

Evaluation Criteria for Automatic Essay Assessment Systems – There is much more to it than just the correlation

Tuomo Kakkonen, Erkki Sutinen

Department of Computer Science and Statistics, University of Joensuu, Finland
tuomo.kakkonen@cs.joensuu.fi

Abstract: Automatic essay grading systems are usually evaluated by comparing the correlation between the grades assigned by the assessment system and a set of human graders. We argue that this method alone is inadequate for evaluating state-of-the-art assessment systems, and define a set of evaluation criteria that covers all the relevant aspects of an essay assessment system.

Keywords: Automatic essay assessment, writing assessment, evaluation, criteria

1. Introduction

Page's [7] pioneering work in the 1960s set the standard in evaluating *automatic essay grading systems*. The evaluation method recommended by Page measures the correlation between the grades assigned by the assessment system and a set of human graders. The extent of agreement between the grades assigned by the assessment system and the human graders measures the *validity* of an assessment system. Systems currently being developed analyze student essays in detail. In addition to assigning a holistic grade, these new systems allocate separate grades for other vital aspects of essay writing such as spelling, grammar and style. The best among these systems are also capable of instructing the learner on how to improve his or her text. The authors of the current paper have coined the term *semi-automatic assessment* systems for systems that are capable of providing automatic assessment that goes beyond "black box" holistic grading. The kind of assessment suggested by this definition incorporates both quantitative or numeric assessment (i.e. the kind that measures the learning outcomes with a numerical grade) and qualitative or verbal feedback (i.e. written assessments and instructions).

It stands to reason that new evaluation methods for essay assessment systems will have to offer means to evaluate the most advanced current systems. It is therefore necessary for evaluation methods to move from a validity-centered methodology to one that emphasizes learning outcomes and processes. The present paper is a step in that direction. It is based on our research and the development of concrete tools for essay assessment, tools such as *Automatic Essay Assessor* and *AntiPlag* plagiarism detection system. As well as a set of evaluation criteria, the framework described in the present paper can be considered as a set of requirements for a successful semi-automatic assessment tool that supports both students and instructors in writing and assessing free-text assignments. Section 2 offers an overview of relevant previous work. Section 3 describes the set of evaluation criteria that we have defined.

2. Previous Work

Modern essay assessment systems are capable of using several different criteria to grade essays. Table 1 summarizes the assessment criteria used by three state-of-the-art systems, namely, *Intelligent Essay Assessor (IEA)* [8], *Criterion* [2] and *IntelliMetric* [3].

Table 1. The assessment criteria in three state-of-the-art essay assessment systems.

System	Criterion	Description
<i>IEA</i>	Mechanics	the number of misspellings, the level of diction (word choice)
	Content	semantic similarity to the course content
	Plagiarism	similarities between student essays
	Style	coherence of the essay, identification of redundant sentences
<i>Criterion</i>	Mechanics	typographical errors
	Usage	agreement errors, verb formation errors, incorrect word use
	Lexical complexity	average word length
	Grammar	missing punctuation
	Style	passive voice sentences, very long or very short sentences, excessive repetition of specific words
	Prompt-specific vocabulary	extent of overlap with other essays
	Organization	identification of discourse elements
	Development	average lengths of discourse elements
<i>Intelli-Metric</i>	Mechanics	punctuation, grammar, spelling, degree of completeness of sentences
	Sentence structure	subject-verb agreement, sentence complexity and variety, readability
	Focus & unity	cohesion of the main ideas
	Organization	identification of certain discourse structures, sequencing of ideas
	Development & elaboration	breadth of content and supporting ideas, vocabulary, choice of words, concepts

Most of the publications that describe an automatic essay grading system, too numerous to list here, present the evaluation results in terms of the correlation between the human and system grades. The article by Yang *et al.* [10] offers a review of methods that can be used to validate the grades of automatic assessment systems. Reports on other types of evaluations are, however, much more rare. [1] used the decrease in errors of grammar, usage, mechanics and style obtained after feedback from Criterion had been implemented as a measure of the effectiveness of the automatic feedback. [5] measured accuracy in detecting off-topic essays. Warschauer and Ware [9] summarized a few examples of evaluations that relied on methods other than grading accuracy.

In their recent work, Haley *et al.* [4] proposed a framework for evaluating automated assessment systems. While we regard this work as a step in the right direction, we nevertheless feel that it concentrates too much on specific types of assessment systems, namely, ones based on *Latent Semantic Analysis*. A consequence of this reliance is that it over-emphasizes the technical aspects of the systems (such as the type of preprocessing method and the term weighting scheme that was used).

We use the definitions from our earlier work on domain of evaluation of syntactic parsers [6]. An *evaluation framework* consists of the following four components. *Criteria* identify the set of characteristics of a system that are being measured. The preciseness of the holistic grades assigned by a system is a commonly used criterion in evaluating assessment systems. *Metrics* are the means that are used for observing the performance of a system in terms of each of the criteria. The most commonly utilized metric in the evaluation of automatic assessment systems is the comparison between the grades given by the system and a set of human graders. We define *measure* as the way in which the results of an evaluation are quantified. Preciseness is most often measured in terms of the Pearson

correlation between the system and human grades. Several types of *evaluation materials* are needed to carry out practical evaluations. A collection of essays with grades assigned by a set of human graders is, for example, needed to evaluate an automatic grading system.

3. The evaluation criteria for automatic essay assessment systems

3.1 Introduction

The validity of an essay assessment system depends on the ability of such a system to assign reliable holistic grades to essays. It was because of this that we made preciseness in assigning holistic grades one of the evaluation criteria in our framework. The remaining criteria (analytical assessment, learner and instructor support, and training materials) are depicted in Figure 1. Each of these evaluation criteria is described in the following sections.

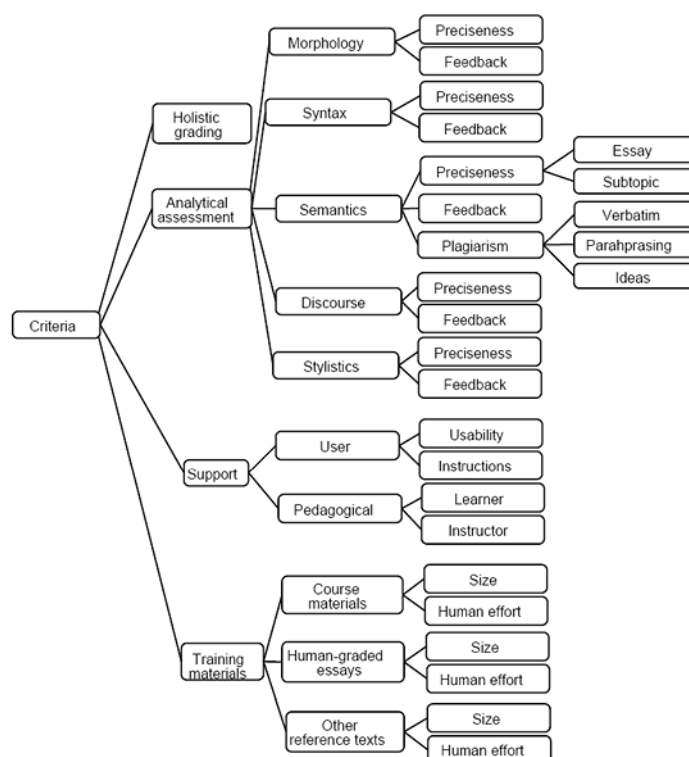


Figure 1. The evaluation criteria.

3.2 Analytical assessment

Analytical assessment, in contrast to holistic assessment, is based on making a separate consideration of the different aspects of an essay such as content and writing style. While it might have seemed logical to base the evaluation criteria on the assessment criteria used in existing assessment systems (see Section 2.1), the criteria used in existing systems appear in most cases to be rather vaguely defined. We used instead the well-known linguistic definition of the levels of knowledge of language as our basis for arriving at a well-defined and defensible set of criteria (see Table 2).

Although we defined *style* as one of the five sub-criteria in analytical assessment, it is accorded a special status among the criteria. While style is also a concept in linguistics, it is a more abstract criterion than the four criteria introduced in Table 2. A consideration of the table also makes it clear that because style is difficult to define, it is difficult to locate in one of the categories. Style is implicated in several levels of language knowledge. To evaluate

the style of a text, one would need to consider, for example, the extent of the vocabulary, aptness of word choice and the writer's skill in manipulating both grammar and discourse. Table 2. The linguistically motivated evaluation criteria and their correspondence to the assessment criteria in the state-of-the-art assessment systems.

Type of knowledge	Function	IEA	Criteria	IntelliMetric
<i>Morphology</i>	How words are constructed from basic units	Mechanics	Mechanics, usage, lexical complexity, style	Mechanics
<i>Syntax</i>	How words can be put together to form sentences	-	Grammar, usage, style	Sentence structure
<i>Semantics</i>	Meaning of words and sentences	Content, plagiarism, style	Prompt-specific vocabulary	Development and elaboration, focus and unity
<i>Discourse</i>	How the interpretation of a given sentence is affected by its preceding sentences	Style	Organization, development	Organization, focus and unity

We then divided our five analytical assessment criteria into two further sub-criteria: preciseness and feedback. The *preciseness* sub-criterion refers to the accuracy with which a system is able to carry out an assessment in terms of the criterion. In order to achieve high preciseness from morphology criterion, an assessment system has to be capable of detecting a high percentage of the spelling mistakes in essays. The *feedback* criterion refers to a systems' ability to provide instructional feedback. An assessment system has got to be able, for example, to provide a learner with coherent feedback on how better to organize her essay if it is to achieve a high feedback score in discourse criteria.

The content is clearly the most vital factor in determining the quality of an essay. Consequently, semantics is the most important of the analytical assessment criteria. If an essay is well written but incoherent or irrelevant to the topic, it should not obtain a high score. To reflect its importance, we divided the semantics criterion into two more refined categories. Firstly, we separated the preciseness sub-criterion into preciseness on essay and preciseness on subtopic levels. *Essay-level preciseness* refers to a system's ability to make judgments about the correctness and relevance of an essay as a whole. *Subtopic preciseness* refers to the ability of a system to score the various themes that appear in an essay. In addition to preciseness and feedback, semantics criterion utilizes a third sub-criterion, namely, *plagiarism*. We regard it is absolutely essential for an automatic essay assessment system to be able to identify and counter this phenomenon that has become (partly because of the widespread use of the Internet and computers) an ever-increasing problem in education throughout the world.

3.3 Learner and instructor support

Learner and instructor support refers to *usability* properties such as the ease with which a system can be used. We use the term *user support* to describe the quality of user documentation and instructions. We employ the term *user support* for this aspect. Automatic essay assessment systems tend to focus on the result rather than the process. Consequently, the focus has been on summative rather than on formative and diagnostic assessment. The *pedagogical support* criterion in our framework is concerned with the level of support for formative and diagnostic assessment in a system. Since the support that learners and instructors need are distinctively different, we divided pedagogical support into the two distinct criteria of *learner* and *instructor support*. For example, a functionality that

allows students to resubmit an essay after considering the feedback is an example of a system feature that supports both an iterative writing process and a formative evaluation.

3.4 Training materials

All existing essay assessment tools need reference materials before they can determine a set of grades for essays that need to be evaluated. The exact composition of the reference materials varies from one system to another. A criterion for comparing assessment systems is the amount of *training materials* that are needed for this purpose in conjunction with the amount of human effort that is needed for composing such materials. We divided the *training materials* criterion into the three further categories of *course materials*, *human-graded essays* and *other reference texts*. Course materials may consist, for example, of text passages from the course textbook and from lecture notes. Other reference materials might include, for example, a learner's personal learning diaries or some external sources of information.

4 Conclusion

We have proposed a set of criteria for evaluating all the relevant aspects of (semi-) automatic assessment systems for free-text responses. Our framework contains four sets of criteria: holistic grading, analytical assessment, user support and training materials. By defining these criteria we have, in addition to defining a standard for evaluating state-of-the-art essay assessment systems, defined a set of requirements for the next-generation semi-automatic writing assessment and assistant tool.

The most severe limitation of this work (which is partly attributable to a limitation on space) is the lack of precisely defined evaluation metrics and measures for each of the criteria. Such definitions represent the most important direction that future work in this kind of research will take. It will also be important to consider the types of evaluation materials that will be needed by each of the criteria.

References

- [1] Attali, Y. (2004). Exploring the feedback and revision features of the Criterion service. *Paper presented at the National Council on Measurement in Education Annual Meeting*. San Diego, California, USA.
- [2] Attali, Y. & Burstein, J. (2006). Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment*, 4(3).
- [3] Elliot, S. (2003). IntelliMetric: from here to validity. In Shermis, M. D., & Burstein, J. C. (Eds.). *Automated Essay Scoring: A Cross Disciplinary Approach*. Mahwah, New Jersey, USA: Lawrence Erlbaum Associates.
- [4] Haley, D. T., Thomas, P., De Roeck, A. & Petre, M. (2007). Seeing the whole picture: evaluating automated assessment systems, *ITALICS*, 6(4), 203-224.
- [5] Higgins, D., Burstein, J., & Attali, Y. (2006). Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering*, 12(2), 145-159.
- [6] Kakkonen, T. (2007). *Framework and Resources for Natural Language Parser Evaluation*. PhD Dissertation, Department of Computer Science and Statistics, University of Joensuu, Finland.
- [7] Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan*, 47(1): 238-243.
- [8] Streeter, L., Pstoka, J., Laham, D., & MacCuish, D. (2003). The credible grading machine: automated essay scoring in the DoD. Paper presented at *Interservice/Industry, Simulation and Education Conference*. Orlando, Florida, USA.
- [9] Warschauer, M., & Ware, P. D. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research*, 10(2), 1-24.
- [10] Yang, Y., Buckendahl, C., Juskiewicz, P., & Bhola, D. (2002). A review of strategies for validating computer automated scoring. *Applied Measurement in Education*, 15(4), 391-412.