interspeech 2008 incorporating SST 08

Open-Vocabulary Spoken-Document Retrieval Based on Query Expansion Using Related Web Documents

Makoto Terao, Takafumi Koshinaka, Shinichi Ando, Ryosuke Isotani and Akitoshi Okumura

Common Platform Software Research Laboratories, NEC Corporation, Japan

m-terao@cd.jp.nec.com

Abstract

This paper proposes a new method for open-vocabulary spoken-document retrieval based on query expansion using related Web documents. A large vocabulary continuous speech recognition (LVCSR) system first transcribes spoken documents into word sequences, which are then segmented into semantically cohesive units (i.e., stories) using a text segmentation technique. Given a text query word, Web documents containing the query word are first retrieved. Each retrieved Web document can be regarded as an expanded form of the original query word. Spoken documents relevant to the query word are then retrieved by searching for the stories with the LVCSR result similar to the previously obtained Web documents. Experimental results show that the proposed method is quite effective in retrieving spoken documents such as broadcast news programs with out-of-vocabulary (OOV) queries. In addition, the proposed method is also useful for ranking retrieval results with in-vocabulary (IV) queries.

Index Terms: Spoken document retrieval, Large vocabulary continuous speech recognition, Query expansion, Web documents, Text segmentation

10.21437/Interspeech.2008-568

1. Introduction

A large number of spoken documents such as broadcast programs are recently available in digital form, which has created a strong need for efficient access to their archives. A simple approach to spoken document retrieval is to prepare a full transcription into word sequences of spoken documents using a large vocabulary continuous speech recognition (LVCSR) system, and to search LVCSR results for a given text query word. Unfortunately, however, LVCSR results usually contain recognition errors, and such errors degrade retrieval performance. This problem is more significant if the query word is out-of-vocabulary (OOV) in the LVCSR system since OOV words never appear in LVCSR results and retrieval with an OOV query word inevitably results in failure. In this paper we propose a sophisticated open-vocabulary spoken-document retrieval method which is highly robust with respect to OOV queries.

Most conventional methods for open-vocabulary spoken document retrieval utilize subword-based recognition results, such as phoneme sequences. A given query word is transformed into a subword sequence, following which it is matched with subword-based recognition results in order to obtain retrieval results.

Iwata et al. have shown that using statistical distances between subword models in the matching process results in good performance [1]. Nishizaki et al. have proposed a method for using subword-based indices for OOV queries and word-based indices for in-vocabulary (IV) queries, in order to avoid the drawback that subword-based retrieving generally performs less well than word-based in IV queries [2]. Hori et al. have applied confusion networks to the representation of multiple hypotheses for subword-based and word-based recognition results in order to achieve robust matching for both OOV and IV queries with compact indices [3]. Since these subword-based methods do not depend on a specific word dictionary, as does an LVCSR system, any word can be used as a query word.

Because these methods utilize the acoustic features of spoken documents, however, irrelevant speech segments will be retrieved if the pronunciation of words in those segments is too similar to that of the query word, even if those segments are semantically quite different from the query. For example, a spoken word “hair” may be incorrectly retrieved for a query word “hare”. As a result, the precision rate of the retrieval tends to be insufficient. In addition, retrieval performance of such methods using acoustic features often degrades if there is background noise in speech signals.

This paper proposes a new method for open-vocabulary spoken-document retrieval based on query expansion using external knowledge, such as Web documents. The proposed method utilizes linguistic features that can be obtained from the LVCSR system for making open-vocabulary spoken document retrieval possible. Section 2 below presents query expansion using Web documents as the basic idea behind the proposed method, and Section 3 describes the method’s overall framework. Finally, Section 4 presents experimental results.

2. Query expansion using web documents

The basic idea behind the proposed method is the introduction of query expansion using Web documents.

Although LVCSR cannot produce OOV words, it has a great advantage in spoken document retrieval over subword-based recognition because natural language processing techniques can be applied to LVCSR results.

More specifically, correctly recognized in-vocabulary words can be useful for retrieval with an OOV query word. If a given text query word is expanded in the form of context words, defined as words that often co-occur with the query word, spoken documents relevant to the query word can be correctly retrieved by searching LVCSR results for these context words even if the query word is OOV in the LVCSR

Accepted after peer review of full paper

system. This is because many of context words for the query word will appear in spoken documents relevant to the query word and many of them are considered to be correctly recognized.

In this study, Web documents are introduced to obtain such context words for the query word. We limit Web documents to those from sites, such as news sites, within which each Web document may be assumed to be devoted to a single topic. Each Web document containing the original query word will be then regarded as an expanded form of that query word.

Figure 1 shows an example of the LVCSR results for a news story about a robot named "PaPeRo." "PaPeRo" is OOV word and has been misrecognized as "popular". For that reason, it would be impossible to retrieve this news story with a simple text query "PaPeRo." Figure 2 shows an example of a Web document containing the query word "PaPeRo." This Web document is devoted to the single topic of "PaPeRo," and words it contains, such as "robot," "speech recognition," "microphone," etc., are considered to be context words for the query word "PaPeRo." These same words also can be found in the LVCSR results for the news story about "PaPeRo," as may be seen in Figure 1. That is, searching LVCSR results using these context words, the news story about "PaPeRo" can be retrieved for an original OOV query word "PaPeRo." Note that although LVCSR results often contain some errors, there will ordinarily be enough correctly recognized words to serve as context words for subsequent queries.

We have developed personal *robot* **popular** …
… *speech recognition* … *image recognition* …
… *microphone*s … … *voice* detection …
… *robot* … … *face* recognition …

Figure 1: Example of an LVCSR result
for a news story about "PaPeRo."
(The word "PaPeRo" is misrecognized as "popular."
Underlined words are correctly recognized.)

NEC has developed personal *robot* **PaPeRo** …
The latest *speech recognition* technology enables …
Three *microphones* are used so that the *robot* can detect *voices*, … **PaPeRo** also has the ability to recognize peoples' *face*s with *image recognition* ...

Figure 2: Example of a Web document
containing a query word "PaPeRo."
(Underlined words are considered to be context words
for "PaPeRo.")

3. Framework of the proposed method

Figure 3 shows the overall framework for the proposed method, using the example of retrieving a news story about "PaPeRo" with an original OOV query word "PaPeRo." The framework is separated into two parts: indexing and retrieving. A description of each is provided below.

3.1. Indexing

Many spoken documents, such as broadcast news programs, contain a variety of stories. The role of the indexing part is to segment, in advance of the retrieving part, such spoken documents into separate stories with an LVCSR result, each of which story deals with a single topic. For example, if the spoken document is a broadcast news program, it would be segmented into separate news stories, as may be seen in Figure 3. The stories obtained here become units for subsequent retrieving.

Indexing is composed of a LVCSR step and a text segmentation step. The LVCSR system first transcribes the spoken documents into word sequences, which are then segmented into semantically cohesive units (i.e., stories) using a text segmentation technique. While there are a number of different methods for text segmentation, text segmentation itself is not the focus of this paper, and we do not go into them in detail here.

3.2. Retrieving

The role of the retrieving part is to retrieve the segments of spoken documents relevant to a given text query word by searching through the stories obtained in the indexing part.

Retrieving is composed of a query expansion step and a search step. Given a query word, Web documents containing the query word are first retrieved in the query expansion step. Since each retrieved Web document is regarded as an expanded form of the given query word, each Web document should deal with only one topic. Online news sites are a typical example of a good source for Web documents that meet this requirement.

The stories with the LVCSR result similar to the obtained Web documents are then searched for in the search step to retrieve spoken documents relevant to the query word. Since all stories and Web documents are likely to deal with a single topic, this search works well.

Let sim(w,s) be the similarity between a certain Web document w and a certain story s. sim(w,s) is calculated as a cosine similarity between the document vector of Web document w and that of story s. Functional words such as prepositions, auxiliary verbs, etc. are removed from documents and tf or tf-idf is used as term weighting method in creating these vectors. sim(w,s) represents the search score for story s with the Web document w. A final search score for story s with given query word q is then calculated as

$$\mathrm{score}(q,s) = \sum_{w \in W} \mathrm{sim}(w,s) + \alpha \qquad (1)$$

where W is the set of Web documents retrieved for the given query word q (i.e., W is the set of Web documents containing the query word q), α is a positive constant number if the LVCSR result for the story s includes query word q, otherwise $\alpha = 0$.

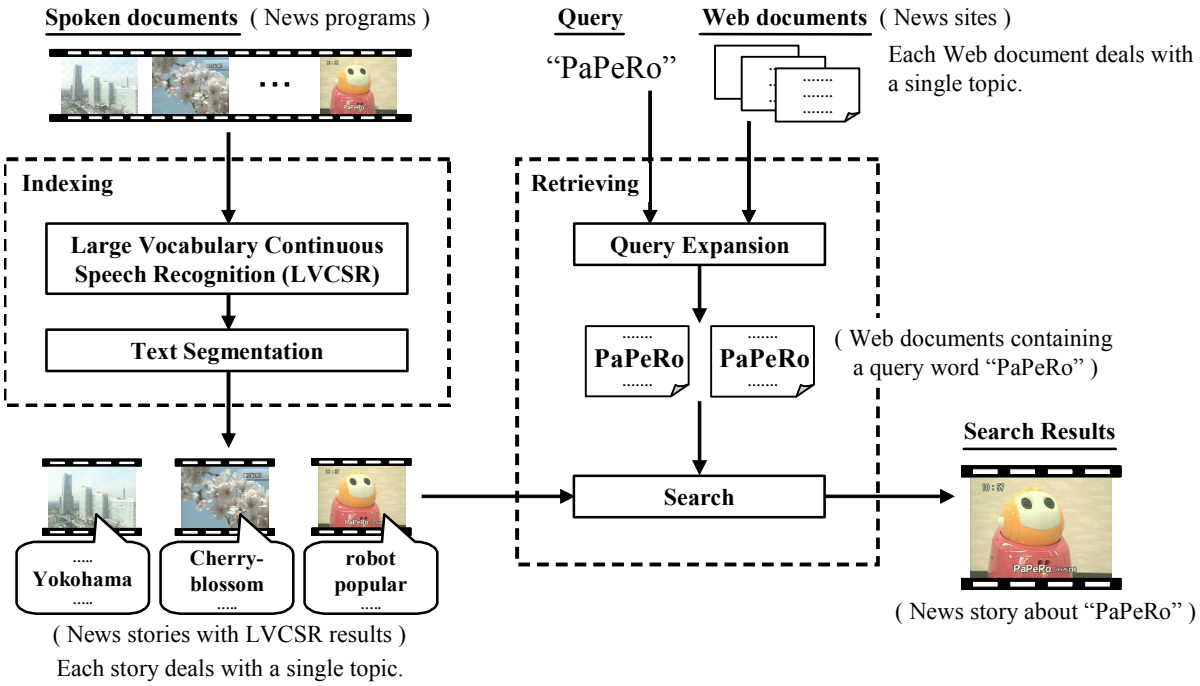


Figure 3: Framework of the proposed method.

Only the first term on the right-hand side of the Eq. (1) would be relevant in an OOV query, while both terms would be meaningful in an IV query. The first term would yield a high score if the story s is relevant to query word q, regardless of whether q is an OOV query or an IV query. That is, the proposed method can be useful not only for making retrieval with OOV query words possible but also for ranking retrieval results obtained in IV queries.

4. Evaluations

We experimentally evaluated the performance of our spoken document retrieval method on Japanese broadcast news programs, which included a large amount of spontaneous speech and various kinds of noise and background music.

4.1. Conditions

The combined time length of the programs was approximately 20 hours, in which totally 860 news stories were contained. The LVCSR system we used employed a 2-pass search strategy based on speaker clustering and an unsupervised speaker adaptation technique [4]. The average word error rate was 37%. We used an HMM-based text segmentation technique [5] to segment LVCSR results automatically into stories. The accuracy of story boundary detection was 61.5 % in F-measure. We also prepared 3254 Web documents from news sites in advance for query expansion. These Web documents appeared during the same general time period as did the broadcast of the news programs.

9 out-of-vocabulary query words and 20 in-vocabulary query words were selected for the evaluation, in consideration of the frequency of word appearance in articles near the top of news sites' article-rankings for the relevant time period, so that these query words represent popular topics during the time period.

We defined the segments of news programs to be retrieved for a certain query word (i.e., correct answer segments) as stories in which the query word was actually spoken at least once. For evaluating retrieval performance, we used the recall and precision measures defined as

$$recall = \frac{Total\ time\ len.\ of\ \ correctly\ retrieved\ segments}{Total\ time\ len.\ of\ \ manually\ retrieved\ segments} \quad (2)$$

$$precision = \frac{Total\ time\ len.\ of\ \ correctly\ retrieved\ segments}{Total\ time\ len.\ of\ \ automatically\ retrieved\ segments} \quad (3)$$

4.2. Results

Figure 4 shows retrieval performance of the proposed method for OOV query words. The x-axis is recall, and the y-axis is precision. Individual dots represent varying search score thresholds. If a search score calculated with Eq. (1) for a certain story is above the threshold, that story is then retrieved. In this way, the total length for retrieved segments can be controlled.

Fig 4(a) shows the result of an experiment comparing term weighting methods in creating document vectors under the condition that the news programs are manually segmented. The solid and dashed lines represent the retrieval performance of the proposed method using tf-idf and tf term weighting method respectively. Although OOV query words never appear in LVCSR results, the solid line shows that the proposed method using tf-idf weighting achieves high performance, e.g., a precision rate of approximately 70% when the recall rate is 70%. This performance is quite sufficient for practical use, such as in broadcast news retrieval applications. On the other hand, tf weighting results in lower performance than tf-idf weighting, which suggests that how to calculate the similarity between each Web document and each story from

spoken documents is essentially important for the proposed method. Tf-idf weighting is employed in the rest of evaluations.

Figure 4(b) shows the result under more practical condition where the automatic segmentation technique [5] is applied to the news programs instead of manual segmentation. The dashed line represents performance using automatic segmentation, while the solid line represents performance using manual segmentation and is the same as the solid line in Figure 4(a). Figure 4(b) indicates that although using automatic segmentation results in lower performance than does using manual segmentation, retrieval performance is still sufficient for practical use. This poorer performance, due to segmentation error, results first of all from the degradation which segmentation error inevitably produces in retrieval performance, as may be seen in Eq. (2) and (3). Further, segmentation error also degrades the accuracy of similarity calculations in the search step because obtained stories will not be limited to single topics.

Figure 5 shows retrieval performance for IV query words. α was set to 10, which was empirically determined. The solid and dashed lines represent, respectively, performance of the proposed method using manual segmentation and automatic segmentation. With IV query words, direct searching of LVCSR results for the query word is feasible, and the performance of such a conventional search is shown as "baseline". Both lines show that the proposed method makes it possible to control retrieval performance, as represented on the recall-precision curve, by varying the threshold value. Precision will reach, for example, nearly 100% with a high threshold, while recall will reach a maximum of roughly 93% with a low threshold when using manual segmentation. That is, the proposed method also turns out to be useful for ranking retrieval results.

5. Conclusions

An open-vocabulary spoken-document retrieval method based on query expansion using related Web documents has been proposed. An LVCSR system first transcribes spoken documents into word sequences, which are then segmented into stories. Given a text query word, Web documents containing the query word are first retrieved. Spoken documents relevant to the query word are then retrieved by searching for the stories with the LVCSR result similar to the previously obtained Web documents. The effectiveness in retrieving spoken documents using OOV query words and in ranking retrieval results obtained in IV queries has been demonstrated in an experiment employing broadcast news programs. The next step for this study will be to apply the proposed method to retrieving spoken documents other than those found in broadcast news programs.

6. References

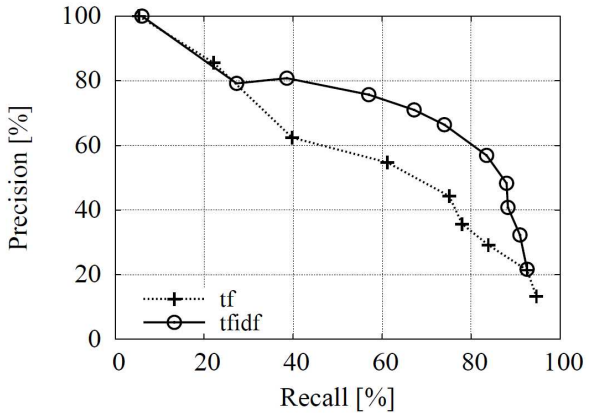
[1] K. Iwata, Y. Itoh, K. Kojima, M. Ishigame, K. Tanaka, and S. Lee, "Open-vocabulary spoken document retrieval based on new subword models and subword phonetic similarity," INTERSPEECH2006, pp.325-328, 2006.

[2] H. Nishizaki and S. Nakagawa, "Robust spoken document retrieval methods for misrecognition and out-of-vocabulary keywords," Systems and Computers in Japan, pp.44-53, 2004.

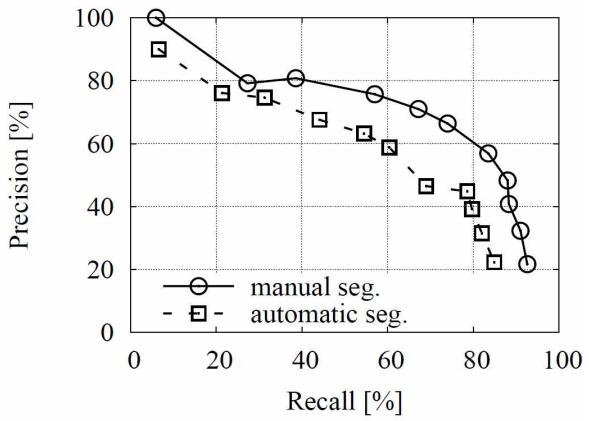
[3] T. Hori, I. L. Hetherington, T. J. Hazen, and J. R. Glass, "Open-vocabulary spoken utterance retrieval using confusion networks," ICASSP2007, pp.73-76, 2007.

[4] K. Shinoda and C. H. Lee, "A structural Bayes approach to speaker adaptation," IEEE Trans. Speech Audio Processing, pp.276-287, 2001.

[5] T. Koshinaka et al., "An HMM-based text segmentation method using variational Bayes approach and its application to LVCSR for broadcast news," ICASSP2005, pp.485-488, 2005.



(a) Comparison of term weighting methods under the condition of manual segmentation.



(b) Comparison of manual and automatic segmentation.

Figure 4: Retrieval performance for OOV queries.

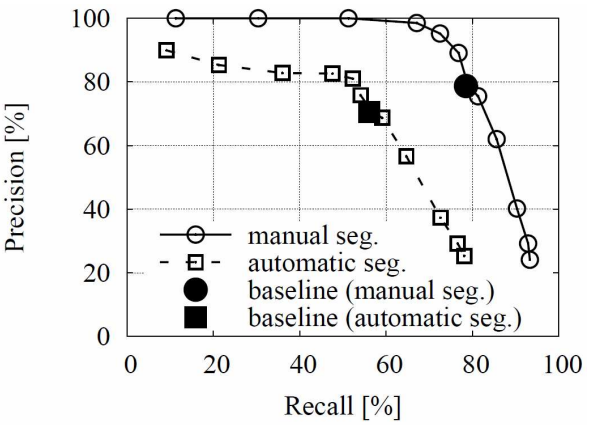


Figure 5: Retrieval performance for IV queries.