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

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

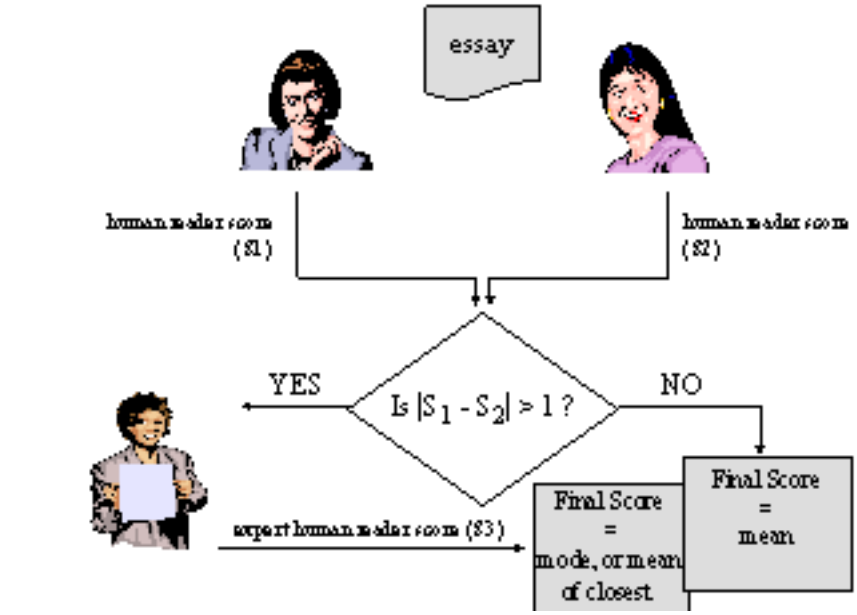
- *E-rater* Goals and Methods
- *E-rater* Application to L2 Writing Samples

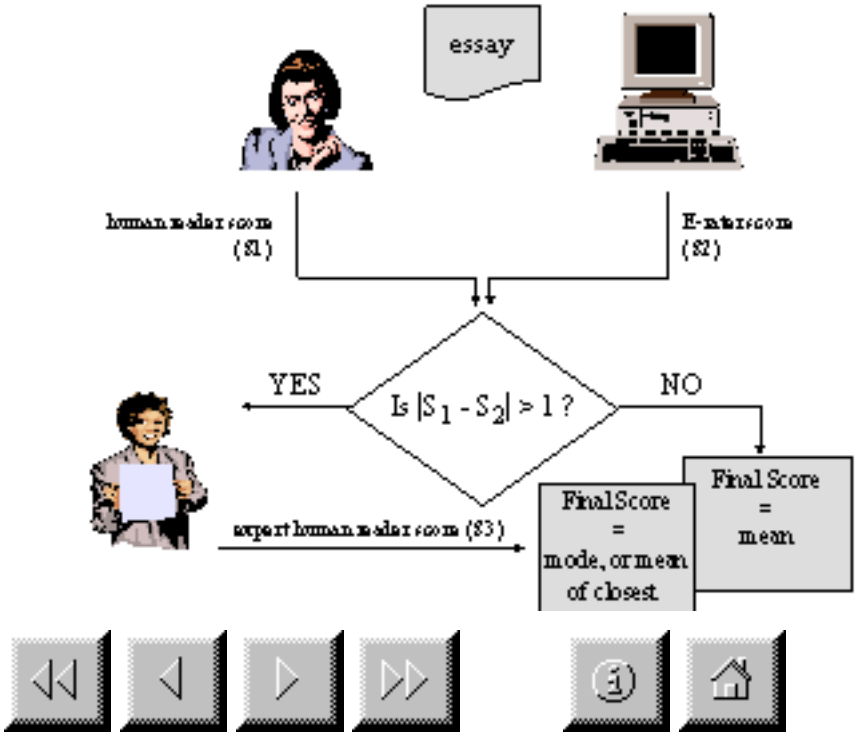
The *e-rater*™ System

- 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


essay
human reader score (S1)
human reader score (S2)
YES
Is $|S_1 - S_2| \geq 1$?
NO
expert human reader score (S3)
Final Score = mode, or mean of closest
Final Score = mean


essay
human rater score (S1)
E-rater score (S2)
YES
Is $|S_1 - S_2| \geq 1$?
NO
expert human rater score (S3)
Final Score = mode, or mean of closest
Final Score = mean

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    NP    |dt The (The)
1            |nn author (author)
2          |bez is (be)
3          |vbg making (make)
4       NP      |jj several (several)
5               |nns assumptions (assumption)
6            SC   COMP  |wdt that (that)
7                       |md should (should)
8                       |be be (be)
9                       |vbn questioned (question)
10 U    |per .(.)
```

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

New Paragraph:

...

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

- eg_initPARALLEL = *also*
- eg_initCLAIM_THAT = *that*
- eg_auxSPECULATE = *may*

...

Sentence 3: It is conceivable that other programs such as arts, music or social sciences will be most affected by their drop in high school population.

...

- eg_devSAME_TOPIC = *It*
- eg_devCLAIM_THAT = *that*
- eg_devDETAIL = *such_as*

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

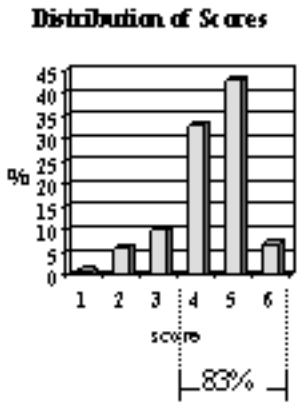
- **Scoring Essay Responses**
  - **Use** Weighted Predictive Feature Set In Formula for Score Prediction

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



Overall Exact + Adjacent Agreement

Prompt	n=	%	Pearson r	GDF Score		E Score	
				Mean	S.D.	Mean	S.D.
TWE1	562	91.1	.667	4.16	.974	4.08	1.041
TWE2	576	93.4	.718	4.16	.936	4.07	.989
Mean		**92.3**	**.693**	**4.16**	**.955**	**4.08**	**1.015**

- Difference in mean score between GDF and *e-rater* was statistically significant ($F_{(1,1128)} = 5.469$, $p < .05$).

- No significant main effect for Prompt, and no interactions between Prompt and the other factors.

TWE1
Exact & Adjacent Agreement

Language Group	n=	%	Pearson r	GDF Score		E Score	
				Mean	S.D.	Mean	S.D.
Arabic	146	89.0	.645	3.83	.973	3.67	.947
Chinese	153	88.2	.543	4.09	.884	4.12	1.00
Spanish	131	92.4	.644	3.96	.986	3.70	.915
US-English	97	96.9	.632	4.96	.624	4.93	.814
Non-US English	35	91.4	.544	4.31	.900	4.51	.981

TWE 2
Exact + Adjacent Agreement

Language Group	n=	%	Pearson r	GDF Score		E Score	
				Mean	S.D.	Mean	S.D.
Arabic	151	96.4	.783	3.85	.959	3.70	.909
Chinese	139	91.0	.707	3.92	.957	4.04	1.03
Spanish	138	93.5	.616	4.07	.845	3.69	.733
US-English	103	92.0	.519	4.83	.613	4.95	.759
Non-US English	45	93.3	.465	4.68	.732	4.60	.780

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
- χ^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