Criterion℠ Online Essay Evaluation: An Application for Automated Evaluation of Student Essays

Jill Burstein
Educational Testing Service
Rosedale Road, 18E
Princeton, NJ 08541
jburstein@ets.org

Martin Chodorow
Department of Psychology
Hunter College
695 Park Avenue
New York, NY 10021
martin.chodorow@hunter.cuny.edu

Claudia Leacock
Educational Testing Service
Rosedale Road, 18E
Princeton, NJ 08541
cleacock@ets.org

Abstract

This paper describes a deployed educational technology application: the *Criterion*℠ Online Essay Evaluation Service, a web-based system that provides automated scoring and evaluation of student essays. *Criterion* has two complementary applications: *E-rater*®, an automated essay scoring system and *Critique* Writing Analysis Tools, a suite of programs that detect errors in grammar, usage, and mechanics, that identify discourse elements in the essay, and that recognize elements of undesirable style. These evaluation capabilities provide students with feedback that is specific to their writing in order to help them improve their writing skills. Both applications employ natural language processing and machine learning techniques. All of these capabilities outperform baseline algorithms, and some of the tools agree with human judges as often as two judges agree with each other.

1. Introduction

The best way to improve one's writing skills is to write, receive feedback from an instructor, revise based on the feedback, and then repeat the whole process as often as possible. Unfortunately, this puts an enormous load on the classroom teacher who is faced with reading and providing feedback for perhaps 30 essays or more every time a topic is assigned. As a result, teachers are not able to give writing assignments as often as they would wish.

With this in mind, researchers have sought to develop applications that automate essay scoring and evaluation. Work in automated essay scoring began in the early 1960's and has been extremely productive (Page 1966; Burstein et al., 1998; Foltz, Kintsch, and Landauer 1998; Larkey 1998; Elliot 2003). Detailed descriptions of these systems appear in Shermis and Burstein (2003). Pioneering work in automated feedback was initiated in the 1980's with the Writer's Workbench (MacDonald *et al*., 1982).

Criterion℠ *Online Essay Evaluation Service* combines automated essay scoring and diagnostic feedback. The feedback is specific to the student's essay and is based on the kinds of evaluations that teachers typically provide when grading a student's writing. *Criterion* is intended to be an aid, *not a replacement*, for classroom instruction. Its purpose is to ease the instructor's load, thereby enabling the instructor to give students more practice writing essays.

2. Application Description

Criterion contains two complementary applications that are based on natural language processing (NLP) methods. The scoring application, *e-rater*®, extracts linguistically-based features from an essay and uses a statistical model of how these features are related to overall writing quality to assign a holistic score to the essay. The second application, *Critique*, is comprised of a suite of programs that evaluate and provide feedback for errors in grammar, usage, and mechanics, identify the essay's discourse structure, and recognize undesirable stylistic features. See Appendices for sample evaluations and feedback.

2.1. The *E-rater* scoring engine

The *e-rater* scoring engine is designed to identify features in student essay writing that reflect characteristics that are specified in reader scoring guides. Human readers are told to read quickly for a total impression and to take into account *syntactic variety*, use of *grammar*, *mechanics*, and *style*, *organization and development*, and *vocabulary usage*. For example, the free-response section of the writing component of the Test of English as a Foreign Language (TOEFL) is scored on a 6-point scale where scores of 5 and 6 are given to essays that are "well organized," "use clearly appropriate details to support a thesis," "demonstrate syntactic variety," and show "a range of vocabulary." By contrast, 1's and 2's show "serious disorganization or underdevelopment" and may show "serious and frequent errors in sentence structure or usage." (See www.toefl.org/educator/edtwegui.html for the complete list of scoring guide criteria.) *E-rater* uses four modules for identifying features relevant to the scoring guide criteria – syntax, discourse, topical content, and lexical complexity.

2.1.1. *E-rater* features. In order to evaluate **syntactic variety**, a parser identifies syntactic structures, such as subjunctive auxiliary verbs and a variety of clausal structures, such as complement, infinitive, and subordinate clauses.

E-rater's **discourse analysis** module contains a lexicon based on the conceptual framework of conjunctive relations in Quirk et al. (1985) in which cue terms, such as *in summary*, are classified. These classifiers indicate whether or not the term is a discourse development term (*for exam-*

ple and *because*), or whether it is used to begin a new discourse segment (*first* or *second*). *E-rater* parses the essay to identify the syntactic structures in which these terms must appear to be considered discourse markers. For example, for *first* to be considered a discourse marker, it cannot be a nominal modifier, as in *"The first time that I saw her..."* where *first* modifies the noun *time*. Instead, *first* must act as an adverbial conjunct, as in, *"First, it has often been noted..."*

To capture an essay's **topical content**, *e-rater* uses content vector analyses that are based on the vector-space model (Salton, Wong, and Yang 1975). A set of essays that are used to train the model are converted into vectors of word frequencies. These vectors are transformed into word weights, where the weight of a word is directly proportional to its frequency in the essay but inversely related to number of essays in which it appears. To calculate the topical analysis of a novel essay, it is converted into a vector of word weights and a search is conducted to find the training vectors most similar to it. Similarity is measured by the cosine of the angle between two vectors.

For one feature, *topical analysis by essay*, the test vector consists of all the words in the essay. The value of the feature is the mean of the scores of the most similar training vectors. The other feature, *topical analysis by argument*, evaluates vocabulary usage at the argument level. *E-rater* uses a lexicon of cue terms and associated heuristics to automatically partition essays into component arguments or discussion points and a vector is created for each. Each argument vector is compared to the training set to assign a topical analysis score to each argument. The value for this feature is a mean of the argument scores.

While the topical content features compare the *specific* words of the test essay to the words in the scored training set, the **lexical complexity features** treat words more abstractly (Larkey 1998). Each essay is described in terms of the number of unique words it contains, average word length, the number of words with five or more characters, with six or more characters, etc. These numerical values reflect the range, frequency, and morphological complexity of the essay's vocabulary. For example, longer words are less common than shorter ones, and words beyond six characters are more likely to be morphologically derived through affixation.

2.1.2. Model building and score prediction. *E-rater* is trained on a sample of 270 essays that have been scored by human readers and that represent the range of scores from 1 to 6. It measures more than 50 features in all, of the kinds described in the previous section, and then computes a stepwise linear regression to select those features which make a significant contribution to the prediction of essay score. For each essay question, the result of training is a regression equation that can be applied to the features of a novel essay to produce a predicted value. This value is rounded to the nearest whole number to yield the score.

2.2. *Critique* Writing Analysis Tools

The *Critique* Writing Analysis Tools detect numerous errors in grammar, usage, and mechanics, highlight undesirable style, and provide information about essay-based discourse elements. In the following sections, we discuss those aspects of *Critique* that use NLP and statistical machine learning techniques.

2.2.1. Grammar, usage and mechanics. The writing analysis tools identify five main types of errors – agreement errors, verb formation errors, wrong word use, missing punctuation, and typographical errors. The approach to detecting violations of general English grammar is corpus-based and statistical. The system is trained on a large corpus of edited text, from which it extracts and counts sequences of adjacent word and part-of-speech pairs called bigrams. The system then searches student essays for bigrams that occur *much less often* than would be expected based on the corpus frequencies.

The expected frequencies come from a model of English that is based on 30-million words of newspaper text. Every word in the corpus is tagged with its part of speech using a version of the MXPOST (Ratnaparkhi 1996) part-of-speech tagger that has been trained on student essays. For example, the singular indefinite determiner *a* is labeled with the part-of-speech symbol AT, the adjective *good* is tagged JJ, the singular common noun *job* gets the label NN. After the corpus is tagged, frequencies are collected for each tag and for each function word (determiners, prepositions, etc.), and also for each adjacent pair of tags and function words. The individual tags and words are called unigrams, and the adjacent pairs are the bigrams. To illustrate, the word sequence, *"a good job"* contributes to the counts of three bigrams: *a*-JJ, AT-JJ, JJ-NN.

To detect violations of general rules of English, the system compares observed and expected frequencies in the general corpus. The statistical methods that the system uses are commonly used by researchers to detect combinations of words that occur *more frequently* than would be expected based on the assumption that the words are independent. These methods are usually used to find technical terms or collocations. *Criterion* uses the measures for the opposite purpose – to find combinations that occur *less often* than expected, and therefore might be evidence of a grammatical error (Chodorow and Leacock 2000). For example, the bigram for *this desks*, and similar sequences that show number disagreement, occur much less often than expected in the newspaper corpus based on the frequencies of singular determiners and plural nouns.

The system uses two complementary methods to measure association: pointwise mutual information and the log likelihood ratio. Pointwise mutual information gives the direction of association (whether a bigram occurs more often or less often than expected, based on the frequencies of its parts), but this measure is unreliable with sparse data.

The log likelihood ratio performs better with sparse data. For this application, it gives the likelihood that the elements in a sequence are independent (we are looking for non-independent, dis-associated words), but it does not tell whether the sequence occurs more often or less often than expected. By using both measures, we get the direction *and* the strength of association, and performance is better than it would otherwise be when data are limited.

Of course, no simple model based on adjacency of elements is adequate to capture English grammar. This is especially true when we restrict ourselves to a small window of two elements. For this reason, we needed special conditions, called filters, to allow for low probability, but nonetheless grammatical, sequences. The filters can be fairly complex. With bigrams that detect subject-verb agreement, filters check that the first element of the bigram is not part of a prepositional phrase or relative clause (e.g., *My friends in college assume*...) where the bigram *college assume* is not an error because the subject of *assume* is *friends.*

2.2.2. Confusable words. Some of the most common errors in writing are due to the confusion of homophones, words that sound alike. The Writing Analysis Tools detect errors among *their/there/they're, its/it's, affect/effect* and hundreds of other such sets. For the most common of these, the system uses 10,000 training examples of correct usage from newspaper text and builds a representation of the local context in which each word occurs. The context consists of the two words and part-of-speech tags that appear to the left, and the two that appear to the right, of the confusable word. For example, a context for *effect* might be *"a typical effect is found"*, consisting of a determiner and adjective to the left, and a form of the verb "BE" and a past participle to the right. For *affect*, a local context might be "*it can affect the outcome"*, where a pronoun and modal verb are on the left, and a determiner and noun are on the right.

Some confusable words, such as *populace/populous*, are so rare that a large training set cannot easily be assembled from published text. In this case, generic representations are used. The generic local context for nouns consists of all the part-of-speech tags found in the two positions to the left of each noun and in the two positions to the right of each noun in a large corpus of text. In a similar manner, generic local contexts are created for verbs, adjectives, adverbs, etc. These serve the same role as the word-specific representations built for more common homophones. Thus, *populace* would be represented as a generic noun and *populous* as a generic adjective

The frequencies found in training are then used to estimate the probabilities that particular words and parts of speech will be found at each position in the local context. When a confusable word is encountered in an essay, the Writing Analysis Tools use a Bayesian classifier (Golding 1995) to select the more probable member of its homophone set, given the local context in which it occurs. If this is not the word that the student typed, then the system highlights it as an error and suggests the more probable homophone.

2.2.3. Undesirable style. The identification of good or bad writing style is subjective; what one person finds irritating another may not mind. The Writing Analysis Tools highlight aspects of style that the writer may wish to revise, such as the use of passive sentences, as well as very long or very short sentences within the essay. Another feature of undesirable style that the system detects is the presence of overly repetitious words, a property of the essay that might affect its rating of overall quality.

Criterion uses a machine learning approach to finding excessive repetition. It was trained on a corpus of 300 essays in which two judges had labeled the occurrences of overly repetitious words. A word is considered to be overused if it interferes with a smooth reading of the essay. Seven features were found to reliably predict which word(s) should be labeled as being *repetitious.* They consist of the word's total number of occurrences in the essay, its relative frequency in the essay, its average relative frequency in a paragraph, its highest relative frequency in a paragraph, its length in characters, whether it is a pronoun, and the average distance between its successive occurrences. Using these features, a decision-based machine learning algorithm, C5.0 (www.rulequest.com), is used to model repetitious word use, based on the human judges' annotations. Function words were excluded from the model building. They are also excluded as candidates for words that can be assigned a repetition label. See Burstein and Wolska (to appear) for a detailed description.

2.2.4. Essay-based discourse elements. A well-written essay should contain discourse elements, which include *introductory material*, a *thesis statement*, *main ideas*, *supporting ideas*, and a *conclusion*. For example, when grading students' essays, teachers provide comments on these aspects of the discourse structure. The system makes decisions that exemplify how teachers perform this task. Teachers may make explicit that there is no thesis statement, or that there is only a single main idea with insufficient support. This kind of feedback helps students to develop the discourse structure of their writing.

For *Critique* to learn how to identify discourse elements, humans annotated a large sample of student essays with essay-based discourse elements. The annotation schema reflected the discourse structure of essay writing genres, such as *persuasive* writing where a highly-structured discourse strategy is employed to convince the reader that the thesis or position that is stated in the essay is valid.

The discourse analysis component uses a decision-based voting algorithm that takes into account the discourse labeling decisions of three independent discourse analysis systems. Two of the three systems use probabilistic-based methods, and the third uses a decision-based approach to

classify a sentence in an essay as a particular discourse element. Full details are presented in Burstein, Marcu, and Knight (2003).

3. Evaluation Criteria

We have described the computational approaches in the two applications in *Criterion*: *e-rater*, and *Critique* Writing Analysis Tools. In this section we answer the question: *"How do we determine that the system is accurate enough to provide useful feedback?"* by discussing the approach we used to evaluate the capabilities before they were commercially deployed.

The purpose of developing automated tools for writing instruction is to enable the student to get more practice writing. At the same time, it is essential that students receive accurate feedback from the system with regard to errors, comments on undesirable style, and information about discourse elements and organization of the essay. If the feedback is to help students improve their writing skills, then it should be similar to what an instructor's comments might be. With this in mind, we assess the accuracy of *e-rater* scores and the writing analysis feedback by examining the agreement between humans who perform these tasks. This inter-rater human performance is considered to be the gold standard against which human-system agreement is compared. Additionally, where relevant, both inter-rater human agreement and human-system agreement are compared to baseline algorithms, when such algorithms exist. The performance of the baseline is considered the lower threshold. For a capability to be used in *Criterion* it must outperform the baseline measures and, in the best case, approach human performance.

3.1. *E-rater* performance evaluation

The performance of *e-rater* is evaluated by comparing its scores to those of human judges. This is carried out in the same manner that the scores of two judges are measured during reader scoring sessions for standardized tests such as the Graduate Management Admissions Test (GMAT). If two judges' scores match exactly, or if they are within one point of each other on the 6-point scale, they are considered to be in *agreement*. When judges do not agree, a third judge resolves the score. In evaluating *e-rater*, its score is treated as if it were one of the two judges' scores. A detailed description of this procedure can be found in Burstein et al. (1998).

For a baseline, the percent agreement is computed based on the assignment of the modal score to all essays in a particular sample. Typical agreement between *e-rater* and the human resolved score is approximately 97%, which is comparable to agreement between two human readers. Baseline agreement using the modal score is generally 75%-80%.

3.2. *Critique* performance evaluation

For the different capabilities of *Critique*, we evaluate performance using *precision* and *recall*. Precision for a diagnostic *d* (e.g., the labeling of a thesis statement or the labeling of a grammatical error) is the number of cases in which the system and the human judge (i.e., the gold standard) agree on the label *d*, divided by the total number of cases that the system labels *d*. This is equal to the number of the system's hits divided by the total of its hits and false positives. Recall is the number of cases in which the system and the human judge agree on the label *d*, divided by the total number of cases that the human labels *d*. This is equal to the number of the system's hits divided by the total of its hits and misses.

3.2.2. Grammar, Usage, and Mechanics. For the errors that are detected using bigrams and errors caused by the misuse of confusable words, we have chosen to err on the side of precision over recall. That is, we would rather miss an error than tell the student that a well-formed construction is ill-formed. A minimum threshold of 90% precision was set in order for a bigram error or confusable word set to be included in the writing analysis tools.

Since the threshold for precision is between 90-100%, the recall varies from bigram to bigram and confusable word set to confusable word set. In order to estimate recall, 5,000 sentences were annotated to identify specific types of grammatical errors. For example, the writing analysis tools correctly identified 40% of the subject-verb agreement errors that the annotators identified and 70% of the possessive marker (apostrophe) errors. The confusable word errors were detected 71% of the time.

3.2.3. Repetitious use of words. Precision, recall, and the F-measure (the harmonic mean of precision and recall, which is equal to 2 * (precision * recall) / (precision + recall)) were computed to evaluate the performance of the repetitious word detection system. The total sample contained 300 essays where human judges had labeled the words in the essay that they considered repetitious. Of the total sample, the two judges noted repetitious word use in only 74 of the essays, so the results are based on this subset.

A baseline was computed for each of the seven features used to build the final system. Of these, the highest baseline was achieved using the *essay ratio feature* that measures a word's relative frequency in an essay. For this baseline, a word was selected as repetitious if the proportion of that word's occurrences was greater than or equal to 5%. This resulted in precision, recall, and F-measure of 0.27, 0.54, and 0.36, respectively. The remaining six features are described in Section 2.2.3. No single feature reached the level of agreement found between two human judges (precision, recall, and F-measure of 0.55, 0.56, and 0.56, respectively). It is interesting to note that the human judges

showed considerable disagreement in this task, but each judge was internally consistent. When the repetitious word detection system, which combines all seven features, was trained on data of a single judge, it could accurately model that individual's performance (precision, recall, and F-measure of 0.95, 0.90, and 0.93, respectively).

3.2.4. Discourse structure. To evaluate system performance, we computed precision, recall, and F-measure values for the system, the baseline algorithm, and also between two human judges. The baseline algorithm assigns a discourse label to each sentence in an essay based solely on the sentence position. An example of a baseline algorithm assignment would be that the system labels the first sentence of every paragraph in the body of the essay as a *Main Point.*

The results from a sample of 1,462 human-labeled essays indicate that the system outperforms the baseline measure for every discourse category. Overall, the precision, recall, and F-measure for the baseline algorithm are 0.71, 0.70, and 0.70, respectively, while for the discourse analysis system, precision, recall, and F-measure are uniformly 0.85. For detailed results, see Burstein, Marcu, and Knight (2003). The average precision, recall, and F-measure are approximately 0.95 between two human judges.

4. Application Use

Criterion with *e-rater*[1] and *Critique* Writing Analysis Tools was deployed in September 2002. The application has been purchased by over 200 institutions, and has approximately 50,000 users as of December 2002. Examples of the user population are: elementary, middle and high schools, public charter schools, community colleges, universities, military institutions (e.g., the United States Air Force Academy and The Citadel), and national job training programs (e.g., Job Corps). The system is being used outside of the United States in China, Taiwan, and Japan.

The strongest representation of users is in the K-12 market. Within K-12, middle schools have the largest user population. Approximately 7,000 essays are processed through *Criterion* each week. We anticipate increased usage as teachers become more familiar with the technology. Most of the usage is in a computer lab environment.

4.1. *Criterion* User Evaluation

As part of an ongoing study to evaluate the impact of *Criterion* on student writing performance, nine teachers in the Miami-Dade County Public School system, who used *Criterion* in the classroom once a week during the fall, 2002 term, responded to a survey about their experience with *Criterion*. The questions elicited responses about *Criterion*'s strengths, weaknesses and ease of use.

The teacher's responses indicate that *Criterion* provides effective help for students. All of the teachers stated that the strength of the application was that it supplies immediate scores and feedback to students. In terms of weaknesses, the responses primarily addressed technical problems that have since been fixed (e.g., problems with the spell checker). In addition, all of the teachers maintained that learning how to use the system was, by in large, smooth.

This study is being conducted independently by Mark Shermis, Florida International University. Results of the study will be available by Fall of 2003.

5. Application Development and Deployment

The *Criterion* project involved about 15 developers at a cost of over one million dollars. The team had considerable experience in developing electronic scoring and assessment products and services with regard to on-time delivery within the proposed budget. Members of the team had previously developed the Educational Testing Service's Online Scoring Network (OSN) and had implemented *e-rater* within OSN for scoring essays for GMAT

The project was organized into four phases: definition, analysis, development, and implementation. In the *definition phase*, we established the scope and depth of the project based on an extensive fact-finding process by a cross-disciplinary team that included researchers, content developers, software engineers, and project managers. This phase established the high-level project specifications, deliverables, milestones, timeline, and responsibilities for the project. In the *analysis phase* the team developed detailed project specifications and determined the best approach to meet the requirements set forth in the specifications. When necessary, storyboards and prototypes were used to communicate concepts that included interface, architecture, and processing steps. The *development phase* included the construction of the platform used to deliver the service, the development and modification of the tools used by the platform, and the establishment of connections to any external processes. The final *implementation phase* involved full integrated testing of the service, and moving it into a production environment. Extensive tests were run to ensure the accuracy and scalability of the work that was produced.

The *Criterion* interface was developed by showing screen shots and prototypes to teachers and students and eliciting their comments and suggestions. The interface presented one of the larger challenges. A major difficulty was determining how to present a potentially overwhelming amount of feedback information in a manageable format via browser-based software.

[1] An earlier version of *Criterion* with *e-rater* only was released in September 2001, and *e-rater* has been used at Educational Testing Service to score GMAT Analytical Writing Assessment essays since February 1999.

6. Maintenance

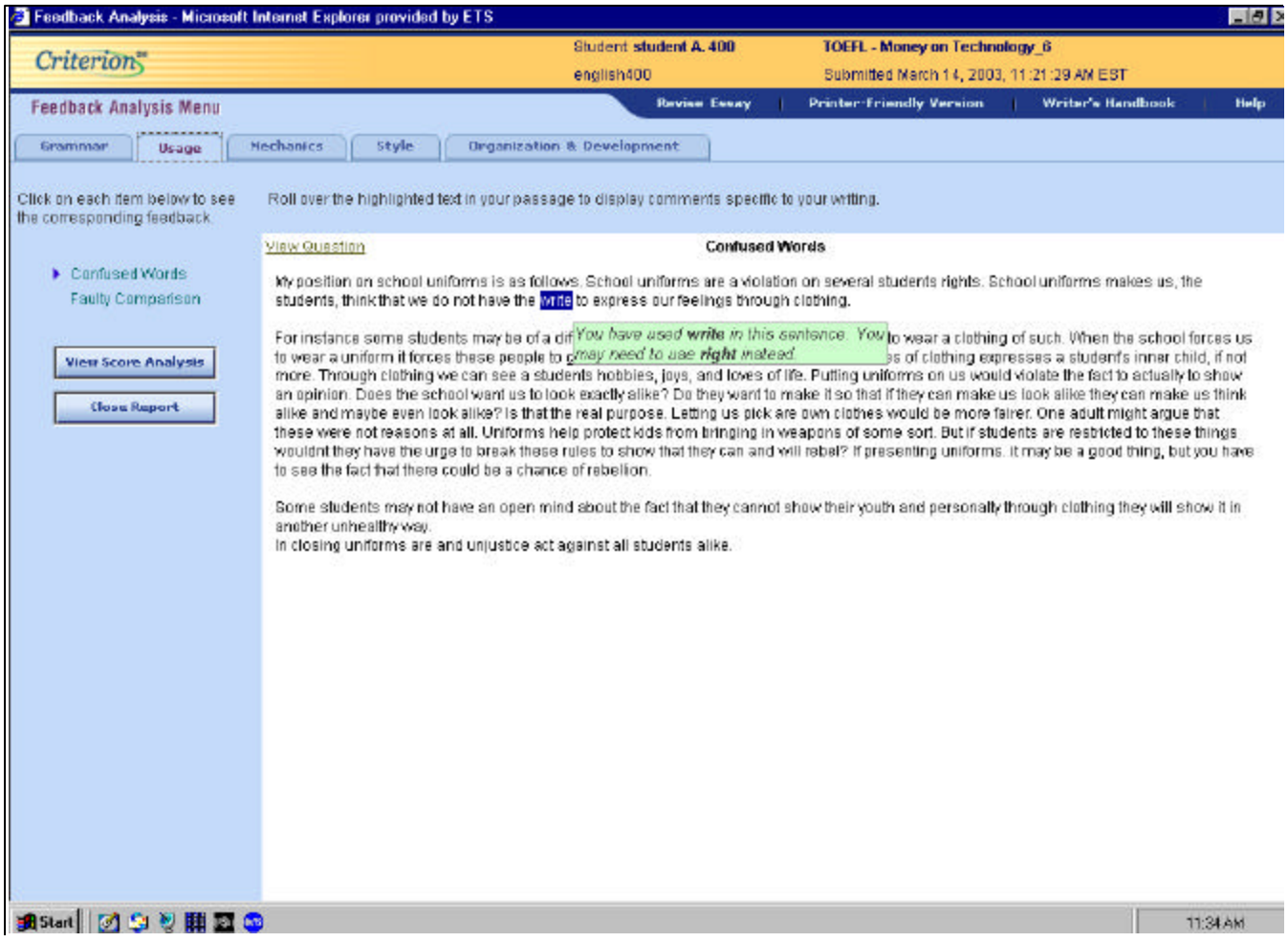
Although a new version of the *Criterion* software is scheduled for release with the start of each school year, interim releases are possible. As new functionality is defined, it is evaluated and a determination is made as to a proper release schedule. *Criterion* was released in September 2002. Because the software is centrally hosted, updates are easily deployed and made immediately available to users. The software is maintained by an internal group of developers.

7. Conclusion

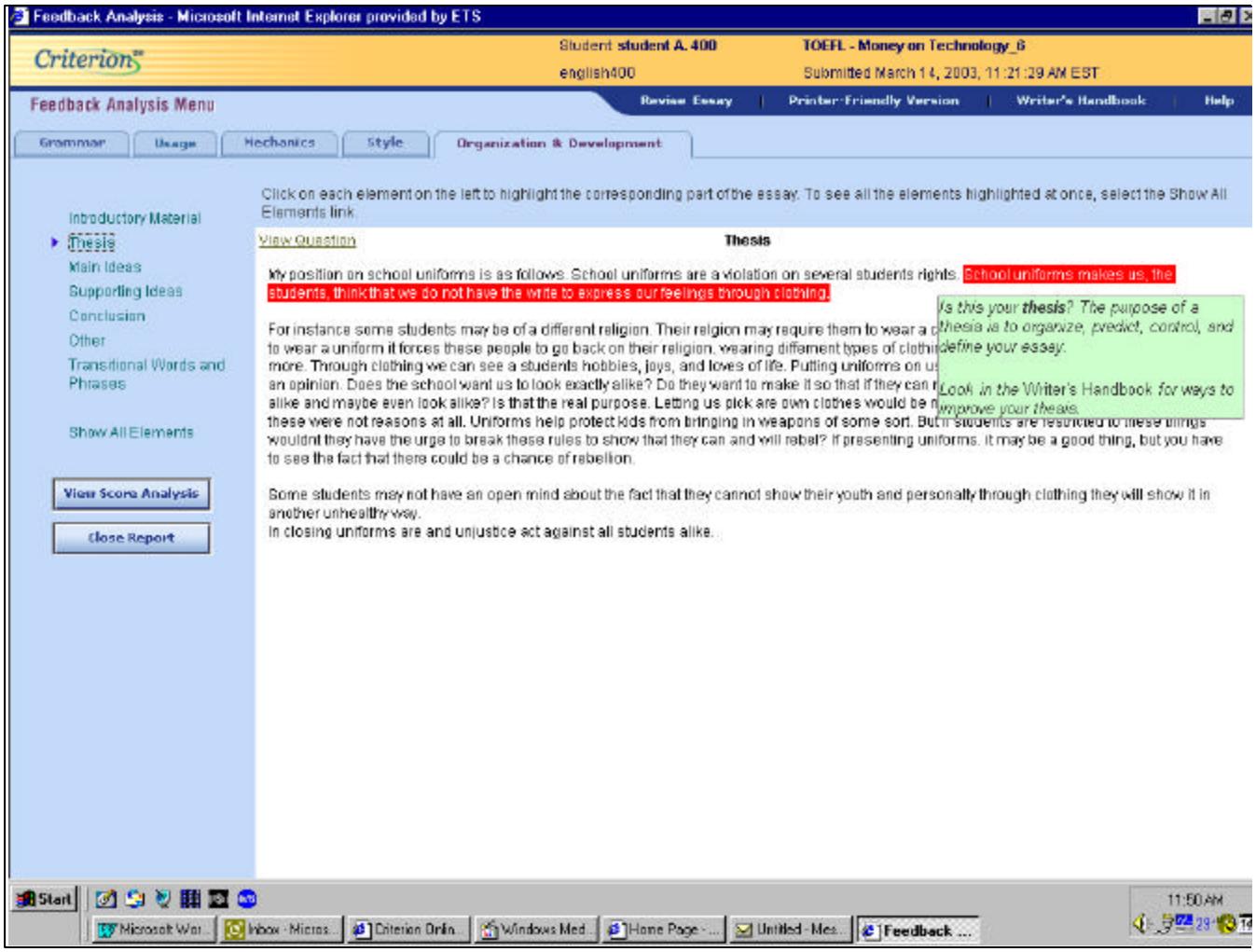
We plan to continue improving the algorithms that are used, as well as adding new features. For example, we hope to implement the detection of grammatical errors that are important to specific native language groups, such as identifying when a determiner is missing (a common error among native speakers of Asian languages and of Russian) or when the wrong preposition is used. We also intend to extend our analysis of discourse so that the quality of the discourse elements can be assessed. This means, for example, not only telling the writer which sentence serves as the thesis statement but also indicating how good that thesis statement is. A newer version of *e-rater* is being developed for *Criterion* 2.0, due to be released in spring 2003. This version incorporates features from the Writing Analysis Tools , such as the number of grammar and usage errors in the essay. These *Critique* features improve *e-rater's* performance, in part, because they better reflect what teachers actually consider when grading student writing.

Acknowledgements: The authors would like to thank John Fitzpatrick, Bob Foy, and Andrea King for essential information related to the application's use, deployment, and maintenance; Slava Andreyev, Chi Lu, and Magdalena Wolska for their intellectual contributions and programming support; and John Blackmore and Christino Wijaya for systems programming. We are especially grateful to Mark Shermis for sharing the teacher surveys from his user evaluation study. The version of *Criterion* described here and the *Critique* Writing Analysis Tools were developed and implemented at ETS Technologies, Inc. Any opinions expressed here are those of the authors and not necessarily of the Educational Testing Service.

Appendix A: Sample *Usage* Feedback



Appendix B: Sample *Organization & Development* Feedback



References

Burstein, J. and Wolska, M. (to appear). Toward Evaluation of Writing Style: Overly Repetitious Word Use in Student Writing. In Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics. Budapest, Hungary, April, 2003.

Burstein, J., Marcu, D., and Knight, K. 2003. Finding the WRITE Stuff: Automatic Identification of Discourse Structure in Student Essays. *IEEE Intelligent Systems: Special Issue on Natural Language Processing* 18(1), pp. 32-39.

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., and Harris M. D. 1998. Automated Scoring Using A Hybrid Feature Identification Technique. *Proceedings of 36th Annual Meeting of the Association for Computational Linguistics*, 206-210. Montreal, Canada

Chodorow, M., and Leacock, C. 2000. An Unsupervised Method for Detecting Grammatical Errors. *Proceedings of the 1st Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, 140-147.

Elliott, S. 2003. Intellimetric: From Here to Validity. In Shermis, M., and Burstein, J. eds. *Automated essay scoring: A cross-disciplinary perspective*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Foltz, P. W., Kintsch, W., and Landauer, T. K. 1998. Analysis of Text Coherence Using Latent Semantic Analysis. *Discourse Processes* 25(2-3):285-307.

Golding, A. 1995. A Bayesian Hybrid for Context-Sensitive Spelling Correction. *Proceedings of the 3rd Workshop on Very Large Corpora*, 39-53. Cambridge, MA.

Larkey, L. 1998. Automatic Essay Grading Using Text Categorization Techniques. *Proceedings of the 21st ACM-SIGIR Conference on Research and Development in Information Retrieval*, 90-95. Melbourne, Australia.

MacDonald, N. H., Frase, L. T., Gingrich P. S., and Keenan, S.A. 1982. The Writer's Workbench: Computer Aids for Text Analysis. *IEEE Transactions on Communications* 30(1):105-110.

Page, E. B. 1966. The Imminence of Grading Essays by Computer. *Phi Delta Kappan*, 48:238-243.

Quirk, R., Greenbaum, S., Leech, G., and Svartik, J. 1985. *A Comprehensive Grammar of the English Language*. New York: Longman.

Ratnaparkhi, A. 1996. A Maximum Entropy Part-of-Speech Tagger. In *Proceedings of the Empirical Methods in Natural Language Processing Conference*, University of Pennsylvania.

Salton, G., Wong, A., and Yang, C.S. 1975. A Vector Space Model for Automatic Indexing. *Communications of the ACM* 18(11): 613-620.

Shermis, M., and Burstein, J. eds. 2003. *Automated Essay Scoring: A Cross-Disciplinary Perspective*. Hillsdale, NJ: Lawrence Erlbaum Associates.