Spoken Document Retrieval Experiments for SpokenDoc-2 at Ryukoku University (RYSDT)

Hiroaki NANJO
Faculty of Science and
Technology, Ryukoku
University, Japan
nanjo@rins.ryukoku.ac.jp

Tomohiro NISHIO
Graduate School of Science
and Technology, Ryukoku
University, Japan
nishio@nlp.i.ryukoku.ac.jp

Takehiko YOSHIMI
Faculty of Science and
Technology, Ryukoku
University, Japan
yoshimi@rins.ryukoku.ac.jp

ABSTRACT

In this paper, we describe spoken document retrieval systems in Ryukoku University, which were participated in NTCIR-10 IR for Spoken Documents ("SpokenDoc-2") task. In NTCIR-10 "SpokenDoc-2" task, there are two subtasks: "spoken term detection (STD) subtask" and "ad-hoc spoken content retrieval (SCR) subtask". We participated in the SCR subtask as team RYSDT. In this paper, our SCR systems are described.

Categories and Subject Descriptors

H3.3 [Information Storage and Retrieval]: Information Search and Retrieval

Keywords

NTCIR-10, ad-hoc spoken content retrieval

Team Name: [RYSDT]
Subtask/Languages: [Spoken Content Retrieval] / [Japanese]
External Resources Used: N/A

1. INTRODUCTION

With the advance of high-speed networks and mass storage, a great deal of audio contents, such as podcasts, TV news, home videos, and lecture/presentation videos, can be easily stored and published. In some universities, lecture videos are actually published in the open domain through web pages. Since such audio contents continue to increase, we need robust methods that can deal with them. Spoken content retrieval (SCR) and spoken term detection (STD), which process such huge amounts of spoken data for efficient search and browsing, are promising and are the most significant tasks of spoken document retrieval.

We have studied both STD [1] and SCR [2] [3] [4]. In NTCIR-10, Spoken Documents ("SpokenDoc-2") task is defined, which covers both STD and SCR. In this paper, our SCR approaches and the results of their applications to NTCIR-10 task are described. Our team name is RYSDT in NTCIR-10 SpokenDoc [5].

2. SPOKEN CONTENT RETRIEVAL SYSTEMS

Spoken content retrieval (SCR) is a process finding the spoken document itself or short portions (passages) of spoken document which are relevant to the query. For the SCR task in NTCIR-10, a search target is Japanese oral presentations in academic conferences and simulated presentations. Since each presentation has a longer duration about 12 minutes to 1 hour, just searching each presentation is not enough since we cannot access a specific scene which we want to know even if the suitable presentations are perfectly retrieved. Therefore, NTCIR-10 SCR defined not only lecture unit retrieval task, but also passage unit retrieval task. We tried to search both lecture unit and passage unit based on an orthodox vector space model (VSM).

2.1 Problem in Vector Space Modeling in Japanese SCR

For a VSM-based SCR, appropriate indexing is significant. Automatic speech recognition (ASR) is performed to make index terms, which essentially contain ASR errors. Therefore, studies of indexing terms that are robust to ASR errors are necessary. In Japanese text, no space is put between words, and word units are ambiguous. Thus, studies of indexing units are also important. Based on this background, we have investigated several indexing units in Japanese SCR [2] including morpheme unit, character n-gram unit, and phone n-gram unit. We have found that morphemes is suitable for indexing unit and baseforms of nouns and verbs are suitable for index terms. For the NTCIR-10 SCR subtask in SpokenDoc-10, we applied above described VSM-based SCR systems.

2.2 Overview of Pseudo Relevance Feedback and its Problems

In information retrieval (IR), users often cannot get documents which they want by their original query since the query does not contain adequate information for their intent. For the problem, in this paper, query expansion (QE) [6] [7] [8] [9], that is, adding some words to an original query is investigated. In SCR, IR target is automatic speech recognition (ASR) results which must contain ASR errors. When a significant word is not recognized correctly in a target document, we cannot find the document by an original query which consists of only the significant word and some trivial words. QE is promising for such a case.

In a VSM-based IR system, a query is denoted by a vector whose each element represents a frequency of its corresponding word. In a VSM, a vector $\mathrm{q_e}$ for an expanded query Q_e is shown in equation (1).

$$\mathrm{q_e} = \mathrm{q} + \alpha \mathrm{q_a} \tag{1}$$

Here, q is a vector for the original query. $\mathrm{q_e}$ is a vector

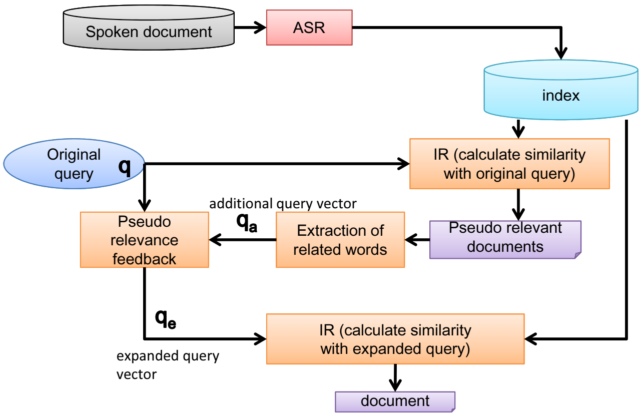
Spoken document
ASR
index
Original query
q
IR (calculate similarity with original query)
additional query vector
q_a
Pseudo relevance feedback
Extraction of related words
Pseudo relevant documents
q_e
expanded query vector
IR (calculate similarity with expanded query)
document

Figure 1: SCR by expanded query based on pseudo relevance feedback

of the expanded query. $\mathrm{q_a}$ is a vector consisting of statistics of new index words to be added. α is a weight. Generally, α less than one is used. In this study, in order to set all elements of $\mathrm{q_e}$ to be integer, QE is performed based on the equation (2) using positive integer β.

$$\mathrm{q_e} = \beta \mathrm{q} + \mathrm{q_a} \tag{2}$$

Relevance feedback is one of the most well-known methods for improving IR accuracy. From IR results obtained by original query q, users select which documents are relevant or not, and modify a query vector $\mathbf{q}$ to $\mathbf{q_e}$. In relevance feedback, users must specify the relevant documents, and it takes much cost. On the other hand, automatic selection of relevant documents for relevance feedback is proposed, namely, pseudo relevance feedback (PRF)[9][10][11]. The overview is shown in Fig. 1. In order to obtain relevant documents automatically, the similarity scores of documents and the original query are used. Actually, documents which have higher similarity scores are regarded as relevant documents.

For longer spoken document retrieval tasks such as lecture retrieval task in NTCIR-10 SpokenDoc-2, PRF does not work well even when relevant documents are correctly retrieved. Extracting relevant words from longer relevant documents is quite difficult since a lot of subtopics, which are not relevant to a given query, are included in a long relevant document and additional words are extracted from such irrelevant subtopics. On the other hands, in a short spoken document retrieval, extracting relevant words from relevant documents is not difficult but retrieving relevant documents itself is difficult. Moreover, in the NTCIR-10 passage retrieval task, longer lectures are not segmented to subtopics. Automatic segmentation to subtopics is difficult. These facts show that the PRF may not work well for the task.

Based on the background, we first construct VSM-based SCR systems and apply orthodox PRF for the NTCIR-10 SCR subtasks; lecture retrieval task and passage retrieval task and show that the orthodox PRF is not effective especially for lecture retrieval task. Then, we propose a PRF method using a special index for extracting relevant words and show that the method works well.

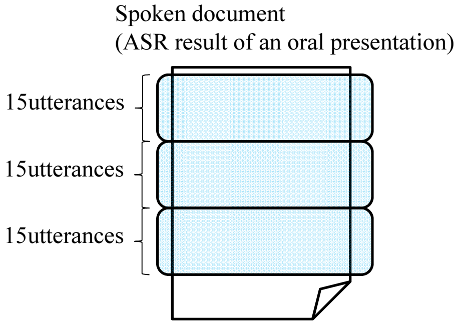

Figure 2: 15-utterances-based unit

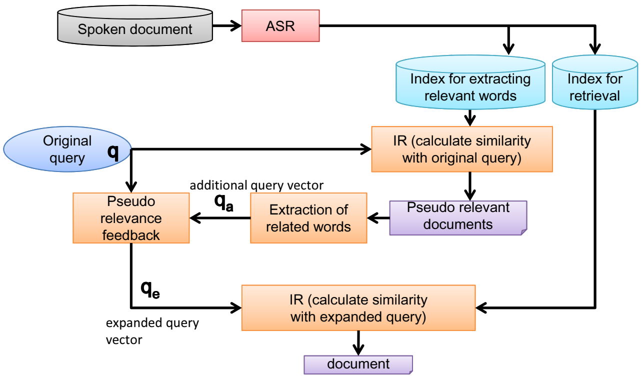
Spoken document
ASR
Index for extracting relevant words
Index for retrieval
Original query
q
IR (calculate similarity with original query)
additional query vector
q_a
Pseudo relevance feedback
Extraction of related words
Pseudo relevant documents
q_e
expanded query vector
IR (calculate similarity with expanded query)
document

Figure 3: Introduction of an Index for Extracting Relevant Words in PRF

2.3 Introduction of an Index for Extracting Relevant Words in PRF

For extracting relevant words in PRF, we propose an introdution of new index which consists of adequate-length spoken document units. As shown in Figure 3, first pseudo relevant documents are retrieved with an original query from the index consisting of adequate length document unit and QE is performed, and then, final results are retrieved with the expanded query with an index consists of retrieval unit.

In longer document unit retrieval task, relevant documents are easily searched but extraction of relevant words is difficult. On the other hand, in shorter document unit retrieval task, relevant documents are not easily searched but extraction of relevant words is not difficult. The main problem is what unit length is suitable for PRF. In the NTCIR-10 subtask, we selected 30, 15, and 10 sequential utterance units and lecture units as a such kind of passage/document, from which we extract pseudo relevant documents and relevant words for PRF.

2.4 Integrating Original Query Similarity into Expanded Query Similarity

Our system have a search strategy that uses both similarity scores between a target document and original/expanded queries. Specifically, for each document, similarities to original/expanded queries are calculated, and the similarities are integrated for IR.

Here, we describe how to integrate original query similarities into expanded query similarities for SCR. Denoting original query and expanded query as Q and Q_e respec-

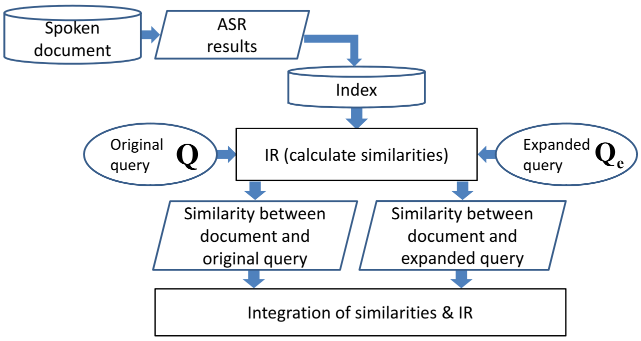
Spoken document
ASR results
Index
Original query Q
IR (calculate similarities)
Expanded query Q_e
Similarity between document and original query
Similarity between document and expanded query
Integration of similarities & IR

Figure 4: Overview of integrating original query similarity and into expanded query similarity

tively, a similarity between a document D_i and Q/Q_e are represented as $\mathrm{SIM}(Q, D_i)$ and $\mathrm{SIM}(Q_e, D_i)$, respectively. Integrating an original query similarity $\mathrm{SIM}(Q, D_i)$ into an expanded query similarity $\mathrm{SIM}(Q_e, D_i)$ is performed as follows (log linear interpolation) with interpolation parameter λ, whose value is between 0.0 and 1.0.

$$\begin{aligned}\mathrm{SIM}'(Q, D_i) &= \lambda \cdot \log \mathrm{SIM}(Q, D_i) \\ &\quad +(1-\lambda) \cdot \log \mathrm{SIM}(Q_e, D_i) \qquad (3)\end{aligned}$$

Then, IR is performed based on the integrated similarity $\mathrm{SIM}'(Q, D_i)$. The overview is shown in Fig. 4.

2.5 SCR System based on Vector Space Model

2.5.1 Indexing Unit and Terms

In our system, Japanese morpheme is defined as an indexing unit. As for indexing terms, baseforms of verbs and nouns are suitable for index terms. To extract such indexing terms from the automatically transcribed lecture texts (ASR results), we performed morphological analysis with Japanese morphological analysis system ChaSen Ver2.2.1 with ipadic-2.4.1. Some character strings are not defined by the ChaSen and are regarded as unknown words, which are regarded as noun in this work. Then, only nouns and verbs (baseform) recognized by the ChaSen are used for index terms.

2.5.2 Retrieving Algorithm

We constructed SCR system based on a VSM and indexing units based on nouns and verbs. In VSM, queries and documents to be retrieved are represented by vectors whose elements correspond to each index term frequencies in each query/document, and vector distance is used as a query-document similarity score. According to the similarity scores, SCR systems output documents. In this work, as a similarity measure between vectors $\mathrm{SIM}(Q, D)$, SMART[12] is used. Specifically, a similarity between a given query Q and a document D_i is given by the equation (4).

$$\mathrm{SMART}(Q, D_i) = \sum_{k=1}^{m} (q_{t_k} \cdot d_{i,t_k}) \qquad (4)$$

where,

$$d_{i,t_k} = \begin{cases} \dfrac{\dfrac{1+\log(\mathrm{tf}_{i,t_k})}{1+\log(\mathrm{avtf})}}{(1 \text{ - slope}) \cdot \mathrm{pivot} + \mathrm{slope} \cdot \mathrm{utf}_i} & \text{if } t_k > 0 \\ 0 & \text{otherwise} \end{cases} \qquad (5)$$

$$q_{t_k} = \begin{cases} \dfrac{1+\log(\mathrm{qtf}_{t_k})}{1+\log(\mathrm{avqtf})} \log \dfrac{N}{n_{t_k}} & \text{if } \mathrm{qtf}_{t_k} > 0 \\ 0 & \text{otherwise} \end{cases} \qquad (6)$$

Here, $t_k (1 \le k \le m)$ represents a k-th index term (word), and m is a size of index terms. tf_{i,t_k} represents a term frequency of t_k in a document D_i. avtf represents an average of the number of the terms in each document. pivot represents an average of the number of unique term in each document. utf_i represents the number of unique terms in D_i. slope represents an interpolation coefficient (0.2).

qtf_{t_k} represents a term frequency of t_k in Q. avqtf represents the average of the number of terms in Q. N represents the total number of documents to be retrieved. n_{t_k} represents the number of documents in which term t_k is included.

In this work, the Generic Engine for Transposable Association (GETA)[13] is used for constructing VSM-based SCR systems.

3. SUBMITTED SCR SYSTEMS

3.1 Lecture Retrieval Systems

First, the results for lecture retrieval task are described. In a lecture retrieval task, search target is an each lecture (total 2702). Each lecture transcription (ASR results) is regarded as a document and VSM-based SCR system is constructed.

Denoting original query as Q, expanded query by PRF from lecture unit as $Q_e(lec)$, expanded query by PRF from 30 unit as $Q_e(30)$, expanded query by PRF from 15 unit as $Q_e(15)$, and expanded query by PRF from 10 unit as $Q_e(10)$, we constructed nine SCR systems as follows. Here, D is lecture unit (document) to be retrieved.

1. SCR uses both $SIM(Q, D)$ and $SIM(Q_e(10), D)$
2. SCR uses both $SIM(Q, D)$ and $SIM(Q_e(15), D)$
3. SCR uses both $SIM(Q, D)$ and $SIM(Q_e(30), D)$
4. SCR uses both $SIM(Q, D)$ and $SIM(Q_e(lec), D)$
5. SCR uses only $SIM(Q_e(10), D)$
6. SCR uses only $SIM(Q_e(15), D)$
7. SCR uses only $SIM(Q_e(30), D)$
8. SCR uses only $SIM(Q_e(lec), D)$
9. SCR uses only $SIM(Q, D)$: baseline

In this work, all parameters (number of pseudo relevant documents/words, the weight for additional words on PRF β in eq.(2), and combination weight λ in eq.(3)) were estimated so that higher 11ptAP was achieved at the NTCIR-9 SpokenDoc dry&formal-run task. We regard the set of the pseudo documents as a query, and based on the equation

(6), words which have higher q_{t_k} are selected for additional words $\mathbf{q_a}$.

For constructing indices, we used "REF-MATCHED" transcription which is given by NTCIR-10 organizers.

3.2 Results

For each query, we tried to retrieve 1000 documents. The results are listed in Figure 5. Mean average precision (MAP) of the baseline system (system 9) is 0.364.

Conventional PRF (system 8), which first retrieves lectures and extracts additional words from them, significantly degraded IR performance. Propopsed PRF (system 5 to 7), which first retrives 10/15/30-utterance-based document units and extracts additional words, achieved comparable performance with baseline system. Combination of baseline IR results and PRF IR results (system 1 to system 4) worked well and improved IR performance. Especially, when we used an index which consists of 10-utterance-based unit for extracting additional words (system 1), we achieved the highest MAP 0.378. Averaged precisions for each query of system 1 and 9 are also shown in Figure 5.

3.3 Passage Retrieval Systems

Next, the results for passage retrieval task are described. In a passage retrieval task, search target is a short part (utterance sequences of arbitrary length) in lectures. Although participants are requested to detect start and end points of such parts, determining such arbitrary length passages is time consuming. Here, we used a uniformly automatically segmented unit as a passage unit instead of such arbitrary length unit. Actually, as shown in Figure 2, we just divided oral presentation speech from its beginning into several segments which consist of 15/10 sequential utterances, which is introduced in the Japanese SCR test collection [14]. We regarded each 15/10 sequential utterance as a passage/document, and VSM-based SCR system is constructed.

Denoting original query as Q, expanded query by PRF from 15 unit as $Q_e(15)$, and expanded query by PRF from 10 unit as $Q_e(10)$, we constructed five SCR systems searches 15 sequential utterance unit and three SCR systems searches 10 sequential utterance unit as follows. Here, $D(15)$ and $D(10)$ is 15/10 sequential utterance unit (document) to be retrieved.

1. SCR uses both $SIM(Q, D(15))$ and $SIM(Q_e(10), D(15))$
2. SCR uses both $SIM(Q, D(15))$ and $SIM(Q_e(15), D(15))$
3. SCR uses only $SIM(Q_e(10), D(15))$
4. SCR uses only $SIM(Q_e(15), D(15))$
5. SCR uses only $SIM(Q, D(15))$

6. SCR uses both $SIM(Q, D(10))$ and $SIM(Q_e(10), D(10))$
7. SCR uses only $SIM(Q_e(10), D(10))$
8. SCR uses only $SIM(Q, D(10))$

Also, all parameters were estimated so that higher 11ptAP was achieved at the NTCIR-9 SpokenDoc dry&formal-run task. We regard the set of the pseudo documents as a query,

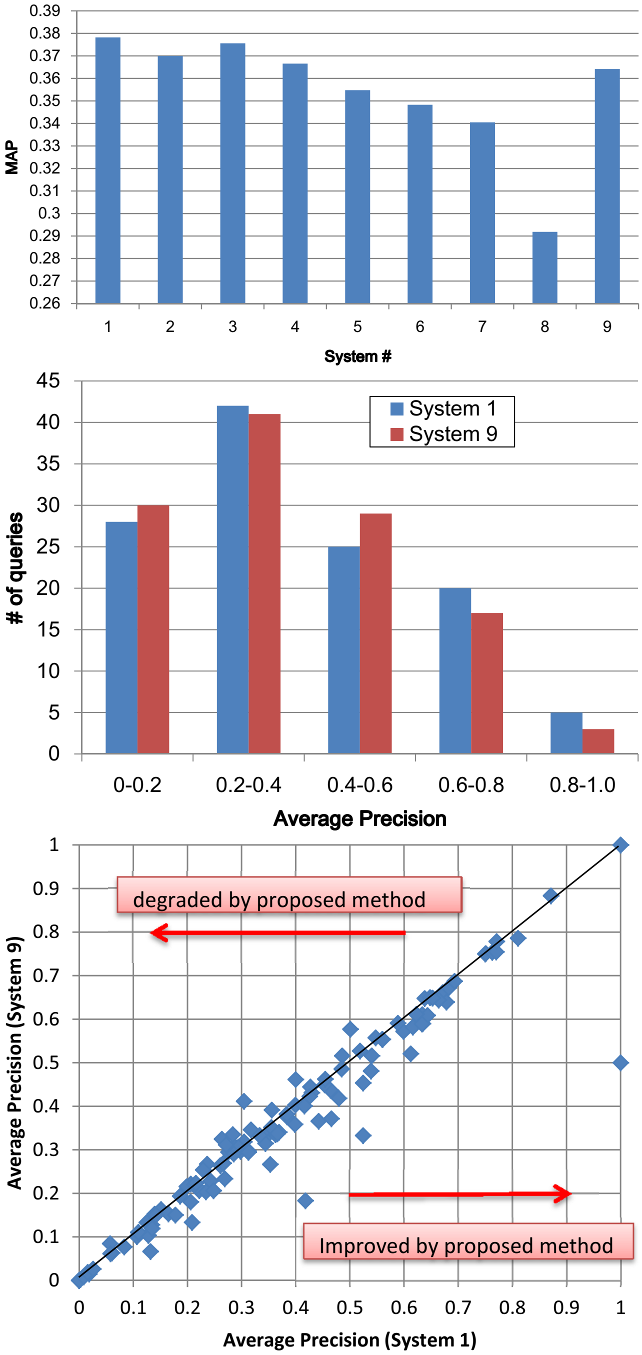

Figure 5: SCR performance for lecture retrieval task

and based on the equation (6), words which have higher q_{t_k} are selected for additional words $\mathbf{q_a}$.

For constructing indices, we used "REF-MATCHED" transcription which is given by NTCIR-10 organizers.

3.4 Results

3.4.1 *15-utterance-based Passage Retrieval*

Above described 15-utterance-based unit is regarded as a document, and for each query, we tried to retrieve 1000 documents.

The results are listed in Figure 6 system 1 to system 5. uMAP, pwMAP, and fMAP of baseline system (system 5) are 0.100, 0.074, and 0.069, respectively. Our system always try to output 1000 of 15-sequential-utterances (total max 15000 utterances), therefore, it is difficult to achieve higher uMAP and fMAP. In our system, retrieved document is a uniformly divided segment, and we did not consider that the center of each segment is relevant to the query. Therefore, it is difficult to achieve higher pwMAP.

Conventional PRF (system 4), which first retrieves 15-utterance-based document and extracts additional words and propopsed PRF (system 3), which first retrives 10-utterance-based document and extracts additional words, achieved slightly higher or comparable performance with baseline system (system 5). As same as the lecture retrieval results, combination of baseline IR results and PRF IR results (system 1 and 2) worked well and improved IR performance. Especially, when we used an index which consists of 10-utterance-based unit for extracting addiional words (system 1), we achieved the highest uMAP, pwMAP, and fMAP.

3.4.2 *10-utterance-based Passage Retrieval*

Next, 10-utterance-based unit is regarded as a document, and for each query, we tried to retrieve 1000 documents.

The results are listed in Figure 6 system 6 to system 8. uMAP, pwMAP, and fMAP of baseline system (system 8) are 0.097, 0.096, and 0.074, respectively. Our system always try to output 1000 of 10-sequential-utterances (total max 10000 utterances), performances of system 8 (baseline) are almost the same with system 5 (15-utterance baseline),

PRF (system 7), which first retrieves 10-utterance-based document and extracts additional words achieved slightly higher or comparable performances than/with baseline system (system 8). Also, combination of baseline IR results and PRF IR results (system 6) worked well, and IR performances are improved.

4. CONCLUSIONS

We participated in NTCIR-10 Spoken Documents ("SpokenDoc") task as a team "RYSDT". In this paper, our SCR systems, which are participated in NTCIR-10 SpokenDoc SCR subtask, were described. Vector space model (VSM) based SCR systems with pseudo relevance feedback (PRF) are evaluated. It is effective preparing a special index to retrieve pseudo relevant document from which additional words are extracted, especially for lecture retrieval task (longer document retrieval task).

5. ACKNOWLEDGMENTS

The work was partially supported by KAKENHI.

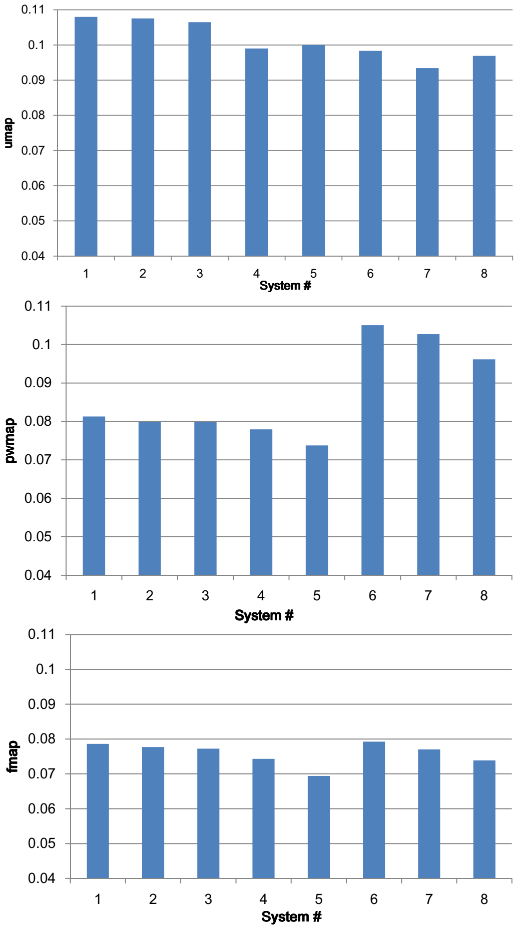

Figure 6: SCR performance for passage retrieval task

6. REFERENCES

[1] Kazuyuki Noritake, Hiroaki Nanjo, and Takehiko Yoshimi. Image processing filters for line detection-based spoken term detection. In the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011), pages 2125–2128, 2011.

[2] Koji Shigeyasu, Hiroaki Nanjo, and Takehiko Yoshimi. A study of indexing units for japanese spoken document retrieval. In 10th Western Pacific Acoustics Conference (WESPAC X), 2009.

[3] Hiroaki Nanjo, Yusuke Iyonaga, and Takehiko Yoshimi. Spoken Document Retrieval for Oral Presentations Integrating Global Do cument Similarities into Local Document Similarities. In Proc. Interspeech (INTERSPEECH 2010), pages 1285–1288, 2010.

[4] Hiroaki Nanjo, Kazuyuki Noritake, and Takehiko Yoshimi. Spoken Document Retrieval Experiments for SpokenDocat Ryukoku University (RYSDT). In NTCIR-9, 2011.

[5] Tomoyosi Akiba, Hiromitsu Nishizaki, Kiyoaki Aikawa, Xinhui Hu, Yoshiaki Itoh, Tatsuya Kawahara, Seiichi Nakagawa, Hiroaki Nanjo, and Yoichi Yamashita. Overview of the NTCIR-10 SpokenDoc-2 Task. In NTCIR-10, 2013.

[6] Tomoyosi Akiba and Koichiro Honda. Effects of Query Expansion for Spoken Document Passage Retrieval. In Proc. Interspeech, pages 2137–2140, 2011.

[7] Young-In Song, Kyoung-Soo Han, So-Young Park, Sang-Bum Kim, and Hae-Chang Rim. Simple Weighting Techniques for Query Expansion in Biomedical Document Retrieval. In IEICE transactions on information and systems, pages 1873–1876, 2007.

[8] Yoojin Chung. Parallel Information Retrieval with Query Expansion. In IEICE transactions on information and systems, pages 1593–1595, 2004.

[9] Hiroko Mano and Yasushi Ogawa. Term selection in automatic query expansion using pseudo-relevance feedback. In IPSJ SIG Notes, pages 121–128, 2001.

[10] Koichi Kise, Markus Junker, Andreas Dengel, and Keinosuke Matsumoto. Effectiveness of passage-based document retrieval for short queries. In IEICE transactions on information and systems, pages 1753–1761, 2003.

[11] Tetstuya Sakai, Yoshimi Saito, Tomoharu Kokubu, Makoto Koyama, and Toshihiko Manabe. High-precision search via question abstraction for japanese question answering. In IPSJ SIG Notes, pages 139–146, 2004.

[12] Amit Singhal, Chris Buckley, and Mandar Mitra. Pivoted Document Length Normalization. In Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 21–29, 1996.

[13] Akihiko Takano, Yoshiki Niwa, Shingo Nishioka, Makoto Iwayama, Toru Hisamitsu, Osamu Imaichi, and Hirofumi Sakurai. Information Access Based on Associative Calculation. In SOFSEM 2000: Theory and Practice of Informatics, Lecture Notes in Computer Science, pages 15–35, 2000.

[14] Tomoyosi Akiba, Kiyoaki Aikawa, Yoshiaki Itoh, Tatsuya Kawahara, Hiroaki Nanjo, Hiromitsu Nishizaki, Norihito Yasuda, Yoichi Yamashita, and Katunobu Itou. Construction of a test collection for spoken document retrieval from lecture audio data. Journal of Information Processing, 17:82–94, 2009.