Segmented spoken document retrieval using word co-occurrence information

Kensuke Hara, Hiroaki Taguchi, Koudai Nakajima, Masanori Takehara, Satoshi Tamura, Satoru Hayamizu

Introduction

- Spoken document retrieval (SDR) is attracting attention in searching news shows and movies.
- In SDR, mis-recognized words have bad effects.
- TF-IDF values of mis-recognized words sometimes become inappropriate.
- The cosine similarity is widely used for comparison. But the cosine similarity treats words having the same meaning (e.g. ASR and speech recognition) as different ones.
- To overcome these issues, pointwise mutual information (PMI) is employed.
  - PMI represents a relationship between two words.
  - To reject mis-recognized words, PMI is used to compute a contextual coherency of a word.
  - For query-document comparison, PMI is used to consider the similarity of different words.

Flows of our proposed methods

A. Building dynamic document collection

1. Word rejection

Automatic transcription

Static document collection

Web query expansion

Dynamic document collection

TF-IDF

TF-IDF

Top 5 terms

B. SDR using query model and LDA (Hasegawa, 2012)

Target document $d$

Static document collection $c$

Dynamic document collection $W$

Query $q$:

$P(w_i | \theta_d) = \frac{1}{\sum_{j=1}^{V} P(w_j | \theta_d)} \sum_{j=1}^{V} P(w_j | \theta_d) P(w_i | w_j)$

$P(q | \theta_d, \mu, \nu) = \sum_{c \in C} \int_{\mu} d \nu \int_{\nu} d \mu P(w_i | \theta_d) P(w_i | \theta_c) P(w_i | \theta_w) + \int_{\mu} d \nu \int_{\mu} d \nu \int_{\nu} d \mu P(w_i | \theta_w) + \int_{\nu} d \mu \int_{\nu} d \nu \int_{\mu} d \mu P(w_i | \theta_w)$

2. Weighting web pages

$P(w_i | \theta_w) = \frac{1}{\sum_{j=1}^{V} P(w_j | \theta_w)} \sum_{j=1}^{V} P(w_j | \theta_w) P(w_i | w_j)$

C. SDR using vector space model

Target document

Static document collection

Dynamic document collection

TF-IDF $v_d$

Expanded TF-IDF

$v_e = \beta v_d + (1 - \beta) v_w$

TF-IDF $v_w$

Query-document comparison with the cosine similarity or

PMI similarity.

Details of four techniques

1. Mis-recognized word rejection
   - Mis-recognized words have less relationship to a document.
   - The relationship between a word $w$ and a document $d$ is computed by $\text{sumPMI}(w)$ used for rejection with a threshold.

   $\text{sumPMI}(w) = \sum_{w \in d} \text{PMI}(w, w_i)$

2. Web page weighting using LDA
   - A web-page weight is determined by the cosine distance between a query topic vector and a web page topic vector.
   - A topic mixture ratio vector is computed using LDA.

3. Document comparison using PMI
   - A word similarity $R(w_1, w_2)$ is computed using PMI.

   $R(w_1, w_2) = \begin{cases} 1 & (w_1 = w_2) \\ \{ \text{PMI}(w_1, w_2) \} & (w_1 \neq w_2) \end{cases}$
   - It can consider a similarity of different words.

4. Segmented document retrieval method
   - Linear combination of a similarity for a segment $\text{Sim}_c$ and a similarity for a whole document $\text{Sim}_d$.

   $\text{Sim} = \alpha \cdot \text{Sim}_c + (1 - \alpha) \cdot \text{Sim}_d$

Experimental condition (Formal-run)

<table>
<thead>
<tr>
<th>Subtask</th>
<th>Slide-Group-Segment retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatic transcription(Query)</td>
<td>REF-WORD-MATCH</td>
</tr>
<tr>
<td>Automatic transcription(Target)</td>
<td>REF-WORD-MATCH</td>
</tr>
<tr>
<td>LDA training data</td>
<td>Mainichi newspaper corpus 2007-2008</td>
</tr>
<tr>
<td>Static document collection</td>
<td>Manual transcription</td>
</tr>
<tr>
<td>Linear combination parameters</td>
<td>$\alpha = 0.6, \beta = 0.9$ (tuned in the dry-run evaluation)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Results and discussion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1</td>
</tr>
<tr>
<td>Method 2</td>
</tr>
<tr>
<td>Method 3</td>
</tr>
<tr>
<td>Method 4</td>
</tr>
<tr>
<td>Method 5</td>
</tr>
</tbody>
</table>

1. Rejecting mis-recognized words
2. Document comparison using PMI
3. PMI scores between general terms tend to be high.
4. Considering whole contents is important for segmented SDR.
5. SDR using query model and LDA is useful
6. The query model is useful for not only short text queries but also long spoken queries.