Detection of gaming in automated scoring of essays with the IEA

Research Report

Karen E. Lochbaum

Mark Rosenstein

Peter Foltz

Marcia A. Derr

Lochbaum, K. E., Rosenstein, M., Foltz, P. W., & Derr. M. A. (2013). Detection of gaming in automated scoring of essays with the IEA. Paper presented at *the National Council on Measurement in Education Conference (NCME*), San Francisco, CA, April.

April 2013

About Pearson
Everything we do at Pearson grows out of a clear mission: to help people make progress in their lives through personalized and connected learning solutions that are accessible, affordable, and that achieve results, focusing on college-and-career readiness, digital learning, educator effectiveness, and research for innovation and efficacy. Through our experience and expertise, investment in pioneering technologies, and promotion of collaboration throughout the education landscape, we continue to set the standard for leadership in education. For more information about Pearson, visit http://www.pearson.com/.

About Pearson's Research Reports
Our network **mission** is to spark innovation and create, connect, and communicate research and development that drives more effective learning. Our **vision** is students and educators learning in new ways so they can progress faster in a digital world. Pearson's research papers share our experts' perspectives with educators, researchers, policy makers and other stakeholders. Pearson's research publications may be obtained at: http://researchnetwork.pearson.com/.

Abstract

In addition to the standard test security issues, automated scoring provides new opportunities for students to deliberately misrepresent their ability. Gaming of essays can take many forms including repetition of words and sentences, incorporation of context irrelevant words, phrases, or sentences, plagiarism, and the insertion of "malicious" sequences of characters such as HTML web page markup language, which may be aimed at causing scoring failures. It can also involve a series of sophisticated words arranged in nonsensical ways that confuse a statistical language model into valuing the writing as sophisticated instead of as gibberish. Computer-based approaches to detecting gaming have advantages in that they can sometimes detect subtle statistical patterns in language and plagiarism, which are imperceptible to humans. However, computers may be less sensitive than human scorers to other aspects of writing, such as certain grammar patterns and language features. This talk will describe a general framework used to detect gaming within essays by the Intelligent Essay Assessor™. The development of features, by analyzing aspects of the topic content elaboration, language structure, coherence, and length and their use in detecting gaming, will be described. The talk will describe some of the tradeoffs between full transparency of scoring and detection methods versus obscuring some level of specificity in the algorithms to impede gaming.

Detection of gaming in automated scoring of essays with the IEA

Scoring student essays requires examining the features embodied in the writing and considering how these features contribute to measures of the students' performance. A student essay encompasses a wide range of features indicating multiple dimensions of skills. The goal in scoring is to measure the construct relevant features of the student performance while not being influenced by potential construct-irrelevant features. Human scorers are typically trained to detect these construct-relevant features through exposure to rubrics, sample essays, anchor essays, as well as by possessing years of extensive experience in language and domain knowledge which aids in distinguishing relevant from irrelevant features. In operational scoring, continual statistical oversight of the human scorers is necessary to ensure that human foibles, such as fatigue, drift or halo effects do not contaminate the scoring process, many of which automated systems are not susceptible. Automated scoring of writing takes a similar approach to measuring performance by detecting and appropriately combining sets of content-relevant features to generate performance measures. While these automated systems have proven to be effective at providing reliable and valid scores (e.g., Shermis & Bernstein, 2003), it remains critical that such systems are designed and continually updated to be as immune as current understandings allow to influences from construct irrelevant features.

We take the view that gaming of writing is the deliberate injection of construct-irrelevant features in order to influence measures of performance. While it would be nice to assume that all students will make a "good-faith" effort, it always remains possible that students will try to game the system in ways that attempt to misrepresent their target skills. These challenges to accurate scoring require approaches to detect construct-irrelevant strategies while minimizing false alarms

on construct-relevant behavior. This paper provides an overview of what can constitute gaming of automated scoring of writing. It provides an outline of different approaches to gaming automated scoring systems and illustrates examples of some methods that can be applied to detect and/or mitigate gaming behavior. It concludes with a discussion of a software architecture for detecting gaming that is implemented within a current operational system.

Outliers in automated scoring

"The intuitive definition of an outlier is: an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism (than the one under observation)." (Bol'shev, 2002)

In the automated scoring of writing, systems are typically trained on a few hundred to thousands of essays that have been previously scored by humans, called the training set. The system analyzes multiple features in the essays and learns to associate those features to the human scores. These features can measure such aspects of writing as the quality of the content, lexical sophistication, grammar, mechanics, style, organization, and development within essays. The resulting model details how these features can be combined to provide overall measures of the student performance, which is then tested on another set of human scored essays, called the test set, to help ensure that the model generalizes to essays beyond the original training set. The training essays also provide a baseline that characterizes the range of acceptable bounds for those features. Any new essay that deviates greatly in its pattern of values across this set of features may be considered an outlier.

There are many reasons why an essay could be considered an outlier. It could be due to a student deliberately gaming the system. However, it could also be due to an essay being off-topic, highly unusual or creative, or because the original training set did not sufficiently characterize the full range of acceptable responses. In these cases, however, because a scoring model is trained on a particular range of features, essays that fall beyond that range or have other unusual properties may cause instabilities in the model, resulting in inaccurate characterization of student performance (see Foltz et al., 2013, paper this session). Thus, we take the view that while gaming is one potential underlying mechanism that can cause detectable outliers in essay scoring, there are additional types of outliers that also need to be detected. From the perspective of a student gaming the system though, the goal is the easy generation of construct irrelevant features that fall below a threshold of detection. It should be noted that in our detection approach, we don't always deliberately distinguish between gaming and other outlier types since all need to be detected.

Parallels to cryptography

Before beginning a discussion of our architecture and strategy to detect gaming, we note that protecting automated essay scoring systems from gaming has parallels to the cryptography security literature. There is a fairly deep connection between the statistics and algorithms behind code breaking, automated scoring and the detection of gaming in automated scoring. Many of the same underlying statistical regularities of language that facilitate breaking codes constitute a significant class of features that allow automated scoring. Deviations from these regularities often indicate outliers, possibly identifying responses that were generated from "a different

mechanism" of which one of those mechanisms is gaming. An instance of this general pattern is Markov models, which are now a widely used general statistical tool, but were introduced in a paper on modeling sequences of letters in Russian literature (Markov, 1913).

In cryptography, two popular security models are "Security through obscurity", in essence attempting to keep details of the cryptographic algorithm hidden, contrasted with "The enemy knows the system." The naming of the first security model is to our knowledge unattributed, but can be considered part of a larger "defense in depth strategy", though in some circles it is a term of derision. The second security model description derives from Claude Shannon's (1949) analysis of cryptographic systems, and is a restatement and possible rediscovery of condition two from Kerckhoffs' 1883 paper on desirable attributes of a cryptographic system: "It must not be required to be secret, and it must be able to fall into the hands of the enemy without inconvenience" (Kerckhoffs, 1883).

In automated scoring, the content of the items or prompts are equivalent to the cryptographic key and must remain secret, to avoid asymmetric information giving a subset of test-takers an unfair advantage. This level of security is not a subject of this paper. Our intent, through robust models and detection techniques, is to make it as difficult as possible to game the scoring system. We subscribe to a defense in depth strategy of overlapping redundant detection measures. Therefore, while behind the scenes we assume "the enemy knows the system", in talks and papers like these, just as in credit card fraud detection, we keep portions of our methods and techniques to detect gaming veiled.

Approaches to detecting gaming

In developing techniques for the detection of gaming, it is critical to examine the different ways that a system can be gamed as well as which approaches can best be applied detection. While our previous paper (Foltz et al., 2013) focused on approaches to make the scoring model robust over the range of acceptable inputs, this paper focuses on how to protect the scoring model by using additional methods that uncover responses that are not within the expected limits. We are especially interested in those essays with indications that the responses potentially misrepresent a writer's skills. Such protections can be rule-based or statistically based. Rule-based approaches are implemented by incorporating methods to detect patterns derived from known attacks or generalizations from known examples of specific features of what would constitute expected and inappropriate input. Statistically based approaches use analyses of large numbers of examples of normal behavior to generalize the extent of acceptable and unacceptable input. These examples can be drawn from essay samples or data derived from large corpora (for example, Google n-grams (Michel et al., 2011)). Below, we describe different general categories of gaming, ordered approximately from less sophisticated to more sophisticated attacks. For each, we provide commentary on methods and for some describe analyses performed on the methods.

Malicious characters or inserting code

A web-based infrastructure introduces security issues in that user input, markup and signaling are all carried in the same communications stream. At least as far back as the early 1960s infamous "blue box" hacks on the U.S. phone system (Lapsley, 2013) in-band signaling

introduces risks that must be managed in order to ensure correct operations of systems built on this type of infrastructure. Specifically the input of malicious characters such as HTML-based content or malicious scripting code has been the basis of many attacks (there are any number of texts on this subject, and a reasonable starting point is the Wikipedia article on "Code Injection"). This type of attack can be achieved through a variety of methods including intercepting and accessing client web requests or insertion of targeted markup into web-based text boxes. Thus a key requirement to avoid such gaming is for scoring applications to have an appropriately paranoid attitude toward all user input, and especially testing and filtering the input stream to remove all possible dangerous special characters. This sanitizing can include recognizing and filtering out HTML tags, HTML escape sequences as well as limiting the character set. These types of attacks should not be a problem for a well-engineered system, but must be considered, and preferably alarmed and logged at the system level to detect systematic attacks on the scoring system.

Length

A common piece of advice in writing essays for standardized tests, whether automatically or human scored is to write a lot. Length, as measured by for instance the number of words, is strongly correlated with human scoring of essays. Students who write more, typically have more content and ideas. They also have more opportunity to show the quality of their performance (e.g., Chodorow & Burstein, 2004). However, since length is typically not directly used as a feature for scoring, it is critical to ensure that padding responses with content irrelevant information or repetition does not influence scoring. Gaming of length can be partially

controlled by providing limits on the sizes of essays submitted. In addition, it is critical to be able to partial out contributions from the length of the writing from measures of content or amount of information. Our previous paper (Foltz et al., 2013), provides additional details on how these approaches can be implemented.

Unusual language and bags of words

Increasing the length of essays without further adding greater construct-relevant content is generally considered to be "padding". This type of gaming can be achieved by adding randomly chosen words or groups of larger content words, which are not written in any proper syntactic form. A powerful feature for dealing with a portion of the padding problem is derived from statistical language models (for an introduction, see for instance, Jurafsky & Martin, 2009). These approaches compute how likely the occurrence of a pattern of words is based on a large corpus of English text. One measure is called entropy, with larger values indicating more unusual or unlikely values. Computing entropy derived from a statistical language model over windows of words in a response gives a measure of how "unusual" the text is. For instance, Table 1 illustrates different entropy values for more or less unusual combinations of words.

Text	Entropy
"We took the dog for a walk"	4.56
"over and over"	6.08
"Our architecture is divided between statistical tests"	11.39
"the the the the the the"	15.91
"octopus octopus octopus octopus octopus octopus"	20.71

Table 1. Entropy values for different text samples.

Repetition and coherence

A related strategy is to pad an essay by repeating words, short phrases, sentences or whole paragraphs over and over. By using entropy validations, the systems can detect test-takers repeating terms or series of terms (possibly using cut and paste). The system can further detect this pattern of padding through special purpose code that detects this type of behavior more generally, since while statistical language models are quite powerful, some repeated sequences are common enough in English corpora to not be detectable, such as:

Text	Entropy
“a a a a a a”	6.61

Table 2. Entropy value for a text

While, a string of a’s may initially seem an unlikely sequence, there are texts with words presented as “a a a a a a rrrrggghhhh”, and indeed, Google n-grams has a count of 2,525,552 of occurrences of the string “a a a a” across their large English corpus. Repetition of well-formed English phrases or sentences will have reasonable entropy values, and thus detection requires other approaches. The repeated sentence type of padding is usually evidenced by an unusual level of semantic redundancy, which is detectable using a range of coherence features based on Latent Semantic Analysis (see Foltz, Kintsch, and Landauer, 1998). These features detect both high and extremely low levels of coherence, which can indicate repetition, even if there are slight changes in the repeated sentences.

Foreign language

A general assumption for automated scoring is that essays will be written in the language for which the system is trained. We have encountered examples where students interject non-English text, as can occur with English Language Learners dropping back to their first language. In such cases, although it may not be a deliberate attempt to game the system, it needs to be detected as being an outlier since the scoring model in most cases was trained with essays in a single language and may not be able to generalize to mixes of other languages. We have found character n-gram text characterization techniques, such as those suggested by Cavnar and Trenkle (1994), to be quite powerful at detecting these types of issues. These character n-gram techniques in which distributions of n-grams are analyzed at the character level (including punctuation and normalized white space) and compared to distributions from English corpora have proven quite successful.

Plagiarism

Plagiarism can encompass using essays from other students or using text derived from external sources. Detection of plagiarism is one area where the power of computer processing wins out over human-based methods. Computers are able to compare millions of samples of text with each other quite efficiently to detect similarities. While plagiarism can be the word-for-word usage of other texts, we have incorporated semantic-based approaches which can detect when a student makes subtle substitutions, maintaining the overall meaning, but hiding the direct word overlap. As an example, in one analysis, our system detected 7 cases of plagiarism out of

520 student essays in scoring for a major university. All essays had also been scored by human raters, although the raters had not detected any of the plagiarism.

Content irrelevant text

While a number of the examples above involve gaming through creation of text that is not well formed (e.g., bags of words), students can also generate content irrelevant, but sophisticated text, often by selecting text from the web or other sources. We incorporate content-based analysis methods to detect how well any particular essay covers the expected content of a prompt. The ability to detect whether an essay is relevant to the context of the prompt or not varies depending on whether the prompt elicits a very wide range of different contents or a more narrow range. As an example, we illustrate analyses of the breadth of essay content as a basis to detect whether submitted essays are context irrelevant.

A sample of approximately 100 student essays from each of 62 writing prompts, so approximately 6,200 essays total (mean number of essays per prompt 105.7, sd 19.2) were used in this analysis. All prompts were selected from WriteToLearn (Landauer, Lochbaum & Dooley, 2009), an online grade 4-12 reading and writing formative assessment educational tool. The mean length (word count) of the essays per prompt was 311.2 (sd 78.7). The counts of prompts by the assigned grade level for the prompts were:

Grade level	4	5	6	7	8	9	10	11	12
Count	8	6	4	16	5	0	7	8	8

Table 3. Distribution of prompts by grade level.

We compared the range of semantic similarities of essays written in response to each prompt. The similarity of two essays is computed by determining the sematic vector for each essay and taking the cosine between the vector representations of the essays (e.g., Landauer, Laham & Foltz, 2001). For each prompt, we computed all pairwise cosines between the essays, and summary statistics on the cosines. The distribution of the mean cosine similarity for essays written to each prompt is shown in Figure 1.

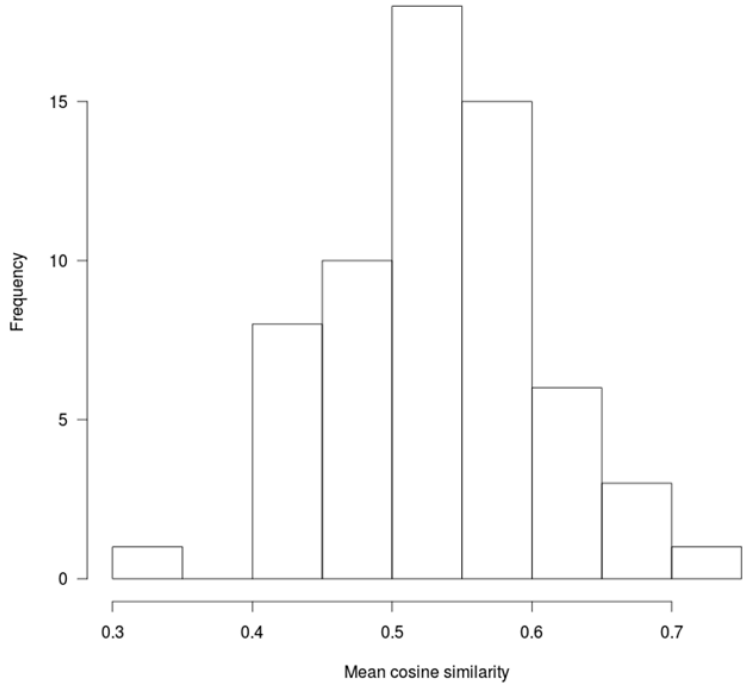

Figure 1. Histogram of mean cosine similarity of essays for 62 prompts.

The cosines present a fairly symmetric distribution exhibiting prompts with a wide range of semantic similarity among their essays. There is a small, but statistically significant correlation between grade level and degree of similarity ($r=0.34$, $t = 2.9$, $df = 60$, $p = 0.005$), but given the wide range of motivations for creating prompts, this may just be an artifact specific to

this set of prompts. For instance, the minimum and maximum extreme similarity cases with mean cosine similarities of 0.34 and 0.70 are both 10th grade prompts. The prompts for these two are shown in the next Figure. The first prompt is semantically broad while the second fairly carefully delimits the overall semantics of a correct response. The fact that there is little similarity among the first set of essays, and strong similarity among the second is a confirmation that our intuitions about the range of responses to these prompts are matched by the quantitative cosine measure.

Compare and Contrast Historical and Current Events (mean cosine 0.34)

Time can make a big difference. For example, even though both involve space travel, the moon landing in 1969 was different from a trip to the International Space Station that might take place today. Write an essay in which you compare and contrast an historical event and a current event.

Computer Training in High School (mean cosine 0.70)

Many people believe that training and experience in computer use are critical for the success of all students in high school now, perhaps more important than anything else they will learn. Take a position on that issue and write an op-ed piece explaining that position.

Figure 2. Two essay prompts. The first elicits essays with low semantic cosine similarity and the second elicits essays with high cosine similarity.

The next figure shows a histogram of the standard deviation of the cosine similarities across the prompts.

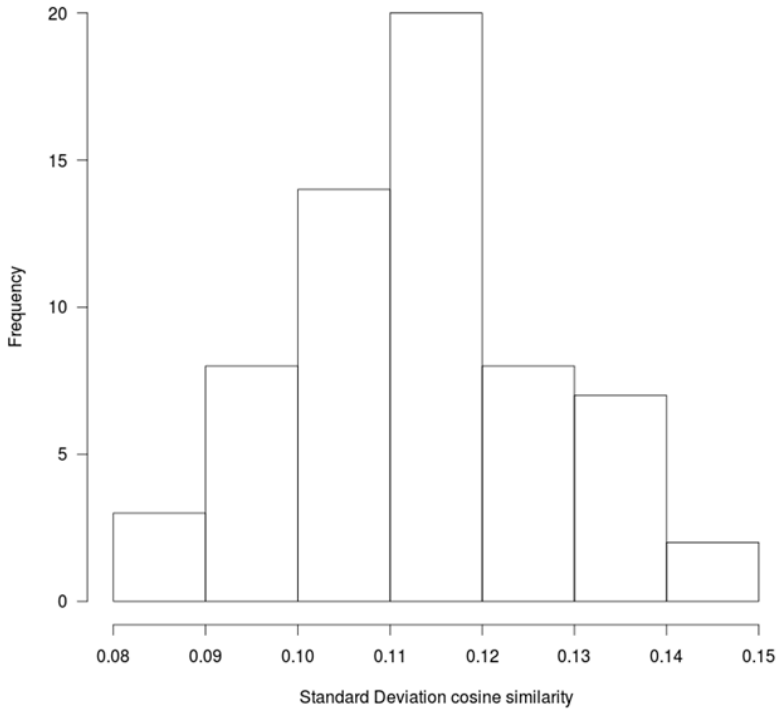

Figure 3. Histogram of standard deviation for cosine similarity of essays for 62 prompts.

The standard deviations show a fairly symmetric distribution, with a reasonably tight range. Of greater interest than the standard deviations taken alone is to consider the relationship between the mean cosine similarity and the standard deviations of the similarities for the prompts, as is shown in the scatterplot in the next Figure. The green line is the linear regression fit, and the red curve is a locally weighted regression model of the data.

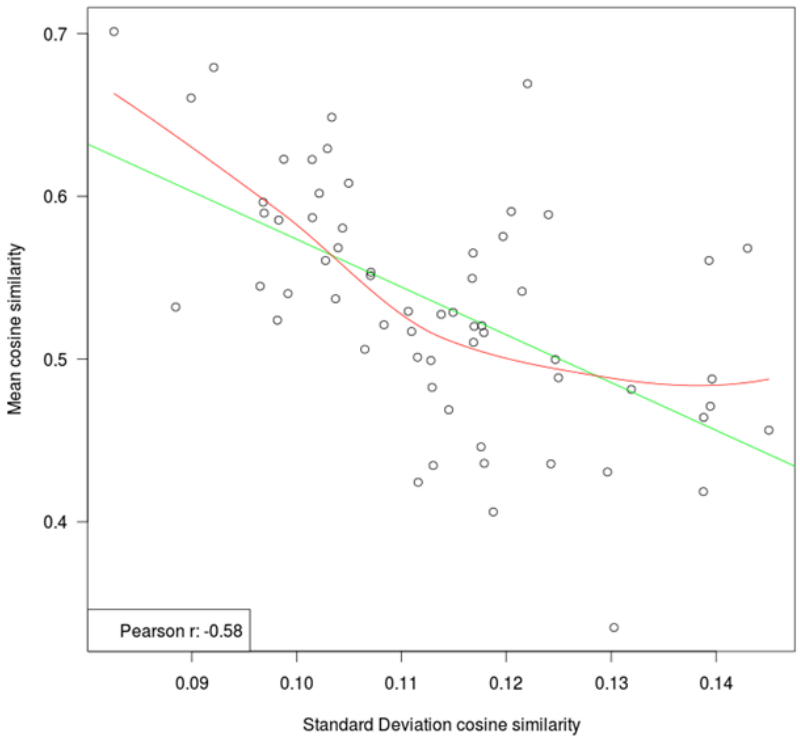

Figure 4. Plot of mean cosine similarity versus the standard deviation for each of 62 prompts. The green line is the linear regression fit, and the red curve represents a locally weighted regression, non-parametric representation of the relationship.

There is a moderately strong negative correlation (-0.58) between mean cosine similarity and its standard deviation. This can be explained in that a prompt that elicits responses that are more semantically similar, also tends to focus the essay writers around that topic, decreasing the semantic variability, while a prompt that allows a wide variety of responses will likely exhibit lower similarities between the essays, but a wider range of cosine similarities resulting in more variability. A concrete example of this relationship is shown in Figure 5, which superimposes the histograms representing the distributions of cosines from the two prompts shown in Figure 2.

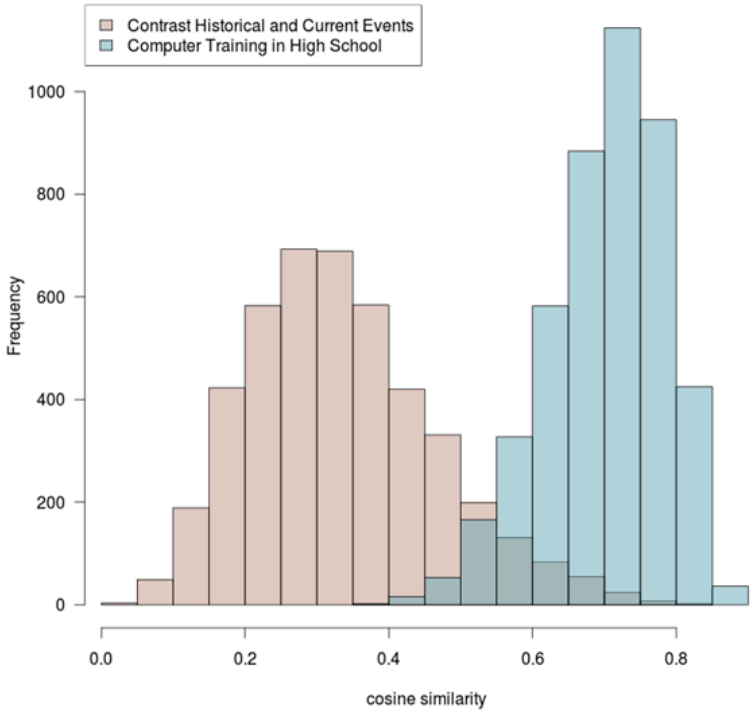

Figure 5. Histogram of cosine similarities for the two prompts shown in Figure 2. The prompt with lower overall similarity also exhibits wider variability than the more narrowly focused prompt.

We see that the "Computer" prompt, besides giving rise to essays that are semantically similar, also causes those essays to be narrowly focused, with a narrower range of cosine values. On the other hand, the expansive "Compare and Contrast" prompt generates a wide range of semantically diverse topics, with typically low cosine similarity among the essays, and the cosines exhibiting a wider range of similarities. Overall, the results indicate that some prompts are quite amenable to detecting off-topicness, while others may be more amenable to context-irrelevant gaming since the potential range of the topic is very wide.

Outliers by system errors rather than human strategy

It should be noted that not all instances of irregularities in text entry arise from test-taker misbehavior. As automated scoring moves beyond high stakes, rigidly controlled interfaces by linking scoring to other vendor applications such as formative systems, where there may be less control over the user-interface, outliers can be caused by the text generated by the web browser itself or due to software processing on the front-end text before being sent to scoring. For instance, this can cause lines being truncated or line breaking hyphenation inappropriately introduced into essays. Similarly, loss of white space at paragraph boundaries can lead to tokens that, at the scoring engine, look like two sentences that are concatenated, thereby hiding the intent of the writer to indicate paragraph structure. These kinds of errors can be detected by regular expression matching as well as avoided through good software practices.

Outlier detection in practice

The discussion above describes a number of approaches to gaming and techniques that can be applied for detection. Within the Intelligent Essay Assessor (see, Foltz et al., 2013,; Landauer, Laham & Foltz, 2003), we use an architecture to assess essays and identify those that are seen as outliers. Depending on the degree to which an essay is an outlier, as well as depending on a client's needs, an outlier essay can be flagged for human scoring, ignored, or scored and returned with a warning and details about the reason for being flagged. In formative contexts, where there is more emphasis on providing timely feedback, criteria can be less stringent and feedback can be given along with a warning that the essay is unusual or may not be scored as accurately. In higher-stakes scoring, the criteria for detecting outliers can be more

stringent and essays can be directed to human scorers for review. The architecture is divided between statistical tests and programmatic (for example, pattern matching) tests. The approach is flexible in that any time a new class of gaming responses is identified, we can reify that class of gaming responses and develop and incorporate new methods for detecting those responses.

Conclusions

Computer-based technology does not enjoy the same wealth of experience a human scorer brings to the scoring task and there is the risk that there are circumstances where the technology does not know what it doesn't know. As described in this paper there are techniques to mitigate these risks as well as operating procedures such as requiring human back-reads on some percentage of the essays to help detect and control the cases of the unknown unknowns. On the other hand, a computer has great strengths in being able to compare student writing simultaneously against many examples and look at multiple features together in order to generate predictions of outliers. Detecting gaming leverages the strengths of the computer to detect outliers across a wide range of cases, some of which have been illustrated in the paper. Analyses of this overall approach have shown that the detection methods are successful at detecting outliers that may prove more difficult for accurate scoring. For example, Foltz et al., (2012) showed that scores provided by IEA for the "flagged" essays have lower agreement rates to the human scorers. The results indicate that the IEA's detection methods are effective at determining essays that likely will not be scored as accurately and should be reviewed by human scorers.

For accurate automated scoring, responses must be similar to the responses used in training the scoring model. Models should be built using techniques robust to the expected deviations from the training set and the characteristics of the feature set. But outliers that can't be accurately scored need to be detected. Good faith responses can appear as outliers for many reasons, such as novel correct responses or – much more likely – going wrong in strikingly unique ways. However, gaming requires catching responses that are intentionally created in an attempt to fool scoring in order to achieve inappropriate scores. Our approach has been to incorporate a range of features that can be tuned to different assessment needs as well as to continually evaluate and incorporate techniques that further refine the performance.

References

Bol'shev L.N. (2002). Errors, theory of. In Hazewinkel, Michiel (Ed.). Encyclopedia of Mathematics. Kluwer Academic Publishers. http://www.encyclopediaofmath.org/index.php?title=Errors,_theory_of&oldid=28547

Cavnar WB and Trenkle JM. (1994). N-Gram-Based text categorization. In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval. 161-175.

Chodorow, M., & Burstein, J. (2004). Beyond essay length: Evaluating e-rater's performance on TOEFL essays (Research Report 73). Princeton, NJ: Educational Testing Service.

Code Injection. (n.d.). In Wikipedia. Retrieved March 30, 2013, from http://en.wikipedia.org/wiki/Code_Injection.

Foltz, P. W., Kintsch, W. & Landauer, T. K. (1998). The measurement of textual coherence with Latent Semantic Analysis. Discourse Processes, 25(2&3), 285-307.

Foltz, P. W., Rosenstein, M. and Lochbaum, K. E. (2013). Improving performance of automated scoring through detection of outliers and understanding model instabilities. Paper presented at the National Council on Measurement in Education Conference, San Francisco. April.

Foltz, P. W., Rosenstein, M., Lochbaum, K. E., & Davis, L. (2012). Improving reliability throughout the automated scoring process. Paper presented at the National Council on Measurement In Education Conference, Vancouver, BC, April.

Foltz, P. W., Streeter, L. A., Lochbaum, K. E., & Landauer, T. K (2013). Implementation and applications of the Intelligent Essay Assessor. *Handbook of Automated Essay Evaluation*, M. Shermis & J. Burstein, Eds. Pp. 68-88. Routledge, NY. NY.

HMTL Code Injection and Cross-Site Scripting. Downloaded 3/14/13 from http://www.technicalinfo.net/papers/CSS.html.

Jurafsky, D. and Martin, J.H. (2009). Speech and Language Processing. Upper Saddle River, NJ: Pearson.

Kerckhoffs, A. (1883). "La cryptographie militaire" Journal des sciences ilitaires, vol. IX, pp. 5-83, January 1883, pp. 161-191, February 1883.

Landauer, T. K., Laham, R. D. & Foltz, P. W. (2003) Automated Scoring and Annotation of Essays with the Intelligent Essay Assessor. In M. Shermis & J. Bernstein, (Eds.).

Automated Essay Scoring: A cross-disciplinary perspective. Mahwah, NJ: Lawrence Erlbaum Publishers.

Landauer, T. K, Lochbaum, K. E., & Dooley, S. (2009). A new formative assessment technology for reading and writing. Theory into Practice. 48: 44-52.

Lapsley, P. (2013). Phreaking Out Ma Bell. IEEE Spectrum. February.

Loader C. (1999). Local Regression and Likelihood. New York: Springer.

Loader, C. (2013). locfit: Local Regression, Likelihood and Density Estimation. R package version 1.5-9. http://CRAN.R-project.org/package=locfit

Markov, A.A. (1913). An example of statistical study on text of "Eugeny Onegin" illustrating the linking of events to a chain. In: Izvestija Imp.Akad.nauk, serija VI, T.X, N3, c.153. (in Russian).

Michel, J., Shen, Y., Aiden, A.P., Veres, A., Gray, M.K., The Google Books Team, Pickett, J.P. Hoiberg, D. Clancy, D. Norvig, P. Orwant, J. Pinker, S. Nowak, M.A. Aiden, E.L. (2011). Quantitative Analysis of Culture Using Millions of Digitized Books. Science. Vol. 331, no. 6014, pp. 176-182.

Peticolas, F., electronic version and English translation of La cryptographie militaire. http://petitcolas.net/fabien/kerckhoffs/ retrieved 2013-03-01.

Shannon, C. (1949). Communication Theory of Secrecy Systems, Bell System Technical Journal 28(4).

Shermis, M. & Bernstein, J. (2003). Automated Essay Scoring: A cross-disciplinary perspective. Mahwah, NJ: Lawrence Erlbaum Publishers.