A User Study on Snippet Generation: Text Reuse vs. Paraphrases

Motivation

Ancillary Copyright
- Snippets reuse text from publishers’ web pages
- Search engines profit from reuse snippets
- Publishers demand compensation

Paraphrase Snippets

The Vanishing Nile: A Great River Faces a Multitude of Threats …

The Nile River is under assault on two fronts - a massive dam under construction upstream in Ethiopia and rising sea levels leading to saltwater intrusion downstream.

The Vanishing Nile: A Great River Faces a Multitude of Threats …

There are two major issues facing the health of the Nile River. Upstream there is a dam being constructed in Ethiopia. Downstream there are rising sea levels causing saltwater intrusion.

Experiment

Crowdsourcing Paraphrase Snippets
- 150 queries from the TREC Web tracks 2009–2011
- Top-5 search results (reuse snippets) of each query by Google
- Paraphrase each of the 750 reuse snippets by two different workers on Amazon’s Mechanical Turk

Snippet Preference
- 5 workers for each pair of reuse / paraphrase snippets
- Worker recruitment: > 80% acceptance rate and at least 100 successful assignments
- Each worker judged at most two pairs
- Rejected if workers spent insufficient time, too much time
- Rejected if they failed to provide sensible explanations for their judgments
- Resulting in 4,235 individual workers and 7,500 accepted annotations

Snippet Usefulness
- Collect 3 relevant and 3 irrelevant web pages of the queries
- Topics: 29 queries from ClueWeb12
- Workers judged their relevant based on
  1. Reuse snippet
  2. Paraphrase snippet
  3. Title and URL
  4. Reuse snippet only
  5. captcha-style snippet

Result

Distribution of judgments

<table>
<thead>
<tr>
<th>Assessment</th>
<th>Judgments</th>
<th>absolute</th>
<th>relative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reuse better</td>
<td>2,731</td>
<td>36.41%</td>
<td></td>
</tr>
<tr>
<td>Paraphrase better</td>
<td>2,652</td>
<td>35.36%</td>
<td></td>
</tr>
<tr>
<td>Both good</td>
<td>1,537</td>
<td>20.49%</td>
<td></td>
</tr>
<tr>
<td>Both bad</td>
<td>580</td>
<td>7.74%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>7,500</td>
<td>100.00%</td>
<td></td>
</tr>
</tbody>
</table>

Average scores of reuse and paraphrase

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Reuse</th>
<th>Paraphrase</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>3.06</td>
<td>2.97</td>
<td>0.51</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>3.31</td>
<td>2.58</td>
<td>0.00*</td>
</tr>
<tr>
<td>Non-Wikipedia</td>
<td>2.75</td>
<td>2.85</td>
<td>0.31</td>
</tr>
<tr>
<td>all</td>
<td>3.05</td>
<td>2.94</td>
<td>0.43</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>3.18</td>
<td>2.64</td>
<td>0.01*</td>
</tr>
<tr>
<td>Non-Wikipedia</td>
<td>2.77</td>
<td>2.82</td>
<td>0.58</td>
</tr>
</tbody>
</table>

1260 out of 750 pages are Wikipedia
* significant (p < 0.05)

F-scores of the snippet usefulness experiment

<table>
<thead>
<tr>
<th></th>
<th>Reuse</th>
<th>Paraphrase</th>
<th>No snippet</th>
<th>Snippet only</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-score</td>
<td>67.64</td>
<td>64.61</td>
<td>63.65</td>
<td>60.16</td>
<td>50.00</td>
</tr>
</tbody>
</table>

Discussion

- On average, the reuse snippets have 1.9 sentences and 41.1 words; the paraphrase snippets have 2.2 sentences and 40.5 words.
- No statistically significant difference between reuse and paraphrase snippets
- Users significantly prefer reuse snippets over paraphrases on Wikipedia results, which is not the case for non-Wikipedia results
- Wikipedia snippets have higher writing quality, and it may have been difficult for the average AMT worker to compete with that
- Reuse snippets are significantly better than showing only snippets
- The combination of snippet (reused or paraphrased), title, and URL is crucial to identify relevant web pages

Future Work

Develop an automatic snippet paraphrase model