Essay Assessment with Latent Semantic Analysis

Technical Report CSRI–440

Tristan Miller
Department of Computer Science
University of Toronto
E-mail: psy@cs.toronto.edu

6 May 2001

Abstract

Latent semantic analysis (LSA) is an automated, statistical technique for comparing the semantic similarity of words or documents. In this paper, I examine the application of LSA to automated essay scoring. I compare LSA methods to earlier statistical methods for assessing essay quality, and critically review contemporary essay-scoring systems built on LSA, including the *Intelligent Essay Assessor*, *Summary Street*, *State the Essence*, APex, and *Select-a-Kibitzer*. Finally, I discuss current avenues of research, including LSA's application to computer-measured readability assessment and to automatic summarization of student essays.

1 Introduction

Few would deny the importance of practice in the development of good writing skills. Like playing a musical instrument, writing is something that cannot be taught by directions or example alone. Practicing writing, and receiving constructive criticism on these attempts, is an integral part of the learning process. Furthermore, the quality of an essay is regarded as one of the best measures of the author's knowledge of the topic. Writing essays requires more thought than many other forms of testing, such as multiple-choice exams, since the students must construct their own coherent answers and justifications therefor. Writing tests not only students' prowess with language, but also their ability to synthesize and analyze information; to find new connections between ideas and to explain their significance.

Unfortunately, assessing student writing and providing thoughtful feedback is extremely labour-intensive and time-consuming. Instructors are often faced with the difficult decision of assigning fewer writing assignments or marking them less thoroughly. Until recently, little thought was given to the idea of automating the essay-scoring process. Early computerized writing assessors focused on mechanical properties—grammar, spelling, punctuation—and on simple stylistic features, such as wordiness and overuse of the passive voice. However, syntax and style alone are not sufficient to judge the merits of an essay.

Enter latent semantic analysis (LSA), a relatively new statistically based technique for comparing the *semantic* similarity of texts. In this report, I survey the use of LSA as a tool for measuring the comprehensibility, coherence, and comprehensiveness of student essays. The next section examines the feasibility of previous computer-based techniques for scoring essays. Following this is a description of how LSA is used to assess essays, and a discussion of some current LSA-based systems. In the final section, I discuss how LSA-based essay scoring is being used in related applications such as readability assessment. I also touch on some future areas of research for LSA in general, and for LSA in education technology in particular.

2 Previous work

2.1 Early readability measures

The notion that some subjective property of a composition can be measured through statistical analysis is not new. As far back as 900 C.E., Jewish scholars reasoned that the more often a word was used generally, the more likely it was to be familiar to readers. By counting the occurrences of words in the Talmud, they produced word frequency lists with which they could roughly assess the readability of any document (Abram, 1981; Taylor and Wahlstrom, 1986).

Interest in readability measures was renewed in the 1920s. At first, only simple word counts were used, but by the 1930s and 1940s research had broadened to stylistic factors such as prepositional phrases and sentence length (Standal, 1987). Dale and Chall (1948) developed one of the most reliable and enduring formulas for predicting readability. It is calculated as follows:

$$\mathit{score} = 0.0496s + 0.1570w + 3.6365$$

where, for a 100-word passage, s is the average sentence length and w is the number of words not on Dale's list of 3,000 familiar words. The score predicts the precise reading level from 4th grade (≤ 4.9) to college graduate (≥ 10.0). The same year, Flesch (1948) devised a readability formula which uses only average word length and average sentence length. Despite being somewhat less reliable than the Dale-Chall formula, it was rapidly popularized as it obviated the need to memorize a word list.

2.2 *PEG*

Because of the time involved in counting words, sentences, and stylistic features, early statistical discourse analysis focused on devising accurate formulas which used the fewest possible factors. With the advent of the computer, however, researchers were emancipated from the tedium of manually compiling complicated statistics. Moreover, the increasing availability and power of computers prompted many to think that machines could soon play a important role in evaluating student writing. The advantages to automated essay scoring in particular were obvious: essays could be marked more quickly, cheaply, and consistently than ever before.

In 1965, Ellis Page developed *Project Essay Grade*, or *PEG* (Page, 1966), the first serious attempt at scoring essays by computer. Recognizing the impossibility of directly measuring the qualitative characteristics which mark a good essay, Page set out to find measurable correlates instead. He expressed this important distinction by coining two terms: *trin*, an intrinsic variable of interest, and *prox*, an independent variable which approximates one or more trins. For example, the proxes of word length variance or proportion of uncommon words may correlate positively with the trin of diction (aptness of word choice). For any given trin, multiple regression analysis is performed on a randomly drawn sample of human-graded essays to determine the extent to which the proxes predict the human scores. The derived weights for each prox may be adjusted to maximize their power in multivariate prediction. The score for the trin in a previously unseen essay can then be predicted with the standard regression equation

$$\mathit{score} = \alpha + \sum_{i=1}^{k} \beta_i P_i$$

where α is a constant and $\beta_1, \beta_2, \ldots, \beta_k$ are the weights (i.e. regression coefficients) associated with the proxes $P_1, P_2, \ldots, P_k$.

Results Page's initial findings were encouraging. When assessing the trin of overall essay quality, *PEG* correlated as well with the human graders as they did with each other ($r = 0.50$). Furthermore, *PEG* could be counted on to consistently assign the same grade to the same essay, unlike the typically erratic ($r = 0.81$) repeat gradings by the same human (Finlayson, 1951).[1] An interjudge correlation of 0.50 is rarely acceptable for important tests, of course—high-stakes essays are usually scored by two independent judges so that individual inaccuracies and biases are suppressed.[2] Judicious choice of proxes and associated β weights has allowed *PEG* to achieve a correlation (shrunken multiple $R = 0.869$) about as high as that of five human judges amongst each other (Page, 1994). This result is even more impressive considering that for large-scale testing programs, routine use of more than two human judges is prohibitively expensive.

Disadvantages *PEG* does have its drawbacks, however. For instance, the system needs to be trained for each essay set used. Users must have access to a large, representative sample of pregraded essays—Page's training data was typically several hundred essays comprising 50–67% of the total number. Also, the scoring method is exclusively relative—the β weights resulting from the multiple regression analysis are applicable only to essays from the same population on which *PEG* was trained.

Another criticism often levelled at *PEG* is that it is susceptible to cheating. While Page and Petersen (1995) acknowledge this as a potential problem, they believe that flagging unusual essays for human inspection would be a minor addition to *PEG*. Furthermore, the proxes used in recent versions are so fine-tuned as to minimize the impact of certain types of cheating. For example, it is well known even to students that essay length correlates with essay quality. There is only so much one can write, however, before becoming off-topic or redundant. For this reason, not essay length but some nth root of essay length is usually used as a prox, since it flattens rapidly as essay length increases. Students who artificially inflate the length of their papers in hopes of fooling the computer may be wasting their time, though a clever cheater might find other ways to manipulate his or her score.

By far the most serious criticism of *PEG*, though, is its use of indirect measures. Skeptics charge that because *PEG* considers only surface features, it cannot reliably judge meaning or coherence. Indeed, even after thirty years of development, *PEG* scores content only as well as a single human grader.[3] While promising, this is not good enough for real-world applications. Page continues to improve *PEG* by seeking proxes more directly related to essay qualities of interest; the latest versions use information from grammar parsers, part-of-speech taggers, and other natural language processing (NLP) tools that did not exist in 1965. Business interests, however, compel him not to disclose details of these improvements.

2.3 *e-rater*

Also worthy of note is *e-rater* (Burstein et al., 1998b,a), an essay-scoring system developed in the mid-1990s by Educational Testing Service. Perhaps attributable to ETS's 1994 collaboration with Page, *e-rater*'s basic technique is identical to that of *PEG*, right down to its use of proxes and regression analysis. Like recent versions of *PEG*, *e-rater* uses NLP tools to extract writing features more fine-grained than simple surface traits, but unlike Page, ETS has been more forthcoming with the details. For example, Miltsakaki and Kukich (2000a,b) explain how the program employs centering theory to measure textual coherence: the syntactic role of referents is tracked across successive sentences, allowing detection of abrupt shifts in topicality. The number of rough shifts is then incorporated into *e-rater*'s scoring model.

Vector-space model Of particular relevance is *e-rater*'s use of a vector-space model to measure semantic content. Originally developed for information retrieval, the vector-space model starts with a co-occurrence matrix where the rows represent terms and the columns represent documents. Terms may be any meaningful unit of information—usually words or

[1] Page and Petersen (1995) cite an abysmal human re-mark consistency of 0.72, but this figure is nowhere to be found in either of their references (Coffman (1971) and Hopkins et al. (1990)).

[2] The Spearman-Brown formula, $\hat{r} = \frac{nr}{1+(n-1)r}$, predicts the increased reliability when one uses n times as many judges with mean correlation $r > 0$.

[3] To assess essay content, *PEG* counts keywords and their synonyms, which must be manually compiled for each essay set.

short phrases—and documents any unit of information containing terms, such as sentences, paragraphs, articles, or books. The value in a particular cell may be a simple binary 1 or 0 (indicating the presence or absence of the term in the document) or a natural number indicating the frequency with which the term occurs in the document. Typically, each cell value is adjusted with an information-theoretic transformation so that it more properly reflects the importance of the term within the document. Document vectors are then compared with each other using some mathematical measure of similarity, such as the cosine coefficient or the Spearman rank coefficient.

In *e-rater*'s case, each "document" of the cooccurrence matrix is the aggregation of pregraded essays which have received the same grade for content. The rows are composed of all words appearing in the essays, minus a stop list of some function words. After an optional information-theoretic weighting, a document vector for an ungraded essay is constructed in the same manner. Its cosine coefficients with all the pregraded essay vectors are computed. The essay receives as its topicality score the grade of the group it most closely matches.

Advantages One of the biggest advantages *e-rater* has over *PEG* is its modular design. Feature identification is divided into three independent modules, one each for syntax, discourse, and topicality analysis. One obvious benefit of this arrangement is the relative ease with which the system can adapt to new data sets—though the topicality module must be retrained for each new essay topic, the other two do not. Another benefit is that *e-rater* can more easily pinpoint the problems with essays and thus provide better feedback; by contrast, *PEG* performs best when it evaluates only holistic essay quality. Also, Burstein and Marcu (2000) explain how *e-rater*'s modularity allows it to be readily adapted to applications such as automatic summarization and scoring of short-answer tests.

Disadvantages Because *e-rater* shares many of *PEG*'s features, it also shares many of its shortcomings. The system requires a sample of pregraded essays on which to be trained, and can therefore make only relative comparisons. Its topicality assessment is (presumably) more sophisticated than *PEG*'s, but even so, *e-rater* is much less successful at judging content than overall essay quality. While agreement with humans on holistic quality is around 0.89, agreement on content may be as low as 0.69. This may be explained in part by an inherent flaw of the vector-space model: it does not account for synonyms. For example, consider a class who are asked to write reports on German shepherds. Students who consistently refer to the dogs as "Alsatians" will be unfairly penalized by *e-rater* if the training essays do not use that term.

3 LSA-based measurement

3.1 LSA described

Latent semantic analysis (Deerwester et al., 1990; Landauer et al., 1998) is a technique originally developed for solving the problems of synonymy and polysemy in information retrieval. Its basic assumption is that every document has an underlying semantic structure, and that this structure can be captured and quantified in a matrix. LSA is unusual among NLP techniques in that it makes no use of human-constructed parsers, taggers, dictionaries, semantic networks, or other tools. The input is simply a collection of documents separated into words or meaningful terms.

LSA is based on the vector-space model discussed previously, but it extends the model in a very important way. Specifically, it exploits *singular value decomposition*, a well-known proof in linear algebra which asserts that any real-valued rectangular matrix can be represented as the product of three smaller matrices of a particular form. The first of these matrices has rows corresponding to the rows of the original matrix, but with m linearly independent columns. The third matrix has columns corresponding to those of the original, but with m rows, also orthogonal. In the middle is a diagonal $m \times m$ matrix of singular values; its purpose is to scale the factors in the other two matrices such that when the three are multiplied, the original is perfectly recomposed. Things get more interesting, however, when *fewer* than the necessary number of factors are used. When one deletes from the scaling matrix one or more of the smallest singular values, the reconstructed matrix is a least-squares best fit to the original.

Taken in the context of a term-document co-occurrence matrix, this means that many terms may appear with greater or lesser frequency in the reconstructed matrix than they did originally. In fact, certain terms may appear at least fractionally in documents they never appeared in at all before. The apparent result of this smearing of values is that the approximated matrix has captured the latent transitivity relations among terms, allowing for identification of semantically similar documents which share few or no common terms withal.[4] The usefulness of this property becomes apparent when one considers that two people will use the same word for a well-known referent less than 20% of the time (Furnas et al., 1983).

3.2 Application to essay scoring

Though LSA was not created as an educational tool, it was not difficult to see how its prowess at judging semantic similarity might be applied to essay scoring. Since essays are typically designed to assess a student's knowledge, and since this knowledge comes from reading a text, the degree of similarity in meaning between the essay and the text should be a good measure of essay quality.

Experiments To test this hypothesis, Foltz (1996) compared essay scores assigned by humans to those assigned by LSA. Four history graduate students were enlisted to grade 24 essays that were based on 21 short texts on the Panama Canal. The graders were instructed to study the source material and then score the essays as they normally would for an undergraduate history class. They were furthermore asked to select from the source texts the ten sentences they considered most relevant to the essay topic.

LSA was then employed to measure the semantic similarity between each essay and the source material. This was done by comparing each sentence in each essay to all sentences in the original texts. The test sentence was given as its score the cosine between it and the most closely matching source sentence. The essay's final grade was its mean sentence score. An alternative essay grade was computed in the same way using only the ten key sentences chosen by the grader.

The results were promising—the mean statistically significant ($p < 0.05$) correlations between the two LSA measures and individual humans were 0.515 and 0.608. These compare favourably with the mean statistically significant human intercorrelation of 0.584. Moreover, the experiment suggests that a significant proportion of the variance in student essays is attributable to the degree of semantic overlap with certain key concepts the graders have in mind.

In a subsequent study by Landauer et al. (1997), LSA grading was performed with a method similar to that used by *e-rater*: each ungraded essay was compared holistically (as opposed to sentence-by-sentence) by cosine with a set of pregraded essays, and received as part of its score the mean grade of the closest ten pregraded essays. The other part of its score was determined by the essay's vector length, which may be interpreted as the breadth (quantity) of information relevant to the topic.[5] It was found that LSA correlated as well with two professional graders combined as the two did with each other ($r = 0.77$). The same study also tested an absolute scoring method which made no use of pregraded essays: the essays were scored with their cosine (again holistically) with a short text on the same topic written by an expert. LSA's performance was almost as good as with the first method ($r = 0.72$).

Analysis LSA's correlation with human graders may seem high considering that it measures only semantic similarity—syntax and morphology are completely ignored. This is not so surprising, however;

[4] Likewise, terms may be compared by examining their vectors across documents. Terms may be judged semantically similar even though they never occur in the same text together.

[5] By contrast, the cosine coefficient measures *depth* (quality) of content—that is, how detailed a treatment the essay gives. The reader is referred to Rehder et al. (1998) for an explanation of the various LSA measures.

empirical studies (Burstein et al. (1998a); Landauer et al. (2000)) have shown that human graders give much greater weight to content than to style or mechanics. Perhaps this is because, as Landauer et al. (1997) posit, syntax is more a convenience than a necessity for the transmission of ideas. Even so, treating an essay as simply a bag of words can mask important errors in logic. For instance, a student writing on the human heart might consistently confuse "left" and "right" when referring to ventricles. With a human grader, this mistake might carry a heavy penalty; LSA would fail to notice it altogether. (To be fair, neither *PEG* nor *e-rater* could catch such an error either.)

Also because of LSA's ignorance of syntax or morphology, it cannot judge most matters of mechanics and style (e.g. spelling, grammar, clichés, tense shifts). Textual coherence, however, is a stylistic feature LSA is particularly well-suited for measuring. In a coherent, fluid document, adjacent sentences in the same paragraph should be semantically similar, while any two paragraphs should exhibit less similarity the further they are apart. By comparing the vectors of pairs of sentences or paragraphs, LSA can pinpoint shifts in topicality far more effectively than simple term-term overlap measures (Foltz et al., 1998). How LSA compares to *e-rater*'s sophisticated, syntax-driven coherence analysis has yet to be established, though it is certainly possible to make some predictions. Take the following two equivalent sentence pairs:

- *She called out to the mastiff. The dog swivelled its ears and looked up.*
- *She called out to the mastiff. It swivelled its ears and looked up.*

Given the strong semantic association between *mastiff* and *dog*, a sufficiently trained LSA would have no trouble finding coherence in the first pair. On the other hand, the hyponymically ignorant *e-rater* would fail to find any common referent. The situation is reversed with the second pair, where only a syntax-aware system would see the link between *it* and its antecedent. Clearly, adding anaphora resolution to LSA would make for an interesting study.

One great advantage LSA has over its predecessors is its ability to make absolute as well as relative comparisons—student papers can be compared to each other, or to a single authoritative source such as an expert's essay or a collection of textbooks. LSA has effectively obviated the expense of pregrading hundreds of essays as training data; the benefits of cost-effective computer grading might now be realized for smaller-scale operations. Of course, this argument presupposes that the training data is available in electronic form, which is not always the case.

LSA has a couple of other disadvantages of note. First, the process of singular value decomposition is computationally expensive—decomposing a large matrix can take days of computer time. This problem is aggravated by the second disadvantage, that determining the number of dimensions by which to reduce the scaling matrix is somewhat of a black art. Too little a reduction reconstructs the original matrix too faithfully to capture any latent semantic information; too large a cut renders the matrix too noisy to be useful. The optimal dimensionality must be determined empirically. Once a suitable degree of reduction has been discovered, however, two documents can be compared in time linear to the number of terms.

3.3 Systems using LSA

Unlike *PEG* and *e-rater*, LSA is not a complete essay-scoring system by itself. However, it has proved to be an excellent tool for assessing content, and has therefore been incorporated into a number of contemporary essay-scoring systems. I now present and compare some of these systems.

3.3.1 *Intelligent Essay Assessor*

The *Intelligent Essay Assessor* (*IEA*) (Foltz et al., 1999; Landauer et al., 2000), developed by Knowledge Analysis Technologies, is an LSA-based system which combines essay scoring with tutorial feedback. Essays are assessed on content, which is handled by LSA; mechanics, which is handled by corpus-statistical measures; and style, to which both types of measures contribute. There are also components for validation and plagiarism detection, which flag papers that are nonsensical, copied or paraphrased from other essays, conspicuously seeded with buzzwords, or otherwise unusual.

Because of the massive computing and memory requirements imposed by LSA, *IEA* is currently offered as a Web-based application only. Students enter their essays into an online form and, a few seconds after submission, receive an estimated grade along with suggestions for revisions. The mechanics module indicates misspelled words and grammatical errors, the style module comments on redundant sentences and organizational issues, and the content module identifies material which is irrelevant to the topic. Custom-built *IEA* interfaces, such as the one available at `http://psych.nmsu.edu/essay/`, can even identify exactly which subtopics received inadequate treatment and point students to relevant textbook chapters.

IEA assigns its scores for content on the basis of the essay's similarity to a standard such as a textbook, or to other student essays. Interestingly, the comparison essays need not be pregraded; given at least two hundred unmarked essays, *IEA* can align them on a continuum of quality and use that as a basis for its grading. If pregraded essays are used, only half as many are required for optimal calibration. Whatever the method used, *IEA*'s scores for content tend to correlate highly with humans ($r = 0.83$), while holistic scores are slightly higher (up to $r = 0.90$). Performance for style and mechanics are considerably worse, with correlations of 0.68 and 0.66, respectively.

IEA's greatest success, though, lies in its reception by students. By repeatedly submitting and revising their writing, students can improve their essays until they are satisfied with them. An *IEA* study at New Mexico State University found a mean increase of 7 points (out of 100) over an average of three essay revisions. More importantly, 98% of the students involved in the study expressed satisfaction with the system and a desire to use it again for other courses.

Further details and two trial versions of *IEA* are available on the company's website, `http://www.knowledge-technologies.com/`.

3.3.2 ***Summary Street*** and ***State the Essence***

The mandate of *State the Essence* (Kintsch et al., 2000) is a bit narrower in scope than that of *IEA*: it was designed to improve elementary school students' summarization skills, helping them mediate the conflict between concision and comprehensiveness. Like *IEA*, though, *State the Essence* is interactive, encouraging students to adopt a submit-revise cycle using system-provided feedback. Its approach is content-driven—treatment of grammar and most stylistic features is lacking. After an initial spell-check, LSA is used to measure topic coverage, irrelevancy, and redundancy. The first is computed by taking the cosines of the essay with each section of the source text; sections with particularly low cosines are deemed to have inadequate coverage. Similarly, irrelevancy is measured by figuring the cosines of the source text with each sentence of the essay. Redundant sentences are identified by computing cosines among each pair of sentences in the essay. A total word count is also taken as a measure of concision. When an essay is submitted, the program returns a numeric grade for overall topic coverage along with comments on length and particular problem areas. Students can revise and resubmit as often as they like; once they are satisfied with the feedback, they submit their papers to the teacher for complete grading.

Initial trials of *State the Essence* indicated a number of areas for improvement. Overall correlation with humans was inconsistent (variously $r = 0.32, 0.88, 0.64$), and there was no evidence that use of the program resulted in increased writing skills or learning. More seriously, students tended to forget or ignore the fact that the program was evaluating content only; preoccupation with the numerical score incited many students to abandon good writing style in favour of increasing their score by the cheapest means possible. They received heavy penalties for organization and mechanics upon human grading.

Hypothesizing that the bulk of the problem lay in the feedback mechanism, Kintsch et al. (2000) revised the system so that the numeric scores were displayed graphically. Content for each section is now represented as a simple bar whose length indicates coverage. Changes were also made to how and when advice on redundancy and relevancy is presented. Trials of the new version, renamed *Summary Street*, found some improvement in the way students used the program, but no gain in writing ability when tested on easy texts. For difficult texts, however, teacher-assigned grades for summaries developed with the system were significantly higher than those written without it ($t(50) = 2.32, p.02$). This has prompted its developers to consider further trials with more advanced students, such as college undergraduates, who often deal with difficult source material.

An online trial of *Summary Street* is available at `http://lsa.colorado.edu/summarystreet/`.

3.3.3 APex

APex (Dessus and Lemaire, 1999; Dessus et al., 2000) is an interactive learning environment developed at the Université Pierre Mendès France in Grenoble. Available as a Web-based application, it uses LSA to assess student essays on topic coverage, discourse structure, and coherence. The system was originally developed with French-language text, though apart from the standard stop list, the LSA module required little modification.

The topicality module of APex differs from that of *IEA* and *Summary Street* in that the partition of the source text into topics is much more fine-grained. For calibration, the teacher must identify *notions*—short passages of text which exposit a certain key concept—and the topic or topics to which each notion belongs. For an essay on a given topic, APex computes the cosine coefficient between the essay and each relevant notion, the average of which forms the final score. (This method hearkens back to the key-sentence technique used by Foltz (1996).) This content score has good correlation with human scores for content ($r = 0.64$) and overall essay quality ($r = 0.59$).

Using notions, APex is able to construct an outline view of an essay, the purpose of which is to aid the student in planning his discourse and highlighting areas of concern. The outline is produced by having LSA find and print each essay paragraph's closest corresponding notion. If no notion correlates above a certain threshhold, the paragraph is flagged as potentially irrelevant. The completed outline allows for easy identification of possibly repetitious sections by the student himself. Following Foltz et al. (1998), APex also performs coherence analysis by comparing adjacent sentence pairs and reporting abrupt topic shifts.

3.3.4 *Select-a-Kibitzer*

Select-a-Kibitzer (Wiemer-Hastings and Graesser, 2000) is an interactive tutor which, like *Summary Street*, views writing as a constraint satisfaction problem. To assist students in finding a happy medium between the conflicting goals of content, concision, clarity, etc., *Select-a-Kibitzer* features an array of anthropomorphic agents, or *kibitzers*. Each kibitzer acts as a critic for a particular discourse feature, be it stylistic, grammatical, or semantic. Kibitzers implementing LSA assess and give counsel on coherence and topicality in a manner similar to *IEA* and *Summary Street*.

Besides its novel user interface, *Select-a-Kibitzer* breaks new ground in the area of automatic summarization. Like APex, *Select-a-Kibitzer* generates outlines of essays, but it does so without reference to a source text. The program uses clustering methods on the LSA semantic space to identify discrete topical chunks in the corpus. For each chunk, the program selects as an archetypical sentence the one that compares best with all other sentences in the chunk. An outline of key points is then produced by printing the selected sentences in order of appearance in the essay.[6] This outline gives the student an idea of the essay's progression of ideas, something particularly useful for beginning writers.

4 Future directions

Rather than seeing LSA essay scoring as its own end, some researchers are seeking further applications for the technique. For example, LSA essay-grading has enjoyed particular success in matching readers to

[6]Use of an aggregation tool (Joanis, 1999) might help turn the disfluent outline into a coherent summary.

texts, long the exclusive domain of readability scores (e.g. Flesch (1948)). The underlying theory is that the ability to learn from a text depends on the match between the background knowledge of the reader and the difficulty of (i.e. level of knowledge contained in) the text. With this in mind, Wolfe et al. (1998) had 106 university undergraduates and medical students write essays on the human heart before and after reading a randomly assigned text on the heart. The texts ranged in difficultly from elementary school to medical school level. For each student, LSA was employed to compute the similarity between the medical school text and the post-essay, and the medical school text and the pre-essay; the difference of these measures the amount of knowledge learned. The cosine between each pre-essay and the corresponding text was also taken. As expected, learning was highest when the text was neither too similar to the pre-essay (too easy) nor too dissimilar (too hard).

Other work focuses on remedying perceived or real flaws with LSA. Though LSA seems to perform well enough without regard for word order (Landauer et al., 1997), some believe it could be made even better by the addition of syntax sensitivity. For example, perhaps the semantic association between a noun and its adjectival modifiers could be made more apparent if LSA were apprised of their syntactic relationship rather than just their pattern of co-occurrence across contexts. Thus far, however, experiments which force LSA to consider syntax (e.g. by marking noun phrases as subject or object, as in Wiemer-Hastings (2000)) have only decreased correlation with human judges.

Finally, it is necessary to acknowledge the work of Rehder et al. (1998), who have investigated a number of technical issues in LSA essay grading. Besides finding the minimum essay length at which LSA is effective and proving that even non-technical words contribute significantly to semantic similarity, they have written extensively on the problem of directionality in high-dimensional semantic space (i.e. determining which of two cotopical yet semantically dissimilar documents contains more knowledge). They have discussed how to find an optimal training corpus, which remains an open question. And of course, their extensive treatment of similarity measures was noted earlier in this paper.

5 Summary

In this paper, I have discussed a number of automated essay-scoring tools currently under development. Their potential for alleviating human workload is tantalizing, as essays are among the most widespread and highly regarded forms of student assessment. Essay questions have been incorporated into several standardized testing programs, such as the GMAT, TOEFL, and SAT, where consistency of grading is paramount.

Though all the systems presented here correlate well with humans—many as well as several humans—when it comes to assessing content, those which incorporate LSA consistently outperform those which do not. This is because LSA is able to divine interword relationships at a much deeper level than is possible with simple co-occurrence measures. LSA cannot evaluate superficial mechanical and syntactic features, but most of these are easily separable from content and can therefore be processed by a separate module within the essay-scoring program.

I have also noted LSA's unique ability to train on a single expert text instead of a set of pregraded essays, which further reduces the component of human labour. I have reviewed some of the faults of LSA, including its inability to grasp logical fallacies and its severe computing resource demands. Finally, I have even suggested some avenues of further research, such as resolving anaphora before coherence assessment, and using LSA in conjunction with aggregation tools to produce fluent document summaries.

Acknowledgments

The author acknowledges the support of the Natural Sciences and Engineering Research Council of Canada.

References

Marie J. Abram. Readability: its use in adult education. *Lifelong Learning: The Adult Years*, 4(5):

8–9, 30, January 1981.

Jill Burstein, Karen Kukich, Susanne Wolff, Chi Lu, and Martin Chodorow. Enriching automated scoring using discourse marking. *Proceedings of the Workshop on Discourse Relations and Discourse Marking, 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, August 1998a.

Jill Burstein, Karen Kukich, Susanne Wolff, Chi Lu, Martin Chodorow, Lisa Braden-Harder, and Mary Dee Harris. Automated essay scoring using a hybrid feature identification technique. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, pages 206–210, August 1998b.

Jill Burstein and Daniel Marcu. Benefits of modularity in an automated scoring system. *Proceedings of the Workshop on Using Toolsets and Architectures to Build NLP Systems, 18th International Conference on Computational Linguistics*, August 2000.

William E. Coffman. Essay examinations. In Robert L. Thorndike, editor, *Educational Measurement*, pages 271–302. American Council on Education, Washington, DC, 2nd edition, 1971.

Edgar Dale and Jeanne S. Chall. A formula for predicting readability. *Educational Research Bulletin*, 87:11–20, 1948.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society For Information Science*, 41: 391–407, 1990.

Philippe Dessus and Benoît Lemaire. A*Pex*, un système d'aide à la préparation d'examens. *Sciences et Techniques Éducatives*, 6(2):409–415, 1999.

Philippe Dessus, Benoît Lemaire, and Angélique Vernier. Free-text assessment in a virtual campus. *Proceedings of the 3rd International Conference on Human-Learning Systems*, 2000.

Douglas S. Finlayson. The reliability of the marking of essays. *British Journal of Educational Psychology*, 21(2):126–134, June 1951.

Rudolf Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32:221–233, 1948.

Peter W. Foltz. Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments, and Computers*, 28(2):197–202, 1996.

Peter W. Foltz, Walter Kitsch, and Thomas K. Landauer. The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25 (2&3):285–307, 1998.

Peter W. Foltz, Darrell Laham, and Thomas K. Landauer. Automated essay scoring: Applications to educational technology. *Proceedings of the ED-MEDIA 1999 World Conference on Educational Multimedia, Hypermedia, and Telecommunications*, 1999.

George W. Furnas, Thomas K. Landauer, L. M. Gomez, and Susan T. Dumais. Statistical semantics: Analysis of the potential performance of keyword information systems. *Bell System Technical Journal*, 62(6):1753–1806, 1983.

Kenneth D. Hopkins, Julian C. Stanley, and B. R. Hopkins. *Constructing and Using Essay Tests*, pages 193–223. Prentice-Hall, Englewood Cliffs, NJ, 7th edition, 1990.

Eric J. S. Joanis. Review of the literature on aggregation in natural language generation. Technical Report CSRG-398, University of Toronto, Department of Computer Science, Toronto, Canada, September 1999.

Eileen Kintsch, Dave Steinhart, Gerry Stahl, the LSA Research Group, Cindy Matthews, and Ronald Lamb. Developing summarization skills through the use of LSA-based feedback. *Interactive Learning Environments*, 8(2):87–109, August 2000.

Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. An introduction to latent semantic analysis. *Discourse Processes*, 25(2&3):259–284, 1998.

Thomas K. Landauer, Darrell Laham, and Peter W. Foltz. The Intelligent Essay Assessor. *IEEE Intelligent Systems*, 15(5):27–31, September/October 2000.

Thomas K. Landauer, Darrell Laham, Bob Rehder, and M. E. Schreiner. How well can passage meaning be derived without using word order? A comparison of latent semantic analysis and humans. *Proceedings of the 19th Annual Conference of the Cognitive Science Society*, pages 412–417, 1997.

Eleni Miltsakaki and Karen Kukich. Automated evaluation of coherence in student essays. *Proceedings of the Workshop on Language Resources and Tools in Educational Applications, 2nd International Conference on Language Resources and Evaluation*, 2000a.

Eleni Miltsakaki and Karen Kukich. The role of centering theory's rough-shift in the teaching and evaluation of writing skills. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, 2000b.

Ellis B. Page. The imminence of grading essays by computer. *Phi Delta Kappan*, 47:238–243, January 1966.

Ellis B. Page. Computer grading of student prose, using modern concepts and software. *Journal of Experimental Education*, 62(2):127–142, 1994.

Ellis B. Page and Nancy S. Petersen. The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan*, 76:561–565, March 1995.

Bob Rehder, M. E. Schreiner, Michael B. W. Wolfe, Darrell Laham, Thomas K. Landauer, and Walter Kitsch. Using latent semantic analysis to assess knowledge: some technical considerations. *Discourse Processes*, 25(2&3):337–354, 1998.

Timothy C. Standal. Computer-measured readability. *Computers in the Schools*, 4(1):123–132, 1987.

Maurice C. Taylor and M. W. Wahlstrom. Readability as applied to an ABE assessment instrument. *Adult Literacy and Basic Education: An International Journal for Basic Education*, 10(3):155–170, 1986.

Peter Wiemer-Hastings. Adding syntactic information to LSA. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, pages 989–993, 2000.

Peter Wiemer-Hastings and Arthur Graesser. Select-a-Kibitzer: A computer tool that gives meaningful feedback on student compositions. *Interactive Learning Environments*, 8(2):149–169, August 2000.

Michael B. W. Wolfe, M. E. Schreiner, Bob Rehder, and Darrell Laham. Learning from text: matching readers and texts by latent semantic analysis. *Discourse Processes*, 25(2&3):309–336, 1998.