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

GIFU UNIVERSITY

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.

$$P(w_i|\theta_d;\mu,\nu)=\frac{|d|}{|d|+\mu+\nu}P(w_i|\theta_d)+\frac{\mu}{|d|+\mu+\nu}P(w_i|\theta_C)+\frac{\nu}{|d|+\mu+\nu}P(w_i|\theta_R)$$

- θ_d: A unigram model using a target document d.
- θ_C: A unigram model using a target document collection.
- θ_R: 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.

$$P(w_i|\theta_d;\mu,\nu)=\frac{|d|}{|d|+\mu+\nu}P(w_i|\theta_d)+\frac{\mu}{|d|+\mu+\nu}P(w_i|\theta_C)+\frac{\nu}{|d|+\mu+\nu}P(w_i|\theta_{KC})$$

$$P(w_i|\theta_{KC})=\left(\gamma P_{KN}(w_i|\theta_{KN})+(1-\gamma)P_{CH}(w_i|M,d)\right)$$

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

- $P_{CH}(w_i|M,d)$: A probability based on a cache model.
- $P_{KN}(w_i|\theta_{KN})$: A probability based on an N-gram model.
- M: $|M|$ terms appeared just before a target document d.
- θ_{KN}: An N-gram model using Kneser-Ney smoothing for a target document d.
- 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.

$$S'(i)=\sum_{n=-L}^{L}w_n S(i+n) \qquad w_n=\frac{1}{|n|+1}$$

- $S(i)$: 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)

Task	SpokenQuery&Doc-2
Sub task	SQ-SCR SGS retrieval
Query	K-REF-WORD-MATCH
Target	K-REF-WORD-MATCH
The number of queries	Formal-run: 80
The number of target documents	2807
Static document collection	Target documents
Dynamic document collection	Research papers in ASJ (published 2005 to 2014)

μ and ν in the Dirichlet smoothing	Method 1 and 4: $\mu=320$ Method 2 and 5: $\mu=320,\ \nu=80$ Method 3 and 6: $\mu=320,\ \nu=10$
$\lvert M\rvert$ in the cache model	$\lvert M\rvert=100$
N in the N-gram model	$N=5$
γ for linear interpolation	$\gamma=0.25$

4. Experiments

(Ⅰ.) MAP results

- We tested the following six retrieval methods using NTCIR-12 SpokenQuery&Doc-2 Formal-run data.

	(A.) Research papers	(B.) Cache model and N-gram model	(C.) Neighboring documents	MAP
Method 1				0.197
Method 2	○			0.193
Method 3		○		0.215
Method 4			○	0.242
Method 5	○		○	0.239
Method 6		○	○	0.252

(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.