Segmented spoken document retrieval using word co-occurrence information

Kensuke Hara
Gifu University
1-1 Yanagido Gifu
Gifu 501-1193 Japan
hara@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 shows several approaches for NTCIR-11 SpokenQuery&Doc [1]. This paper proposes several schemes to use word co-occurrence information for spoken document retrieval. Automatic transcriptions of spoken documents usually contain mis-recognized words, making the performance of spoken document retrieval significantly decrease. The cosine similarity to measure a document similarity must be investigated for spoken documents. It is also difficult to retrieve a segmented document having few terms. To cope with these problem, we utilize Pointwise Mutual Information (PMI). We compute a recognition confidence for each term appeared in a transcription to drop mis-recognized words. We also investigate a PMI-based document comparison approach. Furthermore, a segmented-document retrieval method is also proposed. Experiments were conducted to evaluate these methods using NTCIR-11 test sets.

Team Name

Laboratoire de Professeur Chat Noir

Subtasks

SQ-SCR (Japanese)

Keywords

Error rejection, Document similarity, PMI, Segmented document, Query model

1. INTRODUCTION

In recent years, contents including speech information such as news show and movie have been more and more increasing, and the demand of retrieving these contents has become higher. In this paper, we call these contents as spoken documents. With meta-information such as title, abstract and keyword, spoken documents can be retrieved by using text processing techniques. But adding meta-information requires much costs and human resources. Therefore, Spoken Document Retrieval (SDR) is usually conducted using automatically transcribing speech data obtained by speech recognition.

There are a lot of works related to SDR. For example, Asami et al. used the co-occurrence information in the SDR task [2]. They assumed that mis-recognized words were incoherent in the transcription, and formulated contextual coherence as arithmetic mean of Pointwise Mutual Information (PMI). By rejecting lower coherence transcriptions, they improved the SDR accuracy. The study of Chen et al. is another example of SDR work using a query modeling technique [3]. In the study, the similarity between a query and a document is computed by KL-Divergence. They also tried to overcome a small vocabulary problem for a query by employing a relevance model, and SDR accuracy was improved.

We have developed an SDR method that employs a query model, extended Dirichlet smoothing, and web expansion techniques [4]. Applying our method to NTCIR-9 and NTCIR-10 SpokenDoc test sets, we found our approach is much successful and effective. In these tasks, spoken documents were retrieved according to queries consisting of several keyword terms in short texts, however, it is not clarified our method would be still useful for long or spoken queries. There is another issue about SDR. In the document retrieval literature including our works [4, 5], Term Frequency (TF) and Inverse Document Frequency (IDF) as well as cosine similarity have been widely used. However, a TF-IDF-based approach has several disadvantages; for example, a TF-IDF value for mis-recognized word sometimes becomes an inappropriate score. The cosine distance should be also investigated for spoken queries and documents.

To handle the issues described above, we firstly try to reject mis-recognized words by using word co-occurrence information. We proposed a scheme to extract keywords from spoken transcriptions by using PMI [6]. In the work, we computed coherence measures for possible word pairs in a transcription. If a word has lower PMI scores, each which indicates a coherence between the word and the another term in the transcription, the word must be mis-recognized word and should not be a keyword. Similarly, in this paper, we attempt to remove recognition errors using the technique. Because lower coherence words can be considered as mis-recognized words, the SDR accuracy might be improved by rejecting these mis-recognized words. Note that our proposing method can be incorporated not only in a probabilistic language model such as a query model [4], but also in a vector space model e.g. [7]. This technique is described in Section 2.2 and 4.

Secondly, this paper investigates the cosine distance. In order to measure a distance between a query and a document, we tested a new scheme using PMI instead of the cosine distance. This method formulates a similarity score by summing up all PMIs each which is computed between a word in a spoken query and a word appeared in a document. Since PMI represents a coherence between two terms, if the result for a query and a document has a higher value, we could consider that they must be similar. This method is introduced in Section 2.3 and 5.

Thirdly, this paper studies a document retrieval method for a part of document. In the SDR domain, there are demands to find a paragraph or segment, such as a presentation slide used in a lecture. But a conventional SDR scheme suffers from the low accuracy because such segments have only few words. To overcome this issue, we try to compute the similarity considering not only a query-document similarity but also a query-segment similarity. Section 6 shows the detail of this technique.

This paper is organized as follows. Section 2 shows the flow of proposed SDR schemes. Our previous method proposed in NTCIR-10 is briefly described in Section 3. PMI as well as a proposed error rejection method is introduced in Section 4. Section 5 describes a proposed query-document comparison approach. A new document retrieval using query-document and query-segment similarities is appeared in Section 6. Experimental condition and results are presented in Section 7, and Section 8 concludes this study.

2. FLOW OF DOCUMENT RETRIEVAL

We studied SDR methods in this paper by combining several techniques. The representative methods are introduced in this section:

1. query model using error rejection method
2. TF-IDF using web query expansion, error rejection method, and proposed comparison method using PMI

The flows of these methods are described in following subsection.

2.1 Query model based SDR

Figure 1 displays the flow of document retrieval method using query model. At first, nouns in a target document are extracted by analyzing the document transcription morphologically. These nouns are screened the by proposed word rejection method, and words considered to be mis-recognized words are dropped.

As described, the TF-IDF is available only if common words are appeared in the query. In most cases, the words in the target document do not appear in the query, but terms having the same or similar meanings lie in the query instead. To overcome this issue, the dynamic document collection is obtained by the web query expansion technique[5]. Some web pages in the dynamic document collection contain important information, while some web pages have less information. Therefore, web pages are weighted using LDA(Latent Dirichlet Allocation). Then, probabilities for a word are computed from the automatic transcription from the target document, static document collection, and dynamic document collection, respectively. Using these probabilities, a similarity score between the query and each target document is computed by the query model. The details of the techniques are described in Section 3.

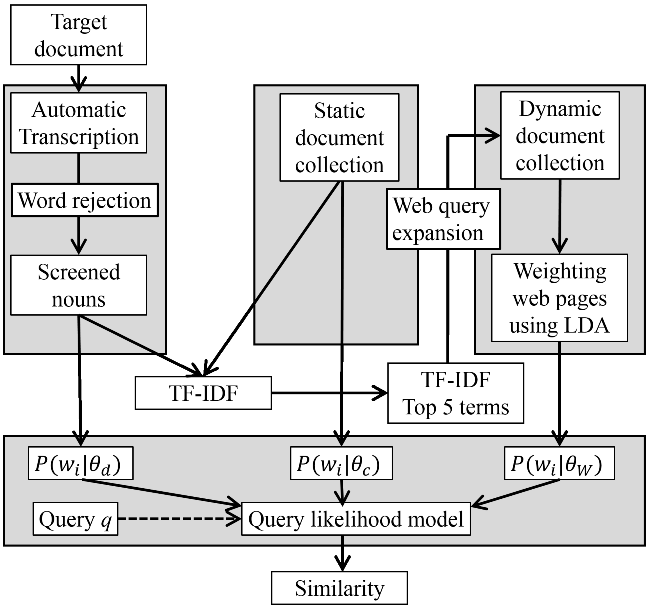

Figure 1: A flow of the query model using the error rejection method

2.2 Proposed word rejection method

Automatic transcription usually contains mis-recognized words, and it causes decrease of SDR accuracy. To handle this issue, we try to reject mis-recognized terms by considering coherence of each term. Coherence measure is computed as sumPMI[6], which is our proposed feature. Words having smaller sumPMI scores is incoherent for the document, and they are usually mis-recognized words. In this study, we try to improve the accuracy of document retrieval by rejecting words having smaller sumPMI. Section 4 presents how to compute sumPMI scores.

2.3 Vector space model SDR

Figure 2 shows the whole aspect of our proposed spoken document retrieval, and figure 3 describes the detailed flow of feature extraction shown in Figure 2. For a target document and a spoken query, expanded TF-IDF vector are computed. Computing simple TF-IDF from the automatic transcription and building the dynamic document collection are same as the query model based SDR. Another TF-IDF is calculated from the dynamic document collection, and an extended TF-IDF vector is by the linear combination of the TF-IDF from the target document and that one from the dynamic document collection. Finally, comparison between TF-IDFs of the query and each target document is conducted using PMI. Section 5 describes more information about this comparison method.

3. QUERY-MODEL-BASED SDR

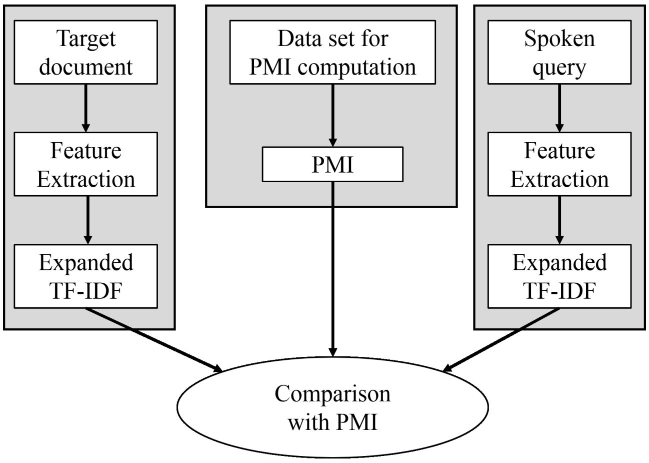

Figure 2: A Flow of spoken document retrieval using PMI

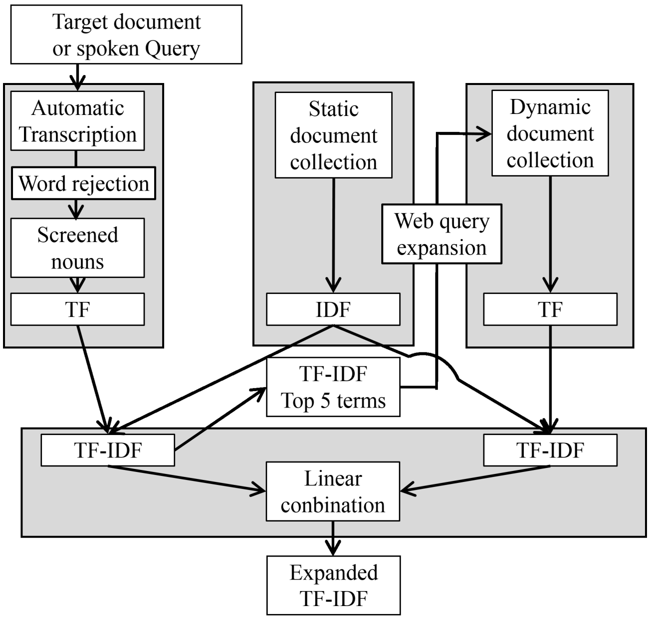

Figure 3: A flow of feature extraction to obtain an expanded TF-IDF vector

This section briefly mentions our SDR method proposed in NICIR-10, employing a query model as well as Dirichlet smoothing and web expansion technologies [4].

3.1 Query model

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

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

In Eq.(1), $P(q)$ can be treated as constant because $P(q)$ is independent of any document. And $P(d)$ 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. In this study, a query likelihood $P(q|\theta_d)$ is estimated by a unigram language model θ_d as:

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

where $w_i \in V = \{w_1, w_2, ..., w_{|V|}\}$ is a term in a given query, and $C(w_i, q)$ is a count for w_i in q. $P(w_i|\theta_d)$ is accomplished by using a relative frequency of each term:

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

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

3.2 Dirichlet smoothing

In the query modeling, a smoothing technique is usually employed to avoid the zero probability problem. We employed an extended Dirichlet smoothing approach using both a static document collection and a dynamic document collection:

$$\begin{aligned} P(w_i|\theta_d; \mu, \nu) &= \frac{|d|}{|d| + \mu + \nu} \cdot P(w_i|\theta_d) \\ &+ \frac{\mu}{|d| + \mu + \nu} \cdot P(w_i|\theta_C) \\ &+ \frac{\nu}{|d| + \mu + \nu} \cdot P(w_i|\theta_W) \end{aligned} \tag{4}$$

where μ and ν are smoothing parameters for a static document collection C and a dynamic document collection W, respectively.

3.3 Web query expansion

The dynamic document collection introduced in the last subsection is obtained utilizing the query expansion technique [5]. At first, a TF-IDF score is computed for each word appeared in speech recognition results. Secondly, query terms for web expansion are chosen based on the TF-IDF values. In this study, we used top-five TF-IDF words. Retrieval queries to get web pages are constructed using all the possible query term triplets, and then web retrieval is conducted. For each query, the same number of web sites are obtained. In this work, we used top-three pages. We finally obtained 30 web pages for each document.

3.4 Weighting web pages

Some web pages contain important information, while some pages have less information. Therefore, weighting a web page is effective. In our method, a weighting score for each web page is obtained using Latent Dirichlet Allocation (LDA) [8]. At first, a topic mixture ratio vector $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, ..., \gamma_z)^\top$ for each web page as well as a target document is respectively computed applying LDA (z is the number of latent topics). Secondly, a cosine distance $\delta(p, d)$ of the topic mixture vectors for a web page p and a candidate document t in the document set is calculated. Thirdly, the distance between the web page p and the target document collection $D = \{d_1, d_2, ..., d_{|D|}\}$ is calculated as:

$$\delta(p, D) = \frac{1}{|D|} \sum_{m=1}^{|D|} \delta(p, d_m) \tag{5}$$

The value $\delta(p, D)$ is used to weight the corresponding web page. Finally, a probability $P(w_i|\theta_W)$ in Eq.(4) to observe a term w_i in the dynamic document collection W is formulated as:

$$P(w_i|\theta_W) = \frac{\sum_{j=1}^{|W|} \delta(p_j, D) \cdot C(w_i, p_j)}{\sum_{j=1}^{|W|} \sum_{k=1}^{N_j} \delta(p_j, D) \cdot C(w_k, p_j)} \tag{6}$$

where N_j is the total number of terms in a web page p_j.

4. MIS-RECOGNIZED WORD REJECTION

In this section, PMI is summarized before our mis-recognized term rejection scheme is introduced.

4.1 PMI

PMI is a measure that represents a strength of relationship between two events. PMI is calculated as:

$$\begin{aligned} PMI(x, y) &= \log \frac{P(x, y)}{P(x)P(y)} \qquad (7) \\ &= \log \frac{f(x, y) \cdot K}{f(x)f(y)} \qquad (8) \end{aligned}$$

where $P(x)$ is an occurrence probability of word x, $f(x)$ is the number of occurrences of word x. $P(x, y)$ is a co-occurrence probability of words x and y, $f(x, y)$ is the number of co-occurrences of words x and y. $P(x)$, $P(y)$, and $P(x, y)$ are calculated using a data set for PMI computation; sometimes the data set corresponds to the document itself. The stronger relationship between x and y is, the larger $PMI(x, y)$ is. In this research, we use nouns in the same document as words x and y. When PMI is applied to a document, it is needed to divide the document into frames (windowing). Frames each including N content words (e.g. nouns) are extracted from the document every M content words, that is, window size and window shift are N and M, respectively. Note that in Eq.(8) K is the number of frames in the data set.

PMI has two problems. One is that PMI does not measure the relationship of words which do not co-occur, and another is that PMI has an excessively large value when x and y rarely occur. In order to solve these problems, we employ the smoothed PMI [2]. This method uses t-test to examine whether $f(x)$ and $f(y)$ are large enough or not. The t-score $t(x, y)$ tests whether the difference between $P(x, y)$ and $P(x)P(y)$ is significant or not. The smoothed PMI is calculated as:

$$PMI(x, y) = \begin{cases} \log \dfrac{\hat{f}(x, y) \cdot K}{f(x)f(y)} & \text{if } t(x, y) > \theta \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

$$\hat{f}(x, y) = \begin{cases} f(x, y) & \text{if } f(x, y) > 0 \\ \dfrac{N_1}{N_0} & \text{otherwise} \end{cases} \tag{10}$$

$$t(x, y) = \frac{\left| \hat{f}(x, y) - \dfrac{f(x)f(y)}{K} \right|}{\sqrt{\hat{f}(x, y)}} \tag{11}$$

where N_0 is the number of word pairs which do not co-occur in any frames in a document, N_1 is the number of word pairs which co-occur only once in the document. θ is a threshold of t-test, which is determined according to the significance level. In this paper, the significance level is set to 5% ($\theta = 1.65$). Also, PMI is normalized into [-1, 1].

4.2 A word rejection approach

According to the previous work [2], mis-recognized words which appear in a recognized hypothesis is likely to be incoherent. Therefore, we introduce the following coherence score which shows how the word is related with the hypothesis d:

$$sumPMI(w) = \sum_{w_i \in d} PMI(w, w_i) \tag{12}$$

Here, w and w_i are words appeared in the recognized hypothesis, and d is a recognition result of a whole spoken document. As mentioned in the last subsection, PMI means the relationship between two words. Thus, $sumPMI$ shows the relationship between the word w and the spoken documents d. The words having lower $sumPMI$ are incoherent in the spoken document, and must be unimportant or mis-recognized words. In this study, words having negative $sumPMI$ values are dropped.

5. QUERY-DOCUMENT COMPARISON

In many studies, the cosine distance has been conventionally chosen. Nevertheless, sometimes a serious problem occurs if there is no common word in comparing documents. Another issue arises, that is, words having the same or similar meaning (e.g. speech recognition and ASR) are treated as completely different ones. We employed smoothing and query expansion techniques to handle these problems. Our approach was empirically successful, however, the smoothing did not consider any similarity of words. Therefore, this paper proposes a comparison method computing relationships between words in a given query and ones in each document, using PMI. In the proposed scheme, a similarity between a query q and a document d is represented as:

$$sim(v_q, v_d) = \frac{\sum_{a_i \in q} \sum_{b_j \in d} v_q(a_i) v_d(b_j) R(a_i, b_j)}{\sum_{a_i \in q} \sum_{b_j \in d} v_q(a_i) v_d(b_j)} \tag{13}$$

$$R(w_1, w_2) = \begin{cases} 1 & \text{if } (w_1 = w_2) \\ PMI(w_1, w_2) & \text{otherwise} \end{cases} \tag{14}$$

Here, a_i and b_j are words where $q = \{a_1, a_2, ..., a_N\}$ and $d = \{b_1, b_2, ..., b_M\}$. In Eq.(13), $v_q(a_i)$ indicates a magnitude of word a_i in a query term vector v_q, and $v_d(b_j)$ is also a word magnitude in a document term vector v_d.

Table 1: Experimental condition

Subtask	Slide Group retrieval
Query	REF-WORD-MATCH
Target	REF-WORD-MATCH
LDA training data	Mainichi newspaper corpus 2007-2008
Static document collection	Manual transcription
Smoothing parameter	$\mu = 4000, \nu = 50$

6. SDR FOR SEGMENTED DOCUMENT

As mentioned in Section 1, there is a demand to identify which part of document is most suitable for a query. Retrieving a part of document is simply achieved by regarding the part of document as one independent document. In such the case, however, the SDR accuracy sometimes decreases because we ignore the contents of the original whole document. In this research, we try to improve the accuracy of retrieving a segment of document by taking a relationship between a query and an original document including the segment into account. The expanded similarity S for a segment of document is calculated as:

$$S = \alpha \cdot S_c + (1 - \alpha) \cdot S_d \quad (15)$$

where S_c is a similarity for a segment, S_d is a similarity score for a whole document including the segment, and α is a weighting factor $(0 < \alpha < 1)$.

In addition, we also implement a query expansion technique using web-retrieved data. A extended query vector $\boldsymbol{q}_e$ is obtained from a linear combination of an original query vector $\boldsymbol{q}_o$ and a term vector obtained by web retrieval $\boldsymbol{q}_w$ as:

$$\boldsymbol{q}_e = \beta \cdot \boldsymbol{q}_o + (1 - \beta) \cdot \boldsymbol{q}_w \quad (16)$$

7. EXPERIMENTS

7.1 Experimental condition

To evaluate our proposed query-model-based method and to investigate the PMI-based comparison as well as the combination of query-segment and query-document similarities, we conducted experiments under the condition of SQ-SCR task in NTCIR-11 SpokenQuery&Doc [1]. Experimental condition is shown in Table 1. For the target documents and the static document collection for smoothing, we used the provided automatic transcription and manual transcription, respectively. Note that we simply employed the automatic transcription provided by the organizer as it was, that is, no improvement for the transcription was done. Because presentation slides in NTCIR-11 data (equivalent to segmented documents) having less than 100 characters might be meaningless, these slides are rejected from the target documents. Retrieved results were evaluated by Mean Average Precision (MAP) score.

7.2 NTCIR-11 dry-run evaluation

In order to optimize the parameter α and β appeared in Section 6, we conducted a preliminary SDR experiment using NTCIR-11 SpokenQuery&Doc dry-run data. Since no correct result for the dry-run test set is available when we conducted this experiment, we manually made the unofficial correct data for five documents in the data set. So we only used the five documents in the experiment. In this examination, we used the expanded TF-IDF and the cosine similarity using the retrieval described in Section 6. We changed the parameter α from 0.0 to 0.9 by 0.1 under the condition $\beta = 1.0$. The parameter β is also tested under the condition $\alpha = 0.0$. Figure 4 and Figure 5 show MAP scores according to α and β respectively. According to Figure 4 and Figure 5, the SDR scheme using $\alpha = 0.6$ and $\beta = 0.9$ achieved the best performance. we chose $\alpha = 0.6$ which achieved the best performance in the following experiment. It is also obvious that the SDR performance is improved by our retrieval technique using not only query-segment similarity but also query-document one.

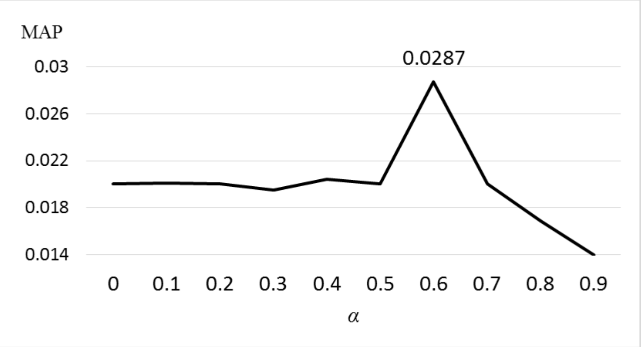

Figure 4: MAP scores of the proposed method using query-segment and query-document similarities, according to the parameter α.

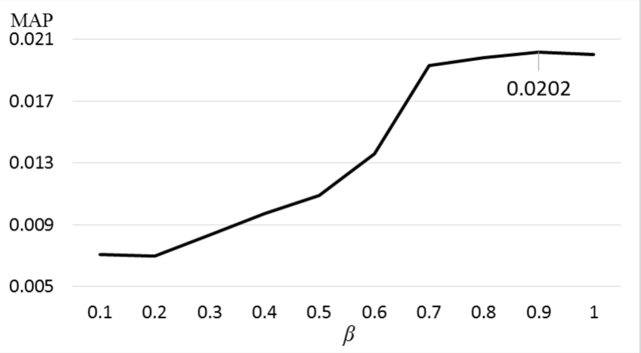

Figure 5: MAP scores of the proposed method using query-segment and query-document similarities, according to the parameter β.

Next, to evaluate our proposed recognition error rejecting method in Section 4, we compared the web-expanded TF-IDF with/without rejecting lower $sumPMI$ terms. The cosine similarity is used in this experiment. Table 2 shows the result. Unfortunately, rejecting lower $sumPMI$ terms could not contribute to the improvement. It is caused maybe because PMI scores between general terms appearing in any documents tend to be high, compared to the other terms. In contrast, PMIs for some keywords and characteristic terms become smaller, then as a result, sometimes they might be wrongly dropped. So we have to refine our method so as to avoid this phenomenon, for example by discounting the PMI score for general terms. Note that the score may be fixed

Table 2: MAP scores with/without the recognition error rejecting method explained in Section 4.

Web-expanded TF-IDF with our rejection method	0.0142
Web-expanded TF-IDF without our rejection method	0.0200

Table 3: MAP scores of our four approaches for NTCIR-11 formal-run data.

	Sec.3	Sec.5	Sec.6	MAP
Method 1	x		x	0.161
Method 2	x			0.114
Method 3			x	0.143
Method 4		x	x	0.047

because our SDR program used in the dry-run evaluation had several bugs.

7.3 NTCIR-11 formal-run evaluation

Using NTCIR-11 SpokenQuery&Doc formal-run data, we tested the following four retrieval methods.

1. Our query-model-based method employing the technique introduced in Section 6,
2. Our original query-model-based method for NTCIR-10,
3. A method using web-expanded TF-IDF vectors and the cosine similarity employing the technique introduced in Section 6,
4. A method using original TF-IDF vectors and our proposed PMI similarity described in Section 5, employing the technique introduced in Section 6.

In this experiment, the parameters α and β are determined as $\alpha = 0.6$ and $\beta = 0.9$. In the cosine similarity method, the number of web pages in the dynamic document collection was set to 100 for each query, slide, and lecture. In the SDR method using the query model, the number of web pages used for each query was 100, but the dynamic document collection was not used in each slide and lecture. In the PMI similarity method, we used simple TF-IDF vectors instead of web-expanded TF-IDF vectors, due to the computational and retrieving loads. Table 3 shows the results computed by the evaluation tool (evalsqscr ver.2.02) distributed from NTCIR-11 SpokenQuery&Doc task organizer. Note that the SDR results computed by these four methods are different from those ones submitted to NTCIR-11 SpokenQuery&Doc task organizer, due to the bugs described in the last section. Comparing the results of Method 1 and 2, we can know that it is useful to use both query-document and query-segment in the retrieval results, as shown in the last subsection. The MAP scores using the cosine similarity (Method 3) and those using our PMI similarity (Method 4) indicate that the proposed PMI similarity is not useful for SDR. This must be also due to the corruption by general terms. Therefore, we also try to reduce the influence caused by general terms in near future.

8. CONCLUSION

This paper proposes several techniques for SDR. The first scheme is to remove mis-recognized words using PMI information, by introducing a measure score called *sumPMI*. The second one is to compute a similarity between a query and a document, by computing a linear combination of PMI scores between terms in a query and ones in a target document. The third approach is to find a segmented document, which is obtained from a query-document similarity and query-segment similarity. Also, we try to apply our query-model-based method to a spoken-query task in addition to the above three proposals.

Experiments were conducted using the NTCIR-11 SpokenQuery&Doc dry-run and formal-run data sets. As a result, the segmented-document retrieval technique improves the SDR performance, and a method based on our query-model system and the technique achieved the best performance. We can conclude that it is useful to use a query-document information when retrieving a segmented document. On the other hand, we found further issues about our first and second schemes based on co-occurrence information.

As our future work, we should improve our methods using PMIs by removing the influence caused by general terms. To deal with these terms, using IDF or taking stop words into account might be useful. In the error rejecting method, we have to optimize the threshold to improve the performance. And the comparison method using PMIs will be improved by incorporating the conventional cosine scale.

9. REFERENCES

[1] T. Akiba, N Hiromitsu, H Nanjo, and Gareth J. F. Jones. Overview of the ntcir-11 spokenquery&doc task. In Proc. NTCIR11, December 2014.

[2] T. Asami, N. Nomoto, S. Kobashikawa, Y. Yamaguchi, H. Masataki, and S. Takahashi. Spoken document confidence estimation using contextual coherence. In Proc. INTERSPEECH2011, pages 1961–1964, August 2011.

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

[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] 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.

[6] K. Hara, H. Sekiya, T. Kawase, S. tamura, and S. Hayamizu. Confidence estimation and keyword extraction from speech recognition result based on web information. In Proc. APSIPA2013, October 2013.

[7] S. Tsuge, K. Ichikawa, N. Kitaoka, K. Takeda, and K. Kita. Spoken content retrieval using distance combination and spoken term detection using hash function for ntcir10 spokendoc2 task. In Proc. NTCIR10, pages 597–603, June 2013.

[8] D. M. Blei et al. Latent dirichlet allocation. Machine Learning Research, 3:993–1022, 2003.