Japanese Spontaneous Spoken Document Retrieval Using NMF-Based Topic Models

Xinhui Hu, Hideki Kashioka, Ryosuke Isotani, and Satoshi Nakamura

National Institute of Information and Communications Technology, Japan
{xinhui.hu,hideki.kashioka,ryosuke.isotani,
satoshi.nakamura}@nict.go.jp

Abstract. In this paper, we propose a document topic model (DTM) which is based on the non-negative matrix factorization (NMF) approach, to explore Japanese spontaneous spoken document retrieval. Each document is interpreted as a generative topic model, belonging to many topics. The relevance of a document to a query is expressed by the probability of a query word being generated by the model. Different from the conventional vector space model where the matching between query and document is at the word level, the topic model complete its matching in the concept or semantic level. So, the problem of term mismatch in the information retrieval can be improved, that is, the relevant documents have possibilities to be retrieved even if the query words do not appear in them. The method also benefit the retrieval of spoken document containing "term misrecognitions", which is peculiar to the speech transcripts. By using this approach, experiments are conducted on a test collection of corpora of spontaneous Japanese (CSJ), where some of the evaluating queries and answer references are suited to retrieval in semantic level. The retrieval performance is improved by increasing the number of topics. When the topic number exceeds a threshold, the NMF's retrieval performance surpasses the tf-idf-based vector space model (VSM). Furthermore, compared to the VSM-based method, the NMF-based topic model also shows its strongpoint in dealing with term mismatch and term misrecognition.

Keywords: **s**poken document retrieval, non-negative matrix factorization, document topic model.

1 Introduction

The search and retrieval of a document is generally conducted by matching keywords in the query to those in the target documents. When the keywords are found in a document, the document is regarded to be relevant to the input query. A fundamental problem of information retrieval (IR) is *term mismatch*. A query is usually a short and incomplete description of the user's information need. Users and authors of documents often use different terms to refer to the same concepts and this produces an incorrect relevance ranking of documents with regard to the information need expressed in the query.

For spoken document retrieval (SDR), it faces a new problem besides the term mismatch. The SDR is generally carried out by using textual approaches to speech

G.G. Lee et al. (Eds.): AIRS 2009, LNCS 5839, pp. 149–156, 2009.

transcripts. The transcripts are generally obtained by utilizing automatic speech recognition systems. However, because of the limitation of current speech recognition technology, the transcript produced by the speech recognition process always contains errors. In SDR, terms misrecognized will not match the query and the document representations. Naturally, this hinders the effectiveness of the SDR system in a way similar to the term mismatch. Here, we call this problem as the term misrecognition.

Advanced users need tools that can find underlying concepts and not just search for keywords appearing in the query. It is widely acknowledged that the ability to work with text on a semantic basis is essential to modern information retrieval systems. Topic models are very popular for presenting the content of documents. Recently, researches on these aspects are becoming booming. The probabilistic latent topic modeling approaches, such as probabilistic latent semantic analysis (PLSA) [1] have been demonstrated effective in the tasks of spoken document retrieval. Chen [2] proposed a word topic model (WTM) to explore the co-occurrence relationship between words, as well as the long-span latent topical information, for language modeling in spoken document retrieval and transcription, and verified that the WTM is a feasible alternative to the existing models, i.e. PLSA.

Non-negative matrix factorization (NMF) [3] is also an approach in latent semantic space. It is a type of dimension reduction technique, and has distinct features of preserving the original data as well as the non-negative of the original data. Different from the other similar decomposition approaches such as singular value decomposition (SVD) [4], the NMF uses non-negativity constraints; the decomposition is purely additive; no cancellations between components are allowed, so they lead to a parts-based representation. Also, the NMF computation is based on a simple iterative process, it is therefore advantageous for applications involving data sparseness, like large vocabulary speech recognition. It is regarded to be suitable for finding the latent semantic structure from the document corpus and to identify document clusters in the derived latent semantic space. We adopt the NMF-based document topic model (DTM) approach for spontaneous spoken document retrieval (SDR) in this study. Since the approaches of latent semantic indexing are based on the semantic relations, a relevant document can be retrieved even if a query word does not appear in that document. So this feature can be used to compensate for the speech recognition errors. In this study, the focuses are mainly on dealing with the term misrecognitions, investigating the effectiveness of this DTM for SDR. The comparisons are conducted between this model and the conventional vector space model (VSM), since we presently limit on investigating the difference between the semantic matching and the keyword matching.

The rest of this paper is organized as follows: In Section 2, based on our previous work, we briefly introduce how to build the term-document matrix stochastically using N-best sequence. In Section 3, we describe the document topic model for information retrieval, and explain the method to construct the topic model by using the factorized matrices of the NMF, and show how to compute relevance of target document to the retrieving query using this topic model. In Section 4, the experimental setups and results are reported, highlighting the comparison between the proposed method and the conventional tf-idf-based VSM. Finally, in Section 5, we present our conclusions, discuss the characteristics of NMF in retrieval, especially when dealing with the term misrecognitions.

2 Term-Document Matrix Built on N-Best

The system presented here operates in two phases combining speech-based processing and text-based processing.

In the speech-based processing phase, the spoken documents are transcribed by an automatic speech recognizer (ASR). The transcription of the ASR is in the form of an N-best list, in which the top N hypotheses of the ASR results are stored in the recognition result lattice. The reason to select the N-best is that it needs less computation and less memory than the original lattice in search a recognition hypothesis. The usage of N hypotheses is to utilize those correct term candidates hidden in other hypotheses, and to compensate the effectiveness of term misrecognitions.

In the text-based processing phase, the term-document matrix used for NMF is built on an updated tf-idf-based vector space model (VSM). In *tf-idf*-based VSM, term frequency *tf*, which is defined as the number of a term occurs in a document and the inverse document frequency *idf*, are the two fundamental parameters. For the N-best, we introduce a stochastic method to compute these two parameters. This method is described as follows:

Let *D* be a document modeled by a segment of the N-Best. P(w|o,D) is defined as the posterior probability or confidence of a term *w* at position *o* in *D* in order to refer to the occurrence of *w* in the N-Best.

The *tf* is evaluated by summing the posterior probabilities of all occurrences of the term in the N-Best. Furthermore, we update it with Robertson’s 2-Poisson model as follows.

$$tf(D,w) = \frac{tf'(D,w)}{tf'(D,w) + \frac{length(D)}{\Delta}} \tag{1}$$

Where the *tf’* is the conventional term frequency, and is defined as follows:

$$tf'(w,D) = \sum_{D} \sum_{i=1}^{N} K(i) * P(w \mid o_i, D) \tag{2}$$

$$K(i) = (N+1-i) \Big/ \sum_{t=1}^{N} t \tag{3}$$

The length(D) is the length of document D, ⊿ is the average length of the whole document set. Similarly, the *idf* is calculated on the basis of the posterior probability of *w*, as shown in the following equation:

$$idf(w) = \log(N_D / \sum_{D \in C} O(w,D)) \tag{4}$$

Here,

$$O(w,D)=\begin{cases}1, & \textit{if } tf'(w,D)>0.5\\ 0, & \textit{otherwise}\end{cases} \tag{5}$$

N_D is the total number of documents contained in the corpus. C is the entire document set of the corpus.

The term-document matrix A for NMF is finally built by using *tf idf*. By using the processing of NMF, a topic model is constructed, and is used for computing the relevance of target documents to the input query.

3 NMF-Based Document Topic Model for Spoken Document Retrieval

3.1 Document Topic Model and Information Retrieval

In information retrieval (IR), the relevance measure between a query Q and a document D can be expressed as *P(D|Q)*. By applying the Bayes theorem, it can be transformed into:

$$P(D \mid Q)=\frac{P(Q \mid D)P(D)}{P(Q)} \tag{6}$$

With the invariability of *P(Q)* over all documents, and assuming that document probability *P(D)* has a uniform distribution, ranking the documents by the *P(D|Q)* can be realized using *P(Q|D)*, the probability of query *Q* being generated by the document *D*. If the query *Q* is composed of a sequence of terms (or words) $\mathrm{Q}=\mathrm{w}_1\mathrm{w}_2\ldots\mathrm{w}_{\mathrm{Nq}}$, the *P(Q|D)* can be further decomposed as a product of the probabilities of the query words generated by the document :

$$P(Q \mid D)=\prod_{w_i} P(w_i \mid D) \tag{7}$$

Each individual document *D* can be interpreted as a generative document topic model (DTM), denoted as M_D, and is embodied with K latent topics. Each latent topic is expressed by the word distribution of the language. So two probabilities are associated with this topic model: the probability of a latent topic given a document and the probability of word in a latent topic. So the probability of a query word w_i generated by *D* is expressed by

$$P_{DTM}(w_i \mid M_D)=\sum_{k=1}^{K} P(w_i \mid k)P(k \mid M_D) \tag{8}$$

Where $P(w_i \mid k)$ denotes the probability of a query word w_i occurring in a specific latent topic *k*, and $P(k|M_D)$ is the posterior probability of the topic *k* generated by the document model M_D.

Therefore, considering on the equation (7) and (8), the likelihood of a query Q generated by D is thus represented by

$$P_{DTM}(Q \mid M_D) = \prod_{w_i} \left[\sum_{k=1}^{K} P(w_i \mid k) P(k \mid M_D) \right] \tag{9}$$

In this study, we compare the retrieval performance of the NMF with the conventional vector space vector (VSM) where the similarity between the query Q and document D is computed by following equation:

$$sim(D,Q) = \frac{D \bullet Q}{|D||Q|} \tag{10}$$

3.2 Link NMF to Topic Model

Let ***A*** be the matrix produced in section 2 to stand for relationships among the terms and documents, with dimension m×n. Let S be the sum of all elements in ***A*.** Then $\overline{A} = A/S$ forms a normalized table to approximate the joint probability *p(w,d)* of term *w*, and document *d*.

NMF is a matrix factorization algorithm [3] that finds the positive factorization of a given positive matrix. Assume that the given document corpus consists of *K* topics. The general form of NMF is defined as:

$$\overline{A} \approx GH \tag{11}$$

The matrix factorization of $\overline{A}$ will result in an approximation by a product of two non-negative matrices, *G* and H with dimension m×k and dimension k×n respectively.
So from the equation (11), the joint probability *p(w,d)* can be expressed by

$$p(w,d) = \overline{A} = \sum_{k=1}^{K} G_{w,k} \cdot H_{k,d} \tag{12}$$

To normalize *G* by $\alpha_k = \sum_w G_{w,k}$, H by $\beta_k = \sum_d H_{k,d}$, and define $p(k) = \alpha_k \beta_k$, the *p(w,d)* can be rewritten as:

$$p(w,d) = \sum_{k=1}^{K} \frac{G_{w,k}}{\alpha_k} \frac{H_{k,d}}{\beta_k} p(k) = \sum_{k=1}^{K} \hat{G}_{w,k} \hat{H}_{k,d} \, p(k) \tag{13}$$

Each entity $g_{w_i,k} (1 \le k \le K)$ of the matrix $\hat{G}$ accounts for the probability of a word w_i that would be generated by a latent topic *k*., that is $P(w_i \mid k)$ of the equation (9).

Each entity $h_{k,n}$ of the matrix $\hat{H}$ accounts for the probability of document D by a latent topic k, that is $P(M_D \mid k)$.

The $P(k|M_D)$ of equation (8) can be obtained by using Bayes theorem

$$p(w,d) = \sum_{k=1}^{K} \frac{G_{w,k}}{\alpha_k} \frac{H_{k,d}}{\beta_k} p(k) = \sum_{k=1}^{K} \hat{G}_{w,k} \hat{H}_{k,d} \, p(k) \tag{14}$$

Based on the above equations, the equation (9) for relevance can be computed by the matrices G and H, the factorized matrices of the NMF.

4 Experiments

4.1 Experimental setups

A test collection of CSJ is used for evaluation. The CSJ (Corpus of Spontaneous Japanese) is the result of a Japanese national project on ‘Spontaneous Speech Corpus and Processing Technology ‘ [5]. It contains 658 hours of speech consisting of approximately 7.5 million words. The speech materials were provided by more than 1,400 speakers of various ages. About 95% of the CSJ corpus is devoted to spontaneous monologues, such as academic presentations and public speaking, including manual transcriptions. This test collection is developed by the Japanese Spoken Document Processing Working Group [6], with the aim of evaluating the retrieval of spoken document retrieval systems. This collection consists of a set of 39 textual queries, the corresponding relevant segment lists, and transcriptions by an automatic speech recognition (ASR) system, allowing retrieval of 2702 spoken documents of the CSJ. The large vocabulary continuous ASR system use an engine in which the acoustical model is trained by a corpus in the domain of travel [8], but the language model is trained by the manually-built transcript of the CSJ corpus. The word accuracy of the recognition system is evaluated as 60.5%. Because the criteria in determining relevant is not merely dependent on query’s keywords, the semantic content needs to be taken into consideration. For examples, for keyword sequence of “ペット 効用 目的” which corresponds to query text “[HN101801] ペットを飼うことの効用または目的について述べている箇所を探したい(*search the utterances about the purposes or effects of raising pets*”, keyword “ペット(pet)” appears in all of its answer files, but “効用(effect), “目的(purpose)” are misrecognized or only appear in just 2 files with low *tf*. That means that when using the VSM, these answers can be hit mainly dependent on the first keyword “ペット(pet)”.

For the data structure of test transcript, three types of transcripts of the same spoken documents are used for evaluations.

(1) N-best (here 10-best is used, denoted as *nbst*).
(2) 1-best (denoted as *1bst*).
(3) Manual transcript (denoted as *tran*).

The mean average precision (**MAP**) is used as the performance measure in this study.

4.2 Experimental Results

4.2.1 Retrieval Performance with Topic Number

Figure 1 shows the MAPs of retrieving *nbst* using the proposed NMF-based model, and conventional tf-idf-based VSM in different topics number. As the topic number increases, the *nbst's* MAP increases nearly monotonously. After the topic number is over a threshold (here it is 700), the MAP of the NMF-based model surpasses the VSM. Therefore it can be concluded that the NMF mainly functions in the high dimensional semantic space.

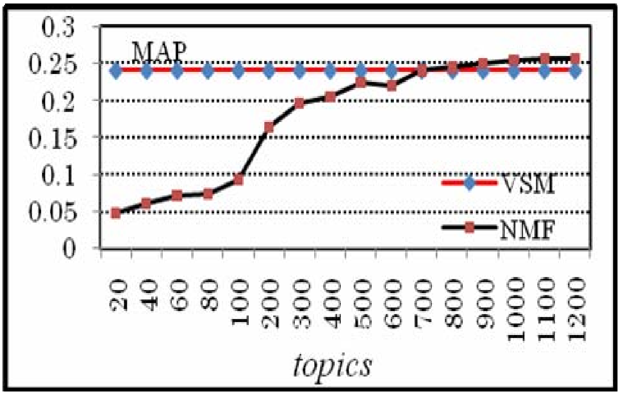

Fig. 1. MAPs in different topic numbers

4.2.2 Effectiveness on different Data Type

Table 1 shows the retrieval performance for different data types using the NMF and VSM methods. In this experiment, the number of topics was selected to be 1000.

For all of 3 data types, the NMF method proved to be superior to the VSM. For example, the improvement of NMF to the VSM is 5.5% for the N-best transcripts. Although the improvement is not so large, the significance of the NMF is that its retrieval is on the semantic level, so it has a potential ability to deal with the problem of misrecognition. Meanwhile, the performance of systems that use the N-best are better than the those that use the *1-bst*, they show the same characteristics as in other research on lattice-based spoken document retrieval that the N-best can search or retrieve correct speech segment by utilizing multiple recognition hypotheses even if the 1-best one is incorrect [7].

Table1. Retrieval performance of different data type (for NMF, dimension=1000)

	1bst	nbst	tran
NMF	0.240	0.255	0.285
VSM	0.233	0.241	0.253

5 Conclusions

In this paper, we proposed a NMF-based document topic model to explore the Japanese spoken document retrieval. By experiments on the CSJ spoken corpus, the retrieval performance of the NMF-based topic model is found to be steadily improved with the increases in the number of topics. When the topic number becomes sufficiently large, the NMF-based model outperforms the conventional tf-idf-based VSM. However, this fact also reveals that the merit of NMF-based topic model for retrieval is conditional on the number of topics.

We show that as in the case of the VSM, the N-best is also effective to compensate for the misrecognition for the proposed NMF-based model. Moreover, its improvement (6.2%) from 1-best to N-best is also larger than the VSM(3.4%). This achievement is due to the characteristics of topic model – matching at the topic level. By analyzing individual queries, the retrieval improvement mainly happens in those containing misrecognition or no keyword exists in documents. For instance, for query [HN101801] mentioned in above, its MAP is changed from 0.051 (VSM) to 0.128 (NMF).

In future work, the comparison of the NMF-based topic model to other topic models such as PLSA, LDA will be analyzed in detail.

Acknowledgments. This work was partly supported by a Grant-in-Aid for Scientific Research on Priority Areas in Japan as a part of Cyber Infrastructure for the Information Explosion Era, under Grant No. 19024074.

References

1. Hoffmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the SIGIR 1999, pp. 50–57 (1999)
2. Chen, B.: Word topic models for spoken document retrieval and transcription. ACM Transactions on Asian Language Information Processing (TALIP) 8(1), 1–27 (2009)
3. Lee, D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: NIPS, vol. 13 (2001)
4. Kita, K., Tuda, K.H., Sisibori, M.: Information Retrieval Algorithms. Kyoritu Press (2003)
5. Maekawa, K.: Corpus of spontaneous Japanese: its design and evaluation. In: SSPR 2003 (2003)
6. Akiba, T., Aikawa, K., Itoh, Y., Kawahara, T., Nanjo, H., Nishizaki, H., Yasuda, N., Yamashita, Y., Itou, K.: Test collection for spoken document retrieval from lecture audio data. In: Proceedings of the LREC 2008 (2008)
7. Saraclar, M., Sproat, R.: Lattice-based Search for Spoken Utterance Retrieval. In: Proc. of HLT-NAACL, pp. 129–136 (2004)
8. Nakamura, S., Markov, K., Nakaiwa, H., Kikui, G., Kawai, H., Jitsuhiro, T., Zhang, J., Yamamoto, H., Sumita, E., Yamamoto, S.: The ATR multilingual speech-to-speech translation system. IEEE Trans. on Audio, Speech, and Language Processing 14(2), 365–376 (2006)