Rubric-based Automated Japanese Short-answer Scoring and Support System Applied to QALab-3

Only Evaluation-method Task

Tsunenori Ishioka (Japan)
Kohei Yamaguchi (Japan)
Tsunenori Mine (Japan)
Outline

- Status of Written Test
- Specification of our System
- The Performance Evaluation
- Summary

© edsurge.imgix.net
Writing test

☐ Essays
  ■ No model answer; 300~ words
  ■ Rhetoric, logics, contents
  ■ Practical application: e-rater, Intellimetric, and Jess (for Japanese)

☐ Short-answers
  ■ Model answer(s); 1~2 statement(s)
  ■ Semantic identity / Recognizing textual entailment
  ■ Technical difficulty
  ■ AES: Not applied for high-stakes test(s)
Support system for short written tests

- Accurate language understanding is almost impossible
  - Agreement to scoring rubric
  - Automated Scoring by rubric
  - Classification scoring by ML
  - Overwriting the score ← Human rater

- Simple scoring rubric description
  - Automatic creation of scoring screen
System Components

1. Scoring by Professional human raters; DB; Scoring model by RF
2. Automatic creation of scoring screen
3. Scoring via Web; Results in DB
4. Reconstruction of new learning model
古代エジプトは、ナイル川を中心に長らく独立王朝が栄えたが、アレクサンドロス大王などの征服を受ける。プトレマイオス朝エジプトのクレオパトラは、

part1 2 古代エジプトは、ナイル川を中心に、古王国から新王国まで、長らく独立王朝が栄えた。

part2 2 古代エジプトは、アケメネス朝やアレクサンドロス大王の征服を受けた。

ナセルは、スエズ運河の国有化、アラブ連合共和国の合邦など、多くの事績をあげた。

lack 1 -5 アクティウムの海戦  mandatory words; When missing, reduce the points

vol /2 540  volume operation max [- min]
## Performance Statistics

A full mark is **20 points**

Professional evaluations were **all zero**.

<table>
<thead>
<tr>
<th>Issue</th>
<th>$\hat{\chi}$</th>
<th>$\frac{\sum(x - x_0)}{n}$</th>
<th>$\frac{\sum(x - x_0)^2}{n}$</th>
<th>Ref. all our predicted values($n=18\sim19$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>0,0,0,2</td>
<td>0.50</td>
<td>1.00</td>
<td>0 × 11, 2, 8, 14, 15 × 4</td>
</tr>
<tr>
<td>C</td>
<td>0,0,0,0</td>
<td>0.00</td>
<td>0.00</td>
<td>0 × 13, 3, 9, 12 × 2, 18 × 2</td>
</tr>
<tr>
<td>G</td>
<td>0,0,0,3</td>
<td>0.75</td>
<td>2.25</td>
<td>0 × 10, 2, 3, 7, 8, 9, 19 × 4</td>
</tr>
<tr>
<td>L</td>
<td>5,0,0,4</td>
<td>2.25</td>
<td>10.3</td>
<td>0 × 9, 4, 5, 8, 9, 11, 12, 14 × 2, 19 × 2</td>
</tr>
<tr>
<td>P</td>
<td>0,4,0,4</td>
<td>2.13</td>
<td>9.06</td>
<td>0 × 8, 4, 4.5 × 6, 5 × 2, 7.5, 9</td>
</tr>
</tbody>
</table>

$\chi^2 = \sum \frac{(x - x_0)^2}{n}$
Comments on Evaluation Indicators ($\rho$ or $\tau$)

- How to predict professional evaluations well
  - When $x_0 \equiv 0$, indices based on the correlation are inappropriate. \( \therefore \sigma = 0 \)
  - $\rho$ or $\tau$ with only 4 data has almost no meaning; the D.F. is only 2 (=4-2).
  - The residential errors are natural and proper.
Conclusion

- Our system can show a certain degree of validity
  - Returned a score close to zero

- Our technique
  - Based on the scoring rubric
  - Considering **superficial** and **semantic** aspects (LSI)
  - Sufficiently suitable
Thank you for your attention.