Improving automated oral testing: identifying features and enhancing speech recognition

The BYU PSST Research Group
http://psst.byu.edu
lonz@byu.edu
Acknowledgements

• Deryle Lonsdale, Ray Graham, and Dan Dewey

• Jeremiah McGhee, Aaron Johnson, Ross Hendrickson, Meghan Eckerson, Malena Weitze, Ben Millard, Kevin Cook, Ranjan Dhungel, Peter McClanahan

• BYU Office of Research and Creative Activities, Center for Language Studies
Elicited imitation (EI) testing

• NNS oral proficiency test for English
• Subjects repeat isolated sentences of varying complexity (60 / testing session)
• Responses are recorded and scored, usually at syllable (σ), item levels
• Rationale: subjects can’t process linguistic vocabulary, structures they don’t know yet
EI advantages: can be

- Administered to multiple learners at the same time
- Administered in a computer lab
- Administered with less cost/time
- Scored by a reasonably proficient speaker of English
- “A reasonable measure of global proficiency” (Bley-Vroman & Chaudron, 1994)
EI testing so far

• Developed about 280 sentences
  – Varying length, complexity, features
• 1500 tests administered to about 1050 subjects since Fall 2006
• Random sampling of subjects also given other tests (ECT speaking, OPI, oral placement, LAT speaking)
• Relatively simple application, standard language lab setting
Sample EI sentences

Discriminate well

• Perhaps he works there.
• Had you ever flown that high before?
• Good cars will never break down.
• When she went to Las Vegas, did she like the shows that she saw?
• If her heart were to stop beating, we might not be able to help her.

Don’t discriminate well

• Have you slept?
• Maybe she likes cats.
• We eat cookies.
• How do good children play baseball?
• Chris has yelled louder than ten sheep.
• He should have walked away before the fight started.
Assessing lexical features

• New items engineered to investigate lexical complexity
  – Lexical density, lexical difficulty (frequency, morphological composition)
  – Characteristics reflect:
    • 6 frequency bands (ranges)
    • 5 $\sigma$-count bands (4-6, 7-9, 10-12, 13-15, 16+)

• Items scored, IRT analysis

• Factors: $\sigma$-count (.73), frequency (.08), lexical density/complexity (.02)
Lexical features: interaction
Assessing syntactic features

• New items engineered to investigate role of syntactic features
  – 44 features from L2 acquisition studies, OPI test guidelines
    • Tense, modality, aspect, transitivity, articles, agreement, contractions, possessives, etc.

• Items scored, IRT analysis

• Factors: σ-count (.68), 3rdPersAgr (.02), negation, imperfect, copula… (≈.01 each)
Discriminative value vs. $\sigma$ count
Scoring since last workshop

- GUI tool developed to help human annotators score EI test items
- Deployed on the Web
- 1250 tests, 810 subjects, 44 graders
- \( \approx 55,500 \) items graded, \( \approx 596,000 \) σ
- Avg. time: \( \approx 50 \) sec./item, 72 items/hr.
Scoring an EI item

Grade Item: 71052

http://psst.byu.edu/EI/Grader/audio/1/9/4/0/563671733/3024.mp3

Insertions
- Word/Syllable Transposition(s)
- Word/Syllable Insertion(s)
- Audio Clipping
- Sneezes/Coughs/Background Noise

Questions or Comments
Sneezing and background conversations

Set all Zero
Set all One
Set Remaining 0
Set Remaining 1
New Item
Submit
Agreement among scorers

• $\approx 175,000$ σ double-graded (so far)

• 91% agreement (raw %, Robinson’s A)

• IRR: 0.82 (Krippendorff’s $\alpha$, Cohen’s $\kappa$, mean of bivariate rank correlations, …)

• Rater bias coefficient: 0.576, $\chi^2=362$

• Exploring team-wise analysis, arbitration, viability of single scoring
ASR and automatic EI scoring

• Discussed at length in last year’s workshop, LREC 2008
• Sphinx, WSJ
• Correlations of between 0.85 and 0.88 with human scores
• Have now trained up acoustic model from EI native model utterances, evaluating
• Ongoing:
  – Develop NNS accented English acoustic models? w/r t L1? gender? both?
  – Fluency measures: pauses, filled pauses, restarts
Holistic evaluations

• Two ways of looking at EI scoring from a holistic perspective
  – Impressionistic ranking of overall intelligibility, grammaticality, fluency
  – Taking into consideration demographic information on test subjects
    • L1, age, scores for reading, writing, listening comprehension, etc.

• Machine learning
  – TiMBL (provides feature rankings)
Data mining for ranking items (1)

- Assumption: OPI score is gold standard
- Rank EI items for predictive value
  - Better item more predictive of OPI score?
  - Attributes: EI score, Item ID, Student ID
  - Label: OPI score
- 34 students, 2600 items
- 80%/20% training/testing split
- WEKA linear regression (default values)
Data mining for ranking items (2)

- Some feature selection on full set of items
- Reduced dataset of 15 items: lowest sum squared error, more predictive of OPI score
- LR model: $\text{OPI} = S_{1008} \times 2.3414 + 3.0987$
- Item 1008: Had he ever played games well?
- Next: full range of data scored so far

<table>
<thead>
<tr>
<th>Metric</th>
<th>All 60 sentences</th>
<th>15 selected sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation coefficient</td>
<td>.4468</td>
<td>0.9225</td>
</tr>
<tr>
<td>Mean absolute error</td>
<td>1.5518</td>
<td>0.6277</td>
</tr>
<tr>
<td>Root mean squared error</td>
<td>1.668</td>
<td>0.7286</td>
</tr>
</tbody>
</table>
Summary

- Ongoing work on several fronts:
  - Administering tests
  - EI item development w/r/t targeted feat’s
  - Human scoring, annotation
  - Data analysis (machine learning, data mining)
  - ASR scoring of EI responses
  - Other language EI tests (Japanese, French)

- Ultimate goal: online, adaptive testing tool for assessing proficiency levels