The Effect of Expanding Relevance Judgements with Duplicates

Gaurav Baruah, Adam Roegiest, Mark D. Smucker
University of Waterloo

**BACKGROUND:** The TREC 2013 Temporal Summarization Track (TST) had systems return relevant sentences between two points in time from a time-ordered document stream.

**OBSERVATION:** There are many duplicates of judged and unjudged sentences in the query time periods.

<table>
<thead>
<tr>
<th>EVALUATION POOL</th>
<th>EXACT DUPLICATES</th>
</tr>
</thead>
<tbody>
<tr>
<td>#sentences</td>
<td>#relevant</td>
</tr>
<tr>
<td>9,113</td>
<td>2,635</td>
</tr>
</tbody>
</table>

**QUESTION:** Does the inclusion of duplicate sentences affect the evaluation of the TST?

**EXPERIMENT:** Expand the original judged set with duplicates and compare system performance over the TST metrics for the 28 runs submitted to the track.

**RESULTS:**

![Graphs showing Expected Latency Gain and Latency Comprehensiveness](image)

- **Expected Latency Gain:**
  - Rank Correlation Kendall’s $\tau = 0.899$.
  - Statistically significant score change for 13 runs.

- **Latency Comprehensiveness:**
  - Rank Correlation Kendall’s $\tau = 0.942$.
  - Statistically significant score change for 12 runs.

**CONCLUSION:** The inclusion of duplicate sentences yields a high correlation with the original rank order, but it changes the scores of many systems. We recommend adding duplicate sentences to the judged set, which could help enhance test collection reusability.

**ACKNOWLEDGEMENT:** Support provided in part by a SIGIR 2014 Student Travel Grant.