Effects of Query Expansion for Spoken Document Passage Retrieval

Tomoyosi Akiba[1], *Koichiro Honda*[1]

[1]Department of Computer Science and Engineering
Toyohashi University of Technology
akiba@nlp.cs.tut.ac.jp

Abstract

One of the major challenges for spoken document retrieval is how to handle speech recognition errors within the target documents. Query expansion is promising for this challenge. In this paper, we apply relevance models, a type of query expansion method, for the spoken document passage retrieval task. We adapted the original relevance model for passage retrieval. We also extended it to benefit from massive collections of Web documents for query expansion. Through our experimental evaluation, we found that our relevance model successfully improved the retrieval performance. We also found that using Web documents was effective when the transcription of the target documents had a high word error rate.

Index Terms: spoken document retrieval, query expansion, relevance models

1. Introduction

Traditionally, human beings have used spoken language mainly for communication. However, advances in speech recognition technologies will make it possible to use spoken language, in addition to written language, as a medium for storing and transmitting knowledge. In practice, audio data such as broadcast news, lectures, and podcasts is increasingly available via the Internet. However, these audio data sources are difficult to reuse because efficient searches within them is much more difficult than for textual material. Spoken document retrieval (SDR) [1] is a promising technology for solving these problems.

Because confirming the content of a spoken document requires playing back its audio data, browsing speech data is much more difficult and time-consuming than browsing textual data. Therefore, it is desirable for users to be able to go directly to the relevant portion of the document rather than play the whole document. For this reason, we focus on passage retrieval of spoken documents, which returns variable-length segments in the spoken documents as search results. Previously, we defined the passage retrieval task that searches for fixed-length segments in the spoken documents that have been segmented automatically in advance [2]. In this paper, we redefine the retrieval task as one that searches for variable-length segments within the spoken documents.

The most straightforward method for SDR is simply to use automatic transcriptions of the target spoken documents for indexing and then to apply a text-based document retrieval method. However, recognition errors in the target documents significantly degrade the information retrieval (IR) performance. In particular, because words that are out of vocabulary (OOV) for the recognition dictionary of the large-vocabulary continuous speech recognition (LVCSR) decoder do not appear in the transcribed text, a query constructed from such words will never match any document in the target collection.

To overcome the problem of recognition errors in SDR, both document and query expansion are known to be effective [3, 4]. While document expansion expands each document in the target collection by using the words related to it, query expansion expands the submitted search query. Considering our passage retrieval task, document expansion appears to be not feasible, as it needs to extend vast amounts of passages from all the documents in the collection. Furthermore, considering the spoken document retrieval setting, as the passages include recognition errors, the effect of the expansion would be limited. For these reasons, we decided to apply query expansion to our task.

In this paper, we apply relevance models, a query expansion method, for the spoken document passage retrieval task. We adapted the original relevance model for passage retrieval. We also extended it to benefit from massive collections of Web documents for query expansion. Through our experimental evaluation, we found that our relevance model successfully improved the retrieval performance. We also found that using the Web documents was effective when the transcription of the target documents had a high word error rate (WER).

The remainder of this paper is organized as follows. Section 2 describes our passage retrieval task and the methods introduced for it. Section 3 explains relevance models proposed in [5] and our extensions to them. In Section 4, we evaluate the proposed method by comparison with conventional document retrieval methods. Finally, we conclude and describe future work in Section 5.

2. Passage Retrieval for Spoken Document

2.1. Test Collection

We used the Corpus of Spontaneous Japanese (CSJ) test collection [6] for evaluating our retrieval methods. The target document collection is 2702 lectures selected from the CSJ [7]. This amounts to more than 600 hours of speech, which is comparable to the TREC SDR test collection [1]. Along with the speech data, the manual transcriptions are also included in the CSJ.

The test collection contains 39 queries about information described in part of a lecture. The relevance of such queries is judged against segments of varying length from the lectures, called *passages*. Relevant passages are assigned to one of two classes, "relevant" or "partially relevant", according to their degree of relevance.

Our passage retrieval task in this paper is just the same as the primary task of this test collection, i.e. finding the relevant variable-length passage from the target document collections. We use "relevant" degree throughout this paper.

Copyright © 2011 ISCA

28 – 31 August 2011, Florence, Italy

2.2. Retrieval Methods for Passage Retrieval

The primary retrieval task specified for the test collection is a type of passage retrieval, which differs from a conventional retrieval task where the target unit of retrieval is predefined and fixed, such as an article in a newspaper. Therefore, we extended our retrieval system with the two specific methods designed for passage retrieval.

2.2.1. Using the Neighboring Context to Index the Passage

Passages from the same lecture may be related to each other in the passage retrieval task, whereas the target documents are considered to be independent of each other in a conventional document retrieval task. In particular, the neighboring context of a target passage should contain related information. It would seem appropriate for the passage retrieval task to use the neighboring context to index the target passage [2]. A similar method was applied in TREC SDR TRACK [8].

Normally, a passage D is indexed by its own term frequencies $TF(t, D)$ of the terms $t \in D$. This can be extended to use the neighboring context for indexing. For the context $context_n(D)$, the preceding n utterances and the following n utterances are used. Therefore, we use

$$TF_{ext}(t, D) = \beta TF(t, D) + TF(t, context_n(D)), \quad (1)$$

where β is introduced to specify the relative importance of D and $context_n(D)$.

In our implementation, an utterance is used for D, and n and β are set to 7 and 5 respectively through the preliminary experiments. We refer this method as to *context indexing*.

2.2.2. Penalizing Neighboring Retrieval Results

In applying context indexing, neighboring passages are liable to be retrieved at the same time as they share the same indexing words. This is not adequate from the perspective of retrieval systems because such systems output many redundant results.

For this reason, we penalize a retrieval result that is neighbor to another result that have been output previously. In practice, the retrieved passage is discarded from the output list, if there are other results already retrieved within an n-utterances neighborhood of it.

3. Query Expansion for SDR

3.1. Relevance Models

Levrenko and Croft [5] proposed *relevance models* as an information retrieval model. They define the relevance class R to be the subset of documents in a collection $\mathcal{C}$, which are relevant to some particular information need, i.e. $R \subset \mathcal{C}$. A relevance model is the probability distribution $P(w|R)$, where $w \in V$ is a word in a vocabulary V. $P(w|R)$ is estimated from a given query Q as follows.

$$P(w|R) \quad \approx \quad P(w|Q) = \frac{P(w, Q)}{P(Q)} \quad (2)$$

Suppose that Q consists of a sequence of words $q_1 \cdots q_k$ and that both $q_1 \cdots q_k$ and w are sampled identically and independently from a unigram distribution $P(w|R)$. Assuming a sampling process where a document D is sampled from $\mathcal{C}$ at first, then words are sampled from D, $P(w, Q)$ is obtained as follows.

$$P(w, Q) = \sum_{D \in \mathcal{C}} P(D) P(w, Q|D) \quad (3)$$

Because we assume that w and $q_1 \cdots q_k$ are sampled independently and identically, the joint probability $P(w, Q|D)$ can be expressed as follows:

$$P(w, Q|D) = P(w|D) \prod_{i=1}^{|Q|} P(q_i|D). \quad (4)$$

By substituting equation (4) into equation (3), the following estimate is obtained:

$$P(w, Q) = \sum_{D \in \mathcal{C}} P(D) P(w|D) \prod_{i=1}^{|Q|} P(q_i|D). \quad (5)$$

Suppose that $P(D)$ is distributed uniformly, $P(w|R)$ is estimated as follows:

$$P(w|R) = \frac{1}{P(Q)} \sum_{D \in \mathcal{C}} P(w|D) \prod_{i=1}^{|Q|} P(q_i|D), \quad (6)$$

where $P(Q)$ is constant with respect to Q.

Then, $P(w|R)$ is used to rank the documents $D \subset \mathcal{C}$ by using the Kullback–Leibler divergence between the distributions $P(w|R)$ and $P(w|D)$:

$$H(R||D) = -\sum_{w \in V} P(w|R) log P(w|D). \quad (7)$$

Relevance models can be seen as an implementation of pseudorelevance feedback, which is a sort of query-expansion technique using the target document collection, i.e. the query Q is expanded with the related words in the collection $\mathcal{C}$ through the estimation of the relevance model $P(w|R)$.

3.2. Extending Relevance Models to Context Indexing

Applying relevance models directly to our passage retrieval, specifically the context-indexing method described in Section 2.2.1, is problematic. Because context indexing uses neighboring utterances to index a document (an utterance), several neighboring documents share the same index words. This makes the estimated $P(w|R)$ inaccurate.

In order to deal with this problem, no context-expanded documents, i.e. a set of utterances, are used in the estimation of $P(w|R)$, but then context-expanded documents are ranked using $P(w|R)$. Namely, $P(w|R)$ is estimated as follows:

$$P(w|R) = \sum_{D \in \mathcal{C}} P(w|D_{nc}) \prod_{i=1}^{|Q|} P(q_i|D), \quad (8)$$

where D and $\mathcal{C}$ are an utterance and a set of utterances, respectively. Then, the context-expanded documents $\tilde{D} \subset \tilde{\mathcal{C}}$ are ranked by the following equation:

$$H(R||\tilde{D}) = -\sum_{w \in V} P(w|R) log P(w|\tilde{D}). \quad (9)$$

3.3. Extending Relevance Models using Web

Though relevance models use target documents for query expansion, the world's largest document collection, the World Wide Web, can also be used to enrich its expanded words. Therefore, we enhanced the original relevance models to take advantage of the abundant Web resources.

Firstly, the query Q is submitted to a Web search engine to get a set of Web documents $\mathcal{C}_{web}$[1]. From the documents $\mathcal{C}_{web}$, the Web-based relevance model is estimated as follows:

$$P(w|R_{web}) = \sum_{D \in \mathcal{C}_{web}} P(w|D) \prod_{i=1}^{|Q|} P(q_i|D). \quad (10)$$

Next, the original relevance model, $P(w|R_{orig})$, and the Web-based relevance model $P(w|R_{web})$ are integrated into a new single relevance model $P(w|R)$. We tried two integration methods as follows.

Linear interpolation: the two models are linearly interpolated:

$$P(w|R) = (1-\gamma)P(w|R_{orig}) + \gamma P(w|R_{web}). \quad (11)$$

Document weighting: the Web model is used to weight the target documents:

$$P(w|R) = \sum_{D \in \mathcal{C}} P(w|D) \prod_{q \in R_{web}} \sqrt{P(q|D)P(q|R_{web})}. \quad (12)$$

Finally, the new model $P(w|R)$ is used to rank the (context extended) documents by using equation (7).

4. Experimental Evaluation

4.1. Transcription

The target spoken documents are transcribed by using LVCSR software. Two types of transcriptions are used for comparison. For both systems, the language model and the acoustic model are trained by using the target documents of the SDR system. The difference is whether the training is conducted subject to a closed condition or an open condition (referred as to CLOSED and OPEN, respectively). The WERs for these transcriptions are 21.4% (CLOSED) and 30.8% (OPEN). The details of the transcription processes are described in [9] and [10], respectively.

4.2. Evaluation Metric

Our bound-free passage retrieval needs two tasks to be achieved; one is to determine the boundary of the passages to be retrieved and the other is to rank the relevancy of the passages. To focus only on the latter task, we adopt the following evaluation metric.

For a given query, a system returns an ordered list of passages in decreasing order of confidence. For each returned passage, only utterances located in the center of it are considered for relevancy. If an utterance is included in some relevant passage described in the golden file, basically the returned passage is deemed relevant with respect to the relevant passage and the relevant passage is considered to be retrieved correctly. However, if there exists at least one formerly listed passage that is also deemed relevant with respect to the same relevant passage, the returned passage is deemed not relevant as the relevant passage has been retrieved already. In this way, all the passages in the returned list are labeled by their relevancy. Now, any conventional evaluation metric designed for document retrieval can be applied to the returned list.

[1]Documents returned through the Yahoo API are used as our Web document collection.

We used 11-point average precision (AP) as our evaluation metric, which is obtained by averaging precisions as follows:

$$IP(x) = \max_{x \le R_i} P_i, \quad AP = \frac{1}{11} \sum_{i=0}^{10} IP(\frac{i}{10}),$$

where R_i and P_i are the recall and the precision, respectively, up to the i-th retrieved document. In practice, we retrieved 1000 passages for each query in calculating the AP.

4.3. Compared Methods

The proposed query expansion methods were compared to two baseline retrieval methods that do not expand queries; the vector space model and the query likelihood model.

4.3.1. Vector Space Model

The traditional vector space model with TF–IDF term weighting was used as the retrieval method with pivoted normalization.

4.3.2. Query Likelihood Model

Recently, the effectiveness of the language modeling approach for IR has been reported [11]. For document re-ranking, we use the probability $P(Q|D)$ that a query Q is constructed from a relevant document D:

$$P(Q|D) = \prod_{q \in Q} P(q|D). \quad (13)$$

$P(q|D)$ is estimated by

$$P(q|D) = (1-\gamma)\frac{TF(q,D)}{\sum_t TF(t,D)} + \gamma \frac{TF(q)}{\sum_t TF(t)}, \quad (14)$$

where $TF(q)$ is the global term frequency of a query term q calculated from the target document collection C by

$$TF(q) = \sum_{D \in C} TF(q,D). \quad (15)$$

The $P(Q|D)$ is used to rank the document $D \in \mathcal{C}$. In this paper, the context-expanded document $\tilde{D} \in \tilde{\mathcal{C}}$ is used instead of D.

4.4. Results

4.4.1. Comparison of Retrieval Models

The retrieval performances of the retrieval models compared with manual transcription of the target spoken documents are shown in Table 1. The baseline indexing methods index just the utterance, which corresponds to the BASE column in the table. The results show that the two language modeling retrieval models outperform the traditional vector space model with TF–IDF term weighting. It also shows that the retrieval model using query expansion (relevance model) outperforms the model without it (query likelihood model). We can find similar results in the case of retrieval against the automatic transcription in Table 2.

4.4.2. Effects of Passage Retrieval Methods

In Section 2, we introduced the retrieval methods for passage retrieval, i.e. context indexing (Section 2.2.1) and neighborhood penalty (Section 2.2.2). The two methods are incrementally applied in this order to the BASE indexing method. The results are

Table 1: Retrieval performance compared with manual transcription. (CI: context indexing, NP: neighborhood penalty)

retrieval model	BASE	+CI	+NP
vector space model	0.144	0.137	0.162
query likelihood model	0.161	0.126	0.176
relevance model	0.170	0.166	**0.183**

Table 2: Retrieval performance compared with automatic transcription (CLOSED). (CI: context indexing, NP: neighborhood penalty)

retrieval model	BASE	+CI	+NP
vector space model	0.126	0.122	0.148
query likelihood model	0.140	0.114	0.156
relevance model	0.143	0.133	**0.167**

shown at the column labeled +CI (applying only context indexing) and +PI (applying both context indexing and neighborhood penalty) in Figures 1 and 2.

The results show that applying only context indexing decreases performance. This is because context indexing favors outputting neighboring passages at the same time, which results in decreasing the retrieval performance. However, the results also show that applying both context indexing and the neighborhood penalty at the same time successfully overcomes the harmful influence resulting in the method outperforming the BASE method. These results are consistent among the compared retrieval methods and among manual and automatic transcriptions used as target documents.

4.4.3. Effects of Web Document Expansion

Finally, we evaluated query expansion using Web documents as described in Section 3.3. We applied the two proposed methods, *linear interpolation* and *document weighting*, individually to the relevance model. Here, we also conducted the retrieval experiment against automatic transcription obtained under the OPEN condition. Both context indexing and neighborhood penalty were applied for all the compared methods. Table 3 shows the results.

The results show that the Web expansion does not successfully improve the retrieval performance against manual transcription and automatic transcription obtained under the CLOSED condition. We examined the results and found that the collected Web documents are noisy. The documents returned through a search engine API are diverse and include those that do not relate to the relevant passage in the target collection to be retrieved. Those irrelevant documents seem to disturb the finding of the relevant passage. On the other hand, the results also show that the performance against automatic transcription obtained under the OPEN condition is improved by using the linear interpolation method of the Web expansion. We think this is because the transcription includes more recognition errors, so Web expansion helps to correct the misrecognized words by using related words collected through Web retrieval.

5. Conclusion

In this paper, we applied query expansion, namely using relevance models, for the spoken document passage retrieval task. We adapted the original relevance models for passage retrieval. We also extended it to take advantage of the massive collection of Web documents for query expansion. Through our experimental evaluation, we found that the relevance model could successfully improve retrieval performance. We also found that using Web documents was effective when the transcription of the target documents had a high WER.

Table 3: Effect of the web documents

retrieval model	MANUAL	CLOSED	OPEN
vector space model	0.162	0.148	0.121
query likelihood model	0.176	0.156	0.120
relevance model	**0.183**	**0.167**	0.136
web (linear interpolation)	0.179	0.164	**0.139**
web (document weighting)	0.181	0.165	0.113

In order to improve the performance of our Web extension of relevance models, filtering for noisy Web documents might be necessary. In future work, we will apply Web document-filtering methods to select only the documents most related to the target documents.

6. References

[1] J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees, "The TREC spoken document retrieval track: A success story," in *Proceedings of TREC-9*, 1999, pp. 107–129.

[2] K. Honda and T. Akiba, "Language modeling approach for retrieving passages in lecture audio data," in *Proceedings of International Conference on Language Resources and Evaluation*, 2010, pp. 1525–1530.

[3] M. Terao, T. Koshinaka, S. Ando, R. Isotani, and A. Okumura, "Open-vocabulary spoken-document retrieval based on query expansion using related web documents," in *Proceedings of International Conference on Speech Communication and Technology*, 2008, pp. 2171–2174.

[4] K. Sugimoto, H. Nishizaki, and Y. Sekiguchi, "Effect of document expansion using web documents for spoken documents retrieval," in *Proceedings of the 2nd Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2010, pp. 526–529.

[5] V. Lavrenko and W. B. Croft, "Relevance models in information retrieval," in *Language Modeling for Infromation Retrieval*, W. B. Croft and J. Lafferty, Eds. Kluwer Academic Publishers, 2003, pp. 11–56.

[6] T. Akiba, K. Aikawa, Y. Itoh, T. Kawahara, H. Nanjo, H. Nishizaki, N. Yasuda, Y. Yamashita, and K. Itou, "Developing an sdr test collection from japanese lecture audio data," in *Proceedings of the 1st Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2009, pp. 324–330.

[7] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese," in *Proceedings of International Conference on Language Resources and Evaluation*, 2000, pp. 947–952.

[8] S. Johnson, P. Jourlin, K. Jones, and P. Woodland, "Spoken document retrieval for TREC-9 at cambridge university," in *Proceedings of TREC-9*, 1999.

[9] T. Akiba, K. Aikawa, Y. Itoh, T. Kawahara, H. Nanjo, H. Nishizaki, N. Yasuda, Y. Yamashita, and K. Itou, "Construction of a test collection for spoken document retrieval from lecture audio data," *Journal of Information Society of Japan*, vol. 50, no. 2, pp. 501–513, 2009.

[10] Y. Itoh, H. Nishizaki, X. Hu, H. Nanjo, T. Akiba, T. Kawahara, S. Nakagawa, T. Matsui, Y. Yamashita, and K. Aikawa, "Constructing japanese test collections for spoken term detection," in *Proceedings of International Conference on Speech Communication and Technology*, 2010, pp. 667–680.

[11] W. B. Croft and J. Lafferty, Eds., *Language Modeling for Information Retrieval*. Kluwer Academic Publishers, 2003.