Spelling Mistakes & Typoos: Can Your ITS Handle Them?

Adam M. Renner*
Philip M. McCarthy*
Chutima Boonthum*
Danielle S. McNamara*

*University of Memphis, Psychology / Institute for Intelligent Systems
*University of Memphis, English / Institute for Intelligent Systems
*Hampton University, Computer Science

Intelligent Tutoring Systems

- Provide assessment of user input
- Guided feedback based on user's response
- Many ITSs use conversational dialogue
- NLP for assessment and determines feedback
  - Input matched to benchmark
  - Assessed for similarity
- Assessment limited to proficiency of user
  - High school students or younger
  - Make typing errors/spelling mistakes
- What the student intended

ITS User-Language

- Contains high rate of typographical & grammatical errors
  - Not a new issue in NLP
- Traditional spellchecking not suitable (e.g., MS Word, email)
- ITSs necessitate automatic corrections
  - Why2-Atlas (VanLehn et al., 2002)
  - CIRCSIM-Tutor (Elmi & Evens, 1998)
  - Many more just ignore errors
- NLP tools thought resistant to errors
  - LSA (Landauer et al., 2007) – semantic overlap across two whole texts
  - Short responses?
  - Responses with multiple errors?
  - NLP tools trained on edited text
  - When used in ITS, similarity assessment inevitably affected

Problems with Evaluating User-Language

- Lack of "colloquial" paraphrase corpora
  - Microsoft Research Paraphrase Corpus (Dolan, Quirk, & Brockett, 2004)
  - Only binary rating (is/is not a paraphrase)
- Echo Chamber (Brockett & Dolan, 2005)
- Paraphrase Game (Chklovski, 2005)
- Limitations in "cleaning" ITS input
  - Datasets artificially created (Fossati & Di Eugenio, 2008)
  - Target populations are relatively proficient
    - Why2-Atlas: College undergraduates
    - CIRCSIM-Tutor: 1st year medical students
  - Use lexicons; computationally expensive

User-Language Paraphrase Corpus

- 1998 target sentence/student response pairs
- Paraphrase attempts by high school students
  - During interactions with iSTART (McNamara, Levinstein, & Boonthum, 2004)
- Paraphrases evaluated on widely used computational indices
  - Latent semantic analysis (LSA; Landauer, McNamara, Dennis, & Kintsch, 2007)
  - Entailment (Rus et al., 2007)
  - Type-Token Ratio (TTR; Graesser, McNamara, et al., 2004)
  - Mean Edit Distance (MED; McCarthy et al., 2007)
- Paraphrases also evaluated by trained experts on 10 dimensions w/ Likert ratings

Research Questions

- How are established computational indices affected by the types of errors found in typed user-language?
- Do user errors affect NLP assessment and feedback produced by an established ITS?
- Does correcting user errors improve the capacity for ITS assessment to correspond to human ratings?
iSTART

- High school students (U.S. grades 9-12)
- Reading strategy training
  - Paraphrasing, Elaboration, Making Bridging Inferences, Comprehension Monitoring
- Paraphrase the following:
  - Over two thirds of the heat generated by a resting human is created by organs of the thoracic and abdominal cavities and the brain.
  - A lot of heat made by a lazy person is made by systems of your stomach and thinking box.

Paraphrase the following:

```
Over two thirds of the heat generated by a resting human is created by organs of the thoracic and abdominal cavities and the brain.
```

```
A lot of heat made by a lazy person is made by systems of your stomach and thinking box.
```

iSTART Evaluation Process

- Based on match between paraphrase and target sentence
- Respond to or remove Frozen expression
  - e.g., I think this is saying...
- Word & Soundex matching against benchmark for length, relevance, & similarity
  - Irrelevant (IRR) – too few words match
  - Too short (SH) – response is shorter than specified threshold
  - Too similar (SIM1) – length and word match is close to benchmark
- Word match & LSA cosines for quality
  - Adequate paraphrase (SIM2)
  - Better than a paraphrase (OK)

Detailed formulae – McNamara, Boonthum, et al. (2007)

Soundex

- Compensates for misspellings (Christian, 1998)
- Vowels removed
- Like-sounding consonants mapped onto same symbol
  - e.g., b, f, p, v
- Lexicon-free
- Word frequency problem
  - Students make more mistakes on new or uncommon words

Procedure

- Identified, coded, & corrected all errors
  - Based on validated models of grammar (e.g., Foster & Vogel, 2004)
- Intrarater agreement for subset (n = 200)
  - Kappa = .70, p < .001
  - Single rater coded entire corpus
- 83% of responses contained some form of error
- 52% had some form of spelling error
- 63% of spelling errors were internal to target sentence

Error types & frequencies

<table>
<thead>
<tr>
<th>Error Type</th>
<th>Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spelling (internal)</td>
<td>665</td>
<td>33%</td>
</tr>
<tr>
<td>Spelling (external)</td>
<td>386</td>
<td>19%</td>
</tr>
<tr>
<td>Capitalization</td>
<td>1157</td>
<td>58%</td>
</tr>
<tr>
<td>S-V Agreement</td>
<td>367</td>
<td>18%</td>
</tr>
<tr>
<td>Article agreement</td>
<td>75</td>
<td>4%</td>
</tr>
<tr>
<td>Preposition agreement</td>
<td>53</td>
<td>3%</td>
</tr>
<tr>
<td>Determiner agreement</td>
<td>59</td>
<td>3%</td>
</tr>
<tr>
<td>Spacing</td>
<td>174</td>
<td>9%</td>
</tr>
<tr>
<td>Punctuation</td>
<td>344</td>
<td>17%</td>
</tr>
<tr>
<td>Conjunction agreement</td>
<td>43</td>
<td>2%</td>
</tr>
<tr>
<td>Possessive agreement</td>
<td>71</td>
<td>4%</td>
</tr>
<tr>
<td>Extra/omitted/substitute</td>
<td>230</td>
<td>12%</td>
</tr>
</tbody>
</table>

Results

- Significant effect of error correction on computational similarity indices
  - Partial $\eta^2$ =
    - LSA  .178
    - Entailment  .268
    - TTR  .240
    - MED  .111

- Spelling internal accounts for large portion of variance
  - Adjusted $R^2$ =
    - LSA  .35
    - Entailment  .45
    - TTR  .46
    - MED  .17
Example

Target Sentence:
An increase in temperature of a substance is an indication that it has gained heat energy.

Student response:
Increase in temperature has gained heat energy.

Revised response:
Increase in temperature has gained heat energy.

Results

Table 1: Crosstabulation of iSTART responses to user paraphrases

<table>
<thead>
<tr>
<th>iSTART response – corrected</th>
<th>Too</th>
<th>Too</th>
<th>Similar</th>
<th>Short</th>
<th>Irrelevant</th>
<th>Frozen</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Better</td>
<td>691</td>
<td>45</td>
<td>37</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>777</td>
</tr>
<tr>
<td>Good</td>
<td>12</td>
<td>194</td>
<td>98</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>304</td>
</tr>
<tr>
<td>Too Similar</td>
<td>7</td>
<td>7</td>
<td>527</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>541</td>
</tr>
<tr>
<td>Too Short</td>
<td>11</td>
<td>0</td>
<td>1</td>
<td>205</td>
<td>2</td>
<td>1</td>
<td>221</td>
</tr>
<tr>
<td>Irrelevant</td>
<td>6</td>
<td>0</td>
<td>6</td>
<td>120</td>
<td>7</td>
<td>139</td>
<td>199</td>
</tr>
<tr>
<td>Frozen</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>16</td>
<td>32</td>
</tr>
<tr>
<td>Total</td>
<td>727</td>
<td>245</td>
<td>663</td>
<td>216</td>
<td>122</td>
<td>24</td>
<td>1998</td>
</tr>
</tbody>
</table>

Cramer’s V = .849, p < .001
Marginal Homogeneity (MH) = 5.892, p < .001

Results

Table 2: Separate pairwise comparisons of Paraphrase Quality

<table>
<thead>
<tr>
<th></th>
<th>Original</th>
<th>Corrected</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Diff.</td>
<td>Sep.</td>
<td>Mean Diff.</td>
<td>Sep.</td>
<td>Mean</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frozen</td>
<td>.152</td>
<td>.402</td>
<td>.081</td>
<td>.361</td>
<td>.032</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Too short</td>
<td>-.776</td>
<td>.370</td>
<td>.581</td>
<td>-.922</td>
<td>.299</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Good</td>
<td>-2.071</td>
<td>.366</td>
<td>.001</td>
<td>-2.421</td>
<td>.297</td>
<td>.001</td>
<td></td>
</tr>
<tr>
<td>Too Similar</td>
<td>-1.897</td>
<td>.361</td>
<td>.001</td>
<td>-2.106</td>
<td>.288</td>
<td>.001</td>
<td></td>
</tr>
<tr>
<td>Too Short</td>
<td>-2.118</td>
<td>.209</td>
<td>.001</td>
<td>-1.002</td>
<td>.245</td>
<td>.001</td>
<td></td>
</tr>
<tr>
<td>Irrelevant</td>
<td>-2.107</td>
<td>.196</td>
<td>.001</td>
<td>-2.257</td>
<td>.231</td>
<td>.001</td>
<td></td>
</tr>
<tr>
<td>Better</td>
<td>-2.223</td>
<td>.203</td>
<td>.001</td>
<td>-2.920</td>
<td>.242</td>
<td>.001</td>
<td></td>
</tr>
<tr>
<td>Good</td>
<td>-2.049</td>
<td>.192</td>
<td>.001</td>
<td>-2.187</td>
<td>.231</td>
<td>.001</td>
<td></td>
</tr>
<tr>
<td>Too Similar</td>
<td>-1.168</td>
<td>.115</td>
<td>.001</td>
<td>-1.250</td>
<td>.171</td>
<td>.001</td>
<td></td>
</tr>
<tr>
<td>Good</td>
<td>-1.305</td>
<td>.127</td>
<td>.001</td>
<td>-1.500</td>
<td>.133</td>
<td>.001</td>
<td></td>
</tr>
<tr>
<td>Too Short</td>
<td>-1.121</td>
<td>.111</td>
<td>.001</td>
<td>-1.186</td>
<td>.131</td>
<td>.001</td>
<td></td>
</tr>
<tr>
<td>Too similar</td>
<td>-1.116</td>
<td>.103</td>
<td>.001</td>
<td>-1.249</td>
<td>.187</td>
<td>.331</td>
<td></td>
</tr>
<tr>
<td>Good</td>
<td>-1.258</td>
<td>.083</td>
<td>.001</td>
<td>-1.070</td>
<td>.157</td>
<td>.1</td>
<td></td>
</tr>
<tr>
<td>Too Short</td>
<td>-1.111</td>
<td>.107</td>
<td>.001</td>
<td>-1.023</td>
<td>.198</td>
<td>.254</td>
<td></td>
</tr>
</tbody>
</table>

* Adjustment for multiple comparisons: Bonferroni.

Discussion

ITS feedback algorithms may be optimized if user-language can be filtered prior to processing
- Misclassification OK for motivation
- Accuracy not OK: simple rewording can pass for good paraphrase; paraphrase can pass for better
- Established NLP approaches not as robust to user-language as believed
- Response length not enough to wash out individual errors
- ULPC represents types & amount of errors real students make
- Most variance accounted for by internal misspellings
  - Provides direction for future research
  - Automatic spelling corrections only for words in the benchmark
  - Will be silent & computationally light

Thank you!

We would like to thank:
Vasile Rus
Ben Duncan
John Myers
Rebekah Guess

Research supported by:

IIS-0735682
IES-0850880

Thank you!