14.48, Economics of Education
Prof. Frank Levy
Lecture 21

Guest Lecturer: Jill Burnstein, from ETS.

Jill has a PhD in Linguistics and is interested in natural language processing and computational linguistics.

ETS works on educational measurement and administers tests such as the TOEFL, GRE and GMAT.

One component of K-12 education is to help students master basic writing skills such as grammar, spelling, punctuation, etc. However, writing classes take up a lot of teacher time because of the time-intensive nature of grading pieces of writing. Additionally, statewide tests that have writing components require a large number of grading hours.

In general, on statewide tests and other standardized tests, essays are graded by 2 readers and the scores are averaged. If the 2 scores are different, a 3rd reader is used and the essay is given the most common score.

Criterion is targeted at middle school students. Initial reactions from English teachers were very negative. They complained that essay scoring technology stifled creative writing and could not pick up writing techniques such as allegory, similes, irony, etc. However, this software is not meant for creative writing. It is meant to evaluate the basic, 5 paragraph essay that is the basis of US writing education.

Criterion is consistent in its grading while human scorers are not, especially when they often grade for hours at a time. Some people become more lenient graders when they are tired while others become harsher.

History:

1) PEG, 1960s essay scoring technology was the first of its kind. It was developed by Page and basically transformed essay length into a score and incorporated limited syntactic analysis. It is important to note that human scorers often associate longer essay with better quality and give longer essays higher scores
2) Writer's Workbench (Cherry et al, 1982) was a writing editing tool intended for students. However, it was Unix based and, therefore, had a very limited audience.
3) Intelligent Essay Assessor
4) E-Rater (Burnstein et al, 1998) used natural language processing and analyzed aspects of writing such as grammar, mechanics, usage and vocabulary

Market Readiness:

In the 1960s, there were no computers in American classrooms and essays were done by hand so there was very little use for PEG since no one had the hardware available to use it. However, since the 1960s there has been a huge increase in school computer access. In fact, state assessment are now frequently administered

on computers, making essay scoring technology much more viable to a large market of school, teachers, and students.

Motivations:

1) Making more writing possible in the classroom
2) Allowing students to create electronic writing portfolios
3) Individual performance assessment
4) Classroom assessments

Criterion offers cost savings for large-scale assessments. Additionally, there is a market for practice assessments for the GRE, TOEFL, etc and Criterion is often used by clients practicing for these tests.

Educational and Business Considerations:

1) Assessment type: High stakes? Is the software reliable?
2) Reliability benefits: does Criterion increase consistency for assessments?
3) Cost/Performance: are there actually cost benefits?

Technology Development Considerations:

Qualities of a well-written essay:
-Clearly states the author's position
-Well-organized
-Develops its arguments
-Varied sentence structure
-Good word choice

Mapping Writing Features to Natural Language Processing (NLP) Tools:

Writing Features:	NLP Tools:
-grammar Errors, sentence structures	-part of speech (pos) taggers, syntactic parsers
-vocab usage	-content analysis, pos taggers
-sentence, word level mechanics	-spelling tool, pos taggers
-organization and development of ideas	-discourse analyzers

1st version of E-Reader: 02/99 – 09/04
-Writing relevant features: syntactic structure, discourse structure, content, lexical complexity
-NLP tools
-Scoring analysis

**In a timed situation, there is little room for creativity. Additionally, readers spend approx 3 min on each essay, trying to get a general impression and not read every word.

If a human reader grades consistently, there must be some set of rules that they follow. If these rules can be articulated, then the graders job can be programmed.

The 5 paragraph essay is seen as a starting point for writing, not where everyone should end up. It is the basis for more complex, creative writing.

This service and state assessments are used to bring the bottom performers up to a certain level of competence but can hold back the more advanced/creative writers.

E-reader was first used on GMAT scores, but isn't used on them anymore.

It is important to realize that E-reader and Criterion are meant for good faith writing. You can trick the software if you want to, but that's not the intended use. ETS has developed some methods for detecting poor writing and those trying to trick the system.

Educational Considerations:
Criterion gives students immediate feedback on their writing; they can see their score go up as they improve.

Technology Development Decisions:
Free: Spell check
Not free, already developed: Grammar check, usage, mechanics and style analyzers
Not developed: essay based organization and development analyzers

Organization and Development: Essay based discourse analyzer

Uses a machine learning approach:
-Probable discourse label sequences
-essay sub language: agree, should, would, opinion, because, however
-rhetorical parse info: contrast, elaboration, antithesis
-syntactic structure: infinitive, complement, subordinating clauses

E-rater version 2: 09/04 – the present
None of the features in E-rater correlate strongly with essay length
*Set of 8-10 features relevant to writing standards
-Grammar, usage, mechanics: error types
-Style: repeated words, sentence length, sentence type
-Organization: thesis, main point, support, conclusion
-Content: vocab usage
-E-rater/Human agreement: increased from 50% in the 1st version to 62%, 98% exact+adjacent
-Builds scoring models with multiple regression, human training

This year: Criterion added essay planning templates which gives ETS more data on how people plan and its correlation with performance on the essay.

Criterion offers a less risky environment for students to get feedback on their writing before they have to turn anything in to their teachers.

Results: Between the 1st and last draft that a student submits, there is a 25% error reduction.

See http://ets.org/research/erater.html for more information