IMPROVED SEMANTIC RETRIEVAL OF SPOKEN CONTENT BY LANGUAGE MODELS ENHANCED WITH ACOUSTIC SIMILARITY GRAPH

Hung-yi Lee[1], Tsung-Hsien Wen[2], and Lin-Shan Lee[1,2]

[1]Graduate Institute of Communication Engineering, National Taiwan University
[2]Graduate Institute of Electrical Engineering, National Taiwan University

ABSTRACT

Retrieving objects semantically related to the query has been widely studied in text information retrieval. However, when applying the text-based techniques on spoken content, the inevitable recognition errors may seriously degrade the performance. In this paper, we propose to enhance the expected term frequencies estimated from spoken content by acoustic similarity graphs. For each word in the lexicon, a graph is constructed describing acoustic similarity among spoken segments in the archive. Score propagation over the graph helps in estimating the expected term frequencies. The enhanced expected term frequencies can be used in the language modeling retrieval approach, as well as semantic retrieval techniques such as the document expansion based on latent semantic analysis, and query expansion considering both words and latent topic information. Preliminary experiments performed on Mandarin broadcast news indicated that improved performance were achievable under different conditions.

Index Terms— Random Walk, Document Expansion, Query Expansion, Latent Semantic Analysis

1. INTRODUCTION

Spoken content retrieval will be very important to retrieve and browse multimedia content over the Internet. Substantial effort has been made in spoken content retrieval in recent years, and many successful techniques have been developed [1]. Most works in spoken content retrieval nowadays focused on spoken term detection, for which the goal is simply returning spoken segments that include the query terms. This is insufficient because users naturally prefer that the technologies can return all the objects the users are looking for, regardless of whether the query terms are included or not. This led to extensive recent works on semantic retrieval of spoken content [2, 3, 4, 5, 6, 7]. The techniques for semantic retrieval such as latent semantic analysis and query/document expansion [8, 9] have been widely studied in text information retrieval. Taking ASR transcriptions as the text, these techniques developed for text information retrieval can be directly applied for spoken content retrieval. However, since these techniques were developed for text without errors, the inevitable recognition errors in ASR transcriptions may seriously degrade the performance. On the other hand, it has been found that enhancing the term frequency estimation for spoken documents based on context information may improve the performance of both approaches of language modeling retrieval and query expansion [2].

On the other hand, it may be assumed that the acoustic feature sequences representing different occurrences of the same term may be relatively similar in some way, while very different feature sequences very possibly imply different terms. It has actually been found that acoustic feature similarity between spoken segments used in graph-based re-ranking approach is very helpful in the spoken term detection task [10, 11]. In this approach, given a user query, the retrieval engine first searches through the lattices for the spoken content to produce a first-pass returned list of spoken segments ranked according to the initial scores. Then a graph is constructed for the first-pass retrieved spoken segments, in which each node represents a spoken segment, and the edge weights are the acoustic similarities between the feature sequences corresponding to the query hypotheses in the lattices. Based on the concept that segments strongly connected to more segments with higher scores on the graph may also have higher probabilities of containing the query terms, the scores for the spoken segments thus propagate over the graph, and the detection results are re-ranked accordingly.

In this paper, in order to have more robust semantic retrieval for content, we propose to enhance the expected term frequencies derived from the lattices by acoustic similarity graphs. Experimental results showed that the graph-enhanced expected term frequencies can not only improve the performance of the conventional language modeling retrieval approach, but boost the performance of the document and query expansion techniques for semantic retrieval.

2. LANGUAGE MODELING FOR SPOKEN CONTENT RETRIEVAL

Here we start with the conventional language modeling approach for retrieval in Section 2.1, explain how it is extended to spoken content with lattices in Section 2.2, and presents the proposed graph-based enhancement approach in Section 2.3.

2.1. Conventional Language Modeling Approach for Retrieval

Language modeling approach has been shown to be very effective for information retrieval not only for text, but for spoken content as well [12]. The basic idea for this approach is that both the query Q and the document d can be respectively represented as language models θ_Q and θ_d, and the relevance score function $S(Q, d)$ used for ranking the documents d for query Q is the inverse of the KL divergence between θ_Q and θ_d:

$$S(Q, d) = -KL(\theta_Q|\theta_d). \quad (1)$$

That is, the documents whose language models are more similar to the query language model are more likely to be relevant. In this way, the problem of retrieval is reduced to the estimation of the language models for the queries and documents. To simplify the presentation below, here we assume only word unigram language models are used for documents and queries, although the proposed approach is

not limited to this case. Hence, all language models below refer to unigrams.

Usually the word unigram language model θ_Q for the query Q is estimated based on the words in Q in (2), assuming Q is in text form,

$$P(w|\theta_Q) = \frac{N(w,Q)}{|Q|}, \tag{2}$$

where $P(w|\theta_Q)$ is the probability of generating the word w from the model θ_Q, $N(w,Q)$ the occurrence counts of the word w in Q, and $|Q|$ the total number of words in the query. Now because the documents considered here are spoken, when the documents are transcribed into 1-best transcriptions, the way to estimate the document language model θ_d is exactly the same as the document language model that for text, or same as (2) except all Q's in (2) replaced by the document d. However, as there are inevitable relatively high rate of errors in the 1-best transcriptions, θ_d thus estimated may be very different from the true word distribution of the spoken document. In the next subsection, the way to construct document language models θ_d from lattices is represented.

2.2. Document Model for Spoken Content

Each document d in the spoken archive to be retrieved through is first divided into a total of N spoken segments $\{x_1, \cdots, x_n, \cdots, x_N\}$, each transcribed into a lattice. We first compute the expected counts of each word w in the lattice of each segment x_n:

$$E[w|x_n] = \sum_{u \in W(x_n)} N(w,u) P(u|x_n), \tag{3}$$

where u is a word sequence in the lattice, $W(x_n)$ the set of all possible word sequences in the lattice for x_n, $N(w,u)$ the occurrence count of the word w in u, and $P(u|x_n)$ the posterior probability of the word sequence u derived from the acoustic and language models.

The language model $\theta^l_{x_n}$ for each spoken segment x_n is then estimated in (4) [1],

$$P(w|\theta^l_{x_n}) = \frac{E[w|x_n]}{L_{x_n}}, \tag{4}$$

where L_{x_n} is the expected length for segment x_n,

$$L_{x_n} = \sum_{u \in W(x_n)} |u| P(u|x_n), \tag{5}$$

in which $|u|$ is the number of word arcs in u. All the language models $\theta^l_{x_n}$ for all spoken segments x_n in the document d are then summed weighted by their expected lengths L_{x_n} to form a document model θ^l_d directly from the lattice in (6),

$$P(w|\theta^l_d) = \frac{\sum_{n=1}^{N} L_{x_n} P(w|\theta^l_{x_n})}{L_d}, \tag{6}$$

where L_d is the expected document length $L_d = \sum_{n=1}^{N} L_{x_n}$.

Now θ^l_d is interpolated with a background language model θ_b trained from all the spoken documents in the spoken archive $\mathcal{C}$ to form a smoothed model $\bar{\theta}^l_d$ in (7),

$$P(w|\bar{\theta}^l_d) = a_d P(w|\theta^l_d) + (1 - a_d) P(w|\theta_b), \tag{7}$$

[1] The superscript ^{l} indicates that the language models are directly derived from the lattices.

where

$$P(w|\theta_b) = \frac{\sum_{d \in \mathcal{C}} L_d P(w|\theta^l_d)}{\sum_{d \in \mathcal{C}} L_d} \tag{8}$$

is the probability of observing the word w in the whole archive $\mathcal{C}$. $a_d = \frac{L_d}{L_d + a}$ in (7) is a document dependent interpolation weight, and a is a parameter to be set. In this way, the background model has more influence on the shorter documents [13]. This smoothed model $\bar{\theta}^l_d$ in (7) is used for θ_d in (1) for ranking [12].

2.3. Graph-enhanced Document Model based on Acoustic Similarity

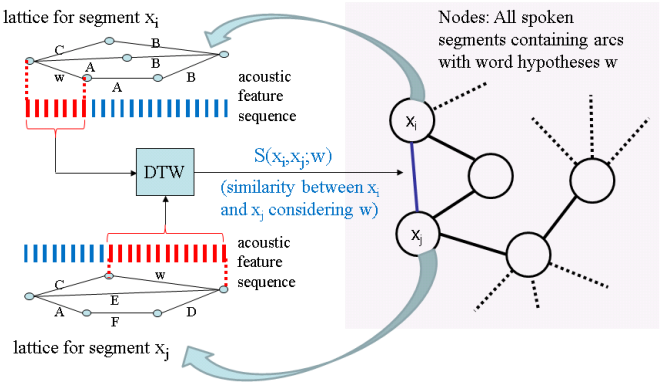

Fig. 1: *The graph constructed for enhancing the expected counts for word w based on the acoustic similarity. Each node in the graph represents a spoken segment containing the word arc w in its lattice, and the edge weights represent the acoustic similarities between the nodes considering the word w.*

Although the document models derived above in (7) directly from the lattices may be better than the ones from the 1-best transcriptions, they unavoidably suffer from the recognition errors and noisy word hypotheses in the lattices. Considering an arc with word hypothesis w in the lattice of a spoken segment, if its corresponding acoustic feature sequence is similar to many other arcs also recognized as word hypothesis w in other spoken segments, this word hypothesis should be more reliable; otherwise it may be less reliable. This concept can be used to enhance the the expected counts $E[w|x]$ in (3) by a graph-based approach as proposed here.

For each word w in the lexicon, we first construct a graph using all the segments in the spoken archive to be retrieved through which contain arcs with word hypotheses w in the lattices, in which each node represents a spoken segment. Such a graph is shown in Fig. 1, and such a graph is produced and all following processes repeated for all words in the lexicon. Dynamic time warping (DTW) is performed between the acoustic feature sequences corresponding to the word hypotheses w in lattices for all segment pairs x_i and x_j on the graph [2]. This yields $d(x_i, x_j; w)$, the DTW distance between x_i and x_j considering the term w. The similarity between x_i and x_j considering w is then

$$S(x_i, x_j; w) = 1 - \frac{d(x_i, x_j; w) - d_{min}}{d_{max} - d_{min}}, \tag{9}$$

[2] If there are multiple arcs whose word hypotheses are w in a lattice, only the one with the highest posterior probability is considered. Although this assumption ignores the case that a word may occur many times in a spoken segment, it resulted in reasonable results in the following experiments.

where d_{max} and d_{min} are the largest and smallest values of $d(x_i, x_j; w)$ for all node pairs on the graph. Equation (9) simply normalizes the DTW distance and transforms it into a similarity score between 0 and 1. Only those node pairs with $S(x_i, x_j; w)$ exceeding a threshold are then connected with an edge with weight $S(x_i, x_j; w)$.

The expected counts $E[w|x_i]$ in (3) for word w in segment x_i can then be enhanced via score propagation on the graph [3],

$$E^g[w|x_i] = (1-\alpha)E[w|x_i] + \alpha \sum_{x_j \in A_i} E^g[w|x_j]\hat{S}(x_j, x_i; w), \tag{10}$$

where $E^g[w|x_i]$ is the graph-enhanced version, α an interpolation weight between 0 and 1, A_i the set of all segments x_j connected to x_i, and $\hat{S}(x_j, x_i; w)$ the normalized edge weight over all edges connected to node x_j on the graph:

$$\hat{S}(x_j, x_i; w) = \frac{S(x_j, x_i; w)}{\sum_{x_k \in A_j} S(x_j, x_k; w)}. \tag{11}$$

In (10) the graph-enhanced expected count $E^g[w|x_i]$ depends on two factors interpolated by α: the original expected count $E[w|x_i]$ in (3) (the first term on the right hand side of (10)) and the score propagation over the graph based on the normalized edge weights $\hat{S}(x_j, x_i; w)$ (the second term on the right hand side). Thus $E^g[w|x_i]$ would be larger if $E[w|x_i]$ is larger, or x_i is strongly connected to more other segments x_j with larger $E[w|x_j]$ on the graph. Equation (10) is actually a random walk problem on the graph, and the random walk theory guarantees that the score propagation over the graph converges and a set of unique solutions of $E^g[w|x_i]$ can be found by the power method [14]. The above process including graph construction and random walk is repeated for all words w in the lexicon off-line, so for all words w in the lexicon the enhanced expected counts $E^g[w|x]$ are available for all segments x in the spoken archive.

The graph-enhanced language model θ_x^g for segment x is then obtained by interpolating the original expected count $E[w|x]$ in (3) and $E^g[w|x]$ in (10) as in (12),

$$P(w|\theta_x^g) = \frac{E[w|x] + \mu E^g[w|x]}{L_x + \mu \sum_w E^g[w|x]}, \tag{12}$$

in which $E[w|x]$ and $E^g[w|x]$ are weighted summed with a parameter μ in the numerator, the summation in the denominator is over all words w in the lexicon, and the normalization with the denominator makes the graph-enhanced scores $P(w|\theta_x^g)$ probabilities of a language model. The language model $P(w|\theta_x^g)$ in (12) for all segments x in the document d are then aggregated to form the language model $P(w|\theta_d^g)$ in the same way as (6). θ_d^g is finally interpolated with the background language model to obtain a smoothed model $\bar{\theta}_d^g$ as in (7), which is in turn used in (1) for ranking.

3. DOCUMENT EXPANSION WITH PROBABILISTIC LATENT SEMANTIC ANALYSIS

The problem of retrieving documents semantically related to the query is that many of such documents may not necessarily contain the query term. For example, if the query is "airplane", but some relevant documents contain the term "aircraft" instead. These relevant documents have very small relevance score $S(Q, d)$ in (1) because the language models for the query and the document are very different if they are directly estimated from the term occurrence counts in the query and the document. This problem can be handled to some extent by incorporating some latent topic analysis approaches. With such approaches, the document with "aircraft" may be found to belong to a certain latent topic regarding "flying vehicle", so we can expand the document model with some terms related to "flying vehicle" (like "airplane") for better representation of the document.

One popular approach for latent topic analysis is the Probabilistic Latent Semantic Analysis (PLSA) [15], which is used in this work. Extension to other latent topic analysis approaches is certainly possible. PLSA uses a set of latent topic variables, $\{T_k, k = 1, 2, ..., K\}$, to characterize the "term-document" co-occurrence relationships in the archive. Given all the spoken documents in the archive, PLSA training yields $P(w|T_k)$, the probability of observing a word w given the latent topic T_k, and $P(T_k|d)$, the mixture weight of topic T_k given the document d. Hence, based on such latent topic analysis the probability of observing a word w given document d can be parameterized by [4]

$$P_{lt}(w|d) = \sum_{k=1}^{K} P(w|T_k)P(T_k|d). \tag{13}$$

The parameters $P(w|T_k)$ and $P(T_k|d)$ is learned using EM algorithm via maximizing the following objective function:

$$L = \sum_{d \in \mathcal{C}} \sum_w P(w|\theta_d) log P_{lt}(w|d), \tag{14}$$

where θ_d can be either θ_d^l in (6) of Section 2.2 or θ_d^g in Section 2.3 for spoken content considered here. Equation (14) can be understood as searching for a set of parameters $P(w|T_k)$ and $P(T_k|d)$ minimizing the KL divergence between the document model and the term distribution in (13) obtained from latent topic analysis for all documents d in the archive $\mathcal{C}$.

To expand the document with semantically related words, we simply adapt the background language model θ_b in (8) for each document d based on its latent topics [9]. This is realized by interpolating the word distribution, $P_{lt}(w|d)$ in (13), based on latent topics with the general background model θ_b in (8) to have a document-expanded background model θ_b^d in (15) which is document dependent,

$$P(w|\theta_b^d) = b_d P_{lt}(w|d) + (1 - b_d)P(w|\theta_b), \tag{15}$$

where $b_d = \frac{L_d}{L_d + b}$ is a document dependent interpolation weight, and b is a parameter. As in (7), this document dependent-expanded background model θ_b^d is then used to smooth the document model θ_d^l in (6) of Section 2.2 or θ_d^g in Section 2.3. Therefore, after smoothed by θ_b^d, the probabilities for the words highly related to the topics in the document d in the document language model are increased.

4. QUERY EXPANSION WITH QUERY-REGULARIZED MIXTURE MODEL

Another popular approach for retrieving semantically related documents is query expansion by automatically adding semantically related terms to the query. The expanded query thus enables the retrieval of documents not containing the original query terms but semantically related to the query. Query expansion is very often realized by the concept of pseudo-relevance feedback (PRF), or assuming the top M documents in the first-pass retrieved results with the

[3] The superscript ^{g} indicates that the expected counts are enhanced by the graph.

[4] The subscript $_{lt}$ indicates the word distribution is obtained from latent topics.

highest $S(Q,d)$ in (1) are relevant (or pseudo-relevant), and taking terms frequently occurring in those pseudo-relevant documents for query expansion. However, since not all pseudo-relevant documents are truly relevant, and not all words in truly relevant documents are semantically related to the query, selecting useful terms for query expansion is not trivial.

4.1. Word-based Query Expansion

Here we borrow the query-regularized mixture model [8] previously proposed for query expansion for text information retrieval. This model assumes that the pseudo-relevant documents are composed of query-related terms and general terms, with a document-dependent ratio of the two. For example, for those irrelevant documents taken as pseudo-relevant, the ratio for the query-related terms to the general ones is low and so on. These document-dependent ratios and which terms are query-related are actually unknown, but can be estimated from the term distributions in the pseudo-relevant documents. With the estimation, these query-related terms form a new query model θ'_Q, which is used to replace θ_Q in (1). A modified version of this model for spoken archive is briefly summarized below.

Suppose the pseudo-relevant document set is $\{d_1, ..., d_m, ..., d_M\}$, where M is the number of documents in the set. Each document d_m is composed of words generated by either the background language model θ_b in (8) (that is, general terms), or the expected query model θ'_Q to be estimated (that is, query-related terms). α_m, the probability of choosing θ'_Q for generating a query-related term in document d_m, is also unknown. It is possible to estimate θ'_Q and α_m for each pseudo-relevant document d_m by maximizing the likelihood of generating these pseudo-relevant documents in (16),

$$F_1(\theta'_Q, \alpha_1, ..., \alpha_M) = \prod_{m=1}^{M} \prod_{w} (\alpha_m P(w|\theta'_Q) + (1-\alpha_m) P(w|\theta_b))^{P(w|\theta_{d_m})}. \quad (16)$$

In (16), the probability of generating the word w in document d_m is formulated as $\alpha_m P(w|\theta'_Q) + (1-\alpha_m)P(w|\theta_b)$. The document model θ_{d_m} can be either $\theta^l_{d_m}$ directly derived from the lattices in (6) of Section 2.2 or the graph-enhanced version $\theta^g_{d_m}$ in Section 2.3. However, the model θ'_Q maximizing (16) may be dominated by the main topics included in the pseudo-relevant documents, which are not necessarily be query-related. To better handle this problem, θ'_Q is "regularized" by the original query model θ_Q in (2), and we define a function $F_2(\theta'_Q)$ as the prior for θ'_Q based on θ_Q,

$$F_2(\theta'_Q) = \prod_{w} P(w|\theta'_Q)^{P(w|\theta_Q)}. \quad (17)$$

$F_2(\theta'_Q)$ will be larger for model θ'_Q closer to θ_Q. θ'_Q and α_m are actually estimated by maximizing the following objective function:

$$F(\theta'_Q, \alpha_1, ..., \alpha_M) = F_1(\theta'_Q, \alpha_1, ..., \alpha_M) F_2(\theta'_Q)^{\lambda}, \quad (18)$$

where λ is a parameter controlling the influence of the prior function $F_2(\theta'_Q)$. The model θ'_Q estimated via maximizing (18) would not be totally drifted away by the pseudo-relevant documents because the function $F_2(\theta'_Q)$ prefers the expanded query model θ'_Q to be similar to the original query model θ_Q.

4.2. Topic-based query expansion

The above query expansion technique is based on words. Here we further extend the approach to a similar version but based on latent topics. Everything is in parallel with the query-regularized mixture model as summarized in Subsection 4.1, but here instead of estimating a word-based query model (or query-related word distribution) θ'_Q, we now seek to estimate a topic-based query model, or actually a query-related *topic distribution* θ^T_Q over the latent topics, $\{P(T_1|\theta^T_Q), ..., P(T_k|\theta^T_Q), ..., P(T_K|\theta^T_Q)\}$ [5], where K is the number of topics. Here we assume the probabilities of observing all words given each latent topic $P(w|T_k)$ are already available (obtained from latent topic analysis such as PLSA). For each query Q, the desired topic distribution θ^T_Q is estimated via maximizing the objective function in (19) completely in parallel to (18),

$$F^T(\theta^T_Q, \alpha^T_1, ..., \alpha^T_M) = F^T_1(\theta^T_Q, \alpha^T_1, ..., \alpha^T_M) F^T_2(\theta^T_Q)^{\lambda}. \quad (19)$$

The formulations of $F^T_1(\theta^T_Q, \alpha^T_1, ..., \alpha^T_M)$ and $F^T_2(\theta^T_Q)$ in (19) are exactly the same as (16) and (17) respectively, except that the word distribution $P(w|\theta'_Q)$ in (16) and (17) is replaced by a word distribution estimated via latent topic, $P_{lt}(w|\theta^T_Q) = \sum_{k=1}^{K} P(w|T_k)P(T_k|\theta^T_Q)$, which is exactly in parallel with (13) except that the document d in (13) is replaced by the topic distribution of θ^T_Q. By maximizing the objective function in (19), the topic-based query model θ^T_Q and the weight parameters α^T_1 to α^T_M can be estimated. This topic-based query model θ^T_Q can be further interpolated with the word-based query model θ'_Q obtained with (18) to have a query model θ''_Q considering both words and topics:

$$P(w|\theta''_Q) = \delta P(w|\theta'_Q) + (1-\delta) P_{lt}(w|\theta^T_Q), \quad (20)$$

where δ is an interpolation weight. The expanded query models θ'_Q obtained in (18) based on words, θ^T_Q obtained in (19) based on latent topics, or the interpolated version θ''_Q in (20) can then be used as θ_Q in (1).

5. EXPERIMENTAL SETUP

In the experiments, we used a broadcast news corpus in Mandarin Chinese as the spoken document archive to be retrieved through. The news stories were recorded from radio or TV stations in Taipei from 2001 to 2003. There were a total of 5047 news stories, with a total length of 198 hours. The story length ranged from 68 to 2934 characters, with an average of 411 characters per story. 163 text queries and their relevant spoken documents (not necessarily including the queries) were provided by 22 graduate students. The number of relevant documents for each query ranged from 1 to 50 with an average of 19.5, and the query length ranged from 1 to 4 Chinese words with an average of 1.6 words, or 1 to 8 Chinese characters with an average of 2.7 characters.

In order to evaluate the retrieval performance of the proposed approaches with respect to different recognition conditions, we used different acoustic and language models to transcribe the spoken documents. As listed below, we used two different recognition conditions for generating the lattices for the spoken archive:

- Lat (A): We used a tri-gram language model trained on 39M words of Yahoo news, and a set of acoustic models with 64 Gaussian mixtures per state and 3 states per model trained on a corpus of 24.5 hours of broadcast news different from the archive tested here. The acoustic features used were MFCC with cepstral mean and variance normalization (CMVN) applied. The character accuracy for the archive was 54.43%.

[5] The superscript ^{T} indicates this is a query model for topic distribution.

- Lat (B): We cascaded Perceptual Linear Predictive (PLP) feature and phone posterior probabilities estimated by a Multilayer Perceptron (MLP) trained from 10 hours of broadcast news different from those tested as the acoustic features. A tri-gram language model trained on 98.5M words of news from several different sources, and a set of acoustic models with 48 Gaussian mixtures per state and 3 states per model trained on the 24.5 hours of broadcast news were used. The character accuracy was 62.13%.

Both Lat (A) and (B) used a 60K-word lexicon, and the beam width for recognition was 100.

For the graph construction in Section 2.3, nodes x_i and x_j are connected if x_i is among the k-nearest neighbors of x_j based on $S(x_i, x_j; w)$, and x_j is among the k-nearest neighbors of x_i, and k = 10 in the experiments. The acoustic features used for recognition were also used to compute the acoustic similarity. We used mean average precision (MAP) as the evaluation measure for the following experiments.

6. EXPERIMENTAL RESULTS

6.1. Basic Language Model Retrieval Approach without Document/Query Expansion

Table 1: *MAP performance yielded by the basic language model retrieval approach without document/query expansion. Columns (a), (b), (c) and (d) respectively correspond to the results based on manual transcriptions, 1-best transcriptions, original lattices and graph enhanced term frequencies respectively. The two rows are for different recognition conditions. The numbers in parentheses are the KL divergence values for the corresponding document language models in columns (b), (c) and (d) with respect to that for column (a).*

MAP	(a) Manual	(b) 1-best	(c) Lattice	(d) Graph-Enhanced
Lat (A)	0.6216	0.4519 (0.3922)	0.4579 (0.3860)	0.4706 (0.3748)
Lat (B)	0.6216	0.4956 (0.3603)	0.5045 (0.3538)	0.5171 (0.3453)

Table 1 reports the results for the basic language model retrieval approach without document/query expansion. The parameter a for evaluating a_d in (7) was set to be 1000. The two rows are the results for the two sets of lattices generated under different recognition conditions. The four columns correspond to the results using different document models θ_d in (1). Columns (a) and (b) are respectively the results based on the manual and 1-best transcriptions. That is, the document language models θ_d used in (1) was estimated based on the word occurrence counts in the manual and 1-best transcriptions, and smoothed by a background model trained on the manual and 1-best transcriptions of all the spoken documents, very similar to that in (8). The results in column (a) serve as the upper bound for the proposed approach. Column (c) is the results when the original document models $\bar{\theta}_d^l$ obtained in (7) of Section 2.2 were taken as θ_d in (1), while column (d) is for the graph-enhanced document models $\bar{\theta}_d^g$ obtained in Section 2.3 with $\mu = 10$. Here the query model θ_Q was estimated as in (2) without expansion. We found that the recognition errors seriously degraded the retrieval performance (columns (b) vs (a)), and the lattices were better than the 1-best transcriptions (columns (c) vs (b)). The proposed graph-based enhancement approach further outperformed the plain lattice approach (columns (d) vs (c)). The number in the parentheses are the KL divergence values for the corresponding language models in columns (b), (c) and (d) with respect to the corresponding language models in column (a). We found that the graph-enhanced language model $\bar{\theta}_d^g$ obtained in Section 2.3 has the smallest KL divergence with respect to the models based on the manual transcriptions. This explains why the proposed graph-based enhancement approach offered improvements.

6.2. Document Expansion

Table 2: *MAP performance yielded by* ***document expansion****. Left and right parts are for different recognition conditions. The results for basic language model approach (columns (c) and (d) in Table 1) are taken as the baseline. K* ***is the number of topics for PLSA.*** *Columns labeled "Lattice" and "Graph-enhanced" are respectively for the results with the original lattices and with graph-based enhancement.*

	(a) Lat (A)		(b) Lat (B)	
	Lattice	Graph-Enhanced	Lattice	Graph-Enhanced
Baseline	0.4579	0.4706	0.5045	0.5171
K=32	0.4855	0.4936	0.5311	0.5402
K=64	0.4912	0.5018	0.5296	0.5391
K=128	0.4860	0.4930	0.5188	0.5313

Table 2 lists the results for document expansion. The parameter b for evaluating b_d in (15) was set to 1000. Left and right parts are respectively for the two different recognition conditions. The results for basic language model approach (columns (c) and (d) of Table 1) are taken as the baseline. The rows are for different values of K, the number of topics in the PLSA analysis used for document expansion. Columns labeled "Lattice" are the results on the original document models, that is, θ_d^l in (6) of Section 2.2 was used for PLSA training in (14), and for interpolating with the document-expanded background model as well. Columns labeled 'Graph-enhanced" are the results with graph-based enhancement, for which θ_d^g obtained from Section 2.3 was used for PLSA training and interpolated with the document-expanded background model. It is clear that the PLSA-based document expansion improved the retrieval performance, and the proposed graph-based enhancement approach can offer extra improvements under all the conditions. The improvements were reasonable because the PLSA model was learned based on the better document models θ_d^g obtained in Section 2.3, and the document-expanded background models were interpolated with these better models.

6.3. Query Expansion

Table 3 lists the results for word-based query expansion introduced in Section 4.1. The value of λ in (18) was set to 10 in the experiments. Left and right parts are for the two recognition conditions. The results for basic language model approach (columns (c) and (d) in Table 1) are taken as the baseline, and they were considered as the first-pass retrieved results for selecting the pseudo-relevant documents. The results are listed for different values of M (the number of pseudo-relevant documents). Columns labeled "Lattice" are for the results using the original lattices, that is, $\bar{\theta}_d^l$ in (7) of Section 2.2 was used for generating the first-pass results, and the expanded query model θ_Q' in (18) of Section 4.1 was estimated based on θ_d^l (θ_d in

Table 3: *MAP performance yielded by the **word-based query expansion** in Section 4.1. The results for basic language model approach (columns (c) and (d) in Table 1) are taken as the baselines, and taken as the first-pass retrieved results for selecting pseudo-relevant documents.* M ***is the number of pseudo-relevant documents.***

	Lat (A)		Lat (B)	
	Lattice	Graph-Enhanced	Lattice	Graph-Enhanced
Baseline	0.4579	0.4706	0.5045	0.5171
M=10	0.4645	0.4757	0.5116	0.5206
M=20	0.4652	0.4792	0.5156	0.5266
M=30	0.4673	0.4811	0.5141	0.5273
M=40	0.4673	0.4813	0.5123	0.5290
M=50	0.4672	0.4818	0.5082	0.5267

Table 4: *MAP performance for word-based and topic-enhanced query expansion with and without document expansion.*

Recognition Conditions	Document Expansion	Query Expansion	Lattice	Graph-Enhanced
Lat (A)	(I) **NO**	word	0.4645	0.4757
		word+topic	0.4682	0.4781
	(II) **YES**	word	0.4953	0.5033
		word+topic	0.4969	0.5037
Lat (B)	(II) **NO**	word	0.5116	0.5206
		word+topic	0.5153	0.5221
	(II) **YES**	word	0.5333	0.5404
		word+topic	0.5342	0.5429

(16) was θ_d^l in (6)). Columns labeled "Graph-enhanced" are the results with the graph-based enhancement approach in Section 2.3, or $\bar{\theta}_d^g$ for first-pass results and θ_d^g for estimating θ'_Q in (18). We found that word-based query expansion outperformed the baseline regardless of the value of M, and clearly the graph-based enhancement improved the performance in all cases. The proposed graph-based enhancement approach outperformed the original lattice version for two reasons: the proposed graph-based enhancement approach provided better first-pass results for query expansion, so more relevant documents were actually included in the pseudo-relevant document set; also matching the expanded query model θ'_Q with $\bar{\theta}_d^g$ obtained in Section 2.3 was better than doing that with $\bar{\theta}_d^l$ in (7) of Section 2.2.

Table 4 lists the results for word-based and word-plus-topic-based query expansion with and without document expansion respectively. M was set to 10 for query expansion, and δ in (20) was 0.9. The upper and lower halfs of the table are respectively for the two different recognition conditions, each with two sections respectively for without (section(I)) and with (section(II)) document expansion. PLSA topic number K was set to 64. Rows labeled "word" are results for word-based query expansion, or θ'_Q in (18) in Section 4.1 was used in (1). Rows labeled "word+topic" are the results for word-plus-topic-based query expansion, or θ''_Q in (20) of Section 4.2 was used. Columns labeled "Lattice" and "Graph-enhanced" are the same as before. We find applying the topic-based query expansion in addition can further improve the word-based version query expansion (rows labeled "word+topic" vs "word"), and the results in Sections (II) are always better than their correspondents in Section (I), or document expansion is additive with query expansion. Last but not least, the proposed graph-based enhancement approach is always helpful (columns labeled "Graph-enhanced" vs "Lattice") for the semantic retrieval techniques tested here, although the enhancement was based on acoustic similarity only, which may not carry semantic information.

7. CONCLUSION

In this paper, we propose to enhance the expected term frequencies for word arcs in the lattices by acoustic similarity graphs. The enhanced term frequencies were then applied on language model retrieval approach, document expansion and query expansion. Improved performance was observed on a corpus of broadcast news in Mandarin Chinese under all tested conditions.

8. REFERENCES

[1] Ciprian Chelba, Timothy J. Hazen, and Murat Saralar, "Retrieval and browsing of spoken content," in *IEEE Signal Processing Magazine 25(3), pp. 39-49*, 2008.

[2] Tsung-Wei Tu, Hung-Yi Lee, Yu-Yu Chou, and Lin-Shan Lee, "Semantic query expansion and context-based discriminative term modeling for spoken document retrieval," in *ICASSP*, 2012.

[3] Hung lin Chang, Yi cheng Pan, and Lin-Shan Lee, "Latent semantic retrieval of spoken documents over position specific posterior lattices," in *SLT*, 2008.

[4] Berlin Chen, Pei-Ning Chen, and Kuan-Yu Chen, "Query modeling for spoken document retrieval," in *ASRU*, 2011.

[5] Xinhui Hu, Ryosuke Isotani, Hisashi Kawai, and Satoshi Nakamura, "Cluster-based language model for spoken document retrieval using NMF-based document clustering," in *Interspeech*, 2010.

[6] Tomoyosi Akiba and Koichiro Honda, "Effects of query expansion for spoken document passage retrieval," in *Interspeech*, 2011.

[7] Ryo Masumura, Seongjun Hahm, and Akinori Ito, "Language model expansion using webdata for spoken document retrieval," in *Interspeech*, 2011.

[8] Tao Tao and ChengXiang Zhai, "Regularized estimation of mixture models for robust pseudo-relevance feedback," in *SIGIR*, 2006.

[9] Xing Wei and W. Bruce Croft, "LDA-based document models for ad-hoc retrieval," in *SIGIR*, 2006.

[10] Hung-Yi Lee, Po-Wei Chou, and Lin-Shan Lee, "Open-vocabulary retrieval of spoken content with shorter/longer queries considering word/subword-based acoustic feature similarity," in *Interspeech*, 2012.

[11] Hung-Yi Lee, Yun-Nung Chen, and Lin-Shan Lee, "Improved speech summarization and spoken term detection with graphical analysis of utterance similarities," in *APSIPA*, 2011.

[12] Tee Kiah Chia, Khe Chai Sim, Haizhou Li, and Hwee Tou Ng, "Statistical lattice-based spoken document retrieval," *ACM Trans. Inf. Syst.*, vol. 28, pp. 2:1–2:30, 2010.

[13] Chengxiang Zhai and John Lafferty, "A study of smoothing methods for language models applied to ad hoc information retrieval," in *SIGIR*, 2001.

[14] Amy N. Langville and Carl D. Meyer, "A survey of eigenvector methods for web information retrieval," *SIAM Rev.*, vol. 47, pp. 135–161, January 2005.

[15] Thomas Hofmann, "Probabilistic latent semantic analysis," in *Proc. of Uncertainty in Artificial Intelligence, UAI99*, 1999.