Speech Indexing Using Semantic Context Inference

Chien-Lin Huang[1,2], Bin Ma[1], Haizhou Li[1] and Chung-Hsien Wu[3]

[1]Human Language Technology Department, Institute for Infocomm Research, A*STAR, Singapore
[2]National Institute of Information and Communications Technology, Kyoto, 619-0288, Japan
[3]Computer Science and Information Engineering, National Cheng Kung University, Taiwan
{clhuang, mabin, hli}@i2r.a-star.edu.sg, chwu@csie.ncku.edu.tw

Abstract

This study presents a novel approach to spoken document retrieval based on semantic context inference for speech indexing. Each recognized term in a spoken document is mapped onto a semantic inference vector containing a bag of semantic terms through a semantic relation matrix. The semantic context inference vector is then constructed by summing up all the semantic inference vectors. Such a semantic term expansion and re-weighting make the semantic context inference vector a suitable representation for speech indexing. The experiments were conducted on 1550 anchor news stories collected from Mandarin Chinese broadcast news of 198 hours. The experimental results indicate that the proposed speech indexing using the semantic context inference contributes to a substantial performance improvement of spoken document retrieval.

Index Terms: speech indexing, semantic context inference, spoken document retrieval

1. Introduction

Speech is the most convenient way for the interaction of human-to-human and human-to-machine. The applications of spoken document retrieval in education, business and entertainment are rapidly growing. The recent attempts include multilingual oral history archives access [1], MIT lecture browsing [2], and the management of National Gallery consisting of speeches, news broadcasts and recordings [3], voice search about spoken dialog, call-routing systems [4], etc. All of them focus on retrieving the information to meet users' requirements. We know that it is not straightforward to directly compare the speech query with the spoken documents in the database. In order to construct an efficient and effective retrieval system, the state-of-the-art spoken document retrieval (SDR) technologies adopt the transcription obtained from automatic speech recognition for indexing. Vector space model [5] and probabilistic models (HMM [6], GMM [7], KL-divergence [8]), rely on certain similarity functions that assume a document is more likely to be relevant to a query if it contains more occurrences of query terms.

The indexing techniques of text-based information retrieval have been widely adopted in spoken document retrieval. However, due to imperfect speech recognition results, out-of-vocabulary, and the ambiguity in homophone and word tokenization, conventional text-based indexing techniques are not always appropriate for spoken document retrieval. The transcription errors may cause undesired semantic and syntactic expression, thus result in an inadequate indexing. Several approaches have been proposed to address these problems with various indexing units such as word, sub-word, phone, and so on. The multi-level knowledge indexing approach considers three information sources including the speech transcription, keywords extracted from spoken documents, and hypernyms of the extracted keywords [9]. Hui et al. applied the *n*-best recognition hypothesis for document expansion prior to the retrieval [10]. Dharanipragada et al. proposed an algorithm consisting of a phone *n*-gram representation stage and a coarse-to-detailed search stage for spotting a word/phone sequence in speech [11]. Another indexing approach called *Particles* was based on syllable-like units [12]. *Particles* are defined as the within-word sequences of characters obtained from the orthographic or phonetic transcription of words. However, these approaches only take into account multiple candidates or types indexing to enhance the retrieved information. The semantic content and semantic relation of speech, which play an important role in the way we perceive the speech transcription and measure their similarity, are not well considered.

In this paper, we study an approach to concept mapping and context expansion of spoken document by introducing the Semantic Context Inference (SCI). First, we need to know the term-by-term associations for inference. We construct a semantic relation matrix that reflects the term-by-term associations using the document-by-term dataset. Then, each recognized term is mapped into a bag of semantic related terms based on the semantic relation matrix. With the semantic term expansion and re-weighting, SCI is expected to alleviate the problems resulting from speech recognition errors. In previous studies, indexing by latent semantic analysis (LSI) [13] took into account of conceptual indexing by projecting the term vector into a lower-dimensional latent semantic analysis (LSA) space. Like LSI, the proposed SCI allows the similarity measure between queries and documents, considering not only the original terms but also the inference concepts. The experimental results indicate that SCI outperforms LSI and the conventional term vector indexing methods.

The rest of this paper is organized as follows. Section 2 presents the proposed speech indexing scheme using the semantic context inference. We describe the experimental setup and report a series of experiments in Section 3. Finally, Section 4 concludes this work.

2. Semantic Context Inference

We proposed the semantic context inference representation by finding the semantic relation between terms, and suggesting semantic term expansion for speech indexing. These associated terms are re-weighted as the new representation on queries and documents for spoken document retrieval.

2.1. Semantic relation matrix

A spoken document database comprises an accumulation of spoken documents from which the document-by-term matrix $\mathbf{W}=[\mathbf{v}_1,\mathbf{v}_2,...,\mathbf{v}_D]$ is derived. The *d*-th spoken document is represented by a row vector of terms $\mathbf{v}_d=[a_1^d,a_2^d,...,a_K^d]$ derived from the statistics of transcription with weighting

terms a_K^d. D denotes the total number of spoken documents for indexing. K is the dimension of the indexing term vector. Due to imperfect speech recognition results and the redundancy transcription, not all of the recognized terms are valid and meaningful. To eliminate those noisy terms, the terms which have low frequency in a document and occur in few documents will be discarded by the following term weighting scheme:

$$a_k^d = \frac{tf(a_k,d)+1}{n_d} \times \log\left(\frac{D}{df(a_k)+1}\right)$$
$$n_d = \sum_k tf(a_k,d) \quad (1)$$

where $tf(a_k,d)$ represents the number of occurrences of a recognized term a_k in the spoken document d; $df(a_k)$ is the number of documents that contain at least one occurrence of the term a_k in the spoken document database. The advantage of term weighting scheme is to provide useful information about how important a term is to a document in the spoken document database.

In order to build a semantic relation matrix for reflecting the term-by-term associations, the semantic context inference starts with the covariance estimation by constructing a term-by-term matrix, $\hat{\mathbf{W}} = \mathbf{W}\mathbf{W}^T$, while T denotes the matrix transposition. In this study, $\hat{\mathbf{W}}$ is a symmetric matrix used to describe co-relations between terms through a collection of documents. The diagonal of the matrix $\hat{\mathbf{W}}$ means the self-terms and shows the highest co-relation scores. We conduct a singular value decomposition (SVD) [14] which finds the optimal projection to explore term co-occurrence patterns as shown in Fig. 1.

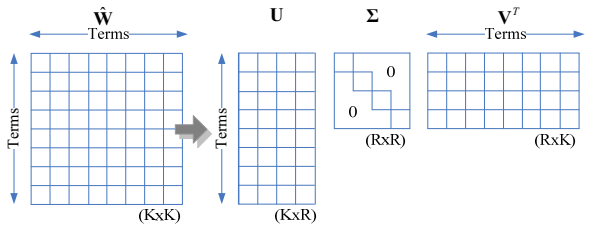

Figure 1: Singular value decomposition.

SVD is commonly used in eigenvector decomposition and factor analysis [15]. We perform SVD of the matrix $\hat{\mathbf{W}}$ as follows

$$\hat{\mathbf{W}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \quad (2)$$

where $\mathbf{U}$ and $\mathbf{V}$ are the left and right singular matrix, respectively. Both $\mathbf{U}$ and $\mathbf{V}$ show the orthogonal characteristics. $\mathbf{\Sigma}$ is the $K \times K$ diagonal matrix whose nonnegative entries are singular K values in a descending order, i.e., $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_K > 0$.

SVD can be used to project all the dimensions of the term vectors onto a latent information space with significantly reduced dimensionality. In this paper, we apply SVD to select the major factors based on a threshold θ.

$$\frac{1}{\bar{\sigma}}\sum_{r=1}^{R}\sigma_r \geq \theta \; ; \quad \bar{\sigma} = \sum_{k=1}^{K}\sigma_k \quad (3)$$

θ was empirically adopted to select the eigenvectors $\hat{\mathbf{U}} = [u_1, u_2, ..., u_R]$ based on the eigenvalues $\hat{\mathbf{\Sigma}} = [\sigma_1, \sigma_2, ..., \sigma_R]$ with the first R dimensions, where $R \leq K$ denotes the projected dimensions of the original term vector in the eigenspace. The associated eigenvalues allow us to rank the eigenvectors according to their usefulness in characterizing the semantic relations between terms [16]. The eigenvectors $\hat{\mathbf{U}}$ are treated as the transform basis in latent semantic analysis [13]. As a result, the semantic relation matrix $\tilde{\mathbf{W}}$ is reconstructed as follows

$$\tilde{\mathbf{W}} = \hat{\mathbf{U}}\hat{\mathbf{\Sigma}}\hat{\mathbf{U}}^T \quad (4)$$

Different from the matrix $\hat{\mathbf{W}}$, the matrix $\tilde{\mathbf{W}}$ removes the noisy factors and captures the most important term-to-term associations. The matrix $\tilde{\mathbf{W}}$, which contains all of the term-to-term dot products is a representation of the co-occurrences and semantic relations among terms. The co-relation scores of the matrix $\tilde{\mathbf{W}}$ are estimated based on the similarity between concepts.

2.2. Semantic context inference for indexing

Each of the recognized term a_k in the spoken document d can be mapped onto a semantic inference vector $a_k \rightarrow \tilde{\mathbf{v}}_k$ through the semantic relation matrix $\tilde{\mathbf{W}} = [\tilde{\mathbf{v}}_1, \tilde{\mathbf{v}}_2, ..., \tilde{\mathbf{v}}_K]$ as illustrated in Fig. 2. The semantic inference vector $\tilde{\mathbf{v}}_k^d$ is actually a representation of the associated terms of term a_k.

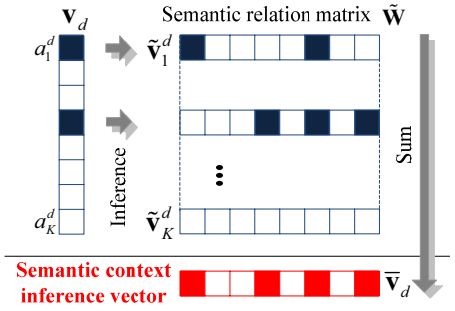

Figure 2: Semantic context inference based on the semantic relation matrix.

By summing up all the semantic inference vectors for the spoken document d, we finally obtain the semantic context inference vector as follows:

$$\bar{\mathbf{v}}_d = \sum_K \tilde{\mathbf{v}}_k^d \quad (5)$$

The semantic context inference vector can be regarded as a re-weighing indexing vector by expanding the indexing terms based on the related terms in the semantic inference vectors $\tilde{\mathbf{v}}_k^d$. We assume that a spoken document is associated with the same topic. We expect such expansion to reinforce the topic. Furthermore, the procedure for deriving the semantic context inference vector is entirely data driven without any pre-defined knowledge, such as WordNet [17] and HowNet [18] which require a pre-defined concept or knowledge database.

The proposed semantic context inference (SCI) is different from the latent semantic analysis (LSI) [13] in that we use different basis, $\hat{\mathbf{U}}$ for LSI and the semantic relation matrix $\tilde{\mathbf{W}}$ for SCI. LSI aims to reduce the data dimensionality to a low-dimensional space, and to project the elements in the

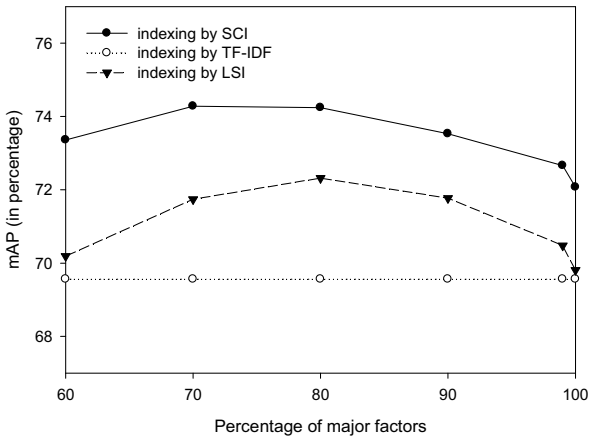

Figure 3: mAP performance on MATBN using different thresholds for selecting major factors in SVD.

document-by-term matrix to the orthogonal axes using the basis $\hat{\mathbf{U}}$, while SCI takes into account the semantic relation matrix $\tilde{\mathbf{W}}$ which shows term-by-term associations.

2.3. Retrieval model

For spoken document retrieval, we adopt the vector space models which have been widely used in information retrieval by offering a highly efficient retrieval with a feature vector representation for a document. The feature vector is estimated by the proposed semantic context inference vector and the cosine measure is applied to estimate the similarity between the query $\overline{\mathbf{v}}_q$ and the spoken document $\overline{\mathbf{v}}_d$ as follows:

$$S(\overline{\mathbf{v}}_q, \overline{\mathbf{v}}_d) = \frac{\sum_{k=1}^{K} \overline{\mathbf{v}}_{k,q} \times \overline{\mathbf{v}}_{k,d}}{\sqrt{\sum_{k=1}^{K} \overline{\mathbf{v}}_{k,q}^2} \times \sqrt{\sum_{k=1}^{K} \overline{\mathbf{v}}_{k,d}^2}} \tag{6}$$

where K is the dimension of the feature vectors. Retrieval results are ranked according to the similarities obtained in the retrieval process.

3. Experiments

This study applied Mel-frequency cepstral coefficients (MFCCs) for speech recognition. Each frame of the speech data is represented by a 36 dimensional feature vector, consisting of 12 MFCCs, along with their deltas, and double-deltas. The features were normalized to zero mean and unit variance for improving discrimination ability [19]. The speech recognition system was based on the hidden Markov model (HMM) and the phonetic structure of Mandarin Chinese with 137 sub-syllables including 100 right-context-dependent INITIALs and 37 context-independent FINALs as the basic units. The decision-based state-tying context-dependent sub-syllable units were used for the acoustic modeling. The number of Gaussian mixtures per HMM state ranged from 2 to 32, depending on the quantity of the training data. Each sub-syllable unit was modeled with three states for the INITIALs and four states for the FINALs. The silence model was a one-state HMM with 64 Gaussian mixtures trained with the non-speech segments [9].

3.1. Evaluation data

The spoken document corpus was acquired from the Mandarin Chinese broadcast news corpus (MATBN) collected by Academia Sinica, Taiwan [20]. The corpus contains a total of 198 hours of broadcast news with the corresponding transcription from the Public Television Service Foundation [21]. 1550 anchor news stories ranging over three years were extracted for experiments. The average news story length was 16.38 seconds with an average of 51.85 words. The speech data in MATBN were recognized by the speech recognition system with the word accuracy of 78.92%. Moreover, the Topic Detection and Tracking collection (TDT2) was also used for this study. 2112 Mandarin Chinese audio news stories from Voice of America news broadcasts (VOA) were used in the experiments. The average document length of TDT2 was 174.20 words. The word accuracy of TDT2 was about 75.49%. For TDT2, the speech recognition transcription was provided by LDC [22].

3.2. Evaluation metrics

To measure the accuracy of retrieved documents and the ranking position of the relevant document, the mean average precision was estimated as follows:

$$\text{mAP} = \frac{1}{N_q} \sum_{i=1}^{N_q} \frac{1}{N_i} \sum_{j=1}^{N_i} \frac{j}{rank_{ij}} \tag{7}$$

where N_q denotes the number of queries, and N_i represents the number of relevant documents contained in the retrieved documents for query i. $rank_{ij}$ denotes the rank of the j-th relevant document for the i-th query q. In this study, in order to evaluate the robustness of speech indexing based on the semantic context inference, the same pool of 164 keyword queries (from two to four Chinese characters) were used for both MATBN and TDT2. The average length of queries was 3.02 Chinese characters. There were 15.71 and 21.20 relevant spoken documents in MATBN and TDT2, respectively.

3.3. Indexing by SCI, LSI and TF-IDF

To remove the noisy factors in eigen decomposition, we set a threshold θ (see Eq. (3)) for keeping the major factors. A θ of higher value indicates that more eigenvectors are used for latent semantic analysis as well as the reconstruction of the semantic relation matrix. The experimental results shown in Fig. 3 were obtained with MATBN broadcast news corpus using different thresholds for the indexing based on LSI and SCI, while the popular term vector indexing (TF-IDF) was used as the baseline which achieved 69.56% mAP. The experiments show that the complete LSA space does not give as good performance as the dimension-reduced LSA space. It also shows that the best results can be achieved when the threshold 80% for LSI and 70% for SCI are selected separately. These results confirm that a better performance can be achieved by removing the noisy factors. The experimental results also confirm that the proposed SCI outperforms both TF-IDF and LSI indexing approaches.

3.4. Evaluation of TDT2 and MATBN

To evaluate the effect of the semantic context inference, the proposed approach for speech indexing was applied on TDT2 and MATBN corpus using both automatic speech recognition results (ASR transcription) and perfect text (text transcription). The experimental results shown in Table 1 indicate that consistent spoken document retrieval improvements have been obtained on TDT2 and MATBN based on SCI indexing, compared with TF-IDF indexing. To understand the upper-bound of spoken document retrieval, the indexing by perfect text transcription was evaluated as the reference. There was a

Table 1: mAP performance of TF-IDF and SCI indexing with ASR and Text transcription of TDT2 and MATBN (in %).

	TDT2		MATBN	
	TF-IDF	SCI	TF-IDF	SCI
ASR transcription	71.92	74.55	69.56	74.28
Text transcription	89.76	90.79	88.76	90.48

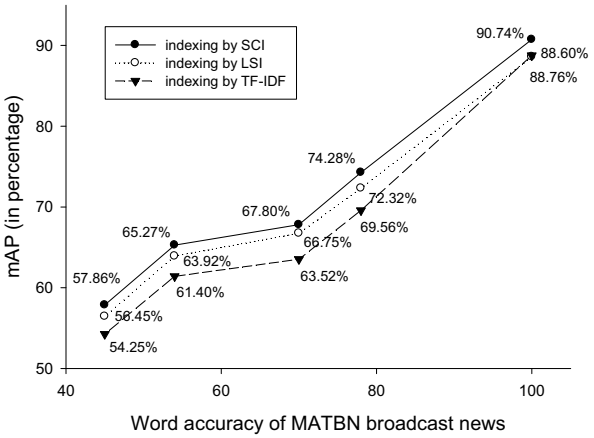

Figure 4: mAP performance of TF-IDF and SCI indexing with different word accuracies.

gap (about 15%~20% mAP) between indexing using ASR transcription and text transcription because of imperfect speech recognition.

3.5. Effect of recognition accuracy for retrieval

Speech recognition performance becomes less predictable under adverse acoustic environments. Figure 4 reports the retrieval results as a function of various speech recognition word accuracies. To study the impact of speech recognition accuracy variance on the semantic context inference, we used different settings of speech recognition system to provide the transcription of different accuracies. We also conducted experiments on the manually transcribed text, called text-based document retrieval. Experiments were conducted on MATBN broadcast news. The proposed semantic context inference approach reports a minor improvement over the conventional word vector retrieval (TF-IDF) method in the text-based document retrieval task, but it consistently outperforms LSI and TF-IDF approaches over a range of different speech recognition accuracies. 4.72% absolutely improvement from 69.56% mAP to 74.28% mAP has been achieved when the word accuracy is at 80%. The results also indicate that SCI offers bigger performance gain over LSI and TF-IDF when the transcription is corrupted with more speech recognition errors. The SCI shows an effective representation which reinforces the related topic in the spoken document.

4. Conclusions

This study presents a novel approach to spoken document retrieval. The proposed semantic context inference explores the latent semantic information and extends the semantic related terms to speech indexing. The semantic context inference vector can be regarded as a re-weighing indexing vector which is a way of query expansion to overcome speech recognition errors. The experimental results show that SCI outperforms the conventional TF-IDF term vector and LSI indexing approaches, and works especially well for speech recognition transcription with errors.

5. References

[1] W. Byrne, D. Doermann, M. Franz, S. Gustman, J. Hajic, D. Oard, M. Picheny, J. Psutka, B. Ramabhadran, D. Soergel, T. Ward, and W.-J. Zhu, "Automatic recognition of spontaneous speech for access to multilingual oral history archives," *IEEE Trans. on Speech Audio Processing*, vol. 12, no. 4, pp. 420–435, 2004.

[2] C. Chelba, T. J. Hazen, and M. Saraclar, "Retrieval and browsing of spoken content," *IEEE Signal Processing Magazine*, vol. 24, no. 3, pp. 39–49, 2008.

[3] J. H.L. Hansen, R. Huang, B. Zhou, M. Seadle, J. R. Deller, Jr, A. R. Gurijala, M. Kurimo, and P. Angkititrakul, "SpeechFind: Advances in Spoken Document Retrieval for a National Gallery of the Spoken Word," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, pp. 712–730, 2005.

[4] Y.-Y. Wang, D. Yu, Y.-C. Ju, and Alex Acero, "An Introduction to Voice Search," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 29–38, 2008.

[5] C. Buckley and J. Walz, "SMART in TREC 8," in *Proc. Eighth Text REtrieval Conf. (TREC-8 "99), NIST Special Publication 500–264*, Voorhees and Harman, eds., pp. 577–582, 2000.

[6] D. R. H. Miller, T. Leek, and R. Schwartz, "A hidden Markov model information retrieval system," in *Proc. ACM SIGIR Conf.*, pp. 214–221, 1999.

[7] C. Zhai and J. Lafferty, "A study of smoothing methods for language models applied to ad hoc information retrieval," in *Proc. ACM SIGIR Conf.*, pp. 334–342, 2001.

[8] T. K. Chia, K. C. Sim, H. Li, and H. T. Ng "A Lattice-Based Approach to Query-by-Example Spoken Document Retrieval," in *Proc. ACM SIGIR Conf.*, pp. 363–370, 2008.

[9] C.-L. Huang and C.-H. Wu, "Spoken Document Retrieval Using Multi-Level Knowledge and Semantic Verification," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2551–2560, 2007.

[10] P. Y. Hui, W. K. Lo, and H. Meng, "Two Robust Methods for Cantonese Spoken Document Retrieval," in *Proc. of the ISCA Multilingual Spoken Document Retrieval Workshop*, pp.7–12, 2003.

[11] S. Dharanipragada and S. Roukos, "A multistage algorithm for spotting new words in speech," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 8, pp. 542–550, 2002.

[12] B. Logan, J.-M. Van Thong, and P. J. Moreno, "Approaches to reduce the effects of OOV queries on indexed spoken audio," *IEEE Trans. on Multimedia*, vol. 7, no. 5, pp. 899–906, 2005.

[13] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman, "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[14] J. R. Bellegarda and K. E. A. Silverman, "Natural Language Spoken Interface Control Using Data-Driven Semantic Inference," *IEEE Trans. On Speech And Audio Processing*, vol. 11, no. 3, pp. 267–277, 2003.

[15] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. ACM SIGIR Conf.*, pp. 50–57, 1999.

[16] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[17] WordNet [Online] http://wordnet.princeton.edu/

[18] HowNet [Online] http://www.keenage.com/

[19] C.-L. Huang and C.-H. Wu, "Generation of Phonetic Units for Mixed-Language Speech Recognition Based on Acoustic and Contextual Analysis," *IEEE Trans. on Computers*, vol. 56, no. 9, pp. 1225–1233, 2007.

[20] CKIP Group, "Analysis of syntactic categories for Chinese," *Institute of Information Science, Academia Sinica, Taipei, Taiwan, CKIP Tech. Rep. 93-05*, 1993.

[21] Public Television Service Foundation [Online] http://www.pts.org.tw/

[22] Project topic detection and tracking, Linguistic Data Consortium [Online] http://projects.ldc.upenn.edu/TDT2/