Spoken document retrieval using neighboring documents and extended language models for query likelihood model

Kazuaki Ogawa, Tatsuaki Murahashi, Hiroaki Taguchi, Kodai Nakajima, Masanori Takehara, Satoshi Tamura, Satoru Hayamizu

1. Overview

- Probabilistic models are employed in document retrieval.
- We aim at retrieving spoken documents corresponding to presentation slides.
- Our methods are based on the query likelihood model and smoothing.

We propose three smoothing methods.

(A.) Using a language model obtained from research papers.
(B.) Using a cache model and an N-gram model.
(C.) Using neighboring documents.

2. Proposed smoothing methods

(A.) Research papers
- We proposed extended Dirichlet smoothing using a dynamic document collection obtained from web pages.
- However, web documents have a lot of noises.

- Instead of web documents, we use a research paper corpus $R$ in ASJ.

$$P(w_i|\theta_R; \mu, \nu) = \frac{|d_i|}{|d_\mu + \mu + |d| + \nu} P(w_i|\theta_d) + \frac{\mu}{|d_\mu + \mu + |d| + \nu} P(w_i|\theta_c) + \frac{\nu}{|d_\mu + \mu + |d| + \nu} P(w_i|\theta_N)$$

$\theta_d$: A unigram model using a target document $d$.
$\theta_c$: A unigram model using a target document collection.
$\theta_N$: A unigram model using research papers.

- We select research papers based on cosine similarities of TF-IDF vectors.
- Employing a corpus related to target documents is effective and helpful.

(B.) Cache model and N-gram model
- We use linear interpolation of a cache model and an N-gram model in the Dirichlet smoothing.

$$P(w_i|\theta_C; \mu, \nu) = \frac{|d_i|}{|d_\mu + \mu + |d| + \nu} P(w_i|\theta_d) + \frac{\mu}{|d_\mu + \mu + |d| + \nu} P(w_i|\theta_c) + \frac{\nu}{|d_\mu + \mu + |d| + \nu} P(w_i|\theta_N)$$

$$P(w_i|M, d) = \frac{1}{|M| + |d|} \left( \sum_{w_j \in M} \delta(w_i, w_j) + \sum_{w_j \in d} \delta(w_i, w_j) \right)$$

$P_C(w_i|M, d)$: A probability based on a cache model.
$P_N(w_i|M, d)$: A probability based on an N-gram model.
$M$: A set of target documents.
$\delta(w_i, w_j)$: A delta function.

We train the N-gram model using $d$ and $M$.
We expect to add contextual information.

(C.) Neighboring documents
- The length of target documents is often short.
- We use neighboring documents of a target document.

$$S(i) = \sum_{n=1}^{l} w_n S(i + n) \quad w_n = \frac{1}{|d| + 1}$$

$S(i)$: A similarity score of a partial document corresponding to $i$-th slide and a query.
$w_i$: A weighting coefficient of the inverse proportion.

3. Experimental condition (Formal-run)

<table>
<thead>
<tr>
<th>Task</th>
<th>SpokenQuery&amp;Doc-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub task</td>
<td>SQ-SCR SGS retrieval</td>
</tr>
<tr>
<td>Query</td>
<td>K-REF-WORD-MATCH</td>
</tr>
<tr>
<td>Target</td>
<td>K-REF-WORD-MATCH</td>
</tr>
</tbody>
</table>

The number of queries:
- Formal-run: 2807

The number of target documents:
- Research papers in ASJ (published 2005 to 2014)

4. Experiments

(I.) MAP results
- We tested the following six retrieval methods using NTCIR-12 SpokenQuery&Doc-2 Formal-run data.

<table>
<thead>
<tr>
<th>(A.) Research papers</th>
<th>(B.) Cache model and N-gram model</th>
<th>(C.) Neighboring documents</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1</td>
<td></td>
<td></td>
<td>0.197</td>
</tr>
<tr>
<td>Method 2</td>
<td>○</td>
<td></td>
<td>0.193</td>
</tr>
<tr>
<td>Method 3</td>
<td></td>
<td>○</td>
<td>0.215</td>
</tr>
<tr>
<td>Method 4</td>
<td>○</td>
<td></td>
<td>0.242</td>
</tr>
<tr>
<td>Method 5</td>
<td></td>
<td>○</td>
<td>0.239</td>
</tr>
<tr>
<td>Method 6</td>
<td>○</td>
<td>○</td>
<td>0.252</td>
</tr>
</tbody>
</table>

(II.) Discussion
- Comparing Methods 1 - 3 to 4 - 6: Methods using neighboring documents are better than the other methods.
- Comparing Methods 1, 4 to 2, 5: The MAP score using research papers slightly decreased.
- We need to reconsider the way to select research papers.
- We obtained the best result in Method 6.
- Employing contextual information has effectiveness for SDR, even if (B.) and (C.) have similar effects.
- We need to improve integration schemes in our approach.

5. Future works

- We should properly deal with unknown words in queries.
- We try to reconsider weighting methods when using neighboring documents.
- It is necessary to investigate smoothing parameters.
- We should investigate more effective integration schemes for our proposed smoothing methods.