Exploring Computerized Lexical Analysis to Predict Calibrated Peer Review Ratings of Student Writing in Chemistry

Kevin C. Haudek
Mark Urban-Lurain
Michigan State University

Arlene A. Russell
University of California – Los Angeles
Acknowledgements

- Automated Analysis of Constructed Response Research Group (AACR)
  - Michigan State University
  - University of California – LA
  - SUNY – Stony Brook
  - University of Colorado – Boulder
  - University of Maine
  - University of Georgia
  - Western Michigan University

This work was supported by the National Science Foundation (Grants DUE 1022653, DUE 1143642, DUE 04-42828, DUE 08-37229 and DUE 08-16660). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.
Overview

- Paper on CD
- Background
  - Computerized text analysis
  - Calibrated Peer Review
- Research Questions
- Methods
- Results
- Future directions
Assessment to Reveal Student Thinking
Background

- Encouraging writing in STEM courses
  - Authentic task for scientists
  - Students must actively construct explanations
  - Better reveals student understanding
- Major barrier: evaluating large number of student responses
- Two approaches:
  - Computerized text analysis
  - Calibrated peer review (CPR)
Computerized Text Analysis

- Software extracts and categorizes terms and phrases in electronic text
  - *Categories* can be pre-defined, based on expert answer, or emergent, through student writing
- Lexical categories used as independent variables to predict expert scoring
  - Scoring models approach human-human IRR
  - Requires large numbers of human scored responses
- More information:
  - www.msu.edu/~aacr/
IBM SPSS Text Analytics

Responses with extracted terms highlighted

During strenuous physical activity, the blood cannot supply enough oxygen to muscles for ATP synthesis; to compensate pyruvate created during glycolysis is reduced to lactic acid for energy production...
AACR Question Development Cycle

- Predictive Model
- Confirmatory Analysis
- Human Coding
- Rubric Development
- Exploratory Analysis
- New Question Design
- Data Collection
- Lexical Resource Development
- Question Revision

The cycle begins with the New Question Design, followed by Data Collection, Lexical Resource Development, Question Revision, Human Coding, Confirmatory Analysis, and Predictive Model. This cycle is designed to ensure that questions are developed and refined through a series of analytical steps.
Calibrated Peer Review (CPR)

- Each student
  - Submits an essay
  - Peer Reviews 3 other essays

- Reviewer Calibration Index (RCI)
  - Each student uses rubric to score 3 known essays
  - Overall score weighted-average of peer scores

- More information:
  - http://cpr.molsci.ucla.edu
Research Questions

- How well do previously developed automated analysis techniques transfer and predict trained peer scoring of longer essays?
  - Previous resources developed for shorter responses about general chemistry

- What improvements can be made in scoring models by including additional information about the text?
  - Limited semantic information and writing-quality evaluation
Methods: CPR Assignment

○ Assignment
  ● Write about bicarbonate buffering capabilities in the bloodstream
    ○ Given learning goals, source materials and seven “guiding prompts” to aid writing
      ● What other pathways does the body use to remove excess protons, carbon dioxide or bicarbonate from the blood?

○ Reviewing
  ● Rubric
    ○ 2 Writing quality items: topic sentence and free of errors
    ○ 9 Content items:
      ● Is the removal of bicarbonate ion through the kidneys discussed?
  ● Assign overall score: 1 - 10
Write a paragraph of the required length on how the bicarbonate system maintains homeostasis in the blood system. In your paragraph be sure to address the issues raised in the "Guidance for Writing Your Text". Remember that you are writing a paragraph, not just answering a list of questions, so you should include a topic sentence and a summary statement.

Your text should also include both a topic sentence and a summary sentence. A topic sentence should introduce the concepts and ideas relevant to the rest of the essay. A summary sentence should summarize the ideas and concepts discussed in the essay.
Methods: Text Analysis

- Applied *term libraries* developed for general chemistry items
- *Word proximity* program to search for common word pairs
- Modeling peer scoring
  - Lexical analysis and word proximity variables in step-wise regression
Results: Response Characteristics

- 2nd semester general chemistry students
  - N = 388
- Mean response character length
  - 1681 ± 91 (std. dev)
- Mean CPR weighted score
  - 6.2 ± 1.6
- Total number of lexical categories = 119
  - 4 of which were used by every response
  - Some categories used infrequently (e.g. ATP, temperature)
- Mean categories / response
  - 43 (range 31-55)
Results: Predictive Scoring Model Using Lexical Categories

- 18 lexical categories
- $R^2 = 0.465$

<table>
<thead>
<tr>
<th>Variable</th>
<th>Beta</th>
<th>Variable</th>
<th>Beta</th>
</tr>
</thead>
<tbody>
<tr>
<td>Character count</td>
<td>0.134</td>
<td>energy</td>
<td>0.106</td>
</tr>
<tr>
<td>increase pH</td>
<td>0.268</td>
<td>glucose</td>
<td>0.103</td>
</tr>
<tr>
<td>ratio</td>
<td>0.267</td>
<td>LeChatelier’s principle</td>
<td>0.086</td>
</tr>
<tr>
<td>bases - general</td>
<td>0.236</td>
<td>strong acid</td>
<td>-0.094</td>
</tr>
<tr>
<td>strong</td>
<td>0.166</td>
<td>Environment</td>
<td>-0.107</td>
</tr>
<tr>
<td>constant</td>
<td>0.160</td>
<td>Help</td>
<td>-0.124</td>
</tr>
<tr>
<td>water</td>
<td>0.154</td>
<td>acids – general</td>
<td>-0.139</td>
</tr>
<tr>
<td>glucose breakdown</td>
<td>0.131</td>
<td>ions</td>
<td>-0.152</td>
</tr>
<tr>
<td>kidneys</td>
<td>0.124</td>
<td>extracellular</td>
<td>-0.153</td>
</tr>
</tbody>
</table>
Results: Predictive Scoring Model Using Lexical Categories

- 18 lexical categories
- $R^2 = 0.465$

<table>
<thead>
<tr>
<th>Variable</th>
<th>Beta</th>
<th>Variable</th>
<th>Beta</th>
</tr>
</thead>
<tbody>
<tr>
<td>Character count</td>
<td>0.134</td>
<td>energy</td>
<td>0.106</td>
</tr>
<tr>
<td>increase pH</td>
<td>0.268</td>
<td>glucose</td>
<td>0.103</td>
</tr>
<tr>
<td>ratio</td>
<td>0.267</td>
<td>LeChatelier’s principle</td>
<td>0.086</td>
</tr>
<tr>
<td>bases - general</td>
<td>0.236</td>
<td>strong acid</td>
<td>-0.094</td>
</tr>
<tr>
<td>strong</td>
<td>0.166</td>
<td>Environment</td>
<td>-0.107</td>
</tr>
<tr>
<td>constant</td>
<td>0.160</td>
<td>Help</td>
<td>-0.124</td>
</tr>
<tr>
<td>water</td>
<td>0.154</td>
<td>acids – general</td>
<td>-0.139</td>
</tr>
<tr>
<td>glucose breakdown</td>
<td>0.131</td>
<td>ions</td>
<td>-0.152</td>
</tr>
<tr>
<td>kidneys</td>
<td>0.124</td>
<td>extracellular</td>
<td>-0.153</td>
</tr>
</tbody>
</table>
Results: Predictive Scoring Model Using Lexical Categories

- 18 lexical categories
- \( R^2 = 0.465 \)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Beta</th>
<th>Variable</th>
<th>Beta</th>
</tr>
</thead>
<tbody>
<tr>
<td>Character count</td>
<td>0.134</td>
<td>energy</td>
<td>0.106</td>
</tr>
<tr>
<td>increase pH</td>
<td>0.268</td>
<td>glucose</td>
<td>0.103</td>
</tr>
<tr>
<td>Ratio</td>
<td>0.267</td>
<td>LeChatelier’s principle</td>
<td>0.086</td>
</tr>
<tr>
<td>bases – general</td>
<td>0.236</td>
<td>strong acid</td>
<td>-0.094</td>
</tr>
<tr>
<td>Strong</td>
<td>0.166</td>
<td>Environment</td>
<td>-0.107</td>
</tr>
<tr>
<td>Constant</td>
<td>0.160</td>
<td>Help</td>
<td>-0.124</td>
</tr>
<tr>
<td>water</td>
<td>0.154</td>
<td>acids – general</td>
<td>-0.139</td>
</tr>
<tr>
<td>glucose breakdown</td>
<td>0.131</td>
<td>ions</td>
<td>-0.152</td>
</tr>
<tr>
<td>kidneys</td>
<td>0.124</td>
<td>extracellular</td>
<td>-0.153</td>
</tr>
</tbody>
</table>
Results: Improving Scoring Using Writing Quality Scores

- Added writing quality rubric items
- 24 variables chosen
  - 17 shared with initial model
- \( R^2 = 0.551 \)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Beta</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rubric – Free of errors</td>
<td>0.233</td>
</tr>
<tr>
<td>Rubric – Topic sentence</td>
<td>0.123</td>
</tr>
<tr>
<td>Character count</td>
<td>0.154</td>
</tr>
<tr>
<td>increase pH</td>
<td>0.254</td>
</tr>
<tr>
<td>ratio</td>
<td>0.236</td>
</tr>
<tr>
<td>bases - general</td>
<td>0.202</td>
</tr>
<tr>
<td>water</td>
<td>0.169</td>
</tr>
<tr>
<td>glucose</td>
<td>0.132</td>
</tr>
<tr>
<td>strong</td>
<td>0.128</td>
</tr>
<tr>
<td>glucose breakdown</td>
<td>0.127</td>
</tr>
<tr>
<td>constant</td>
<td>0.120</td>
</tr>
<tr>
<td>kidneys</td>
<td>0.109</td>
</tr>
<tr>
<td>exchange</td>
<td>0.082</td>
</tr>
<tr>
<td>carbonate ion</td>
<td>0.075</td>
</tr>
<tr>
<td>named chemical compound</td>
<td>-0.081</td>
</tr>
<tr>
<td>function</td>
<td>-0.082</td>
</tr>
<tr>
<td>phosphate buffer</td>
<td>-0.084</td>
</tr>
<tr>
<td>strong acid</td>
<td>-0.097</td>
</tr>
<tr>
<td>lungs</td>
<td>-0.101</td>
</tr>
<tr>
<td>transport</td>
<td>-0.113</td>
</tr>
<tr>
<td>help</td>
<td>-0.123</td>
</tr>
<tr>
<td>ions</td>
<td>-0.124</td>
</tr>
<tr>
<td>acids - general</td>
<td>-0.130</td>
</tr>
<tr>
<td>extracellular</td>
<td>-0.144</td>
</tr>
</tbody>
</table>
Results: Improving Scoring Using Writing Quality Scores
Conclusions

- Existing lexical resources applied in new context
  - Additional semantic information improves model
- Successfully analyzed long, complex, scientific essays
- Writing quality important consideration in peer review
- Model best at predicting “average responses”
  - Need additional data to improve accuracy at ends
Future Directions

- Balancing generalizable scoring models with instructor calibration
  - Unique calibration and evaluation rubrics limits predictive model’s transferability
- Exploring how students use rubrics
  - Which factors effect how student assign overall scores?
Questions

- Mark Urban-Lurain
  - urban@msu.edu

- Automated Analysis of Constructed Response research group (AACR)
  - www.msu.edu/~aacr
  - Related talk: Tuesday, 8:30 – 10:00, Strand 5
    - Michele Weston: *Comparing Formative Feedback Reports: Human and Machine Analysis of Constructed Response Questions in Biology*

- Calibrated Peer Review (CPR)
  - cpr.molsci.ucla.edu