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 speech recognition errors may cause undesired inverse document frequency (TF-IDF) weighted vector space spoken document retrieval [15]–[18].

And relevance feedback, provide good solutions on problems of conventional text-based IR methods, such as document expansion 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 speech recognition errors may cause undesired inverse document frequency (TF-IDF) weighted vector space spoken document retrieval [15]–[18].

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. 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.

2.1. Spoken Term Representation and Speech Recognition

According to the idea of bag-of-words, each term \( v_i \) in a document \( d \) can be represented as an indexing vector \( \mathbf{v}_d = [v_{i_1}^d, v_{i_2}^d, ..., v_{i_K}^d] \). \( K \) 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
Entropy: Shannon’s entropy indicates that the information derivable from outcome \( x \) depends on its probability [29]. A high probability means low information due to the well expected outcome. The amount of information is defined as \( I(x) = \log(1/P(x)) \) which represents uncertainty in the probabilistic framework. \( X \) is a discrete random variable and from a finite set of observations \( x \). One of important properties of an information source is the entropy \( H(X) \) defined as the average information [21],

\[
H(X) = E[I(X)] = \sum_{x} P(x)I(x) = \sum_{x} P(x)\log\frac{1}{P(x)} \tag{6}
\]

\( H(X) \) is the amount of information required to specify what kind of \( x \) 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:

\[
Q(v_i) = -q(v_i)\log q(v_i) \tag{7}
\]

\[
q(v_i) = \frac{df(v_i) + 1}{D} \tag{8}
\]

The final product of the local term frequency \( A(v_i') \) and the new document weight \( Q \) is estimated as follows:

\[
v(v_i') = A(v_i') \times Q(v_i) \tag{9}
\]

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 \( V = [v_1, v_2, ..., v_n] \) is derived from the database. \( v_i \) 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, \( V^T V = W \), while \( T \) means the matrix transposition. \( W \) is a symmetric matrix used to describe co-relations between terms through a collection of documents. Since \( W \) is a symmetric matrix, the eigen-decomposition \( W = U\Lambda U^T \) is used to find the optimal projection and explore term association patterns. \( U = [u_1, u_2, ..., u_k] \) is the singular matrix with the orthogonal characteristic. \( \Lambda = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_K) \) 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 \( \alpha \),

\[
\sum_{i=1}^{K} \lambda_i / \sum_{i=1}^{K} \lambda_i \geq \alpha \tag{10}
\]

\( \alpha \) is empirically set to select eigenvectors \( \hat{U}_{(K \times K)} \) with the first \( R \) dimensions, where \( R \leq K \) denotes projected dimensions of
the original term vector in the eigenspace [30, 31]. The associated eigenvalues allow us to rank the eigenvectors \( \mathbf{U} \) based on their usefulness in characterizing semantic relations between terms.

The term association matrix is reconstructed by \( \mathbf{U}^T \mathbf{X} \mathbf{U} = \mathbf{W} \), 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 \( \mathbf{v}_d = [v_{d1}, v_{d2}, \ldots, v_{dk}] \). The value of term \( v_{dk} \) is 1 if the \( k \)-th term occurs in the document \( d \) and 0 otherwise. The semantic context inference vector \( \mathbf{v}_d \) is then estimated with the term association matrix \( \mathbf{W} \),

\[
\mathbf{Wv}_d = \mathbf{v}_1^T \mathbf{w}_1 + \mathbf{v}_2^T \mathbf{w}_2 + \ldots + \mathbf{v}_k^T \mathbf{w}_k = \mathbf{v}_d. \tag{11}
\]

The semantic context inference vector \( \mathbf{v}_d \) is regarded as a reweighted 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 \( \mathbf{U} \) 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 \( \mathbf{v}_d \) by \( \mathbf{Uv}_d = \mathbf{v}_d \) 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 term association matrix \( \mathbf{W} \) instead of the singular matrix \( \mathbf{U} \), 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 \( \mathbf{v}_q \) and spoken document \( \mathbf{v}_d \),

\[
S_{\text{WD}}(\mathbf{v}_q, \mathbf{v}_d) = \frac{\mathbf{v}_q^T \mathbf{v}_d}{\|\mathbf{v}_q\| \|\mathbf{v}_d\|}. \tag{13}
\]

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

\[
S(\mathbf{v}_q, \mathbf{v}_d) = \beta S_{\text{WD}}(\mathbf{v}_q, \mathbf{v}_d) + (1 - \beta) S_{\text{SL}}(\mathbf{v}_q, \mathbf{v}_d), \tag{14}
\]

The best fusion factor \( \beta = 0.5 \) is experimentally determined.

### 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 \( \alpha \) (see Eq. (10)) for keeping the major factors. A \( \alpha \) 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 1. mAP of comparing TF-IDF, LSA, SCI and selecting major factors on MATBN

<table>
<thead>
<tr>
<th>Type</th>
<th>Corpus</th>
<th>TDT2</th>
<th>MATBN</th>
<th>TDT2</th>
<th>MATBN</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF-IDF</td>
<td></td>
<td>69.56%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \alpha )</td>
<td>0.6</td>
<td>0.7</td>
<td>0.8</td>
<td>0.9</td>
<td>1</td>
</tr>
<tr>
<td>LSA</td>
<td></td>
<td>70.19%</td>
<td>71.74%</td>
<td>72.32%</td>
<td>71.77%</td>
</tr>
<tr>
<td>SCI</td>
<td></td>
<td>73.36%</td>
<td>74.28%</td>
<td>74.24%</td>
<td>73.53%</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Type</th>
<th>TF-IDF</th>
<th>SCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus</td>
<td>TDT2</td>
<td>MATBN</td>
</tr>
<tr>
<td>ASR transcription</td>
<td>71.92%</td>
<td>65.27%</td>
</tr>
<tr>
<td>Manual transcription</td>
<td>89.76%</td>
<td>74.28%</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>%WER</th>
<th>55</th>
<th>45</th>
<th>30</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF-IDF</td>
<td>54.25%</td>
<td>61.40%</td>
<td>63.52%</td>
<td>88.76%</td>
</tr>
<tr>
<td>SCI</td>
<td>57.86%</td>
<td>65.27%</td>
<td>67.80%</td>
<td>90.48%</td>
</tr>
</tbody>
</table>
Table 4. F-score of the top 5, 10, 15, 20, and 25 retrieved documents for different SCI indexing

<table>
<thead>
<tr>
<th>Top-n</th>
<th>SCI_SL</th>
<th>SCI_WD</th>
<th>SCI Fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.36</td>
<td>0.48</td>
<td>0.51</td>
</tr>
<tr>
<td>10</td>
<td>0.50</td>
<td>0.63</td>
<td>0.66</td>
</tr>
<tr>
<td>15</td>
<td>0.53</td>
<td>0.64</td>
<td>0.68</td>
</tr>
<tr>
<td>20</td>
<td>0.53</td>
<td>0.60</td>
<td>0.64</td>
</tr>
<tr>
<td>25</td>
<td>0.51</td>
<td>0.60</td>
<td>0.66</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Type</th>
<th>Word</th>
<th>SCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metrics</td>
<td>mAP</td>
<td>F-score</td>
</tr>
<tr>
<td>TF-IDF</td>
<td>69.56%</td>
<td>0.62</td>
</tr>
<tr>
<td>Okapi BM25</td>
<td>69.97%</td>
<td>0.62</td>
</tr>
<tr>
<td>Entropy</td>
<td>72.07%</td>
<td>0.63</td>
</tr>
</tbody>
</table>

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


