Inherent Biases in Reference-based Evaluation for Grammatical Error Correction and Text Simplification

Leshem Choshen & Omri Abend
School of Computer Science and Engineering & Department of Cognitive Sciences
The Hebrew University of Jerusalem Israel

Reference Based Measures

Number of Valid Corrections
Estimated with crowdsourcing and UnseenEst

<table>
<thead>
<tr>
<th>Frequency</th>
<th>0</th>
<th>0.001</th>
<th>0.01</th>
<th>0.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variants</td>
<td>1351.24</td>
<td>74.34</td>
<td>8.72</td>
<td>1.35</td>
</tr>
<tr>
<td>Mass</td>
<td>1</td>
<td>0.75</td>
<td>0.58</td>
<td>0.37</td>
</tr>
</tbody>
</table>

Perfect Correctors (Humans)

Accuracy, GLEU and $M^2$
Loss and evaluation metrics assign low scores to perfect correctors
Increasing references won’t solve it

GEC Performance

Under-Prediction (Conservatism)
SoTA systems correct an order of magnitude less than humans
In terms of: word changes, sentence splits/merges and word reordering

Systems on par with Humans

RBMs Favor Some (Valid) Corrections
And SoTA favors similar ones
Encourage close-class errors
Discourage open-class errors
Disincentivized to correct-
Even if you know the answer
Precision oriented measures make it worse

What can we do?
Reference-less measures
Beyond n-gram overlap of source/reference
(Semantics)
USim [Choshen & Abend 2018, a]

More in the paper
Significance, methodological contributions, Empirical number of corrections per error type [Choshen & Abend 2018, b]

References
a. Choshen & Abend (NAACL 2018)
Reference-less Measure of Faithfulness for Grammatical Error Correction.
b. Choshen & Abend (ACL 2018)
Automatic Metric Validation for Grammatical Error Correction