Improving Synoptic Querying for Source Retrieval

Šimon Suchomel
Process Overview
Building of Queries

**Keywords-based**
- Pilot query
  - 6 best KW, ChatNoir, Indri
- Collocational Phrasal
  - 3 terms long collocations, Derived from the Pilot, Indri
- Collocational
  - Derived from the Pilot, 2 terms long collocations combined into 6 terms long queries, Chatnoir
- Other Keywords-based
  - Remaining KW, 6 terms long q., Chatnoir

**Paragraph based**
- Paragraph chunking
- One query from each paragraph
- Paragraph position [start, end], inside the document
- 10 terms with highest TF-IDF score from the whole paragraph
- Chatnoir
Queries Scheduling
Method Assessment During Test Phase

- 98 documents
- 32.9 queries per document on average
  - 18.8% directed to Indri, 81.2% to ChatNoir
- Max 100 URLs per one query
- 134,247 unique URLs retrieved in total
- 32,538 URLs downloaded
- 6,392 URLs were relevant
- Master hit as retrieval of an annotated URL
  - 0.45 recall, 5 documents with recall 1, and 12 documents with recall 0
## Query Type Scope

<table>
<thead>
<tr>
<th>Query type</th>
<th>#Queries</th>
<th>#URLs retrieved</th>
<th>Scope Usage</th>
<th>Top Retrieval</th>
<th>Zero Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pilot</td>
<td>183</td>
<td>16341</td>
<td>89.3%</td>
<td>83.6%</td>
<td>1.1%</td>
</tr>
<tr>
<td>Collocational Phrasal</td>
<td>520</td>
<td>34095</td>
<td>65.6%</td>
<td>56.7%</td>
<td>12.3%</td>
</tr>
<tr>
<td>Collocational</td>
<td>311</td>
<td>23188</td>
<td>74.6%</td>
<td>64.3%</td>
<td>2.9%</td>
</tr>
<tr>
<td>Other Keywords-based</td>
<td>101</td>
<td>5367</td>
<td>53.1%</td>
<td>38.6%</td>
<td>8.9%</td>
</tr>
<tr>
<td>Paragraph-based</td>
<td>2109</td>
<td>81788</td>
<td>38.8%</td>
<td>26.8%</td>
<td>18.5%</td>
</tr>
</tbody>
</table>
## Query Type Performance

<table>
<thead>
<tr>
<th>Query type</th>
<th>#Queries</th>
<th>#Relevant URLs</th>
<th>Theoretical Portion</th>
<th>Hits per Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pilot</td>
<td>183</td>
<td>2815</td>
<td>44.0%</td>
<td>15.4</td>
</tr>
<tr>
<td>Collocational Phrasal</td>
<td>520</td>
<td>2974</td>
<td>46.5%</td>
<td>5.7</td>
</tr>
<tr>
<td>Collocational</td>
<td>311</td>
<td>1730</td>
<td>27.1%</td>
<td>5.6</td>
</tr>
<tr>
<td>Other Keywords-based</td>
<td>101</td>
<td>401</td>
<td>6.3%</td>
<td>4.0</td>
</tr>
<tr>
<td>Paragraph-based</td>
<td>2109</td>
<td>2713</td>
<td>42.4%</td>
<td>1.3</td>
</tr>
</tbody>
</table>
Success Rate per SERP Rank

Number of hits

SERP ranking
Source Retrieval Progress Based on 2 Selected Documents
Conclusions

• Usable methodology for source retrieval
• The pilot queries proved to be the best choice for synoptic search
• Paragraph-based queries perform well in position retrieval, but not well enough
• Achieved the highest recall among this year’s softwares