Wei-Fan Chen[1] Matthias Hagen[2] Benno Stein[1] Martin Potthast[3]
[1]Bauhaus-Universität Weimar [2]Martin-Luther-Universität Halle-Wittenberg [3]Leipzig University

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 ...
http://e360.yale.edu/features/vanishing-nile-a-great-river-faces-a-multitude-of-threats-egypt-dam
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 ...
http://e360.yale.edu/features/vanishing-nile-a-great-river-faces-a-multitude-of-threats-egypt-dam
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

Assessment	Judgments	
	absolute	relative
Reuse better	2,731	36.41%
Paraphrase better	2,652	35.36%
Both good	1,537	20.49%
Both bad	580	7.74%
Total	7,500	100.00%

Average scores of reuse and paraphrase

Experiment	Reuse	Paraphrase	p-value
all	3.06	2.97	0.51
Wikipedia[1]	3.31	2.58	0.00*
Non-Wikipedia	2.75	2.85	0.31
all	3.05	2.94	0.43
Wikipedia[1]	3.18	2.64	0.01*
Non-Wikipedia	2.77	2.82	0.58

[1]260 out of 750 pages are Wikipedia
* significant ($p < 0.05$)

F-scores of the snippet usefulness experiment

	Reuse	Paraphrase	No snippet	Snippet only	Random
F-score	67.64	64.61	63.65	60.16	50.00

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