Language model expansion using webdata for spoken document retrieval

Ryo Masumura, Seongjun Hahm, Akinori Ito

Graduate School of Engineering, Tohoku University
{ryo77373,branden65,aito}@spcom.ecei.tohoku.ac.jp

Abstract

In recent years, there has been increasing demand for ad hoc retrieval of spoken documents. We can use existing text retrieval methods by transcribing spoken documents into text data using a Large Vocabulary Continuous Speech Recognizer (LVCSR). However, retrieval performance is severely deteriorated by recognition errors and out-of-vocabulary (OOV) words. To solve these problems, we previously proposed an expansion method that compensates the transcription by using text data downloaded from the Web. In this paper, we introduce two improvements to the existing document expansion framework. First, we use a large-scale sample database of webdata as the source of relevant documents, thus avoiding the bias introduced by choosing keywords in the existing methods. Next, we use a document retrieval method based on a statistical language model (SLM), which is a popular framework in information retrieval, and also propose a new smoothing method considering recognition errors and missing keywords. Retrieval experiments show that the proposed methods yield a good results.
Index Terms: Spoken document retrieval, statistical language models, World Wide Web

1. Introduction

With the development of information and communication technology, we can now access huge amounts of multimedia contents including recorded audio and video. However, is is difficult to perform content-based searches of such data compared with text data. Most search engines provide a function for searching data based on metadata such as titles, tags or text data surrounding the multimedia contents. Large vocabulary continuous speech recognition (LVCSR) is one of the most promising technologies for content-based searches of multimedia contents including human speech. Using LVCSR, we can convert speech into text and search the speech-based content using text-based search techniques.

Recent text retrieval methods are based on the statistical language model (SLM) [1]. These methods use a mathematical framework rather than heuristics such as tf-idf, and have been proved to be more accurate than the classical heuristic retrieval methods.

However, there are a couple of problems when applying text retrieval methods to transcriptions produced by LVCSR. The first problem is recognition errors: automatic transcriptions generated by a speech recognizer contain many recognition errors, and so important words are missing when searching those documents. The other problem is out-of-vocabulary (OOV) words, which are words not included in the dictionary of the speech recognizer. As the recognizer cannot recognize OOV words, those words appearing in the spoken document inevitably become recognition errors. As a result of these two problems, the accuracy of document retrieval for spoken documents using LVCSR is much lower than that for written documents; for research in the field of spoken document retrieval, these problems need to be solved.

Many attempts have been made to solve these problems, such as by using multiple recognition hypotheses [2], topic modeling [3], and document clustering techniques [4]. Although these methods can solve the recognition error problem, they cannot solve the OOV problem because they improve the recognition result within the vocabulary of the speech recognizer.

To deal with the OOV problem, we are developing a method that acquires new words from the World Wide Web [5]. This approach first extracts keywords from the automatic transcription, and then retrieves Web documents using the extracted keywords. The downloaded documents are used for compensating the index generated from the automatic transcription. A similar approach has also been proposed by Sugimoto et al. [6].

Two problems of these Web-based approaches can be pointed out. The first one is that these methods use only a few keywords as representatives of the spoken document, yet it is difficult to express the features of a document using only a few keywords. The second problem is that these works use the classical retrieval method based on a vector space model. It is desirable to use the state-of-the-art document retrieval method based on the statistical language model for exploiting the advances in information retrieval technology.

In this paper, we propose a spoken document retrieval method based on document expansion using documents downloaded from the Web. There are two novel points in this work. First, we create a database downloaded from the Web so that the database contains as many kinds of words as possible, thus increasing the possibility of acquiring OOV words, and the whole transcription of a spoken document is used for choosing data from the database for document expansion. Second, we use a document retrieval framework based on the SLM. We not only apply the existing method for text retrieval, but also propose a new extension of the SLM so that spoken documents can be retrieved with high accuracy.

This paper is organized as follows. Section 2 briefly describes information retrieval based on statistical language models and the problems of using automatic transcriptions including recognition errors. In Section 3, we propose a model expansion method using webdata to solve the problems of the language modeling approach. In Section 4, we carry out a retrieval experiment to verify the effectiveness of the proposed method.

2. Information retrieval based on statistical language models

2.1. Query likelihood model

Information retrieval using statistical language models can be achieved by obtaining a conditional probability $P(D|Q)$, a probability that a document D is generated from the same information source as a query Q. $P(D|Q)$ is calculated using Eq. (1) by Bayes' theorem.

$$P(D|Q) = \frac{P(Q|D)P(D)}{P(Q)} \propto P(Q|D) \tag{1}$$

Here, $P(Q)$ is independent of a document D, and $P(D)$ can be considered to be constant. $P(Q|D)$ is called a query likelihood model [1].

Generally, a multinomial distribution is used as a language model based on a document D [7]. A language model based on multinomial distribution is equivalent to a unigram, which assumes that each word is generated from an information source independently. A query likelihood $P(Q|\theta_D)$ can be written as Eq. (2). Terms on the right hand side of the equation that do not affect the retrieval are omitted.

$$P(Q|\theta_D) \propto \prod_{w \in V} P(w|\theta_D)^{c(w,Q)} \tag{2}$$

Here, θ_D is the parameter set of a language model (document model) that generates document D, $P(w|\theta_D)$ is the production probability of word w, and $c(w, Q)$ is the frequency of word w in query Q. V is the vocabulary that contains all words appearing in all documents. Information retrieval based on SLM is achieved by ranking documents according to the value of $P(Q|\theta_D)$.

When estimating $P(w_i|\theta_D)$ using maximum likelihood (ML) estimation,

$$P_{\mathrm{ML}}(w|\theta_D) = \frac{c(w, D)}{|D|} \tag{3}$$

where

$$|D| = \sum_{w \in V} c(w, D). \tag{4}$$

In this case, when a query Q contains one or more words that do not appear in document D, $P(Q|\theta_D)$ becomes zero. To avoid this zero frequency problem, the probability is smoothed as follows using relative frequency over the entire collection of documents (*Dirichlet smoothing*).

$$P_{\mathrm{MAP}}(w|\theta_D) = \frac{c(w, D) + \mu P(w|\theta_C)}{|D| + \mu} \tag{5}$$

Here, θ_C is a language model of the entire collection of documents C, which is called the collection model. The smoothing parameter μ can be estimated using the leave-one-out likelihood [7].

2.2. Problems of using SLM for spoken document retrieval

In spoken document retrieval, the document model θ_D is estimated from an automatic transcription generated by a speech recognizer. Here, we should consider the following two problems.

The first problem is that the document collection does not necessarily contain all words that appear in the real documents. As explained above, the document collection is a set of automatic transcriptions obtained from a speech recognizer, and contains recognition errors and OOV problems. Therefore, some of the words that appear in the spoken documents are missing in the document collection. To assign a probability to a word that does not appear in the collection, we should perform smoothing also for the probability of the document collection.

The second problem is that some of the important words in the target spoken document are missing. Even if zero probability problems can be avoided by the above smoothing, the retrieval performance deteriorates if probabilities assigned to important words are small. To solve this problem, we need to predict words that are likely to appear in the target document and assign higher probabilities to those words.

3. Model expansion using webdata

3.1. Probability smoothing using global model

When estimating a document model θ_D using Dirichlet smoothing using Eq. (5), the collection model θ_C is usually estimated by ML estimation.

$$P_{\mathrm{ML}}(w|\theta_C) = \frac{c(w, C)}{|C|} \tag{6}$$

Here, $c(w, C)$ is the frequency of word w in document collection C, and $|C|$ is the total occurrence of words in documents collection C. However, as explained above, we need to smooth $P(w|\theta_C)^{ML}$ so that those words that do not appear in the document collection have non-zero probabilities. Therefore, we introduce a prior distribution into the collection model using the Dirichlet smoothing framework. The prior distribution used in this smoothing should be a general distribution estimated from large document collection, therefore we use word frequency distribution in the Web as the prior distribution.

We define a prior distribution of the collection model as a global model θ_G. The global model can be estimated by counting the frequency of all words in all documents in the Web, but this is not realistic. Therefore, we use the hit count of a word (approximate number of Web documents that contain that word) obtained from a Web search engine instead of the word count. Let the Web hit count of word w be $h(w)$. The global model θ_G can be estimated by obtaining $h(w)$ of words in the large vocabulary W ($V \subset W$).

$$P_{\mathrm{ML}}(w|\theta_G) = \frac{h(w)}{\sum_{w' \in W} h(w')} \tag{7}$$

We used dictionaries of morphemic analyzers for composing W (see section 4.2).

The collection model is estimated by introducing $P(w|\theta_G)$ as a prior distribution:

$$P_{\mathrm{MAP}}(w|\theta_C) = \frac{c(w, C) + \eta P_{\mathrm{ML}}(w|\theta_G)}{|C| + \eta} \tag{8}$$

This smoothing parameter η can be estimated by using the leave-one-out likelihood. The collection model is used as the prior distribution of Eq. (5). Using the global model as the prior distribution of the collection model, we can assign nonzero probabilities for all words in W even if a word $w \in W$ is an OOV word (i.e. $w \notin V$).

3.2. Model expansion using a relevant document model

Next, we predict words that are likely to appear in the target spoken document and compensate the query likelihood using the probability of the predicted words. To predict those words, we exploit webdata relevant to the automatic transcription of the spoken document. If we can choose documents relevant to the target spoken document from the Web, we can expect those relevant documents to contain words that appears in the spoken document and are not included in the automatic transcription.

In previous works, the relevant data were downloaded through Web searches using keywords extracted from the transcription of the target spoken document [5, 6]. However, this approach uses only a few keywords to express the topic of the transcription. Considering recognition errors of important topic words in the transcription, we should consider all words appearing in the transcription when choosing relevant documents. Therefore, we propose a new method to determine relevant documents using all words in the transcription.

Determination of relevant documents is carried out in the following three steps. First, we use a Web search engine to gather a large collection of Web documents from which relevant documents are selected. A list of nouns that could be keywords is prepared first, then each of the nouns is used as a keyword for a Web search, and the first n documents are downloaded from the Web. The documents downloaded using one keyword are combined into one large document and associated with that keyword. We denote the downloaded documents as $S = \{s_w | w \in W\}$. Second, similarities between the transcription and the downloaded documents are calculated. The similarity between transcription D and downloaded document $s \in S$ is based on KL divergences.

$$Sim(\theta_D||\theta_s) = \sum_{w \in D} P_{\mathrm{ML}}(w|\theta_D) \log P_{\mathrm{MAP}}(w_i|\theta_s) \quad (9)$$

Finally, we choose N documents among the downloaded documents that have the highest $Sim(\theta_D||\theta_s)$, and combine the selected documents to create a large relevant document R and its language model θ_R.

After determining the relevant document R, the language model of R is calculated. We apply Dirichlet smoothing for calculateing the language model just like the language model for transcriptions.

$$P_{\mathrm{MAP}}(w|\theta_R) = \frac{c(w_i, R) + \mu P_{\mathrm{MAP}}(w_i|\theta_C)}{|R| + \mu} \quad (10)$$

Then the document model and the relevant document model are linearly interpolated [8, 9] and the mixture model is generated.

$$P_{\mathrm{rel}}(w|\theta_D) = \lambda P_{\mathrm{MAP}}(w|\theta_D) + (1 - \lambda) P_{\mathrm{MAP}}(w|\theta_R) \quad (11)$$

4. Retrieval experiments

4.1. Test collection

We use the CSJ test collection [10] as a test set of the retrieval. This test collection consists of the target spoken documents of 2,702 lectures (which are part of the Corpus of Spontaneous Japanese [11]) and 39 retrieval queries along with sets of ID numbers of the relevant documents.

We recognized those 2,702 lectures using two N-gram language models. The training data for the N-grams were 3,302 lectures of the CSJ that includes the test collection. We prepared two vocabularies for training N-grams: the first one was the most frequent 50,000 words, and the second one was anothre 50,000 words excluding words that appear in the queries to simulate OOV words in queries. Then we trained two N-gram language models using the two vocabularies respectively. We denote the automatic transcriptions recognized by the first language model as $C1$, and those by the second one as $C2$. The manual transcriptions of the lectures are denoted as $C0$. Details of the three transcriptions are shown in Table 1. We used 11-point average precision (11pt AP) as an evaluation metric [10].

Table 1: Details of the transcriptions

	$C0$	$C1$	$C2$
Word accuracy (%)	100.00	75.12	69.01
OOV rate (%)	0	0.23	0.52
No. of queries with OOV	0/39	0/39	39/39

Table 2: Retrieval performance using document models

	Collection	*Global*	*Collection* + *Global*
$C0$	0.4823	0.4499	0.4976
$C1$	0.3767	0.3604	0.4012
$C2$	0	0.0062	0.0062

4.2. Training of language models for retrieval

We exploited only noun words (excluding stop words) for language models for document retrieval. The vocabulary of the documents (denoted as V) was the same as the nouns in the vocabulary of the speech recognizer. The vocabulary of the global model (denoted as W) consisted of 288k nouns included in IPAdic [12] and unidic [13], dictionaries for Japanese morphemic analyzer. The numbers of Web documents required for training the global model were obtained from Yahoo! Japan.

When gathering the downloaded documents from the Web, we collected 50 Web documents from the retrieval result of one keyword, and all 288k nouns were used as keywords. Finally, we downloaded 1.44M documents containing 14G words.

4.3. Effect of the global model

We carried out an experiment using three models to investigate the effect of smoothing the collection model using the global model. The difference of the three models was the prior distribution in Eq. (5), as follows.

1. The *Collection* model: $P(w|\theta_C) = P_{\mathrm{ML}}(w|\theta_C)$.
2. The *Global* model: $P(w|\theta_C) = P_{\mathrm{ML}}(w|\theta_G)$.
3. The *Collection*+*Global* model (the proposed method): $P(w|\theta_C) = P_{\mathrm{MAP}}(w|\theta_C)$.

The experimental results are shown in Table 2. These results show that the *Global* model was not as effective as the *Collection* model, but the precision could be improved by combining the two models (*Collection*+*Global*). In $C2$, the *Collection* model could not retrieve anything at all because all nouns in the queries were OOV words. Using the *Global* model, we could present some results even though all the query words were OOV words.

4.4. Effect of relevant document models

We carried out experiments to investigate the effect of mixture modeling using relevant Web documents. In these experiments,

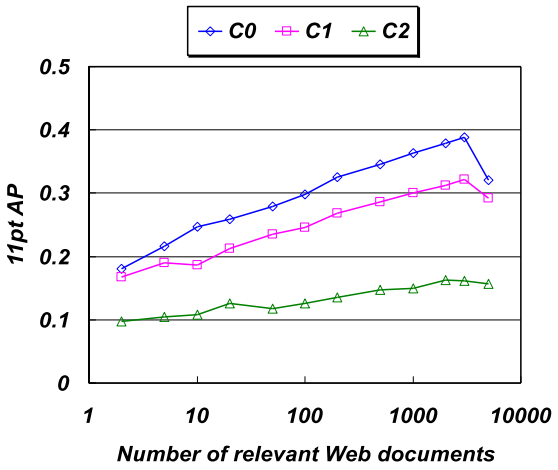

Figure 1: Retrieval performance using relevant document models

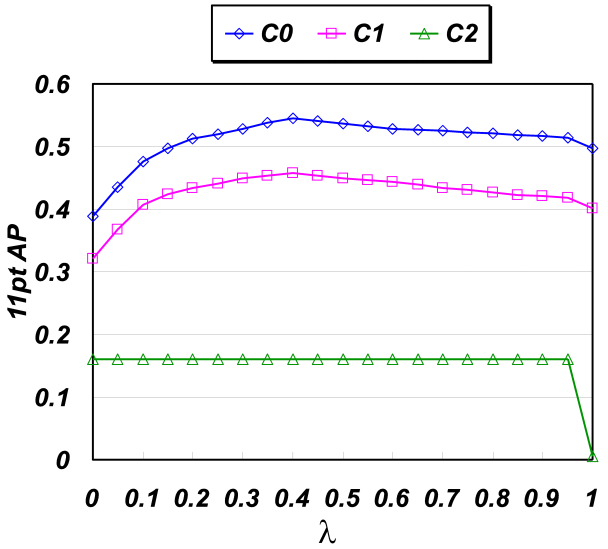

Figure 2: Retrieval performance using mixture models

Eqs. (5) and (8) were used for smoothing the document model and the relevant document model.

In the first experiment, we used only the relevant document model for retrieval (i.e $\lambda = 0$ in Eq. (11)). We changed the number of relevant documents N and observed the precision of retrieval. Figure 1 shows the experimental result. This result suggests that we can retrieve spoken documents using *only* downloaded documents. The maximum 11pt AP was obtained when using 3000 documents.

Next, we investigated the effect of mixing two models according to Eq. (11). The number of relevant documents for creating the relevant model was fixed to 3000. Figure 2 shows the precision with respect to the value of λ. In this graph, $\lambda = 0$ is the result using only the relevant document model and $\lambda = 1$ is that using only the document model.

These results confirm that the precision can be improved by combining the two models. The improvement seemed to be achieved by adding higher probabilities to the relevant words to which the document model gave a low probability. The maximum improvements for $C0$, $C1$ and $C2$ were 0.0468, 0.0564 and 0.1549, respectively, suggesting that the proposed method is more effective for a document collection with higher OOV rate.

5. Conclusions

In this paper, we proposed a model expansion method using webdata to improve retrieval performance for spoken document retrieval based on statistical language models. First, we proposed a new smoothing method using global model based on word frequency distribution of webdata to assign probabilities to missing words in the document collection. Next, we proposed a mixture modeling using a relevant document model based on Web documents relevant to the target spoken document. Using the proposed method, we can assign a larger probability to important words in the target spoken document even if those words do not appear in the automatic transcription. The retrieval experiments showed that the proposed method improved the precision of retrieval.

6. References

[1] J.M.Ponte and W.B.Croft, "A language modeling approach to information retrieval ", In Proc. SIGIR 1998, pp.275-281, 1998.

[2] T.K.Chia, H.Li and H.T.Ng, "A Statistic Language Modeling Approach to Lattice-Based Spoken Document Retrieval", In Proc. Joint Meeting of the Conference on Empirical Methods in Natural Language Processing and the Conference on Natural Language Learning, pp.810-818, 2007.

[3] B.Chen, "Latent Topic Modeling of Word Co-Occurrence Information for Spoken Document Retrieval," In Proc ICASSP, pp.3961-3964, 2009.

[4] X.Hu, R.Isotani, H.Kawai and S.Nakamura, "Cluster-based Language Model for Spoken Document Retrieval Using NMF-Based Document Clustering," In Proc. InterspeechCpp705-708C2010D

[5] R.Masumura, A.Ito, Y.Uno, M.Ito and S.Makino, "Document Expansion using Relevant Web Documents for Spoken Document Retrieval ," In Proc. Int. Conf. on Natural Language Processing and Knowledge Engineering, pp.612-619, 2010.

[6] K.Sugimoto, H.Nishizaki and Y.Sekiguchi, "Effect of Document Expansion using Web Documents for Spoken Documents Retrieval ," In. Proc. APSIPA ASC 2010, pp.526-529, 2010.

[7] C.Zhai and J.Lafferty, "A study of smoothing methods for language models applied to information retrieval," ACM TOIS, vol.22, no.2, pp.179-214, 2004.

[8] X.Wei and W.B.Croft, "LDA-based document models for ad-hoc retrieval ," In Proc. SIGIR 2006, pp.178-185, 2006.

[9] D. Zhou, J. Bian, S. Zheng, H. Zha and C. L. Giles, "Exploring social annotations for information retrieval," Proc. Int. Conf. on WWW, 2008.

[10] T.Akiba, K.Aikawa, Y.Ito, T.Kawahara, H.Nanjo, H.Nishizaki, N.Yasuda, Y.Yamashita and K.Ito, "Test collections for spoken document retrieval from lecture audio data ," In Proc. LREC, 2008.

[11] K. Maekawa, H. Koiso, S. Furui and H. Isahara, "Spontaneous speech corpus of Japanese," Proc. LREC, pp. 947–952, 2000.

[12] M. Asahara and Y. Matsumoto, "IPADIC User Manual," Nara Institute of Science and Technology, Japan, 2002.

[13] Y. Den, J. Nakamura, T. Ogiso and H. Ogura, "A proper approach to Japanese morphological analysis: Dictionary, model, and evaluation," Proc. LREC, 2008.