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

Kazuaki Ogawa
Gifu University
1-1 Yanagido Gifu
Gifu 501-1193 Japan
kazuaki@asr.info.gifu-u.ac.jp

Tatsuaki Murahashi
Gifu University
1-1 Yanagido Gifu
Gifu 501-1193 Japan
murahashi@asr.info.gifu-u.ac.jp

Hiroaki Taguchi
Gifu University
1-1 Yanagido Gifu
Gifu 501-1193 Japan
tag@asr.info.gifu-u.ac.jp

Koudai Nakajima
Gifu University
1-1 Yanagido Gifu
Gifu 501-1193 Japan
nakajima@asr.info.gifu-u.ac.jp

Masanori Takehara
Gifu University
1-1 Yanagido Gifu
Gifu 501-1193 Japan
takehara@asr.info.gifu-u.ac.jp

Satoshi Tamura
Gifu University
1-1 Yanagido Gifu
Gifu 501-1193 Japan
tamura@info.gifu-u.ac.jp

Satoru Hayamizu
Gifu University
1-1 Yanagido Gifu
Gifu 501-1193 Japan
hayamizu@gifu-u.ac.jp

ABSTRACT

This paper proposes several approaches for NTCIR-12 SpokenQuery & Doc-2[1]. Our methods are based on the query likelihood model which is one of the probabilisrtic language models choosing Dirichlet smoothing. We try to improve the performance by using extended language models. First, this paper develops and uses the language model obtained from related research papers. Second, this paper proposes a smoothing method employing the cache model and the n-gram model based on Kneser-Ney smoothing. Finally, this paper proposes a smoothing method using neighboring documents. Experiments were conducted to evaluate these methods using NTCIR-12 test sets.

Team Name

Laboratoire de Professeur Chat Noir

Subtasks

SQ-SCR (Japanese)

Keywords

Information retrieval, Query likelihood model, Dirichlet smoothing, Research papers corpus, Cache model, N-gram model, Neighboring document

1. INTRODUCTION

In recent years, multi media contents such as musics, news shows, and movies have been increasing. Moreover the demand of retrieving these contents has been growing. In this paper, we focus on a speech modarity for retrieval, therefore, we call these contents as spoken documents. Spoken documents are desirable to be retrieve quickly. In order to deal with numerous contents, Spoken Document Retrieval(SDR) is conducted using automatically transcribed speech data obtained by speech recognition. SDR has two tasks called Spoken Contents Retrieval(SCR) and Spoken Term Detection(STD). We focus on the SCR task.

There are a lot of researches related to SCR. The TF-IDF score is used widely in document retrieval including our works[2, 3, 4], however, TF-IDF-based methods have a disadvantage that it is hard to retrieve spoken documents including speech recognition errors. Many researches employ probabilistic models, especially the query likelihood model[5]. For example, Hasegawa et al.[3, 4] proposed a new modeling approach by expanding the query likelihood model. In the method, not only static document collection but also dynamic document collection obtained from web pages were employed in the Dirichlet smoothing.

This paper investigates effectiveness of using neighboring documents and external resources, and proposes several approaches using Dirichlet smoothing. Dynamic documents from web pages often include unnecessary or useless sentences. In order to efficiently utilize external resources, we use research papers for query expansion instead of web pages. Because research papers include many technical terms related to target documents that should be retrieved, exploiting documents much related to the target domain must be useful. In addition, we focus on contextual information. Focusing on one sentence or document, its neighboring ones must have useful information for document retrieval. We thus try to use such the information. This paper proposes extended language models for the query likelihood model. As the extended language models, we use 1) unigram from the research papers, 2) cache model, and 3) n-gram language model.

This paper is organized as follows. Setcion 2 describes the query likelihood model and smoothing techniques. Section 3 introduces our proposed approaches. The experimental conditions and the results are presented in Section 4. Section 5 concludes this study.

2. SDR BASED ON QUERY LIKELIHOOD MODEL

This section describes the query likelihood model, Dirichlet smoothing, and query expansion techniques.

2.1 Query likelihood model

The document retrieval issue can be formulated by estimating $P(d|q)$, where q is a given query and d is a document. Applying the Bayesian theorem, $P(d|q)$ is calculated as follows:

$$P(d|q) = \frac{P(q|d)P(d)}{P(q)} \propto P(q|d) \quad (1)$$

In Eq.(1), $P(q)$ can be regarded as a constant because $P(q)$ is independent of any document. $P(d)$ can be ignored when no previous knowledge can be used. Therefore the document retrieval issue can be resolved by estimating $P(q|d)$. $P(q|d)$ is a probability to generate a query q under the condition that a document d is found. This strategy is called the query likelihood model.

The probability $P(q|d)$ is obtained using a language model in this study. Considering corpus sizes used in this work, we employ a unigram language model. Therefore the query likelihood model $P(q|\theta_d)$ is estimated by the unigram language model θ_d as:

$$P(q|\theta_d) = \prod_{w_i \in V} P(w_i|\theta_d)^{C(w_i,q)} \quad (2)$$

where $w_i \in V = \{w_1, w_2, ..., w_{|V|}\}$ is a term in a given query, and w_i can appear more than once in the query q. $C(w_i, q)$ is the number of w_i in the query q. $P(w_i|\theta_d)$ is calculated by using a relative frequency of each term:

$$P(w_i|\theta_d) = \frac{C(w_i, d)}{|d|} \quad (3)$$

where $|d|$ means the total number of terms in a document d.

2.2 Dirichlet smoothing

In the query likelihood model, smoothing techniques are usually used to avoid the zero probability problem. One of the smoothing ways is the linear interpolation. The linear interpolation is formulated by:

$$P(w_i|\theta_d; \lambda) = \lambda \cdot P(w_i|\theta_d) + (1 - \lambda) \cdot P(w_i|\theta_C) \quad (4)$$

$$P(w_i|\theta_C) = \frac{\sum_{d \in C} C(w_i, d)}{\sum_{d \in C} |d|} \quad (5)$$

where λ is a smoothing parameter ($0 \leq \lambda \leq 1$). In Eq.(4), θ_C is a unigram language model for a static document collection C. $P(w_i|\theta_C)$ is an occurence probability of w_i in the static document collection C. The above method uses fixed smoothing parameters, on the other hand, another smoothing technique, Dirichlet smoothing, performs flexible combination according to the document length[6, 7]. The Dirichlet smoothing is given by:

$$P(w_i|\theta_d; \mu) = \frac{|d|}{|d| + \mu} \cdot P(w_i|\theta_d) + \frac{\mu}{|d| + \mu} \cdot P(w_i|\theta_C) \quad (6)$$

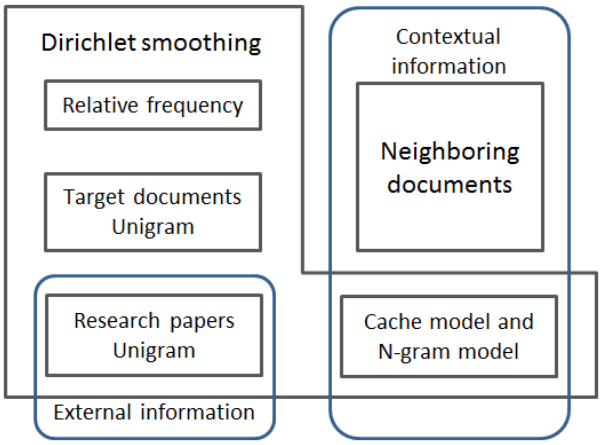

Figure 1: Overview of proposed methods

where μ is a smoothing parameter. The Dirichlet smoothing is empirically better than the linear interpolation, therefore, we employ the Dirichlet smoothing approach.

2.3 Dirichlet smoothing using dynamic documents

In Eq.(6), the target document collection is used as the static document collection C. But if the given query has any terms that do not appear in the static documents, the model θ_d cannot deal with the terms. Therefore Hasegawa et al.[3] proposed to use another document as a dynamic document collection obtained from web pages, in addition to the static document collection. The Dirichlet smoothing was extended using these documents expecting to reduce unknown words. In the method, the extended Dirichlet smoothing is given by:

$$\begin{aligned} P(w_i|\theta_d; \mu, \nu) &= \frac{|d|}{|d| + \mu + \nu} \cdot P(w_i|\theta_d) \\ &+ \frac{\mu}{|d| + \mu + \nu} \cdot P(w_i|\theta_C) \\ &+ \frac{\nu}{|d| + \mu + \nu} \cdot P(w_i|\theta_W) \end{aligned} \quad (7)$$

where $P(w_i|\theta_W)$ is a unigram model for a dynamic document collection W, and ν is a smoothing parameter for W.

3. PROPOSED METHODS

3.1 A research paper corpus

Hasegawa et al. used web documents as the dynamic documents in the extended Dirichlet smoothing[3, 4]. However, the documents from web pages have a lot of noises; some sentences, paragraphs or documents themselves are unnecessary, useless or unstudied. We propose to use a corpus consisting of much relevant documents, in this case, i.e. research papers in ASJ(Acoustical Society of Japan) for the extended Dirichlet smoothing instead of web documents. The corpus of research papers contains lots of professional terms related to the queries and the target documents. In addition, the corpus consists of formatted sentences. We selected 1000 papers from 8336 papers(published in 2005 to 2014) in the corpus based on cosine similarities of TF-IDF vectors between the papers and the queries. Then, selected papers are different at each query. We used selected papers for the

Table 1: Experimental condition.

Task	SpokenQuery&Doc-2
Subtask	SQ-SCR SGS retrieval
Query	K-REF-WORD-MATCH
Target	K-REF-WORD-MATCH
Static document collection	Automatic transcription

smoothing method instead of the dynamic document collection. The formula of the smoothing method using the research papers is shown as:

$$\begin{aligned} P(w_i|\theta_d;\mu,\nu) &= \frac{|d|}{|d|+\mu+\nu}\cdot P(w_i|\theta_d) \\ &+ \frac{\mu}{|d|+\mu+\nu}\cdot P(w_i|\theta_C) \\ &+ \frac{\nu}{|d|+\mu+\nu}\cdot P(w_i|\theta_R) \end{aligned} \quad (8)$$

where R is the dynamic document collection using research papers, θ_R is a unigram model for the corpus, and ν is a smoothing parameter for R.

3.2 Cache model and n-gram model based on Kneser-Ney smoothing

The cache model is based on the local nature of terms that preceding terms are likely to be used again. A probability for the cache model $P_{CH}(w_n|M)$ is calculated as:

$$P_{CH}(w_n|M) = \frac{1}{M}\sum_{m=1}^{M}\delta(w_n, w_{n-m}) \quad (9)$$

$$\delta(w_n, w_{n-m}) = \begin{cases} 1 & (w_n = w_{n-m}) \\ 0 & (otherwise) \end{cases} \quad (10)$$

where $M = \{w_{n-M}, \ldots, w_{n-1}\}$ is M terms appeared just before, and $\delta()$ is the Kronecker's δ function. Typically, the cache model is used for linear interpolation in the n-gram model. This paper uses the n-gram model using Kneser-Ney smoothing. The formula of smoothing method using the linear interpolation both the cache model and the n-gram model is shown as:

$$\begin{aligned} P(w_i|\theta_d;\mu,\nu) &= \frac{|d|}{|d|+\mu+\nu}\cdot P(w_i|\theta_d) \\ &+ \frac{\mu}{|d|+\mu+\nu}\cdot P(w_i|\theta_C) \\ &+ \frac{\nu}{|d|+\mu+\nu}\cdot P(w_i|\theta_{KC}) \end{aligned} \quad (11)$$

$$P(w_i|\theta_{KC}) = \left(\gamma P_{KN}(w_i|w_{n-N+1}^{n-1}) + (1-\gamma)P_{CH}(w_i|w_{n-M}^{n-1})\right) \quad (12)$$

In Eq.(11) and Ep.(12), $P_{KN}(w_i|w_{n-N+1}^{n-1})$ is a probability based on the n-gram model using Kneser-Ney smoothing, $P_{CH}(w_i|w_{n-M}^{n-1})$ is a probability based on the cache model, ν is a smoothing parameter for KC, and γ is a parameter for linear interpolation ($0 \leq \gamma \leq 1$). We expect to add contextual information by using the cache model and the n-gram model.

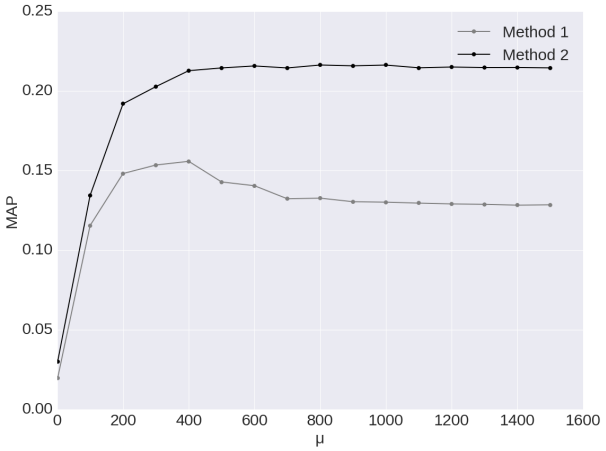

Figure 2: MAP scores of the Method 1 and 2, according to the parameter μ using NTCIR-12 Dry-run data.

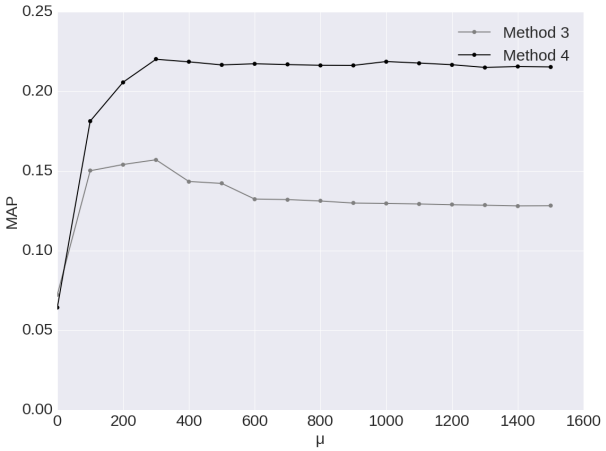

Figure 3: MAP scores of the Method 3 and 4, according to the parameter μ when fixed $\nu = 80$ using NTCIR-12 Dry-run data.

3.3 Neighboring documents

Parallel materials such as presentation slides and related papers are useful informations in Spoken Document Retrieval. In this task, the target documents are devided by the presentation slides. The term length of a target document is often too short to match a query. Therefore we use not only a target document but also its neighboring documents to calculate a similarity between a query and a target document. The formula of the similarity score $S'(i)$ using neighboring documents is as follows:

$$S'(i) = \sum_{n=-L}^{L} w_n S(i+n) \quad (13)$$

$$w_n = \frac{1}{|n|+1} \quad (14)$$

In Eq.(13), $S(i)$ is a similarity score of a partial document corresponding to i-th slide and a query. L is the number of neighborhood slides. w_n is the weighting coefficient of each score. We use an inverse proportion coefficient for weighting. For more details, please refer to [8].

4. EXPERIMENTS

4.1 Experimental condition

To evaluate our proposed methods, we conducted experiments under the condition of SQ-SCR SGS retrieval task in NTCIR-12 SpokenQuery&Doc-2. Experimental conditions are shown in Table 1.

For the target documents and the static document collection, we used the provided automatic transcription by KALDI tool kit. Retrieved results were evaluated by Mean Average Precision(MAP) score.

4.2 NTCIR-12 Dry-run evaluation

In order to determine the hyper parameters, we tested following our retrieval methods using NTCIR-12 SpokenQuery&Doc-2 Dry-run data.

1. Query-likelihood-model-based method using Dirichlet smoothing.
2. Our Method 1 + using neighboring documents.
3. Our Method 1 + using the extended unigram model from the research papers.
4. Our Method 3 + using neighboring documents.

We considered smoothing parameter μ in the static document smoothing. Figure 2 shows the result by various values of μ in our method 1 and 2. In addition, figure 3 shows the result by various values of μ in our method 3 and 4 when $\nu = 80$ experimentally. As a result, we are empirically determined $\mu = 320$ and $\nu = 80$.

4.3 NTCIR-12 Formal-run evaluation

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

1. Query-likelihood-model-based method using Dirichlet smoothing.
2. Our Method 1 + using the extended unigram model from the research papers.
3. Our Method 1 + using the extended language model employing the linear interpolation with the cache model and the n-gram model based on Kneser-Ney smoothing.
4. Our Method 1 + using neighboring documents.
5. Our Method 2 + using neighboring documents.
6. Our Method 3 + using neighboring documents.

The smoothing parameters μ and ν were determined as $\mu = 320$ and $\nu = 80$ in Method 1, 2, 4, and 5. Only Method 3 and 6, $\mu = 320$ and $\nu = 10$ were used according to preliminary experiments. Method 2 and 5 used dynamic documents obtained from research papers as external information, but Method 3 and 6 used only static documents. In addition, the Dirichlet smoothing is extended to use a target document and a part of static documents in the cache model and the n-gram model. Thus the parameter ν in Method 3 and 6 was experimentally determined $\nu = 10$. Moreover, in Method 3 and 6, the parameter M in the cache model, the parameter n in the n-gram model, and the parameter γ for linear interpolation are experimentally determined $M = 100$, $n = 5$, and $\gamma = 0.25$.

Table 2: MAP scores of our approaches for NTCIR-12 Formal-run data.

	$P(w_i\|\theta_d)$	$P(w_i\|\theta_C)$	$P(w_i\|\theta_R)$	$P(w_i\|\theta_{KC})$	neighboring	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

Table 2 shows MAP results computed by the evaluation tool distributed from NTCIR-12 SpokenQuery&Doc-2 Task organizer. Comparing Method 1 to 3 and the others, our proposed methods using neighboring documents were better than the other methods. We can see that it is useful to use neighboring documents.

Next, compared to Method 1 and 4, we found that each MAP score of Method 2 and 5 slightly decreased. It is considered to affect extended smoothing method using the unigram model with the research papers corpus. Then, it is estimated that unknown words and the way to select research papers are the factor to become worse.

We obtained the best result when using Method 6. Consequently, smoothing method using neighboring documents in addition to the linear interpolation with the cache model and the n-gram model is useful on NTCIR-12 Formal-run evaluation. On the other hand, not using neighboring documents, the best result obtained from Method 3 using the cache model and the n-gram model. We can see that it is useful to use the cache model and the n-gram model. However, using neighboring documents to Method 3 is not so effective as using these to Method 1 and 2. Note that using the cache model and the n-gram model has the similar effects as using neighboring documents. Therefore we may need to improve the integration scheme in our approach.

In conclusion, we got better results when using neighboring documents, the cache model, and the n-gram model. Therefore it is found that using contextual information has the effectiveness for Spoken Document Retrieval.

5. CONCLUSION

This paper proposed several techniques for SDR. First, we employed a research papaer corpus as the dynamic document collection in the extended Dirichlet smoothing. Second, we used the linear interpolation with the cache model and the n-gram model based on the Kneser-Ney smoothing in the extended smoothing method. Finally, we utilized neighboring documents by weighting them using the inverse proportion coefficients.

Experiments were conducted using the NTCIR-12 SpokenQuery & Doc-2 Formal-run data sets. As a result, the proposed method is successful, which adopted the extended Dirichlet smoothing using the linear interpolation with the cache model and the n-gram model, further neighboring documents. It also turns out that using contextual information is important for SDR.

Our future works are as follows. It is still considered to exist unknown words in queries. We should deal with these words. We also have to investigate effectiveness of employing

more related papers to queries. Moreover we must try to reconsider weighting methods for neighboring documents. It is also necessary to investigate smoothing parameters. In spite our methods using contextual information are effective, we should try to further improve the performance.

6. REFERENCES

[1] T. Akiba, N.Hiromitsu, H. Nanjo, and Gareth J. F. Jones. Overview of the NTCIR-12 SpokenQuery&Doc-2 task. In *Proc. NTCIR-12*, June 2016.

[2] K. Hara, H. Taguchi, K. Nakajima, M. Takehara, S. Tamura, and S. Hayamizu. Segmented spoken document retrieval using word co-occurrence information. In *Proc. NTCIR11*, pages 395–401, December 2014.

[3] K. Hasegawa, H. Sekiya, M. Takehara, T. Niinomi, S. Tamura, and S. Hayamizu. Toward improvement of SDR accuracy using LDA and query expansion for SpokenDoc. In *Proc. NTCIR9*, pages 261–263, December 2011.

[4] K. Hasegawa, M. Takehara, , S. Tamura, and S. Hayamizu. Spoken document retrieval using extended query model and web documents. In *Proc. NTCIR10*, pages 608–611, June 2013.

[5] B. Chen, K. Chen, P.Chen, and Y.Chen. Spoken Document Retrieval With Unsupervised Query Modeling Techniques. *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*, 20:2602–2612, 2012.

[6] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22:179–214, 2004.

[7] D.J.C. MacKay and L.C.B. Peto. A hierarchical Dirichlet language model. *Natural Language Engineering*, 1:289–307, 1995.

[8] H. Taguchi, S. Tamura, and S. Hayamizu. Expansion of query and documents using relevant documents in spoken document retrieval. In *Proc. 2015 Autumn Meeting, Acoustical Society of Japan*, pages 3–Q–10, September 2015.