SEMANTIC CONTEXT INFERENCE FOR SPOKEN DOCUMENT RETRIEVAL USING TERM ASSOCIATION MATRICES

Chien-Lin Huang and Chiori Hori

National Institute of Information and Communications Technology, Kyoto, Japan
chien-lin.huang@nict.go.jp

ABSTRACT

This study presents a novel approach to semantic context inference based on term association matrices for spoken document retrieval. Each recognized term in a spoken document infers a semantic vector containing a bag of semantic terms from a term association matrix. Such a semantic term expansion and re-weighting make the semantic context inference vector a suitable representation for speech indexing. We consider both words and syllables on term association matrices for semantic context inference. The syllable lattice bigram instead of the single-best speech recognition results and various term weighting schemes have been studied for semantic context inference. Experiments were conducted on Mandarin Chinese broadcast news. The results indicate the proposed approach offers a significant performance improvement of spoken document retrieval.

Index Terms— Semantic context inference, spoken document retrieval, term association matrices

1. INTRODUCTION

Indexing and retrieval of speech are active research topics [1]–[5]. It is highly demanded to identify multimedia content such as video and audio. Since speech is commonly observed in multimedia archives, applications of spoken document retrieval (SDR) are rapidly growing. The popular technologies of SDR adopt the transcription obtained from automatic speech recognition (ASR). Based on results of ASR, indexing techniques of text-based information retrieval (IR) have been widely used in SDR. Due to imperfect speech recognition results and out-of-vocabulary words, the conventional text-based IR techniques are not always appropriate for SDR. In the past decade, several approaches have been proposed to address these problems. For example, indexing by using phone, sub-word, and particles is used to address the problem of out-of-vocabulary words [6]–[9]. The use of ASR lattices or *n*-best lists enhances the diversity of the speech transcription and reduces the adverse effects of ASR errors on retrieval [10]–[13]. The multi-level knowledge indexing approach considers complementary information sources from the speech transcription to improve the retrieval performance [14]. The conventional text-based IR methods, such as document expansion and relevance feedback, provide good solutions on problems of spoken document retrieval [15]–[18].

Typical speech search methods make use of a term frequency-inverse document frequency (TF-IDF) weighted vector space model [19]. Since speech transcription errors may cause undesired semantic and syntactic expression, they result in an adequate indexing. Indexing of bag-of-words is one of the most frequently used methods for information retrieval. However, each "word" or term is considered in isolation, ignoring its neighbors and semantic relation in context. We present a semantic context inference (SCI) for SDR by exploring the association between spoken terms [20]. The contribution of this study is at least threefold. First, we propose methods to construct a term association matrix for semantic context inference. The indexing vector is then mapped into a bag of semantic related terms in an SCI vector based on the term association matrix, which is regarded as a semantic context re-weighting of associated terms. With the term expansion and re-weighting, semantic context inference is expected to alleviate problems resulting from errors of ASR. Second, we investigate three term weighting schemes of TF-IDF [21], Okapi BM25 [22], and entropy on both conventional word and SCI indexing [23]. Last, the score fusion of words and syllable lattice bi-grams is applied to combine multiple information for SDR.

2. SEMANTIC CONTEXT INFERENCE USING TERM ASSOCIATION MATRICES FOR SDR

The semantic context inference is used to explore spoken term association and generate an effective representation of queries and documents. Figure 1 illustrates a diagram of the overall system.

2.1. Spoken Term Representation and Speech Recognition

According to the idea of bag-of-words, each term v_k in a document d can be represented as an indexing vector $\mathbf{v}_d = [v_1^d, v_2^d, ..., v_K^d]$. K is the dimension of the indexing term vector. The phonetic structure of Mandarin Chinese is with 137 sub-syllables including 100 right-context-dependent INITIALs and 37 context-independent FINALs as basic units [24, 25]. Based on basic units, the number of Mandarin Chinese syllables is about 400 (base syllables) without considering tonal information. We extend base syllables to syllable bi-gram pairs for indexing. The *n*-best syllable lattice is used to deal with recognition error problems. To obtain the *n*-best syllable lattice, the two-pass speech decoding strategy is used for speech recognition [26]. First, the syllable-based speech recognizer transcribes the speech into a sequence of ranked *n*-best syllable candidates. The one-stage search algorithm based on the frame-based decoding method is used to obtain the *n*-best syllable candidates. The Viterbi parallel backtracking algorithm is used to determine the most likely syllable boundaries [24]. Second, each syllable segment is re-recognized to generate the *n*-best syllable candidates according to the determined syllable boundaries. A word lattice is constructed in the second pass. The dynamic programming algorithm is applied to search the best

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

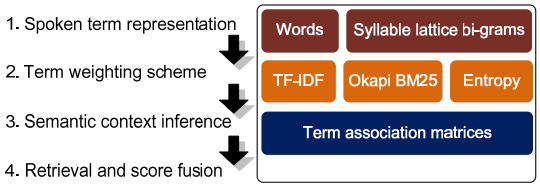

Fig. 1. Diagram of the overall system.

word sequence.

2.2. Term Weighting Scheme

Because of imperfect speech recognition results, not all of the recognized terms are meaningful. The purpose of term weighting schemes is used to eliminate those noisy terms and emphasize semantic terms. We investigate three term weighting schemes and compare the effect of estimating term association.

TF-IDF: One popular term weighting technique is TF-IDF which has been successfully applied in many applications such as spoken document summarization and extraction [12, 27, 28]. TF-IDF is calculated as follows:

$$\mathbf{v}_d(v_k) = A(v_k^d) \times B(v_k) \ . \tag{1}$$

The term frequency (TF) weighting of a term v_k^d is estimated by

$$A(v_k^d) = \frac{tf(v_k^d)+1}{n_d} \ , \tag{2}$$

where $tf(v_k^d)$ represents the number of occurrences of a term v_k^d in the document d. $n_d = \sum_{k=1}^{K} tf(v_k^d)$ is a normalization which indicates the number of terms in d. $B(v_k)$ is the inverse document frequency (IDF) of a term v_k estimated by

$$B(v_k) = \log(\frac{D}{df(v_k)+1}) \ , \tag{3}$$

where $df(v_k)$ means the number of documents that contain at least one occurrence of the term v_k in the database. The benefit of TF-IDF is to provide useful information about how important a term is to a document.

Okapi BM25: Robertson et al. proposed Okapi BM25 in 1994 [22] like TF-IDF. Okapi BM25 uses the same IDF $B(v_k)$, but takes into account the document length for the computation of TF. Okapi BM25 is then defined by

$$\mathbf{v}_d(v_k) = O(v_k^d) \times B(v_k) \ , \tag{4}$$

$$O(v_k^d) = \frac{tf(v_k^d) \times (\kappa+1)}{tf(v_k^d) + \kappa \times (1-b+b \times nl(d))} \ , \tag{5}$$

where $nl(d) = dl(d) / dl_{avg}$ is the normalization document length. $dl(d)$ shows the length of the document d. The constant dl_{avg} means the average document length in the database. κ and b are empirically selected parameters. We choose $(\kappa, b) = (7,\ 0.75)$ based on experiments.

Entropy: Shannon's entropy indicates that the information derivable from outcome x_t depends on its probability [29]. A high probability means low information due to the well expected outcome. The amount of information is defined as $I(x_t) = \log(1 / P(x_t))$ which represents uncertainty in the probabilistic framework. X is a discrete random variable and from a finite set of observations x_t. One of important properties of an information source is the entropy $H(X)$ defined as the average information [21],

$$H(X) = E[I(X)] = \sum_T P(x_t) I(x_t) = \sum_T P(x_t) \log \frac{1}{P(x_t)} \ . \tag{6}$$

$H(X)$ is the amount of information required to specify what kind of x_t has occurred on average. The entropy is investigated for the term weighting in this study. According to TF-IDF, TF is regarded as the weighting of local term frequencies while IDF is regarded as the weighting of global term frequencies. Instead of the conventional IDF, the new document weight Q is inspired by the definition of entropy:

$$Q(v_k) = -q(v_k) \log q(v_k) \ , \tag{7}$$

$$q(v_k) = \frac{df(v_k)+1}{D} \ . \tag{8}$$

The final product of the local term frequency $A(v_k^d)$ and the new document weight Q is estimated as follows:

$$\mathbf{v}(v_k^d) = A(v_k^d) \times Q(v_k) \ . \tag{9}$$

We further apply term weighting schemes of TF-IDF, Okapi BM25, and entropy for estimating term association matrix.

2.3. Term Association Matrix for SCI

To construct the term association matrix for semantic context inference, the document-by-term matrix $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_D]$ is derived from the database. $\mathbf{v}_d$ means the indexing vector which is estimated by the term weighting scheme. D denotes the total number of spoken documents. SCI starts with the covariance estimation of the term-by-term matrix, $\mathbf{V}^T\mathbf{V} = \mathbf{W}$, while T means the matrix transposition. $\mathbf{W}$ is a symmetric matrix used to describe co-relations between terms through a collection of documents. Since $\mathbf{W}$ is a symmetric matrix, the eigen-decomposition $\mathbf{W} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ is used to find the optimal projection and explore term association patterns. $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_K]$ is the singular matrix with the orthogonal characteristic. $\mathbf{\Lambda} = diag(\lambda_1, \lambda_2, ..., \lambda_K)$ is the diagonal matrix whose nonnegative entries are singular K values in a descending order, i.e., $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_K > 0$. The eigen-decomposition is applied to select major factors according to a threshold α,

$$\sum_{r=1}^{R} \lambda_r \ / \sum_{k=1}^{K} \lambda_k \ \geq \alpha \ . \tag{10}$$

A α is empirically set to select eigenvectors $\tilde{\mathbf{U}}_{(K \times R)}$ with the first R dimensions, where $R \leq K$ denotes projected dimensions of

Table 1. mAP of comparing TF-IDF, LSA, SCI and selecting major factors on MATBN

TF-IDF	69.56%				
α	0.6	0.7	0.8	0.9	1
LSA	70.19%	71.74%	72.32%	71.77%	69.81%
SCI	73.36%	74.28%	74.24%	73.53%	72.07%

Table 2. mAP of TF-IDF and SCI indexing with ASR and manual transcription of TDT2 and MATBN

Type	TF-IDF		SCI	
Corpus	TDT2	MATBN	TDT2	MATBN
ASR transcription	71.92%	69.56%	74.55%	74.28%
Manual transcription	89.76%	88.76%	90.79%	90.48%

the original term vector in the eigenspace [30, 31]. The associated eigenvalues allow us to rank the eigenvectors $\tilde{\mathbf{U}}$ based on their usefulness in characterizing semantic relations between terms.

The term association matrix is reconstructed by $\tilde{\mathbf{U}}_{(K\times R)}\mathbf{\Lambda}_{(R\times R)}\tilde{\mathbf{U}}^{T}_{(K\times R)} = \tilde{\mathbf{W}}$, which shows a more robust representation of co-occurrences and semantic relations among terms. In indexing, a document d is represented as an binary weighting vector $\hat{\mathbf{v}}_d = [\hat{v}_1, \hat{v}_2, ..., \hat{v}_K]$. The value of term $\hat{v}_k$ is 1 if the k-th term occurs in the document d and 0 otherwise. The semantic context inference vector $\overline{\mathbf{v}}_d$ is then estimated with the term association matrix $\tilde{\mathbf{W}}$,

$$\tilde{\mathbf{W}}\hat{\mathbf{v}}_d = \hat{v}_1^d\tilde{\mathbf{w}}_1 + \hat{v}_2^d\tilde{\mathbf{w}}_2 + ... + \hat{v}_K^d\tilde{\mathbf{w}}_K = \overline{\mathbf{v}}_d. \quad (11)$$

$$\begin{bmatrix} \tilde{w}_{11} & \cdots & \tilde{w}_{1K} \\ \vdots & \ddots & \vdots \\ \tilde{w}_{K1} & \cdots & \tilde{w}_{KK} \end{bmatrix}\begin{bmatrix} \hat{v}_1^d \\ \vdots \\ \hat{v}_K^d \end{bmatrix} = \begin{bmatrix} \overline{v}_1^d \\ \vdots \\ \overline{v}_K^d \end{bmatrix} \quad (12)$$

The semantic context inference vector $\overline{\mathbf{v}}_d$ is regarded as a re-weighed indexing vector by expanding indexing terms based on related terms.

Early, indexing by latent semantic analysis (LSA) took into account conceptual indexing by projecting the term vector into a lower-dimensional latent semantic analysis space [31, 32]. Both ideas of SCI and LSA are used to explore latent semantic information. However, the semantic representation is explicit in SCI but implicit in LSA. For instance, the eigenvector $\tilde{\mathbf{U}}$ is treated as the transform basis in the traditional indexing by LSA. The indexing vector $\mathbf{v}_d$ is projected into the new indexing vector $\breve{\mathbf{v}}_d$ by $\tilde{\mathbf{U}}\mathbf{v}_d = \breve{\mathbf{v}}_d$ in LSA. The proposed SCI allows the similarity measure between queries and documents, considering not only the original terms but also the semantic terms. We explore the use of the tem association matrix $\tilde{\mathbf{W}}$ instead of the singular matrix $\tilde{\mathbf{U}}$, makes the work different. In this study, the latent semantic information is embedded or represented in term co-relations. Such term co-relations are used for SCI and to alleviate the transcription errors in spoken document retrieval.

2.4. Retrieval and Score Fusion

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 vector representation for a document [33]. The cosine measure is applied to estimate the similarity between query $\overline{\mathbf{v}}_q$ and spoken document $\overline{\mathbf{v}}_d$ vectors,

$$S_{WD,SL}(\overline{\mathbf{v}}_q, \overline{\mathbf{v}}_d) = \frac{\overline{\mathbf{v}}_q\overline{\mathbf{v}}_d^T}{\|\overline{\mathbf{v}}_q\|\cdot\|\overline{\mathbf{v}}_d\|} = \frac{\sum_{k=1}^{K}\overline{\mathbf{v}}_{q,k}\times\overline{\mathbf{v}}_{d,k}}{\sqrt{\sum_{k=1}^{K}\overline{\mathbf{v}}_{q,k}^2}\times\sqrt{\sum_{k=1}^{K}\overline{\mathbf{v}}_{d,k}^2}}, \quad (13)$$

Retrieval results are ranked according to similarities. Both word and syllable indexing are integrated by using the proposed semantic context inference. $S_{WD}(\overline{\mathbf{v}}_q, \overline{\mathbf{v}}_d)$ and $S_{SL}(\overline{\mathbf{v}}_q, \overline{\mathbf{v}}_d)$ represent similarity scores for words and syllable lattice bi-grams.

$$S(\overline{\mathbf{v}}_q, \overline{\mathbf{v}}_d) = \beta\cdot S_{WD}(\overline{\mathbf{v}}_q, \overline{\mathbf{v}}_d) + (1-\beta)\cdot S_{SL}(\overline{\mathbf{v}}_q, \overline{\mathbf{v}}_d), \quad (14)$$

The best fusion factor $\hat{\beta} = 0.5$ is experimentally determined.

Table 3. mAP of TF-IDF and SCI indexing with different word error rate (%WER) on MATBN

%WER	55	45	30	0
TF-IDF	54.25%	61.40%	63.52%	88.76%
SCI	57.86%	65.27%	67.80%	90.48%

3. EXPERIMENTS

Experiments were reported based on Mandarin Chinese broadcast news MATBN database. MATBN was collected by Academia Sinica Taiwan which contained a total of 198 hours of broadcast news [28]. 1,550 anchor news stories ranging over three years were extracted for experiments. The average document length of MATBN is 51.85 words. The word error rate (WER) is 21.05%. We applied two standard evaluation metrics to evaluate the retrieval performance including F-score and mean average precision (mAP) [29]. 164 keyword queries (from two to four Chinese characters) were used. The average length of queries is 3.02 Chinese characters. There are 15.71 relevant spoken documents in MATBN database.

3.1. Semantic context inference for SDR

Experimental results were obtained with MATBN using indexing of TF-IDF, LSA, and the proposed SCI as shown in Table 1. The popular term vector indexing TF-IDF was used as the baseline which achieved 69.56% mAP. To remove the noisy factors in the eigen-decomposition, we set a threshold α (see Eq. (10)) for keeping the major factors. A α of higher value indicates that more eigenvectors are used for latent semantic analysis as well as the reconstruction of the term association matrix. Experiments showed that the complete LSA space did not give as good performance as the dimension-reduced LSA space. The best results can be achieved when thresholds of LSA of 80% and SCI of 70% were selected separately. We applied the best setting on the following experiments. These results confirmed that a better performance can be achieved by removing the noisy factors. In Table 1, results also confirm that SCI outperformed both LSA and baseline TF-IDF indexing methods.

To evaluate the effect of semantic context inference, the proposed approach for spoken document retrieval was applied on TDT2 and MATBN corpus using both ASR transcription and

Table 4. F-score of the top 5, 10, 15, 20, and 25 retrieved documents for different SCI indexing

Top-n	5	10	15	20	25
SCI_SL	0.36	0.50	0.53	0.53	0.51
SCI_WD	0.48	0.63	0.64	0.60	0.55
SCI_Fusion	0.51	0.66	0.68	0.64	0.60

Table 5. Evaluation of TF-IDF, BM25 and Entropy weighting schemes on the conventional word indexing and SCI indexing

Type	Word		SCI	
Metrics	mAP	F-score	mAP	F-score
TF-IDF	69.56%	0.62	74.28%	0.64
Okapi BM25	69.97%	0.62	76.33%	0.65
Entropy	72.07%	0.63	72.85%	0.63

perfect text (manual transcription). 2,112 Voice of America (VOA) Mandarin Chinese news broadcast from Topic Detection and Tracking collection (TDT2) were used as a comparing or evaluation set. The WER of TDT2 was 24.51%. mAP results shown in Table 2 indicated that consistent improvements have been obtained on TDT2 and MATBN based on indexing of SCI compared with TF-IDF. Conventionally, the upper-bound of SDR was the indexing by perfect text which was evaluated as the reference. Due to imperfect speech recognition, there was still a gap (about 15%~20% mAP) between indexing using ASR and manual transcription. Normally, the value of R is smaller than K according to a threshold α . In addition, speech recognition performance becomes less predictable under adverse acoustic environments. To study the impact of speech recognition accuracy variance on SCI, we used different settings of speech recognition system to provide the transcription of different WER as shown in Table 3. SCI consistently outperforms TF-IDF over a range of different WER. It is worthwhile to note that the computation of eigen-decomposition spends a lot of time when we apply a larger vocabulary size. Since we aim at the approach of SCI for improving SDR, the vocabulary size K was 6,880 in this study.

3.2. Fusion of Words and Syllable Lattice Bi-Grams

Besides word information, syllables can also be used for SCI indexing by building the syllable association matrix. Figure 2 illustrated mAP results of indexing by words and syllables on MATBN. First, we compared indexing by using syllables (TF-IDF_SY) and syllable lattice bi-grams (TF-IDF_SL), and found syllable lattice bi-grams outperform syllables only. Since syllables showed less context information than words, the word indexing (TF-IDF_WD) obviously outperformed syllable methods (TF-IDF_SL and TF-IDF_SY). However, the semantic context inference made consistent improvements between TF-IDF_SL and SCI_SL.

F-score was employed to words and syllable lattice bi-grams for the documents ranked at top 5, 10, 15, 20, and 25 in Table 4. The best F-score was achieved on the top 15 retrieved documents. We conducted a score fusion considering words and syllable lattice bi-grams. The combination of SCI_SL and SCI_WD further improved F-score from 0.64 of SCI_WD to 0.68 of SCI_Fusion, which showed words and syllables were complementary for retrieval.

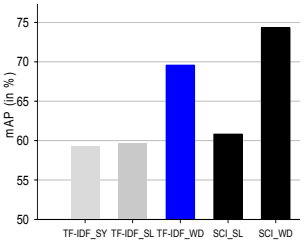

Fig. 2. mAP results on different indexing.

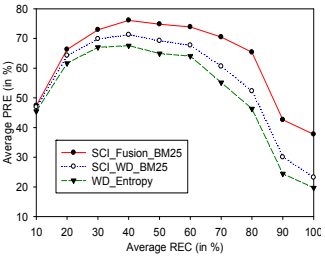

Fig. 3. Recall and precision plots of different indexing.

3.3. Evaluation of Term Weighting Schemes

On above experiments, both word indexing and semantic context inference only adopt the TF-IDF term weighting scheme. Table 5 showed the comparison results of term weighting schemes of TF-IDF, Okapi BM25, and entropy on MATBN. Results were evaluated by mAP and F-score. In the conventional word indexing, the entropy term weighting scheme outperformed TF-IDF and Okapi BM25 as shown on the left of Table 5. Further gains with the proposed SCI have been observed by applying three weighting schemes as shown on the right of Table 5. Interesting observation was that Okapi BM25 with SCI indicated the best performance compared with others.

According to these findings, Figure 3 illustrated the precision versus recall plot for the step-by-step improvements. We are able to see an incremental increasing performance between indexing approaches, word entropy indexing (WD_Entropy), indexing by Okapi BM25 with semantic context inference (SCI_WD_BM25), and the score fusion of words and syllable lattice bi-grams with semantic context inference (SCI_fusion_BM25). The recall and precision plots showed the significant precision gain at SCI_Fusion_BM25 and others. The analytical results indicated and SCI_WD_BM25 is significantly better than WD_Entropy. SCI_fusion_BM25 brought an extra improvement than individual indexing of words or syllable lattice bi-grams.

4. CONCLUSION

In this study, we present a novel approach using the term association matrix for semantic context inference and offer a good treatment of ASR problem on spoken document retrieval. Our strategies are to explore the latent semantic information and extend semantic related terms to speech indexing. The semantic context inference vector can be regarded as a re-weighting indexing vector and finding semantic relation in context to overcome speech recognition errors. Our spoken document retrieval experiments indicate that the proposed semantic context inference outperforms the conventional TF-IDF term vector and LSA indexing approaches, and works especially well for speech recognition transcription with errors. We explore words, syllable lattice bi-grams, and three term weighting schemes for spoken document retrieval. It can be found that the entropy term weighting scheme is useful in conventional word indexing. The Okapi BM25 term weighting scheme with SCI indexing shows significant gain compared with the conventional indexing. The information of word and syllable semantic context inference is complementary. Their score fusion indicates the best performance.

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. Speech and Audio Processing*, vol. 12, no. 4, pp. 420–435, 2004.

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

[3] M. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based Multimedia Information Retrieval: State of the Art and Challenges," *ACM Trans. Multimedia Computing, Communications, and Applications*, vol. 2, no. 1, pp. 1–19, 2006.

[4] 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.

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

[6] K. Ng, "Subword-Based Approaches for Spoken Document Retrieval," Ph.D. dissertation, Mass. Inst. Technol., Cambridge, MA, 2000.

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

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

[9] S. Parlak and M. Saraçlar, "Performance Analysis and Improvement of Turkish Broadcast News Retrieval," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 731–741, 2012.

[10] X. He and L. Deng, "Speech-Centric Information Processing: An Optimization-Oriented Approach," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1116–1135, 2013.

[11] 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.

[12] B. Chen, H.-M. Wang, and L.-S. Lee, "Discriminating capabilities of syllable based features and approaches of utilizing them for voice retrieval of speech information in Mandarin Chinese," *IEEE Trans. Speech and Audio Processing*, vol. 10, no. 5, pp. 303–314, 2002.

[13] L.-S. Lee and B. Chen, "Spoken document understanding and organization," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 42–60, 2005.

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

[15] A. Singhal and F. Pereira, "Document Expansion for Speech Retrieval," in *Proc. ACM SIGIR Conf.*, pp. 34–41, 1999.

[16] 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.

[17] T. Tao and C. Zhai, "Regularized estimation of mixture models for robust pseudo-relevance feedback," in *Proc. ACM SIGIR Conf.*, 2006.

[18] X. Wang, H. Fang, and C. Zhai, "A study of methods for negative relevance feedback," in *Proc. ACM SIGIR Conf.*, 2008.

[19] D. Yu, Y.-C. Ju, Y.-Y. Wang, G. Zweig, and A. Acero, "Automated Directory Assistance System - from Theory to Practice," in *Proc. Interspeech*, pp. 2709– 2712, 2007.

[20] C.-L. Huang, B. Ma, H. Li, and C.-.H Wu, "Speech Indexing Using Semantic Context Inference", in *Proc. Interspeech*, pp. 717–720, 2011.

[21] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.

[22] S. E. Robertson, S. Walker, S. Jones, M.M. Hancock-Beaulieu, and M. Gatford, "Okapi at TREC-3," in *Proc. of the third Text REtrieval Conference (TREC-3)*, pp. 109–126, 1994.

[23] C.-L. Huang, C. Hori, and H. Kashioka, "Semantic Inference based on Neural Probabilistic Language Modeling for Speech Indexing", in *Proc. ICASSP*, pp. 8480–8484, 2013.

[24] C.-H. Wu and Y.-J. Chen, "Multi-keyword spotting of telephone speech using a fuzzy search algorithm and keyword-driven two-Level CBSM," *Speech Communication*, vol. 33, pp. 197–212, 2001.

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

[26] H. Ney and S. Ortmanns, "Progress in Dynamic Programming Search for LVCSR," *Proceedings of The IEEE*, vol. 88, no. 8, pp. 1224–1240, 2000.

[27] C. Hori and S. Furui, "A new approach to automatic speech summarization," *IEEE Trans. Multimedia*, vol. 5, no. 3, pp. 368–378, 2003.

[28] C.-H. Wu, C.-H. Hsieh, and C.-L. Huang, "Speech Sentence Compression Based on Speech Segment Extraction and Concatenation," *IEEE Trans. Multimedia*, vol. 9, no. 2, pp. 434–438, 2007.

[29] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423/623–656, 1948.

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

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

[32] 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.

[33] 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.