Automated Essay Scoring for Nonnative English Speakers
Jill Burstein, ETS
Martin Chodorow, Hunter College of CUNY & ETS
7/1/99

Click here to start

Table of Contents
Automated Essay Scoring for Nonnative English Speakers
Outline of Presentation
The e-rater System
Flowchart of Scoring With Human Raters
Flowchart of Scoring With Human Rater and e-rater
50+ Rubric-Based Features for Scoring
Syntactic Analysis
Discourse Annotation & Partitioning Heuristics
Discourse Annotation
Topical Analysis
Building Models & Scoring
Generalizing to L2
Baseline (chance) Performance
Overall Exact + Adjacent Agreement
TWE1 Exact + Adjacent Agreement
TWE 2 Exact + Adjacent Agreement
Summary of E-rater Performance for Language Groups
E-rater Summary & Further Research

Email:
jburstein@ets.org
mchodorow@ets.org

Home Page:
http://www.ets.org/research/erater.html
Automated Essay Scoring for Nonnative English Speakers

Jill Burstein
Educational Testing Service

Martin Chodorow
Hunter College of CUNY
&
Educational Testing Service
Outline

- *E-rater* Goals and Methods

- *E-rater* Application to L2 Writing Samples
The e-rater™ System

- The NEED
  automate holistic scoring of essays
  0-6 point score scale

- The METHOD
  adapt & combine existing NLP techniques

- The RESULTS
  compare performance with human readers
50+ Rubric-Based Features for Scoring

- Syntactic Structure Features
  - Subordinate, Relative, Infinitive ... clauses

- Topic / Content Features
  - “score” from ranking content words in essay
  - “score” from ranking content words in each essay argument

- Rhetorical / Discourse Structure Features
  - parallel, contrast, evidence ... words that begin or develop arguments
SYNTACTIC ANALYSIS

sentence No. 4
S> The author is making several assumptions that should be questioned.

0 S
1 NP [d: The (The)]
2 NP [m: author (author)]
3 be [z: is (be)]
4 vbg [g: making (make)]
5 NP [i: several (several)]
6 SC COMP [w: that (that)]
7 MD [d: should (should)]
8 be [e: be (be)]
9 NP [n: questioned (question)]
10 U [p: . (.)]
Discourse Annotation & Partitioning Heuristics

- Heuristics For Lexically & Structurally-based Discourse Annotation
- Cue Word Classification-Based Lexicon
- Heuristics To Partition Essays Based on Annotation
- Argument = *a new discussion point in an essay*
Discourse Annotation & Partitioning

Hierarchical...

Sentence 1: It is also assumed that reducing high school enrollment may lead to a shortage of qualified engineers.

... 

Sentence 2: It is conceivable that some programs such as art, music, or social sciences will be more affected by this drop in high school population.

...
Topical Analysis  
(vector-space model of essay similarity)

- 2 classifiers are trained on content words

  Essay-Based Approach  
  Argument-Based Approach

- Predicted scores are used as e-rater features
Building Models & Scoring

- **Building Test Question Models**
  - **Collect** Feature Info from Sample Essays
  - **Generate** Weighted Predictive Feature Set With Regression for Each Test Question

- **Scoring Essay Responses**
  - **Use** Weighted Predictive Feature Set In Formula for Score Prediction
Generalizing to L2

- Samples from 2 Test of Written English Prompts
- Approximately 1700+ Essays
  - 255 For Model-Building
  - 550+ For Cross-Validation
- Arabic, Chinese and Spanish Speakers
- Results
  - Overall e-rater Performance
  - E-rater Performance for Subgroup Languages
Baseline (chance) Performance

- 83% exact or adjacent agreement if modal score 5 is always assigned
- 74% exact or adjacent agreement if scores are randomly assigned in proportion to their relative frequency in the population

![Distribution of Scores](image)
## Overall Exact + Adjacent Agreement

<table>
<thead>
<tr>
<th>Prompt</th>
<th>n</th>
<th>%</th>
<th>Pearson r</th>
<th>GDF Score</th>
<th>E Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean</td>
<td>S.D.</td>
<td>Mean</td>
</tr>
<tr>
<td>TWE1</td>
<td>562</td>
<td>91.1</td>
<td>6.67</td>
<td>4.16</td>
<td>974</td>
</tr>
<tr>
<td>TWE2</td>
<td>576</td>
<td>93.4</td>
<td>7.18</td>
<td>4.16</td>
<td>936</td>
</tr>
<tr>
<td>Mean</td>
<td>92.3</td>
<td>693</td>
<td>4.16</td>
<td>955</td>
<td>4.08</td>
</tr>
</tbody>
</table>

*Difference in mean score between GDF and e-rater was statistically significant (F_{1,1123} = 5.469, p < 0.5).

*No significant main effect for Prompt, and no interactions between Prompt and the other factors.*
## TWE1

### Exact & Adjacent Agreement

<table>
<thead>
<tr>
<th>Language Group</th>
<th>n=</th>
<th>%</th>
<th>Pearson r</th>
<th>GDF Score</th>
<th>E Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>146</td>
<td>89.0</td>
<td>.645</td>
<td>3.83</td>
<td>.973</td>
</tr>
<tr>
<td>Chinese</td>
<td>153</td>
<td>88.2</td>
<td>.543</td>
<td>4.09</td>
<td>.884</td>
</tr>
<tr>
<td>Spanish</td>
<td>131</td>
<td>92.4</td>
<td>.644</td>
<td>3.96</td>
<td>.986</td>
</tr>
<tr>
<td>US-English</td>
<td>97</td>
<td>96.9</td>
<td>.632</td>
<td>4.96</td>
<td>.624</td>
</tr>
<tr>
<td>Non-US English</td>
<td>35</td>
<td>91.4</td>
<td>.544</td>
<td>4.31</td>
<td>.900</td>
</tr>
</tbody>
</table>
**TWE 2**  
**Exact + Adjacent Agreement**

<table>
<thead>
<tr>
<th>Language Group</th>
<th>n=</th>
<th>%</th>
<th>Pearson r</th>
<th>GDF Score</th>
<th>E Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>151</td>
<td>96.4</td>
<td>.783</td>
<td>3.85</td>
<td>9.59</td>
</tr>
<tr>
<td>Chinese</td>
<td>139</td>
<td>91.0</td>
<td>.707</td>
<td>3.92</td>
<td>9.57</td>
</tr>
<tr>
<td>Spanish</td>
<td>138</td>
<td>93.5</td>
<td>.616</td>
<td>4.07</td>
<td>8.45</td>
</tr>
<tr>
<td>US English</td>
<td>103</td>
<td>92.0</td>
<td>.519</td>
<td>4.83</td>
<td>6.13</td>
</tr>
<tr>
<td>Non-US English</td>
<td>45</td>
<td>93.3</td>
<td>.465</td>
<td>4.68</td>
<td>.732</td>
</tr>
</tbody>
</table>
Summary of E-rater Performance For Language Groups

- Two English Groups Scored Significantly Higher Than Normative Speakers
- Interaction of Language Group by Reader was Significant
- $\chi^2$ Analyses Showed No Significant Differences on the Agreement Measure for Language Group
- Effect of Prompt in the Analysis of Agreement for Arabic speakers, where Agreement levels in TWE1 and TWE2 were Significantly Different
E-rater Summary &
Further Research

- E-rater Overall Agreement ~92%
- Operational Agreement ~92%
- Small Main Effect for Reader Across Prompts; e-rater ↓
- Small Reader Effect on Language Group
- Small Effect of Prompt on Agreement for Arabic Speakers
- Larger-Scale Studies with TOEFL Essays