Documents/Summaries as background models for update summary extraction

Maheedhar Kolla, Olga Vechtomova, Charles L.A. Clarke
IR Group
University of Waterloo

April 26, 2007
Users going through search results often get deceived.

Highlighting only shows that the document has the query term(s).

Would it help if we could show them snippets that would highlight the content novel to the docs on previous results page(s)?
Proposed model:

- Overview: Find the key terms and extract the sentences.
- Which terms are key and which are redundant?
- Words’ score in current document is discounted by the weight of the term in the background model.

\[
P(w|d_{new}) = (1 - \alpha)(P(w|d)) - \alpha(P(w|\theta_{old})) \tag{1}
\]

where, \(P(w|\theta_{old})\) - probability of word in background model and \((P(w|d))\) is the Maximum Likelihood estimate of the word in current document.
Maximum Likelihood Estimate

- Probability of a term based using raw term counts:
  
  \[ P_{mle}(w|d) = \frac{tf(w, d)}{\sum_{w_j} C(w_j; d)} \]  

  where \( \sum_{w_j} C(w_j; d) \) is equivalent to the length of the document.

- Drawback: less frequent words are assigned lesser weights.

- Better approximate would be to smooth with collection probabilities, using Cluster Based Document Model:
  
  \[ P(w|d) = \lambda(P_{mle}(w|d) + (1-\lambda)[\beta P(w|cl) + (1-\beta)(P(w|coll))]) \]  

  where

  \[ P(w|cl) = \lambda(\frac{tf(w, cl)}{\sum_{w_j \in cl} tf(w_j; cl)}) + (1 - \lambda)\frac{tf(w, coll)}{\sum_{w_j \in V} tf(w_j, coll)} \]
Not every recurring term redundant.
Care should be taken to avoid topic drift.
Which of the following make a better background model?
- Models constructed from all previous documents.
- Models constructed from only summaries of previous documents.
Document based model: $\theta_d$

- Compute the term probability in each of the previously seen document(s).
- Term’s weight in the background model is then the average of those probabilities.
- Some other methods may be to use the term’s max probability value in the previous document set(s).
- Could use divergence based measures too.
Generic single document summaries are used for model construction.

We used simple lexical term overlap between the sentences to rank the sentences.

Two sentences should have at least 3 word overlap to be considered bonded.

Top two sentences were selected from each document.

Weight of the term in this model is probability computed using the raw counts of the term to the total length of the summaries.
Estimation of the parameters $\lambda$ and $\beta$

- First, we determine the optimal values for $\lambda$ and $\beta$
- Used the initial set of documents— that lacks any previous documents.
- Fixed the values of $\lambda$ and $\beta$ that maximize the ROUGE-2 and ROUGE-SU4 scores.
- The values were found to be 0.4 and 0.6 respectively.
- Values were fixed for experiments to study the effect of $\alpha$ on two background models.
Performance of the background models, at various $\alpha$ values

**Table:** ROUGE-SU4 measures for various values of $\alpha$ on the two different background models [ $\lambda$ and $\beta$ constant in this case]

<table>
<thead>
<tr>
<th>Model</th>
<th>$\alpha = 0.3$</th>
<th>$\alpha = 0.4$</th>
<th>$\alpha = 0.5$</th>
<th>$\alpha = 0.7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_{doc}$</td>
<td>0.10830</td>
<td>0.10514</td>
<td>0.09809</td>
<td>0.08279</td>
</tr>
<tr>
<td>$\theta_{summ}$</td>
<td>0.10630</td>
<td>0.10192</td>
<td>0.09319</td>
<td>0.08449</td>
</tr>
</tbody>
</table>

**Table:** ROUGE-2 measures for various values of $\alpha$ on the two different background models [ $\lambda$ and $\beta$ constant in this case]

<table>
<thead>
<tr>
<th>Model</th>
<th>$\alpha = 0.3$</th>
<th>$\alpha = 0.4$</th>
<th>$\alpha = 0.5$</th>
<th>$\alpha = 0.7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_{doc}$</td>
<td>0.07073</td>
<td>0.06938</td>
<td>0.06089</td>
<td>0.04395</td>
</tr>
<tr>
<td>$\theta_{summ}$</td>
<td>0.06752</td>
<td>0.06433</td>
<td>0.05652</td>
<td>0.04601</td>
</tr>
</tbody>
</table>
Discussion:

- For a given set of $\alpha \beta$ and $\lambda$, both models perform almost the same.
- Both models have similar effect on the variation of the parameter $\alpha$.
- Document clusters were hand-picked - very focused on the topic.
- The values of the $\lambda$ and $\beta$ estimated from the first set were stable across the other summary extraction experiments.
Conclusion and Future Work:

- Proposed method to update the word probability using previously seen docs.
- Further experiments to use standard query-likelihood techniques to rank the documents.
- Induce noise documents into all sets - close approximate to the real-world problem.
- Extensive experiments using several smoothing functions.
Thank you