Higher Education Research
2017; 2(4): 111-116
http://www.sciencepublishinggroup.com/j/her
doi: 10.11648/j.her.20170204.12

An Evaluation of Assessment-Oriented Computer-Based Text Analysis Paradigms

Andrew Aken

Department of Information Systems & Technology, Northeastern State University, Broken Arrow, USA

Email address:
aken@nsuok.edu

To cite this article:
Andrew Aken. An Evaluation of Assessment-Oriented Computer-Based Text Analysis Paradigms. *Higher Education Research.* Vol. 2, No. 4, 2017, pp. 111-116. doi: 10.11648/j.her.20170204.12

Received: August 25 2017; **Accepted**: September 8, 2017; **Published**: October 9, 2017

Abstract: Computer-based text analysis applications have come a long way since Ellis Page's Project Essay Grader [1]. Automated assessment applications have achieved better than human reliability and other methods of assisting assessment have opened up additional venues for utilization in the classroom and beyond. However, a lack of understanding of the differences between the different types of applications and their limitations has made selecting the appropriate application a difficult task. This study will present the most comprehensive examination of different paradigms of computer-based text analysis applications and a new typology for classifying them.

Keywords: Text Analysis, Content Analysis, Natural Language Processing, Latent Semantic Analysis, Peer Review, Automated Essay Scoring, Essay Assessment, Formative Assessment

1. Introduction

In addition to evaluation and interpretation of texts, assessment (including evaluation of student writing performance) is a fundamental decision-making task. Previously, grading written texts was an extremely labor-intensive, repetitious process with unreliable results. The human grader must understand the rubric that is to be applied to the texts and attempt to do so consistently. Support from various expert systems should lead to quicker and more reliable assessments of written texts [2, 3]. Additionally, because students routinely input their essays directly into computers, utilization of computer-based analysis of human texts has become even more feasible than was previously practicable [4].

Computer-based text analysis (CBTA) has been referred to by many names in the literature including Automated Essay Scoring (AES), which is the most prevalent name particularly within education literature. CBTA has also been referred to as Automated Essay Grading (AEG), computerized essay scoring, computer essay grading, computer graded essays, computer-assisted writing assessment, machine scoring of essays, Automated Writing Evaluation (AWE), and essay assessment [5].

Page [1] summarized potential classifications of computer-based text analysis across two dimensions: accuracy and substance.

Table 1. *Dimensions of Essay Grading [1].*

Accuracy \ Substance	I. Content	II. Style
Rating Simulation	I (A)	II (A)
Expert Analysis	I (B)	II (B)

In Page's [1] model, text analysis can be simulated or derived from actual interpretation of the texts for a deeper, more accurate analysis. Likewise, the components of the texts to be analyzed can either be based upon the content to ensure that it matches the requirements for the assignment or the style in which it was written.

However, with the continuously increasing number of text analysis applications currently on the market, Page's typology does not provide the granularity of analysis to effectively differentiate the disparate technologies available. It also fails to classify some of the more recent innovations within this collection of software. This makes comparisons of the different applications based upon their capabilities substantively more complicated.

2. Computer-Based Text Analysis Techniques

Many new applications to assess texts have been developed subsequent to Page's [1] model. Therefore, a more comprehensive classification system is necessary to understand the current state of analysis techniques. In this revised schema, computer-based text analysis tools fall into two categories: automated assessment and machine-assisted analysis. Automated writing assessment requires no human intervention (subsequent to the initial configuration of the prompt) while machine-assisted analysis is dependent upon human interaction to provide an analysis of the text being analyzed.

There are two primary differentiable methods of automated text assessment: those which rely on analysis of the text in isolation (TII) and those which compare the text to an existing corpus (Corpus-Based Analysis or CBA). Within each of these categories there are sometimes subtle and sometimes dramatic different methods of analysis. Within the TII analysis techniques, there are three wholly different approaches. The oldest approach uses parametric analysis of the writing style to assess the texts being analyzed based upon superficial features of the text. Slightly more sophisticated TII techniques analyze the syntax of the texts but do not address the content of the texts. Content analysis techniques attempt to derive the meaning of the texts being analyzed to compare to content-based rubrics. CBA techniques require the existence of a volume of previously evaluated texts to which the text to be analyzed is compared, but are differentiated by the techniques employed to perform the comparison.

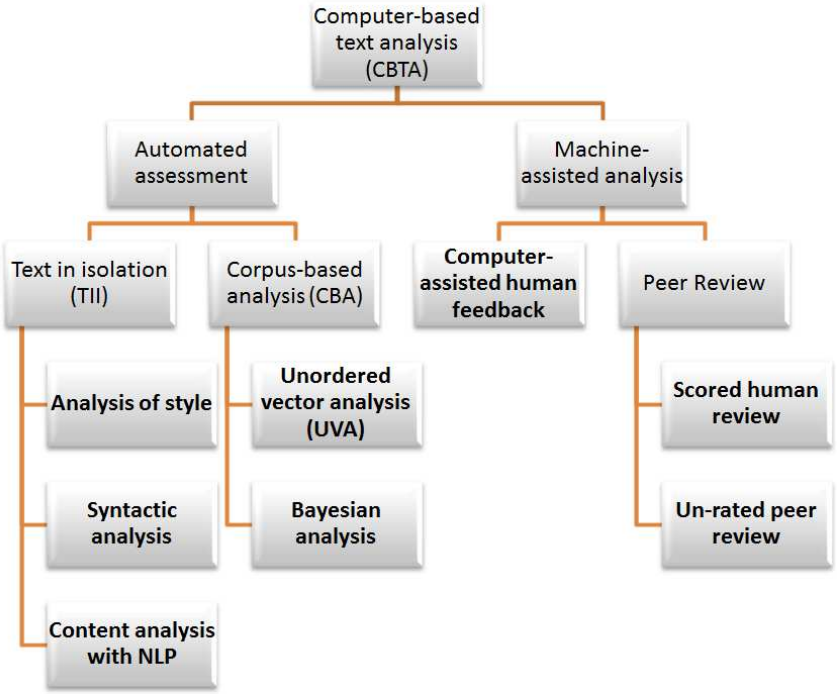

Figure 1. *Hierarchical Typology of Text Analysis Paradigms.*

It should be noted that although the reviewed applications have been classified as belonging to a single text analysis paradigm, they rarely utilize just a single analysis technique. They have, however, been classified based upon the dominant technique utilized to provide an assessment of the text being analyzed. E.g., although many popular automated assessment applications rely on Unordered Vector Analysis techniques for scoring texts, they also may incorporate syntactic analysis and peer review components which may, or may not, also be utilized in the final assessment of the text.

2.1. Text in Isolation

What makes text in isolation (TII) paradigms so attractive is that they do not require the previous human grading of a large number of texts in response to a particular prompt before they can be utilized to assess new texts. The algorithms utilized, however, can range from fairly simplistic measures of superficial elements of the text which happen to correlate highly with well-written texts, to comprehensive analyses which attempt to derive the actual meaning of the written text.

2.1.1. Analysis of Style

The earliest attempts to automate the process of grading essays were based upon the realization by the developers that texts which received high grades had certain qualities in common that were unrelated to the content of the text [1]. Certain qualities of the writing are derived from existing texts which are then turned into an algorithm which looks for those qualities. Examples of some of these qualities include the number of words in the text, the number of unique words in the text (a proxy for the fluency of the author), preposition

counts, relative pronouns or other parts of speech, word length, etc. These derived algorithms are then applied to new texts to determine their assessed score [6]. These types of analysis techniques, however, are unsuitable for assessing factual content in which the words used are very important [7] and do not provide instructive feedback to the authors [8].

Applications utilizing this as the dominant paradigm: Project Essay Grader (PEG), Intelligent Fuzzy Decision Support System Essay Assessment (IFDSSEA)

2.1.2. Syntactic Analysis

Syntactic analysis techniques attempt to break apart the components of the text into their intended parts of speech to determine if they have been written correctly. Applications employing these techniques are designed to improve the structure of writing and may also address elements of style in an attempt to help authors learn to write well. They can be used to assess correctness of grammar, spelling, punctuation, voice, and other elements of style [9]. These types of analysis techniques, however, are unsuitable for assessing factual content in which the words used are very important [7].

Syntactic analysis is also often used as a precursor to more robust analysis techniques since they may often require syntactically correct texts in order to function correctly or to return the best possible results.

Applications utilizing this as the dominant paradigm: Writer's Workbench, Essay Rater, Sentence works

2.1.3. Content Analysis with Natural Language Processing

Content analysis methods for automated assessment of texts are the most computationally intensive and also the most nascent techniques currently available to TII applications. They rely on at least one of several methods of Natural Language Processing (NLP) including linguistic analysis, artificial intelligence, fuzzy logic, semantic networks, and rule-based expert systems. Unlike Corpus-Based Assessment techniques, Content Analysis does not assess texts based upon statistically comparing them to other good and bad essays [10] but will evaluate and identify the main factors in the context of the written text which provides an improved method of information extraction [11].

Content analysis applications function by attempting to capture the meaning of the texts being analyzed to compare to a standard or rubric which has been established by the proctor. The proctor will specify the task in a prompt and then create a rubric which identifies the key features that should be included in an acceptable response along with the relationship between those features [10]. The applications will then use content analysis techniques to determine if the text meets the requirements for content and the appropriate relationships between the required content elements. Unfortunately, comprehensive assessment tools utilizing natural language processing remains primarily an area of research and are not yet fully implemented [12].

Applications utilizing this as the dominant paradigm: SAGrader, C-Rater, Automark

2.2. Corpus-Based Assessment

Unlike the TII paradigms, corpus-based assessment (CBA) algorithms require a typically large set of previously graded texts which will be compared in some manner to the text to be analyzed utilizing text clustering techniques [13]. The general belief in each of these paradigms is that if a text that was scored highly by human graders is similar to a new text, it should also be scored highly. Likewise, if the text is similar to those which had previously been scored poorly, it should also receive a low score. One of the key drawbacks of any CBA algorithm is the high cost of creating a new prompt, since a large number of essays must be human-graded before the CBA algorithm can be used.

2.2.1. Unordered Vector Analysis

Unordered Vector Analysis (UVA) techniques decompose the text to be analyzed into a matrix of important words and the frequency in which they occur in parts of the analyzed text. Ordering of the words based upon where they occur in the text is typically lost except for their initial occurrence [14, 15]. To analyze the following paragraph:

Betty Botter bought some butter. But, she said, the butter's bitter. If I put it in my batter, it will make my batter bitter. But, a bit of better butter is sure to make my batter better. So, she bought a bit of butter better than her bitter butter, and she put it in her batter and the batter was not bitter. So, it was better Betty Botter bought a bit of better butter.

The initial decomposition segmented by sentence structure would produce something similar to the following matrix:

Table 2. *LSA Initial Decomposition.*

	S1	S2	S3	S4	S5	S6
Betty	1					1
Botter	1					1
bought	1				1	1
butter	1	1		1	2	1
she		1			2	
said		1				
bitter		1	1		2	
I			1			
put			1			
batter			2	1	2	
bit				1	1	1
better				2	1	2
make				1		
her					2	

Latent Semantic Analysis (LSA) is one of the UVA algorithms frequently employed in text analysis. In LSA, the next step in the process is to transform the values into their relative importance in the passage and the degree that the word type is utilized within the general context of similar texts. A form of factor analysis is then utilized which reduces the values to the most common traits. In the final correlation matrix, texts with the highest correlations may not have been those that were similar, but which were differentiated from other texts the most.

Also, contrary to many researchers' views [5], LSA is not a type of natural language processing paradigm in that it does

not utilize dictionaries, semantic analysis, grammar [14], or attempt to derive meaning from the original text in any manner. As Landauer, et al. [14], admit: "LSA's 'bag of words' method, which ignores all syntactical, logical and nonlinguistic pragmatic entailments, sometimes misses meaning or gets it scrambled." Consequently, these types of analysis techniques are unsuitable for assessing factual content in which the word order is very important [7].

Applications utilizing this as the dominant paradigm: My Access!, Intelli Metric, Criterion, Write To Learn, Web Grader, Mark IT, Intelligent Essay Assessor, Write Placer

2.2.2. Bayesian Analysis

Bayesian analysis refers to statistical methods that look at qualities of a random population as variables with probability distributions based on Thomas Bayes' probability theorem which involves prior knowledge and accumulated experience [5]. The primary concept behind Bayesian analysis is to identify words or phrases that are most closely associated with previously assessed texts [16]. The two most common Bayesian models used for computer-based text analysis are the Multivariate Bernoulli Model and the Multinominal Model. The Bernoulli method looks at whether or not a specific feature exists in an essay (content), while the Multinominal method checks the use of specific features in an essay (semantics). The Bernoulli method is significantly more computationally intense and runs slowly compared to the Multinominal model [5].

Unlike LSA, the Bernoulli method of Bayesian Analysis reduces the words found in the texts to their root stem (e.g., "spelling", "spells", & "speller" would be reduced to "spell") to increase the likelihood of meaningful matches found in the comparison texts. Bayesian methods also retain the original text order in their decompositions and consequently, word order does matter.

Applications utilizing this as the dominant paradigm: BETSY

2.3. Machine-Assisted Analysis

Machine-assisted analysis techniques do not provide a summative score for the texts being analyzed without human intervention (if at all). They primarily assist a human evaluator to assess the text and provide feedback.

2.3.1. Computer-Assisted Human Feedback

Computer-assisted human feedback systems often provide many of the features of the more automated assessment tools (e.g., semantic and syntactic analysis) but leave the final analysis to a human intervener. These tools are primarily designed to help automate the process of creating feedback and marking up text. They may provide dropdown menus or buttons to insert comments describing common errors that occur in the texts and may insert scorecards into which the grader can input their evaluation. Semi-automated annotation and feedback systems are important for providing timely feedback to students especially in an environment where a large number of texts need to be assessed. Feedback-only tools, however, lack the functionality to automatically provide a quantifiable assessment of the texts being analyzed [12].

Applications utilizing this as the dominant paradigm: Semi-Automatic Grader

2.3.2. Scored Human Review

Scored peer review applications allow reviewers to provide feedback and annotations to the original text along with ratings which may be summarized to provide a final assessment of the texts [12]. These applications typically facilitate information sharing, collaboration, and portfolio management as well. In these systems, it is also possible for the peer reviewers themselves to be rated to improve the quality of the comments [17].

Applications utilizing this as the dominant paradigm: Di GIMIS Online, My Comp Lab, Calibrated Peer Review, Choices, Moodle Workshops, Turnitin Peer Mark & Grade Mark, Blackboard, Marking Assistant.

2.3.3. Un-rated Peer Review

Un-scored peer review applications are generally not developed to perform text analysis but have often been utilized to accomplish those tasks. In applications utilizing these methods, reviewers can markup text and leave annotations for the authors while maintaining the original text intact. These systems do not have the capacity to automatically generate summative assessments based upon the feedback left by others.

Applications utilizing this as the dominant paradigm: Microsoft Word

3. Classification by Intent

Another method of classifying computer-based writing assessment (CBTA) programs is by their intent. Most CBTA programs are developed to assess written text to provide a summary score (summative assessment). Some applications, however, are primarily as a learning environment designed to assist students in learning how to write (formative assessment) as well as possibly providing summative assessment. Although most formative assessment applications also provide summative scores, the applications classified as summative assessment do not include formative assessment capabilities.

As Philips [5] has stated, the primary purpose of formative assessment is to "provide understandable feedback to the learner and to guide instruction by the teacher. To be most effective, the feedback needs to be immediate, detailed and specific. It should be available in individual and group formats for the teacher, and it should suggest further directions for learning". One of the deficiencies of formative assessment applications has been their lack of ability to adequately justify their analyses so that the authors and teachers fully understand the implications of their assessment [5].

Summative assessment, however, "is a one-time assessment for evaluating the skills and knowledge acquired

at a specific reporting point on attained achievement at the student, school, district, and/or province/state level in order inform a decision about the individual or cohort. In many cases, the individual learner never receives feedback about performance or if so, it is in generalized terms, for example the student minimally met, successfully met or exceeded expectations, or the student successfully met or did not meet the assessment standard" [5].

Formative and summative are not distinct types of assessment, but describe the ways in which an assessment is primarily used. The distinction between formative and summative is not clear for all applications. Consequently, a well-defined criteria for the classification is necessary. Therefore, for our classification of applications, those which only provide a holistic score without additional feedback are classified as summative assessment. Otherwise, they can be plausibly utilized for formative assessment purposes.

Applications which are primarily designed for summative assessment include: Intelli metric, BETSY, Web Grader, Write Placer, PEG, Alaska Assessment Project, Intelligent Essay Assessor (IEA), C-Rater, Intelligent Essay Marking System, SEAR, IFDSSEA

Applications which can be utilized for formative assessment include: My Access!, Criterion, Write To Learn, Essay Rater, Calibrated Peer Review (CPR), SA Grader, Glencoe Online Essay Grader, Writer's Workbench, Di GIMS Online, Semi-Automatic Grader, ETIPS, Mark IT, Microsoft Word, My Comp Lab, Sentence works, Choices, Paperless School free text Marking Engine

4. Comparison of Methodologies

Each of the different paradigms discussed have their benefits and limitations. Each of the automated assessment applications has the advantage of unmatched reliability since they score the same essays consistently and are unaffected by the subjective nature of human assessment [6]. They have also consistently been shown to score the texts being assessed similarly to human raters [8, 14, 4, 1, 5, 6, 18] although significant questions remain as to whether this can be equated with validity [5, 6, 18, 19, 10]. Other advantages to automated assessment applications include cost savings, increased turnaround time, accessibility, and better reuse of existing assignments [5]. However, one of the continuing concerns of any of the automated assessment applications is their ability to accurately reflect the writing skill of the author or if the assessment is based upon some other characteristic [5]. Another concern is their inability to appreciate the nuances of language such as humor, irony, creative imagery, idiomatic phrases, metaphors, etc.

Whereas the machine-assisted analysis applications evaluate texts based upon the intrinsic qualities of the text and typically have more face validity. However, external validity is still questionable [5].

Within the corpus-based assessment applications, one of the key limiting aspects is the necessity to acquire or develop a large enough set of previously graded texts [14, 5, 6]. Text in isolation techniques do not, generally, require this costly component. However, with the exception of content analysis techniques, they are ambivalent towards what the authors are saying and focus exclusively on how it is being said.

5. Implications

Although there are benefits and limitations to each of the paradigms, most could substantively contribute to an as-of-yet developed comprehensive computer-based text analysis application. Revision and feedback elements are essential components of formative assessment [8, 12] and peer-reviews of the texts help to improve not only the quality of the texts being analyzed but benefit the reviewers as well [17]. To help improve the style of the writers and better enable automated analysis techniques function successfully, syntactic analysis must also be incorporated [9].

Because content analysis utilizing natural language processing is best suited to extracting the content intended by the author to compare to the rubric designed by the proctor and do not require a corpus of existing texts, this technique is the preferred method for automatic analysis of the texts should a successful implementation be developed [18, 10]. This will be beneficial for both summative and formative assessments.

As Page [1] so succinctly put it:

What is sought is not necessarily the perfect humanoid behavior, but rather those portions of that behavior which, given any current state of the art, will contribute optimally to efficient and practicable improvements in output. Indeed, regardless of the eventual perfection of deep linguistic behavior, for any specific application to essay grading, at any one moment, large portions of such available behavior may be irrelevant, just as it seems that ordinary human language processing does not usually call for our full linguistic effort.

Yet we regard it as eventually important to be able to perform these various kinds of advanced machine analysis when required. Therefore, the eventual uses of the ideal essay analyzer may require analytic capability as deep as may be imagined. Writing out suitable comments for the student, for example, will in some cases tax any system which may be foreseen.

This analysis of the current state of computer-based text analysis techniques contributes to the body of literature on content analysis and essay assessment by providing a succinct, but comprehensive typology of text analysis paradigms which had not previously been developed. This summation of the current state of CBTA applications and their classification based upon underlying analysis paradigms is useful for future research in this area as well as in defining future development of CBTA applications.

It also provides a previously undefined analysis of the different paradigms for empiricists and software developers in order to better understand the ramifications of selecting a particular application or algorithm. This will help them to better know the advantages or limitations of the selected technology. Additionally, with the rapidly expanding artillery

of new tools, this analysis should help to eliminate some of the confusion over what these tools can or cannot do.

References

[1] E. B. Page, "Statistical and Linguistic Strategies in the Computer Grading of Essays," University of Connecticut, Storrs, CT, 1967.

[2] J. Zeleznikow and J. R. Nolan, "Using Soft Computing to Build Real World Intelligent Decision Support Systems in Uncertain Domains," *Decision Support Systems,* vol. 31, no. 2, pp. 263-285, 2001.

[3] P. J. van Vliet, "Scaling Up Student Assessment: Issues and Solutions," *Journal of Higher Education Theory and Practice,* vol. 16, no. 6, p. 32, 2016.

[4] L. S. Larkey, "Automated Essay Grading Using Text Categorization Techniques," in *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval*, 1998.

[5] S. M. Phillips, "Automated Essay Scoring: A Literature Review," Society for the Advancement of Excellence in Education (SAEE), Kelowna, BC, 2007.

[6] S. Valenti, F. Neri and A. Cucchiarelli, "An Overview of Current Research on Automated Essay Grading," *Journal of Information Tehcnology Education,* vol. 2, pp. 319-330, 2003.

[7] D. Callear, J. Jerrams-Smith and V. Soh, "Bridging Gaps in Computerised Assessment of Texts," in *Proceedings of the IEEE International Conference on Advanced Learning Technologies*, Washington, 2001.

[8] S. Dikli, "An Overview of Automated Scoring Essays," *The Journal of Technology, Learning, and Assessment,* vol. 5, no. 1, pp. 3-35, 2006.

[9] N. H. MacDonald, L. T. Frase, P. S. Gingrich and S. A. Keenan, "The Writer's Workbench: Computer Aids for Text Analysis," *IEEE Transactions on Communications,* vol. 30, no. 1, pp. 105-110, 1982.

[10] E. Brent, C. Atkisson and N. Green, "Time-Shifted Online Collaboration: Creating Teachable Moments through Automated Grading," in *Monitoring and Assessment in Online Collaborative Environments: Emergent Computational Technologies for E-learning Support*, A. A. Juan and T. Daradoumis, Eds., Hershey, PA, IGI Global, 2009.

[11] S. W. Chan, "Beyond Keyword and Cue-Phrase Matching: A Sentence-Based Abstraction Technique for Information Extraction," *Decision Support Systems,* vol. 42, no. 2, pp. 759-777, 2006.

[12] M. Shortis and S. Burrows, "A Review of the Status of Online, Semi-Automated Marking and Feedback Systems," in *ATN Assessment Conference 2009*, RMIT University, 2009.

[13] D. G. Roussinov and H. Chen, "Document Clustering for Electronic Meetings: An Experimental Comparison of Two Techniques," *Decision Support Systems,* vol. 27, no. 1-2, pp. 67-79, 1999.

[14] T. K. Landauer, P. W. Foltz and D. Laham, "An Introduction to Latent Semantic Analysis," *Discourse Processes,* vol. 25, no. 2 & 3, pp. 259-284, 1998.

[15] S. Dikli, "Automated Essay Scoring," *Turkish Online Journal of Distance Education,* vol. 7, no. 1, 2006.

[16] L. M. Rudner, V. Garcia and C. Welch, "An Evaluation of the Intelli Metric Essay Scoring System," *The Journal of Tehcnology, Learning, and Assessment,* vol. 4, no. 4, pp. 3-21, 2006.

[17] P. A. Carlson and F. C. Berry, "Calibrated Peer Review and Assessing Learning Outcomes," in *33rd ASEE/IEEE Frontiers in Education Conference*, Boulder, CO, 2003.

[18] H.-C. Wang, C.-Y. Chang and T.-Y. Li, "Assessing Creative Problem-Solving with Automated Text Trading," *Computers & Education,* vol. 51, pp. 1450-1466, 2008.

[19] Y. Attali and J. Burstein, "Automated Essay Scoring With e-rater V. 2," *The Journal of Technology, Learning, and Assessment,* vol. 4, no. 3, pp. 3-30, 2006.