Human error: reference out of scope

Adam Craig\textsuperscript{1,2}, Anousha Āthreya\textsuperscript{1},
and Carl Taswell\textsuperscript{1,3}

\textsuperscript{1}Brain Health Alliance, Ladera Ranch, CA, USA, www.BrainHealthAlliance.org
\textsuperscript{2}Hong Kong Baptist University Center for Nonlinear Studies, Kowloon Tong, Hong Kong
\textsuperscript{3}UC Sand Diego School of Engineering, San Diego, CA, USA

ERROR workshop at eScience 2023-10-10
Authors may avoid citing the work of potential rivals. They may also misrepresent the content of prior work. Peer reviewers and editors may have their own biases or perverse incentives. Institutional ethics committees may care more about avoiding damage to the institute’s reputation than about righting wrongs. See (Taswell et al., 2020, ASIS&T 2020) for a review of these issues. We need an alternative to subjective judgments: Quantify it. In (Craig & Taswell, 2018, ASIS&T-SIGMET 2018), we proposed FAIR Attribution to Indexed Reports (FAIR) Metrics of adherence to good citation practices.
Methods: FAIR Metric counts and ratios

- We calculate the 4 ratios from 4 counts, first defined in (Craig & Taswell, 2018, *BIBM 2018*).
- \( Q \): statements correctly attributed to prior work
- \( M \): statements misrepresenting the content of prior work
- \( P \): statements taken without attribution (potentially plagiarized) from prior work
- \( N \): statements presented as novel and not found in prior work
- In (Craig et al., 2019, *ASIS&T 2019*), we introduced 4 ratio FAIR Metrics, each with a different emphasis.

\[
F_Q = \frac{Q}{Q+P+M}.
\]
\[
F_M = \frac{Q-M}{Q+P+M}.
\]
\[
F_P = \frac{Q-P}{Q+P+M}.
\]
\[
F_N = \frac{Q-N}{Q+P+M+N}.
\]

- Also briefly summarizes the pilot validation study described here.
Methods: ideal automated FAIR Metric calculation

- Start with a test document $T$ and a collection of all prior work $C = C_1, C_2, ..., C_N$.
- Represent all claims made in $T$ and in every member of $C$ as RDF triples.
- Set $Q = M = P = N = 0$, and iterate over all claims in $T$.
- If a claim in $T$ cites a document $C_i$, search $C_i$ for an equivalent claim.
  - If found, increment $Q$. Otherwise, increment $M$.
- If a claim in $T$ does not cite a source, search all documents in $C$ for an equivalent claim.
  - If found, increment $P$. Otherwise, increment $N$.
- Worst-case time complexity is $O(|T| \sum_{i=1}^{\mid C \mid} |C_i|)$ where $|C_i| = \# \text{ of statements in } C_i$, $|T| = \# \text{ of statements in } T$, $|C| = \# \text{ number of documents in } C$, and statement comparison is unit-time.
Methods: limited-scope manual FAIR Metric calculation

- Start with a test document $T$, a small (singleton) set of prior works claimed to be similar $C$.
- List all the claims in $T$ as natural language sentences.
- Set $Q = M = P = N = 0$, and iterate over all claims in $T$.
- If a claim in $T$ cites a document other than $C$, discard it.
- If a claim in $T$ cites $C$, search $C$ for an equivalent claim.
  - If found, increment $Q$. Otherwise, increment $M$.
- If a claim in $T$ does not cite a source, search $C$ for an equivalent claim.
  - If found, increment $P$. Otherwise, increment $N$.
- Worst-case time complexity is $O(|T| \sum_{i=1}^{C} |C_i|)$ where $|C_i| = \#$ of statements in $C_i$, $|T| = \#$ of statements in $T$, $|C| = \#$ of documents in $C$, and statement comparison is unit-time.
- 8 evaluators work independently.
Methods: example case set

- Search Retraction Watch for computer science- or neuroscience-related papers retracted for plagiarism to use as a $T$.
- For each, look up the plagiarized paper to use as $C$.
- Search Google Scholar for a paper on a related topic to use as a second $T$ to compare to $C$.
- 32 found for CS.
- 18 found for Neuro.
Results: divide by 0 error

- It turns out two arbitrarily selected papers in the same field will not necessarily cite each other.
- Ended up with not only $P = 0$ but $M = 0$ and $Q = 0$ for almost all non-plagiarizing test papers.
- This made $F_M$, $F_Q$, and $F_P$ undefined, since their denominators are $Q + P + M$.
- Even the plagiarizing test cases often ended up with $M = Q = 0$, regardless of what sources they actually cited.
- This attempt at calculating FAIR metrics was not very fair.
Methods: limited-scope manual FAIR Metric calculation 2.0

- Start with a test document $T$, a small set of prior works claimed to be similar $C$, and the set of works referenced by $T$, $R$.
- List all the claims in $T$ as natural language sentences.
- Set $Q = M = P = N = 0$, and iterate over all claims in $T$.
- If a claim in $T$ cites a document $R_i$, search $R_i$ for an equivalent claim.
- If found, increment $Q$. Otherwise, increment $M$.
- If a claim in $T$ does not cite a source, search $C$ for an equivalent claim.
- If found, increment $P$. Otherwise, increment $N$.
- Worst-case time complexity is $O(|T| \max(\sum_{i=1}^{\mid C \mid} |C_i|, \max(|R_i|)))$ where $|C_i| = \#$ of statements in $C_i$, $|R_i| = \#$ of statements in $R_i$, $|T| = \#$ of statements in $T$, $|C| = \#$ of documents in $C$, and statement comparison is unit-time.
Results: Seems to work this time

<table>
<thead>
<tr>
<th>Target text</th>
<th>Retracted?</th>
<th>Comparison text</th>
<th>M</th>
<th>N</th>
<th>P</th>
<th>Q</th>
<th>$F_M$</th>
<th>$F_N$</th>
<th>$F_P$</th>
<th>$F_Q$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taswell 2007</td>
<td>no</td>
<td>Mons 2005</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td>22</td>
<td>1.00</td>
<td>0.05</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Uddin 2022</td>
<td>yes</td>
<td>Foster et al. 2019</td>
<td>0</td>
<td>18</td>
<td>18</td>
<td>87</td>
<td>0.83</td>
<td>0.56</td>
<td>0.66</td>
<td>0.83</td>
</tr>
<tr>
<td>Gnat et al. 2022</td>
<td>yes</td>
<td>de Hoog et al. 2017</td>
<td>0</td>
<td>3</td>
<td>10</td>
<td>30</td>
<td>0.75</td>
<td>0.63</td>
<td>0.50</td>
<td>0.75</td>
</tr>
<tr>
<td>Ullah et al. 2018</td>
<td>yes</td>
<td>Sansaniwal &amp; Kumar 2015</td>
<td>31</td>
<td>3</td>
<td>7</td>
<td>2</td>
<td>-0.73</td>
<td>-0.02</td>
<td>-0.13</td>
<td>0.05</td>
</tr>
<tr>
<td>Wilkinson et al. 2016</td>
<td>no</td>
<td>Taswell 2007</td>
<td>6</td>
<td>5</td>
<td>24</td>
<td>28</td>
<td>0.38</td>
<td>0.37</td>
<td>0.07</td>
<td>0.48</td>
</tr>
</tbody>
</table>

- Target: the text for which we are calculating FAIR Metrics.
- Retracted?: Was Target retracted for plagiarism of Comparison?
- Comparison: We are checking the Target for plagiarism of this text.
- Counts: Misquoted; Novel; Potentially Plagiarized; Quoted;
- $F_M = \frac{Q-M}{Q+P+M};$ $F_N = \frac{Q-N}{Q+P+M+N};$ $F_P = \frac{Q-P}{Q+P+M};$ $F_Q = \frac{Q}{Q+P+M}$
Discussion: Limitations of this method

- Can only detect plagiarism where it is already suspected
- Relies on potentially biased judgments of equivalence
- e.g., How much can you summarize and still convey the same idea?
- Claims with a correctly cited source are still Quoted even if copied verbatim from prior work.
- Novel claims in $T$ about “asparagus” are still Novel even if they are otherwise identical to claims in $C$ about ginger.
- 1 sentence $=$ 1 claim? If not, division gets arbitrary.
- If authors reiterate their points, how do we select only unique statements?
- Does the evaluation method unfairly favor a particular style of writing?
- $F_N$ score favors review articles
Conclusion

- Target manual evaluation of FAIR Metrics allows systematic comparison of two papers.
- Is more labor-intensive than traditional peer review.
- Results in a well-organized document that can serve as substrate for peer review of the peer review.
- These semantically formatted manual evaluation records using the PDP-DREAM Ontology will provide an annotated data set against which to validate future AI/automated approaches.
Required references


Athreya, A., Taswell, S. K., Mashkoor, S., & Taswell, C. (2020, September). Essential question: ‘equal or equivalent entities?’ about two things as same, similar, or different. In 2020 Second International Conference on Transdisciplinary AI (TransAI) (pp. 123-124). IEEE.


Contact Info

- ctaswell@bhavi.us
- www.BHAVI.us
- www.BrainHealthAlliance.org