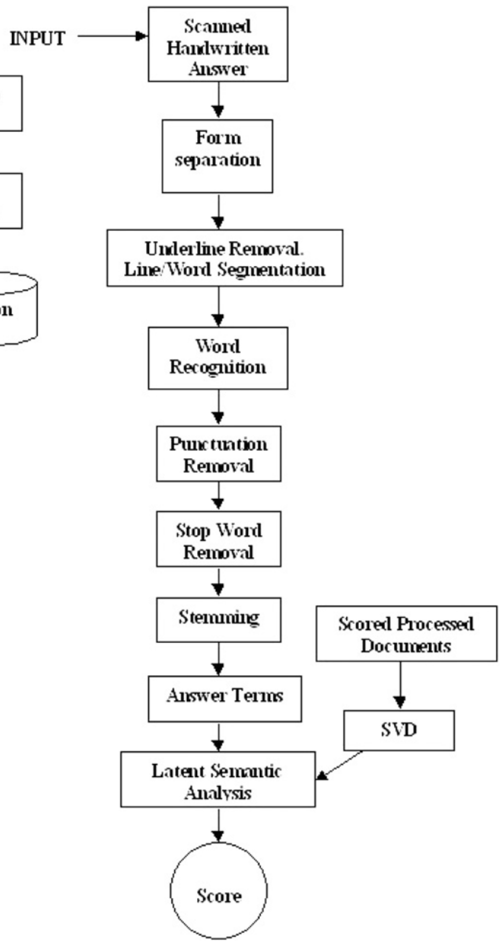


Fig. 11. Operational steps for latent semantic analysis.

After the words in the scanned answer documents are recognized by the OHR system 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 punctuation 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 [25] 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*. The underlying semantics of the training corpus are extracted using LSA and without the use of any other external knowledge. The method captures how the variations in term choices and variations in answer document meanings are related. However, it does not take into consideration the order of occurrence of words. This implies that even if two students have used different words to convey the same message, LSA can capture the co-relation between the two documents. This is because LSA depicts the meaning of a word as an average of the connotation of the documents in which it occurs. It can similarly judge the correctness of an answer document as an average of the measure of correctness of all the words it contains.

Mathematically this can be explained as the simultaneous representation of all the answer documents in the training corpus as points in semantic space, with initial dimensionality of the order of the number of terms in the document. This dimensionality is reduced to an optimal value large enough to represent the structure of the answer documents and small enough to facilitate elimination of irrelevant representations. The answer document to be graded is also placed in the reduced dimensionality semantic space and the by and large term-based similarity between this document and each of those in the training corpus can then be determined by measuring the cosine of the angle between the two documents at the origin.

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. However a general method to determine this optimal dimension is still an open research problem. Currently a brute force approach is adopted. 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 $t \times n$ term-by-document matrix M whose entries are frequencies. SVD or two-mode factor analysis decomposes this rectangular matrix into three matrices [2]. The SVD for a rectangular matrix M can be defined as

$$M = TSD', \tag{4}$$

where prime ($'$) indicates matrix transposition, M is the rectangular term by document matrix with t rows and n columns, T is the $t \times m$ matrix, which describes rows in the matrix M as left singular vectors of derived orthogonal factor values, D is the $n \times m$ matrix, which describes columns in the matrix M as right singular vectors of derived orthogonal factor values, S is the $m \times m$ 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 the dimensionality to a value, say 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 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

$$M_1 = T_1 S_1 D'_1. \tag{5}$$

Standard algorithms are available to perform SVD. To illustrate, a document-term matrix constructed from 31 essays from the *American First Ladies* example shown in Figs. 1 and 2 are given in Table 2. Since the corpus contains 31 documents with 154 unique words, M has dimensions $t = 154$ and $m = 31$.

The first two principal components are plotted in Fig. 12. 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.

Table 2
An example 154×31 term by document matrix M, where M_{ij} is the frequency of the ith term in the jth answer document

Term/Doc	D1	D2	D3	D4	D5	D6	D7	D8	...	D31
T1	0	0	0	0	0	0	0	0	...	0
T2	2	1	2	1	3	1	2	1	...	2
T3	0	1	0	1	3	1	2	1	...	1
T4	0	1	0	0	0	1	0	0	...	0
T5	0	1	0	0	0	0	1	0	...	0
T6	0	1	1	0	0	1	2	0	...	0
T7	0	0	0	0	1	0	0	0	...	0
T8	0	0	0	0	1	0	1	0	...	0
T9	...	...	...	...	...	...	...	...	...	...
T154	0	0	0	0	0	0	0	0	...	0

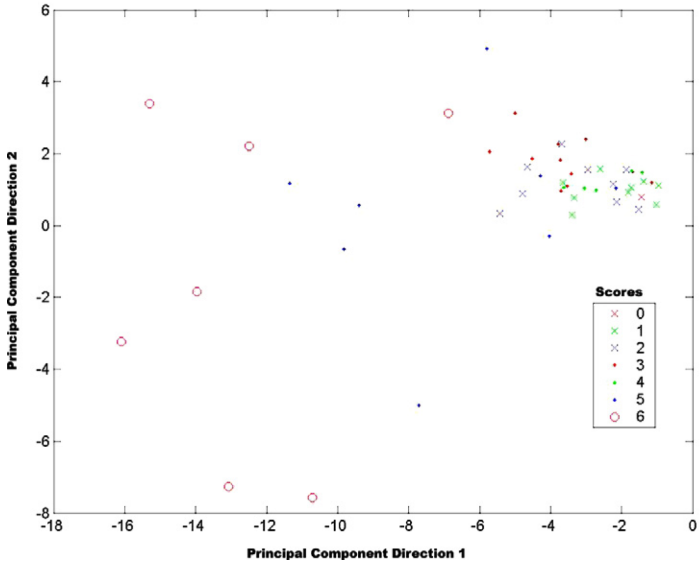


Fig. 12. Projected locations of 50 answer documents in two-dimensional plane.

Data preparation phase. The following steps are performed in the training phase:

1) Answer documents are preprocessed and tokenized into a list of words or terms—using the document preprocessing steps described in Section 3.1.
2) An *Answer Dictionary* is created which assigns a unique file ID to all the answer documents in the corpus.
3) A *Word Dictionary* is created which assigns a unique word ID to all the words in the corpus.
4) An *index* with the word ID and the number of times it occurs (word frequency) in each of the 31 documents is created.
5) A *Term-by-Document Matrix*, M is created from the index, where M_{ij} is the frequency of the ith term in the jth answer document.

Training and validation phases. A set of human graded documents, known as the training set, are used to determine the optimal value of k by employing a leave one out cross validation technique. Each of the queries are passed as validation query vectors and compared with the remaining documents in the training corpus. The following steps are repeated for each document.

1) A vector of term frequencies in the query document is selected as the validation query vector Q.
2) Q is then added as the 0th column of the matrix M to give a matrix M_q.

3) SVD is performed on the matrix M_q, to give the TSD' matrices.
4) Steps 5–10 are repeated for dimension values, $k = 1$ to $\min(t, m)$.
5) Delete $m - k$ rows and columns from the S matrix, starting from the smallest singular value to form the matrix S_1. The corresponding columns in T and rows in D' are also deleted to form matrices T_1 and D'_1 respectively.
6) Construct the matrix M_{q1} by multiplying the matrices $T_1 S_1 D'_1$.
7) The similarity between the query document x (the 0th column of the matrix M_{q1}) and each of the other documents y in the training corpus (subsequent columns in the matrix M_{q1}) are determined by the cosine similarity measure defined as

$$CosineSimilarity = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}}. \tag{6}$$

8) The training documents with the highest similarity score, when compared with the query answer documents are selected and the human scores associated with these documents are assigned to the documents in question respectively.
9) The mean difference between the LSA graded scores and that assigned to the query by a human grader is calculated for each dimension over all the queries.
10) Return to step 4.
11) The dimension with least mean difference is selected as the optimal dimension k which is used in the testing phase.

Testing phase. 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.

4.2. Feature-based approach

An alternative approach to automatic essay scoring is to extract a set of features from the essay. Given a set of human scored essays, the features can be derived from the essays and a classifier can be trained to associate the feature values with a score. Ideally the features themselves are those that have been shown to be effective in studies of reading comprehension as described in Section 2.3.

Some features that can be computed from a transcription of the essay, and which have relevance to score, are: (i) number of words, (ii) number of sentences, (iii) average sentence length, (iv) essay length. Others that can be computed by using an information extraction based approach are: (iv) number of verbs, (v) number of nouns, (vi) number of noun phrases, and (vii) number of noun adjectives [30].

In addition, features can be derived from the answer rubric based on connectivity analysis, i.e., how well concepts are connected in the essay [8]. These include (viii) count of use of "and", "or", "if", "when", "because", etc., (ix) count of bigrams/trigrams from the reading passage, for example in the "Martha Washington" question some of these are: number of mentions of "Washington's role", number of mentions of "different from", and (x) no of uniquely used words.

Once the features of the essay are computed the remaining task is to assign that particular combination of features to a particular score. There are several methods for implementing a classifier based on features, a simple one being an artificial neural network (ANN). The input nodes correspond to the features, the output nodes correspond to each of the possible scores. The design of the ANN for the features described, with four hidden nodes, is shown in Fig. 13. The ANN can be trained to learn weights for each of the connections in the network using a set of scored responses.

While the above feature set is useful for typed or manually-transcribed text, in the case of handwritten input the feature set has to be simpler because of the difficulty of computing them. The modified set of features were: (i) number of words automatically segmented, (ii) number of lines, (iii) average number of character segments in line, (iv) count of Washington's role from automatic recognition, (v) count of differed from, or was different from automatic recognition, (vi) total number of character segments in document, and (vii) count of and from automatic recognition. An example of a handwritten response for the "Martha Washington" prompt along with the ANN features is shown in Fig. 14.

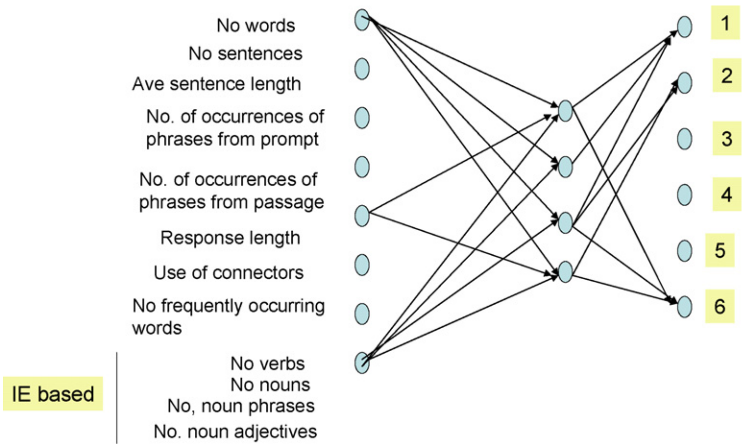


Fig. 13. Artificial neural network to score handwritten essays. are specific to the essay. Examples of phrases used in the features are: Prompt 1—“Washington’s role” and “different from” and Prompt 2—“Nechita’s paintings”.

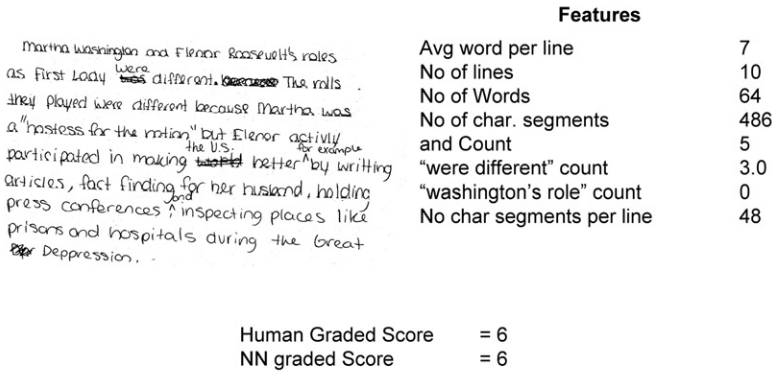


Fig. 14. Example of handwritten response to “Martha Washington” prompt along with ANN features.

Information extraction features

An important role can be played by information extraction (IE) features. IE provides the who, when, where and how much in a sentence. In addition the relationships between entities, and entity profiles including modifiers and descriptions, e.g., “hostess for the nation” can be identified. An example of the result of IE processing for a student essay is given in Fig. 15. This is an output screen of the Semantex™ engine which shows that (i) “Marta Washington” is detected as a person entity, and (ii) “Eleanor Roosevelt” is not a co-reference, i.e., is not the same person. To handle spelling and recognition errors the profiles for “Martha Washington” and “Marta Washington” could be merged by noting spelling and semantic similarities.

5. Evaluation test-bed and performance

The dataset for design and evaluation of the methods described are given here. Two sets of prompts and corresponding responses were used: one from Grade 8 and the other from Grade 5.

1. Prompt 1 (Grade 8): The student was required to read “American First Ladies” (Fig. 1) and respond in writing to the prompt: *How was Martha Washington’s role as First Lady different from that of Eleanor Roosevelt?* There were a total of 300 handwritten responses to this prompt. The score range was 1–6 as indicated by the holistic rubric of Table 1.
2. Prompt 2 (Grade 5): The student was required to read two passages titled “The Languages of Art” and “A Piece of Art” (Fig. 16), both of which concerned the child artist Alexandra Nechita, the second one being the transcript

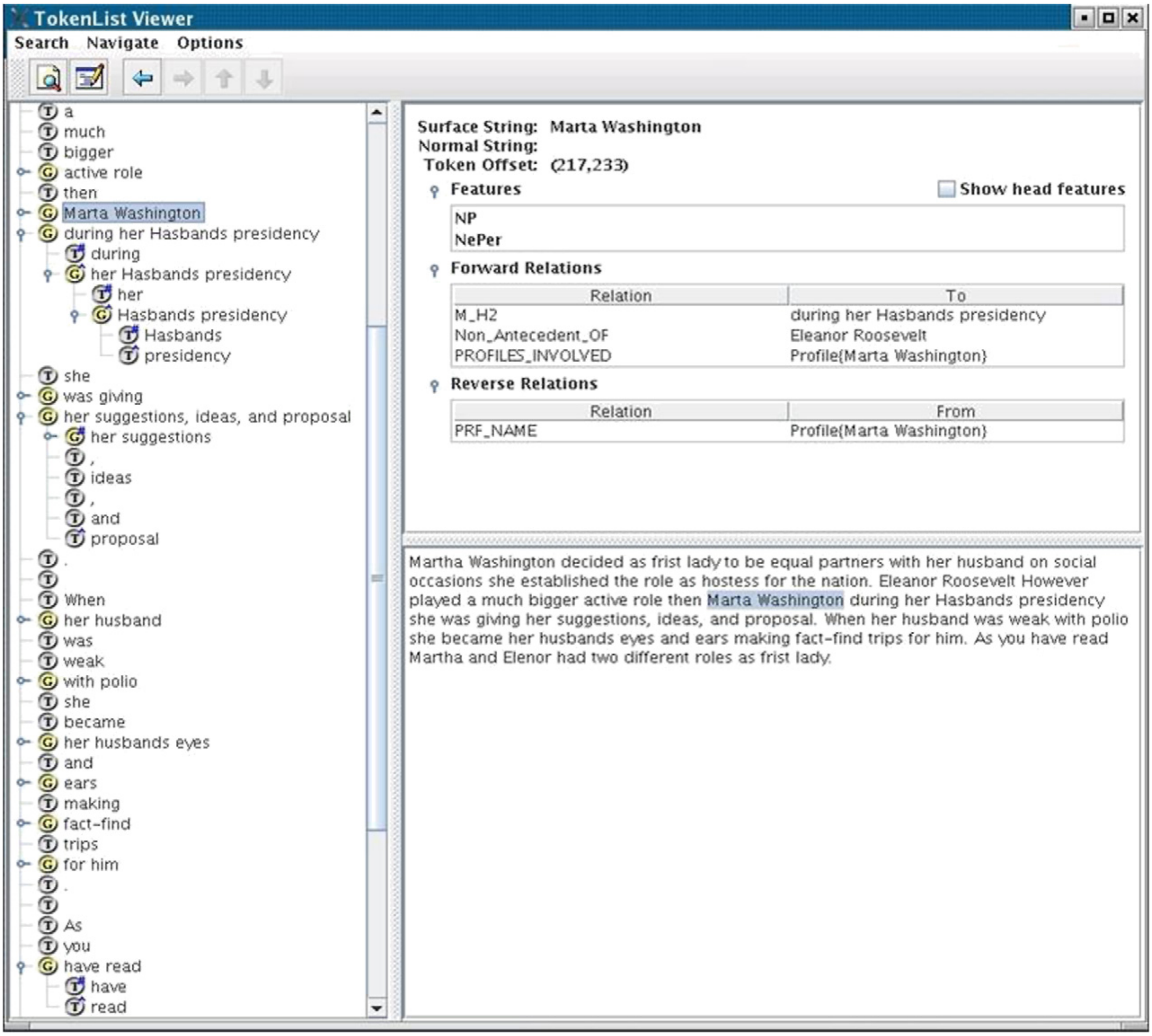
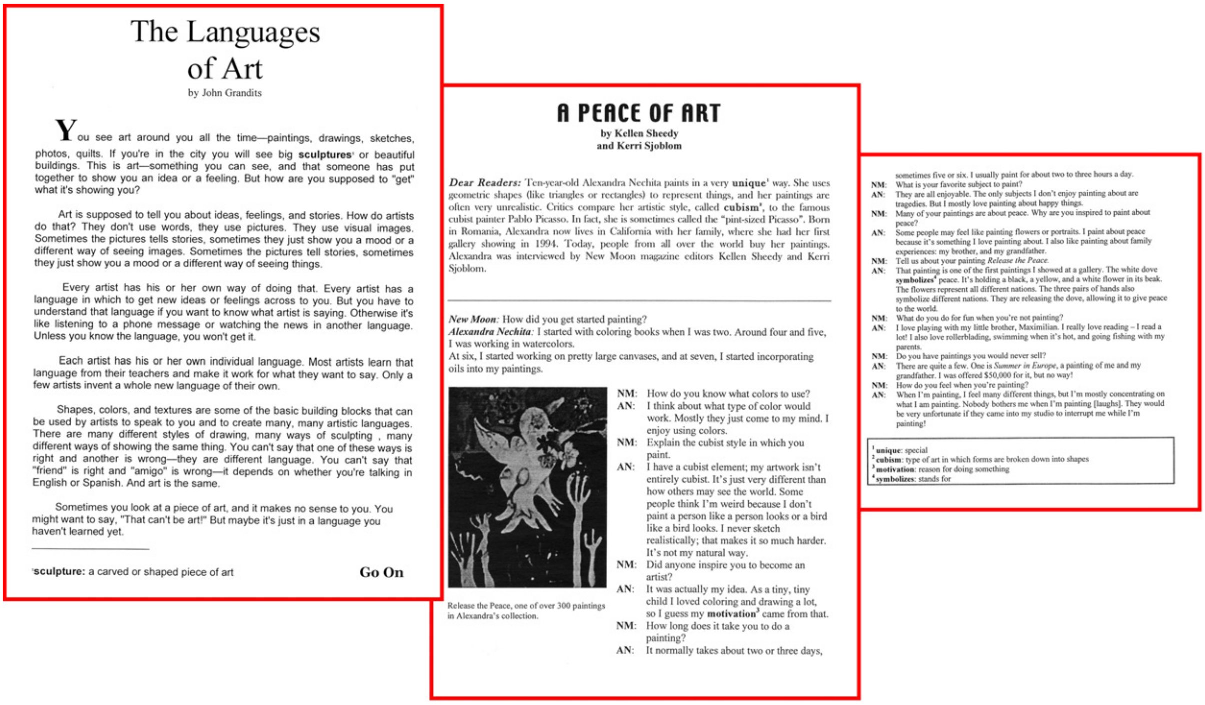


Fig. 15. Information extraction result of processing student essay by Semantex™. The lower right-hand pane shows the essay being processed with "Marta Washington" highlighted. The left-hand pane shows the partial token list that gives the syntactic parse and semantic features where "Marta Washington" is identified as a group (G). The right-hand top pane shows the profile of "Marta Washington" which is identified as a noun phrase with named entity tag of "person" who is different from "Eleanor Roosevelt".

of an interview with Nechita. The test prompt was: *Write a newspaper article encouraging people to attend an art show where Alexandra Nechita is showing her paintings. Use information from BOTH articles that you have read. In your article be sure to include* (i) *Information from Alexandra Nechita*, (ii) *Different ways people find art interesting, and* (iii) *Reasons people might enjoy Alexandra Nechita's painting*. There were a total of 205 handwritten responses to this prompt. The scoring range was 1–4.

Ground truth for the scoring part of the project was created by having each handwritten response scored by two education researchers. There was agreement within one point on 95% of the cases for prompt 1 and 96% of the cases for prompt 2. For the small number of cases where disagreement was greater than one point, full agreement was achieved by a third reading. The final agreed-upon score is referred to as the *gold standard*.



The Languages of Art

by John Grandits

You see art around you all the time—paintings, drawings, sketches, photos, quilts. If you're in the city you will see big **sculptures**[1] or beautiful buildings. This is art—something you can see, and that someone has put together to show you an idea or a feeling. But how are you supposed to "get" what it's showing you?

Art is supposed to tell you about ideas, feelings, and stories. How do artists do that? They don't use words, they use pictures. They use visual images. Sometimes the pictures tells stories, sometimes they just show you a mood or a different way of seeing images. Sometimes the pictures tell stories, sometimes they just show you a mood or a different way of seeing things.

Every artist has his or her own way of doing that. Every artist has a language in which to get new ideas or feelings across to you. But you have to understand that language if you want to know what artist is saying. Otherwise it's like listening to a phone message or watching the news in another language. Unless you know the language, you won't get it.

Each artist has his or her own individual language. Most artists learn that language from their teachers and make it work for what they want to say. Only a few artists invent a whole new language of their own.

Shapes, colors, and textures are some of the basic building blocks that can be used by artists to speak to you and to create many, many artistic languages. There are many different styles of drawing, many ways of sculpting , many different ways of showing the same thing. You can't say that one of these ways is right and another is wrong—they are different language. You can't say that "friend" is right and "amigo" is wrong—it depends on whether you're talking in English or Spanish. And art is the same.

Sometimes you look at a piece of art, and it makes no sense to you. You might want to say, "That can't be art!" But maybe it's just in a language you haven't learned yet.

[1]**sculpture:** a carved or shaped piece of art

Go On

A PEACE OF ART

by Kellen Sheedy
and Kerri Sjoblom

Dear Readers: Ten-year-old Alexandra Nechita paints in a very **unique**[1] way. She uses geometric shapes (like triangles or rectangles) to represent things, and her paintings are often very unrealistic. Critics compare her artistic style, called **cubism**[2], to the famous cubist painter Pablo Picasso. In fact, she is sometimes called the "pint-sized Picasso". Born in Romania, Alexandra now lives in California with her family, where she had her first gallery showing in 1994. Today, people from all over the world buy her paintings. Alexandra was interviewed by New Moon magazine editors Kellen Sheedy and Kerri Sjoblom.

New Moon: How did you get started painting?
Alexandra Nechita: I started with coloring books when I was two. Around four and five, I was working in watercolors.
At six, I started working on pretty large canvases, and at seven, I started incorporating oils into my paintings.

Release the Peace, one of over 300 paintings in Alexandra's collection.

NM: How do you know what colors to use?
AN: I think about what type of color would work. Mostly they just come to my mind. I enjoy using colors.
NM: Explain the cubist style in which you paint.
AN: I have a cubist element; my artwork isn't entirely cubist. It's just very different than how others may see the world. Some people think I'm weird because I don't paint a person like a person looks or a bird like a bird looks. I never sketch realistically; that makes it so much harder. It's not my natural way.
NM: Did anyone inspire you to become an artist?
AN: It was actually my idea. As a tiny, tiny child I loved coloring and drawing a lot, so I guess my **motivation**[3] came from that.
NM: How long does it take you to do a painting?
AN: It normally takes about two or three days, sometimes five or six. I usually paint for about two to three hours a day.
NM: What is your favorite subject to paint?
AN: They are all enjoyable. The only subjects I don't enjoy painting about are tragedies. But I mostly love painting about happy things.
NM: Many of your paintings are about peace. Why are you inspired to paint about peace?
AN: Some people may feel like painting flowers or portraits. I paint about peace because it's something I love painting about. I also like painting about family experiences: my brother, and my grandfather.
NM: Tell us about your painting *Release the Peace.*
AN: That painting is one of the first paintings I showed at a gallery. The white dove **symbolizes**[4] peace. It's holding a black, a yellow, and a white flower in its beak. The flowers represent all different nations. The three pairs of hands also symbolize different nations. They are releasing the dove, allowing it to give peace to the world.
NM: What do you do for fun when you're not painting?
AN: I love playing with my little brother, Maximilian. I really love reading – I read a lot! I also love rollerblading, swimming when it's hot, and going fishing with my parents.
NM: Do you have paintings you would never sell?
AN: There are quite a few. One is *Summer in Europe*, a painting of me and my grandfather. I was offered $50,000 for it, but no way!
NM: How do you feel when you're painting?
AN: When I'm painting, I feel many different things, but I'm mostly concentrating on what I am painting. Nobody bothers me when I'm painting [laughs]. They would be very unfortunate if they came into my studio to interrupt me while I'm painting!

[1] **unique**: special
[2] **cubism**: type of art in which forms are broken down into shapes
[3] **motivation**: reason for doing something
[4] **symbolizes**: stands for

Fig. 16. Two text passages from reading comprehension test for Grade 5, the second of which is an interview transcript.

Ground truth for the recognition part of the project was created by having each of the handwritten responses manually transcribed (MT) into text. Any spelling mistakes were transcribed verbatim. Having the MT responses allows different approaches to be compared, e.g., effect on scoring when there are no OHR errors.

The handwritten responses were divided into training and test sets. For the first prompt, the corpus was divided into 150 training and 150 testing samples with equal numbers for each score. For the second prompt, the corpus was divided into 103 samples for training and 102 for testing, with approximately equal number of samples for each score.

5.1. Training data for LSA

Three textual sources were used to build vocabularies for the learning phase of LSA.

1. Prompt 1 (Grade 8): the vocabulary from 150 (out of 300) responses, along with the words in the reading passage.
2. Prompt 2 (Grade 5): the vocabulary from 103 (out of 205) responses, along with the words in the reading passages.
3. Text book passages: ten long general passages from the text books of Grade 5 and Grade 8. This set was not used for scoring, but merely to increase the vocabulary for the LSA.

The total number of terms by combining all the corpora was 2078.

5.2. OHR performance

For the first prompt, the lexicon for OHR consisted of all words from the 150 training samples, which had a size of 454. The lexicon and the number of word image templates available for each word are shown in Fig. 17. This lexicon is larger than the 276 words in the passage as shown in Fig. 7. The word recognition rate, after combining the analytic and holistic recognition results and then doing trigram contextual processing, was 57%.

For the second prompt, full-fledged OHR was not used. This was because OHR results were poorer with the lower Grade 5 students. Causes for poor handwriting recognition are errors in line segmentation, word segmentation, and lexicon size. Thus the LSA method, which depends only on words, could not be tested in conjunction with OHR.

1790s (1)	1800s (1)	1933 (2)	19th (1)	1st (2)	2 (2)
38000 (2)	a (53)	able (2)	about (4)	accompanying (1)	accomplished (1)
across (1)	act (2)	acted (3)	active (2)	affairs (1)	after (8)
aid (1)	all (5)	allowed (1)	alot (4)	also (24)	although (2)
always (5)	ambassador (6)	amendment (1)	american (1)	amount (1)	an (30)
and (117)	anna (1)	answer (1)	any (3)	anyone (1)	anything (1)
appointed (1)	are (2)	aristocracy (5)	around (2)	articles (5)	as (63)
ass (1)	at (2)	attend (1)	attention (1)	author (1)	awarded (1)
back (2)	basically (2)	be (70)	because (56)	become (6)	before (2)
behind (1)	being (11)	besides (1)	better (1)	between (2)	big (2)
bigger (2)	both (2)	boyz (1)	brought (2)	business (1)	but (11)
buy (1)	by (21)	c (1)	call (7)	came (2)	candidate (2)
candle (1)	care (1)	career (2)	cause (2)	columns (2)	conclusion (1)
conference (14)	considered (2)	constantly (1)	constitution (1)	continued (1)	contributed (1)
control (2)	could (5)	couldnt (2)	country (6)	create (2)	curse (1)
daily (2)	darkness (1)	days (3)	death (2)	decide (5)	decisions (1)
declaration (3)	delano (2)	delegate (4)	depression (3)	did (36)	didnt (15)
died (2)	differ (1)	difference (3)	different (76)	directly (1)	disease (1)
disussion (1)	do (12)	does (1)	doing (5)	dolly (1)	done (1)
dont (2)	during (3)	e (1)	early (1)	earned (1)	ears (6)
easier (1)	economy (1)	education (1)	eleanor (160)	eleanors (2)	elected (2)
election (1)	encountered (2)	equal (32)	er (1)	established (9)	even (7)
events (1)	ever (6)	everybody (1)	everyone (1)	everything (6)	everywhere (1)
exact (1)	example (4)	exploited (1)	express (1)	eyes (7)	fact (2)
fdr (3)	felt (1)	few (1)	fight (3)	finally (1)	finding (2)
first (103)	fist (1)	for (55)	foreign (1)	former (1)	franklin (1)
frequently (1)	frist (2)	from (45)	future (3)	garment (2)	gave (4)
george (4)	get (5)	getting (1)	girls (1)	give (2)	given (8)
go (2)	going (2)	good (2)	got (2)	great (4)	had (24)
half (1)	hand (4)	happen (1)	happy (1)	harry (2)	has (1)
have (10)	he (11)	held (12)	help (20)	helping (2)	her (90)
herself (2)	him (2)	his (7)	hold (2)	hospital (6)	hostess (15)
how (4)	however (2)	huge (2)	human (2)	husban (1)	husband (70)
i (2)	ideas (9)	if (1)	impact (1)	important (4)	in (31)
inaugural (2)	inspect (2)	inspected (2)	inspecting (1)	involved (2)	is (15)
it (7)	james (1)	job (2)	just (13)	kind (2)	kinds (1)
know (2)	known (3)	ladies (4)	lady (65)	late (1)	leave (1)
lecture (2)	less (1)	life (2)	light (3)	like (41)	limited (1)
little (2)	live (1)	lived (2)	lot (3)	loyalty (1)	made (4)
madison (2)	magazines (2)	major (1)	make (3)	making (2)	many (3)
marks (1)	martha (132)	meet (2)	meetings (2)	miles (1)	more (23)
most (2)	mr (2)	mrs (2)	much (5)	my (1)	nation (12)
nations (2)	navy (2)	nearly (1)	need (1)	newspaper (3)	no (2)
normal (2)	not (25)	occasions (6)	of (60)	often (3)	on (22)
one (5)	only (6)	opening (1)	opinion (2)	or (13)	other (7)
out (2)	over (3)	overseas (3)	own (2)	paper (2)	part (3)
particular (1)	parties (1)	partner (23)	parts (1)	past (1)	people (4)
person (2)	pertaining (1)	places (2)	plans (1)	play (2)	played (7)
pointed (2)	polio (5)	politics (2)	presidency (6)	president (29)	presidential (15)
press (14)	prison (5)	probably (1)	problems (1)	propels (1)	proposal (2)
proposals (6)	queen (1)	r (2)	rather (3)	really (5)	reason (1)
reasons (1)	remarkable (1)	responsibility (3)	rest (1)	right (3)	rights (7)
role (89)	roosevelt (107)	royalty (19)	run (1)	said (1)	same (3)
schedule (1)	school (1)	schools (1)	seas (2)	second (2)	see (1)
seem (1)	seemed (2)	self (2)	set (2)	several (1)	shared (1)
she (113)	short (1)	should (2)	show (1)	showing (1)	shows (1)
sick (3)	side (2)	sidekick (1)	signed (1)	significant (1)	since (1)
slavery (1)	small (1)	so (5)	social (10)	soldiers (1)	some (2)
someone (1)	speaker (1)	special (2)	speech (1)	speeches (1)	spoke (1)
spot (2)	standers (1)	standing (1)	started (1)	states (4)	stevenson (1)
still (1)	stood (1)	story (1)	students (2)	such (1)	suggestion (15)
summed (1)	support (1)	supportive (1)	supposed (1)	sweet (2)	take (3)
taking (2)	talk (1)	taylor (2)	text (1)	than (15)	that (30)
thats (2)	the (164)	their (5)	them (2)	then (10)	there (9)
these (2)	they (4)	things (10)	think (3)	this (4)	those (1)
though (3)	time (2)	title (2)	to (124)	together (1)	took (5)
tours (1)	town (1)	travel (6)	traveled (7)	traveling (3)	treat (36)
trips (4)	troops (4)	truman (2)	two (3)	united (9)	universal (2)
unlike (2)	up (7)	us (3)	very (7)	views (1)	visiting (2)
vote (2)	w (2)	want (30)	wanted (24)	wanting (2)	was (136)
washingtons (51)	wasnt (2)	way (4)	ways (2)	weakened (2)	weakness (1)
weather (1)	well (2)	went (3)	were (8)	what (5)	when (11)
where (5)	which (2)	while (3)	who (5)	why (2)	wife (10)
will (1)	with (30)	wives (1)	woman (2)	women (10)	words (1)
work (7)	worked (2)	workers (1)	world (10)	worried (1)	worry (1)
would (8)	write (2)	wrote (3)	year (2)		

Fig. 17. Lexical words for use in OHR for responses from prompt 1. Compiled from 150 student responses, there are 454 words with word counts shown in parentheses.

While the ANN method would also have to do without IE features, the presence of a few key phrases were spotted and used in ANN testing. Such words/phrases can be spotted using an image based method such as word spotting described in Section 3.3.

Table 3
LSA performance: differences between human- and automatically-assigned scores

	Score range	Mean diff.	Diff = 0	Diff ⩽ 1	Diff ⩽ 2	Diff ⩽ 3	Diff ⩽ 4	Diff ⩽ 5
Prompt 1 MT	1–6	1.35	28%	64%	83%	92%	97%	100%
Prompt 1 OHR	1–6	1.58	25%	58%	75%	87%	96%	100%
Prompt 2 MT	1–4	0.96	31%	79%	93%	100%		

5.3. Scoring performance

Both the LSA and feature-based ANN approaches were evaluated. The four scenarios were: (i) manually transcribed with LSA, (ii) OHR with LSA, (iii) manually transcribed with ANN, and (iv) OHR with ANN. Each of the four scenarios was compared to the "gold standard" of human-scored responses. The four cases were evaluated on the two test beds (prompts 1 and 2) mentioned above, with the exception of OHR with LSA on prompt 2, due to extreme poor word recognition on them. It is important to mention here the flexibility of the ANN with respect to the LSA method. The LSA method depends on the entire set of words in the handwritten response to be recognized and thus requires a large lexicon for word recognition. It is a known fact, a larger lexicon implies a poorer word recognition performance and additionally the handwritten responses to prompt 2 (grade 5) were of poorer legibility than those of prompt 1 (grade 8). On the other hand, the ANN method does not require the entire handwritten response to be recognized, but only requires a count of occurrences of few phrases that is useful in making up the features for the ANN. Hence, the lexicon is made up of only those phrases which need to be recognized, resulting in a much better performance of word recognition.

The performance with each scenario can also be compared to a random guess score, which is that of assigning any of the possible scores (1–6 for prompt 1 and 1–4 for prompt 2) randomly to a response. The average difference between a random score and the gold standard for prompt 1 is 2.03 and that for prompt 2 is 1.25. This was evaluated, by using a uniform distribution to model the random score and the expected mean difference was calculated analytically, thereby avoiding the need of statistical tests for significance. Any useful automatic method will have to have a smaller difference than these.

LSA performance

The first set of experiments were performed with the LSA method of scoring. Separate training and validation phases were conducted for the MT and OHR essays. For MT the optimal value of k (best dimension) was determined to be 47 for prompt 1 and 213 for prompt 2. For the OHR essays, the corresponding values for prompt 1 was $k = 50$.

Table 3 summarizes differences between human-assigned scores (the gold-standard) and (i) automatically assigned scores based on MT for prompt 1, (ii) automatically assigned scores based on OHR for prompt 1, and (iii) automatically assigned scores based on MT for prompt 2. The mean differences and percentiles are shown. The interpretation of cell entry for row 1, column 4 is that for prompt 1 with machine transcription, human-assigned and automatically-assigned scores differed by 1 or less 64% of the time. Note that LSA on prompt 2-OHR was not attempted due to poor word recognition.

The LSA method together with OHR performs significantly better than a random guess. It also demonstrates robustness with OHR errors.

ANN performance

A second set of experiments with the ANN method of scoring was performed using both manual transcription and OHR. The ANN score on each response was compared to its human score and the difference determined. With manual transcription (MT), the mean difference between ANN and human scores on the prompt 1 test (150 cases) was 0.79. ANN scores differed from human scores by 1 or less in 82% of the cases. With OHR the mean difference between human and ANN scores was 1.02. In this case 71% of responses were assigned a score $\leqslant 1$, from the true score. For prompt 2, the mean difference with 102 responses for MT and OHR were 0.66 and 0.78, respectively. Table 4 summarizes the results.

Table 4
ANN performance: differences between human- and automatically-assigned scores

	Score range	Mean diff.	Diff = 0	Diff ⩽ 1	Diff ⩽ 2	Diff ⩽ 3	Diff ⩽ 4	Diff ⩽ 5
Prompt 1 MT	1–6	0.79	49%	82%	95%	99%	100%	100%
Prompt 1 OHR	1–6	1.02	33%	71%	94%	99%	100%	100%
Prompt 2 MT	1–4	0.66	44%	89%	100%	100%		
Prompt 2 OHR	1–4	0.78	39%	82%	100%	100%		

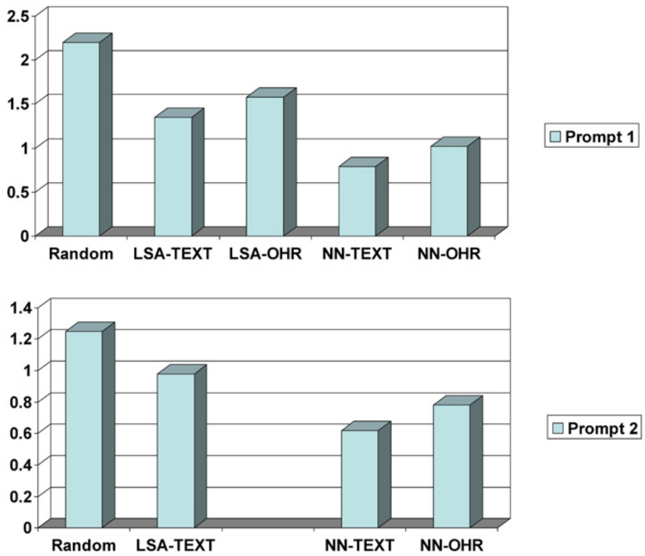


Fig. 18. Differences between human- and automatic-scores for prompts 1 and 2. Results are shown are for: (i) random score assignment, (ii) latent semantic analysis (LSA), and (iii) artificial neural network (NN). For each case mean differences from human scores are shown. LSA-OHR on prompt 2 is absent due to poor word recognition.

5.4. Comparison of LSA and ANN

The ANN numbers in Table 4 are clearly better than the LSA numbers in Table 3. The average differences with the gold standard (human scores) for both methods and both prompts are shown side-by-side in the bar chart of Fig. 18. The differences between human scores and randomly assigned scores are also shown.

With manual transcription, which corresponds to perfect or near perfect OHR, scoring performance is better than with actual OHR. Also ANN performed better than LSA. In particular, the ANN can be effectively used in scenarios where OHR is poor and the LSA method cannot be employed as in the case of prompt 2. The ANN approach does not need the entire handwritten response to be recognized, but only certain words/phrases to populate its features.

The robustness of the ANN method is because it uses features that are both content-independent (e.g., no. of words, no. of sentences, average sentence length) and content-dependent (e.g., count of phrases in response, connectors, nouns, verbs). In fact the ANN method is within a difference of unity from the human score. In testing practice, only if two scores differ by one or more, a third score is used on a response. Thus the ANN method passes the test of being useful as one of two scorers in a practical scenario.

However, there are several caveats concerning the comparison of LSA and ANN. First, we have only conducted a limited testing scenario involving two prompts. Second, some of the features used by the ANN may be coachable, i.e., the test taker can be instructed to use certain phrases, produce a certain length of response, etc. Finally the LSA method described here was trained on a small set of textual passages. In practice LSA systems are trained on a large general corpus of text (or an encyclopedia), plus the training essays, and often use about 300 SVD dimensions (as opposed to 47 and 213 in the experiments described).

6. Summary and discussion

Reading comprehension is an important learning task for children. It is commonly tested with a reading passage and a test prompt that requires the composition of handwritten essays. Handwriting continues to be the standard method of providing responses, as opposed to on-line composition, due to issues such as academic integrity, computer downtime, questions as to how early to introduce keyboarding skills, etc. Given the advances in technologies for automatic essay scoring and handwriting recognition, methods for automatic scoring of handwritten essays can now be explored.

The recognition solution has to contend with not only the standard difficulties of recognizing handwriting but also the writing skills of children. The task involves integrating methods from two very different areas of cognitive science, viz., image/spatial reasoning and computational linguistics. Contextual information is crucial for handwriting recognition, which is available abundantly due to the presence of a reading passage, the prompt as well as potentially a scoring rubric and sample responses. The limited vocabulary and language constructs in a school scenario can also be a positive factor for automatic methods. However poor writing skills and spelling errors are a challenge.

Scoring methods evaluated in this research are LSA and a feature-based approach. The LSA approach has been proven to work well with textual input. However error-free transcription of handwriting-to-text has not been reached by current handwriting recognition technology. The feature-based approach is an alternative scoring solution that can make use of a variety of inputs including image-level features, textual features and features computed by information extraction techniques from textual input.

End-to-end results on a test-bed of handwritten essays on two different prompts using both LSA and feature-based methods show promise. The feature-based methods are usable when recognition rates are very poor. Moreover their scores are within one point of human scored essays—which is a measure of scoring acceptability. State of the art essay scorers such as ETS's e-rater and Person's IEA achieve agreement within one score point for more than 90% of the essays. While the feature-based method appears to perform better than LSA when partial recognition results are available, it may not be a fair test of LSA which in practice is trained on very large text inputs. Also, the feature-based methods need some user input in specifying key phrases which is not needed by LSA. Finally some of the surface features used by the feature-based method may be coachable.

Despite errors in word recognition, scoring performance is promising even if the testing was limited to two prompts. As in other handwriting recognition applications, when evaluation 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 postal address reading where the goal is not so much as to read every word correctly but achieve a correct overall sortation (determine ZIP + 4 Code).

The holistic scoring approaches would need to be extended to analytic scoring, which would attempt to quantify idea development, organization, cohesion, style, grammar, or usage conventions. Such approaches will be more useful for assessing and responding meaningfully to the writing of students to monitor student progress and to provide feedback to guide integrated reading and writing instruction. Language-based methods are likely to play an important role not only in scoring but also in recognition. Information extraction techniques could assist both in front-end handwriting recognition and in back-end essay scoring.

References

[1] M. Arivazhagan, H. Srinivasan, S.N. Srihari, A statistical approach to segmentation of scanned handwritten documents, in: Document Recognition and Retrieval XIV: Proceedings of SPIE, San Jose, vol. 6500, 2007, pp. 6500T-1–6500T-11.
[2] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval, New York, Addison-Wesley, 1999.
[3] R. Bozinovic, S.N. Srihari, A multi-resolution perception approach to cursive script recognition, Artificial Intelligence 33 (2) (1987) 217–255.
[4] J. Burstein, The E-rater scoring engine: Automated essay scoring with natural language processing, in: M.D. Shermis, Y. Burstein (Eds.), Automated Essay Scoring: A Cross-Disciplinary Perspective, Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, 2003.
[5] S.F. Chen, J. Goodman, An empirical study of smoothing techniques for language modeling, in: A. Joshi, M. Palmer (Eds.), Proceedings Thirty-Fourth Annual Meeting of the Association for Computational Linguistics, Morgan Kaufmann Publishers, San Francisco, 1996, pp. 310–318.
[6] J. Collins, Strategies for Struggling Writers, Guilford, New York, 1998.
[7] J. Collins, G.V. Godinho, Help for struggling writers, Learning Disabilities Research and Practice 11 (1996) 177–182.
[8] J.L. Collins, J. Lee, J. Brutt-Gifler, J. Fox, T. Madigan, E. Vosburgh, The writing intensive reading comprehension study, Technical Report, University at Buffalo, June 2006.
[9] R.O. Duda, P.E. Hart, Use of the Hough transform to detect lines and curves in pictures, Communications of the ACM 15 (1972) 11–15.

[10] C.S. Englert, Teaching written language skills, in: P. Cigilka, W. Berdine (Eds.), Effective Instruction for Students with Learning Difficulties, Allyn and Bacon, Boston, MA, 1995, pp. 304–343.

[11] A.C. Graesser, D.S. McNamara, M.M. Louwerse, Z. Cai, Coh-metrix: Analysis of text on cohesion and language, Behavior Research Methods, Instruments, and Computers 36 (2004) 193–202.

[12] C. Huang, S.N. Srihari, Mapping transcripts to handwritten text, in: Proceedings of the 10th International Workshop on Frontiers in Handwriting Recognition, LaBaule, France, Universite de Rennes, 2006, pp. 15–20.

[13] J.J. Hull, Incorporation of a Markov model of syntax in a text recognition algorithm, in: Proceedings Symposium on Document Analysis and Information Retrieval, 1992, pp. 174–183.

[14] T. Ishioka, M. Kameda, Automated Japanese essay scoring system: Jess, in: Proceedings 15th International Workshop on Database and Expert Systems Applications, 2004.

[15] G. Kim, S.N. Srihari, A segmentation and recognition strategy for handwritten phrases, in: Proceedings International Conference on Pattern Recognition, Vienna, Austria, IEEE Computer Society Press, 1996, pp. D510–D514.

[16] W. Kintsch, Comprehension: A Paradigm for Cognition, Cambridge University Press, Cambridge, England, 1998.

[17] T. Landauer, D. Laham, P. Foltz, Automated scoring and annotation of essays with the Intelligent Essay Assessor, in: Automated Essay Scoring: A Cross-Disciplinary Perspective, Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, 2003.

[18] T.K. Landauer, P.W. Foltz, D. Laham, An introduction to latent semantic analysis, Discourse Processes 25 (1998) 259–284.

[19] C. Leacock, M. Chodorow, C-rater: Scoring of short-answer questions, Computers and the Humanities 37 (4) (2003) 389–405.

[20] U. Mahadevan, S.N. Srihari, Parsing and recognition of city, state and ZIP codes in handwritten addresses, in: Proceedings of Fifth International Conference on Document Analysis and Recognition (ICDAR), Bangalore, India, IEEE Computer Society Press, 1999, pp. 325–328.

[21] S. Nicolas, T. Paquet, L. Heutte, Complex handwritten page segmentation using contextual models, in: Proceedings of International Workshop on Document Image Analysis for Libraries (DIAL), Lyons, France, IEEE Computer Society Press, 2006, pp. 46–57.

[22] E.B. Page, Computer grading of student prose using modern concepts and software, Journal of Experimental Education 62 (1961) 127–142.

[23] J. Palmer, R. Williams, H. Dreher, Automated essay grading system applied to a first year university subject—how can we do better? Informing Science (June 2002) 1221–1229.

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

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

[26] T. Raphael, B.W. Kirschner, C.S. Englert, Acquisition of expository writing skills, in: J. Mason (Ed.), Reading and Writing Connections, 1988.

[27] R. Reddy, Three open problems in Artificial Intelligence, Journal of the ACM 50 (1) (2003).

[28] V. Sanjosé, E. Vidal-Abarca, O.M. Padilla, A connectionist extension to Kintsch's construction–integration model, Discourse Processes 42 (2006) 1–35.

[29] R.K. Srihari, S. Ng, C.M. Baltus, J. Kud, Use of language models in on-line sentence/phrase recognition, in: Proceedings of International Workshop on Frontiers in Handwriting Recognition, Buffalo, 1993, pp. 284–294.

[30] R.K. Srihari, W. Li, T. Cornell, C. Niu, InfoXtract: A customizable intermediate level information extraction engine, Journal of Natural Language Engineering (2006).

[31] S.N. Srihari, G. Kim, PENMAN: A system for reading unconstrained handwritten page images, in: Proceedings of Symposium on Document Image Understanding Technology (SDIUT 97), Annapolis, MD, 1997, pp. 142–153.

[32] S.N. Srihari, B. Zhang, C. Tomai, S. Lee, Z. Shi, Y.C. Shin, A system for handwriting matching and recognition, in: Proceedings of Symposium on Document Image Understanding Technology (SDIUT 03), Greenbelt, MD, 2003, pp. 67–75.

[33] S.N. Srihari, E.J. Keubert, Integration of handwritten address interpretation technology into the United States Postal Service Remote Computer Reader System, in: Proceedings of Fourth International Conference on Document Analysis and Recognition (ICDAR 97), Ulm, Germany, 1997, pp. 892–896.

[34] S.N. Srihari, J. Collins, R.K. Srihari, P. Babu, H. Srinivasan, Automatic scoring of handwritten essays based on latent semantic analysis, in: H. Bunke and A. L. Spitz (Eds.), Document Analysis Systems VII, Proceedings of Seventh International Workshop on Document Analysis Systems, Nelson, New Zealand, Springer-Verlag, February 2006, pp. 71–83.

[35] B. Zhang, S.N. Srihari, Word image retrieval using binary features, in: Proceedings of Document Recognition and Retrieval XI, E.H. Smith, J. Hu, J. Allen (Eds.), Proceedings of SPIE, vol. 5296, 2004, pp. 45–53.