SEMANTIC QUERY EXPANSION AND CONTEXT-BASED DISCRIMINATIVE TERM MODELING FOR SPOKEN DOCUMENT RETRIEVAL

Tsung-wei Tu #1, *Hung-yi Lee* *2, *Yu-yu Chou* *, *Lin-shan Lee* #*3

Graduate Institute of Computer Science and Information Engineering, National Taiwan University #
Graduate Institute of Communication Engineering, National Taiwan University *
hpttw0608@gmail.com[1], tlkagkb93901106@yahoo.com.tw[2], lslee@gate.sinica.edu.tw[3]

ABSTRACT

In this paper, we propose a semantic query expansion approach by extending the query-regularized mixture model to include latent topics and apply it to spoken documents. We also propose to use context feature vectors for spoken segments to train SVM models to enhance the posterior-weighted normalized term frequencies in lattices. Experiments on Mandarin broadcast news showed that this approach offered good improvements when applied on spoken documents including relatively high recognition errors.

Index Terms— Semantic Retrieval, Spoken Term Detection

1. INTRODUCTION

In recent years, the information needs of people have clearly gone beyond traditional text-form information. With the ever-increasing bandwidth of the Internet and rapidly falling storage costs, multimedia data such as broadcast programs, lectures and meeting records, and many other video and audio materials are now the most widely accessed network content. Compared to text, however, multimedia/audio content is quite difficult to retrieve and browse; while the speech information included in such content very often indicates its subject or topic area. This underscores the need for efficient technologies for retrieving spoken documents that provide users with easy access to the spoken documents out of the huge amount of such content on the Internet.

Substantial effort has been made in speech information retrieval, and many successful techniques have been developed [1]. Lattice-based approaches that take into account multiple recognition hypotheses have been used to handle the low accuracies of 1-best transcriptions. Many other efficient approaches have also been proposed in recent years. However, most works in speech information retrieval continue to focus on literal term matching, for which the goal is simply returning spoken segments or documents that contain the query terms. This is insufficient because users naturally prefer that the technologies can return all the objects that the user really wants, regardless of whether the query terms are contained or not. There has been some recent works on concept matching in speech information retrieval [2, 3, 4, 5], although concept matching in text information retrieval systems has been widely studied. Taking ASR transcriptions as pure text, most concept match techniques developed for text information retrieval can be directly applied in speech information 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, occurrences of a given term are usually characterized by similar context, while widely-varying contexts typically imply different terms. Hence the use of context information has been proposed to verify the presence of spoken terms in spoken content [6]. Such context information was used to construct Support Vector Machines (SVMs) to classify the desired query term discriminatively.

In this paper, we extend the query-regularized mixture model to develop its topic-based version applied in spoken documents, referred to as the semantic query expansion. We further use context feature vectors for spoken segments to train SVM term models, and use them to enhance the posterior-weighted normalized term frequencies in lattices. Good improvements were observed in initial experiments.

2. BASIC LANGUAGE MODELING RETRIEVAL MODEL

Language Modeling (LM) has been known to be very effective for information retrieval not only for text, but for speech information as well [7]. The basic idea is that if the language model trained with a document has higher probability to generate a given query, we have higher confidence that it is relevant to the query. To simplify the presentation, here we assume word unigram language models only, although the proposed approach is not limited to this case. The language model θ_d for the document d is obtained with maximum likelihood estimation (MLE) based on the term occurrence counts in the document and smoothed using a background model. The language model θ_Q of the query Q is also estimated using MLE based on the terms in the query. The relevance score $S(Q,d)$ for the document d with respect to the query Q is then

$$S(Q,d) = \prod_{q \in Q} P(q|\theta_d)^{P(q|\theta_Q)}, \tag{1}$$

where q is a term in Q, and $P(q|\theta_d)$ and $P(q|\theta_Q)$ are the probabilities given the language models. (1) above is in principle for text information retrieval. For speech information retrieval, we can apply (1) directly on the 1-best transcriptions. However, as there are inevitable relatively high recognition errors in the 1-best transcription, θ_d thus estimated may be very different from the true word distribution of the spoken document. One way to handle this problem is to estimate the probability $P(q|\theta_d)$ from lattices to include many recognition hypotheses.

In the approach proposed here, a spoken document is first divided into spoken segments, and then each spoken segment X is transcribed into a lattice. We first compute the normalized term frequency of each term w in the lattice of a segment X as

$$P(w|X) = \sum_{u \in W(X)} \frac{N(w,u)}{|u|} P(u|X), \tag{2}$$

where u is a word sequence in the lattice, $W(X)$ is the set of all possible word sequences in the lattice for X, $P(u|X)$ is the posterior probability of the word sequence u derived from the acoustic and language models, $|u|$ is the number of word arcs in u, and $N(w,u)$ is the occurrence count of the term w in u. The expected length L_X of the segment X can then be estimated by

$$L_X = \sum_{u \in W(X)} |u| P(u|X). \tag{3}$$

For a spoken document d with N segments $\{X_1, X_2, \cdots, X_N\}$, the distribution of the term w for θ_d is then

$$P(w|\theta_d) = \frac{\sum_{n=1}^{N} L_{X_n} P(w|X_n)}{\sum_{n=1}^{N} L_{X_n}}. \tag{4}$$

$P(w|\theta_d)$ in (4) is then used in (1), which is in fact a weighted posterior probability.

3. SEMANTIC QUERY EXPANSION

The problem for concept matching is that many documents semantically related to the query do not necessarily contain the query term. So the LM retrieval model described above is not able to find semantically related documents, if they do not contain the query term. Query expansion by adding related terms to the query is a common technique to handle this problem. The expanded queries enable the retrieval of additional documents that don't contain the query term but are semantically related to the query.

Here we borrow the query-regularized mixture model [8] originally proposed for text information retrieval for query expansion. Instead of adding extra related terms to the query, this approach directly estimates a new query model θ'_Q from the first-pass retrieved documents, and then simply replaces $P(q|\theta_Q)$ in (1) by $P(q|\theta'_Q)$. This model is very briefly summarized here.

This model assumes that each of the top M documents in the first-pass retrieval results with the highest $S(Q,d)$ in (1) is generated by the interpolation of two language models, the background language model θ_B and the expanded query model θ'_Q. These two models are interpolated with a document dependent weight α_d, so different documents have different weights α_d. The background model θ_B is trained with the entire document archive, and is therefore known. So α_d for the top M documents and the expanded query model θ'_Q are the parameters to be estimated for each query Q. More precisely, the likelihood of generating a document d out of the top M given α_d and θ'_Q is

$$P(d|\alpha_d, \theta'_Q) = \prod_{w \in d} \left(\alpha_d P(w|\theta'_Q) + (1-\alpha_d) P(w|\theta_B)\right)^{P(w|\theta_d)}, \tag{5}$$

where $P(w|\theta'_Q)$, $P(w|\theta_B)$, $P(w|\theta_d)$ are the probabilities for the term w given the language models. Additionally, θ'_Q is "regularized" based on the original query term distribution θ_Q. That is, θ_Q is used as the prior of θ'_Q:

$$P(\theta'_Q) = \prod_{w} P(w|\theta'_Q)^{P(w|\theta_Q)}. \tag{6}$$

According to (6), the closer θ'_Q is to θ_Q, the higher the probability of $P(\theta'_Q)$ is. The parameters α_d and θ'_Q are then estimated by maximizing the following objective function

$$F(\alpha_d, \theta'_Q) = P(\theta'_Q) \prod_{d \in \mathcal{D}} P(d|\alpha_d, \theta'_Q), \tag{7}$$

where $\mathcal{D}$ is the set of the top M documents, and two parts on the right hand side are in (5)(6). In this way the query is expanded based on the first-pass retrieval top M documents. This expanded query model θ'_Q can be used in (1) by directly replacing the original query model θ_Q by θ'_Q. This model is originally for text information retrieval, but can be equally applicable to spoken documents, as long as the probability $P(w|\theta_d)$ in (4) for lattices can be used in (5).

The above query expansion technique is based on words in the documents. Here we further extend the approach to a semantic version based on latent topics. Everything is in parallel with the query-regularized mixture model as summarized above. But here instead of estimating a query dependent language model θ'_Q for word distribution, we now seek to estimate a query dependent language model θ^T_Q for the distribution of K latent topics $T_1, T_2, \ldots T_k$. We assume the probabilities of observing all words given each latent topic $P(w|T_k)$ are available, which are obtained from Probability Latent Semantic Analysis (PLSA) or other latent semantic analysis approaches. For each query the likelihood of each document d given the latent topic distribution θ^T_Q to be estimated, $P(d|\alpha_d, \theta^T_Q)$ in parallel with (5) above, and the prior of θ^T_Q, $P(\theta^T_Q)$ in parallel with (6) above, are exactly the same as (5) and (6) respectively, except that $P(w|\theta'_Q)$ in (5) and (6) is replaced by $P(w|\theta^T_Q) = \sum_{k=1}^{K} P(w|T_k) P(T_k|\theta^T_Q)$, where $P(T_k|\theta^T_Q)$ is the topic distribution given the model θ^T_Q to be estimated. Everything else in (5)(6) including $P(w|\theta_B)$, $P(w|\theta_d)$, $P(w|\theta_Q)$, α_d remain unchanged. The parameters α_d and θ^T_Q are similarly estimated by maximizing the objective function in parallel with (7),

$$F(\alpha_d, \theta^T_Q) = P(\theta^T_Q) \prod_{d \in \mathcal{D}'} P(d|\alpha_d, \theta^T_Q), \tag{8}$$

where $\mathcal{D}'$ is the set of top M' documents in the first-pass retrieval results. This is referred to as semantic query expansion here. With the semantically expanded query model θ^T_Q derived above, the probability $P(w|\theta^T_Q)$ to be used in (1) is then

$$P(w|\theta^T_Q) = \sum_{k=1}^{K} P(w|T_k) P(T_k|\theta^T_Q). \tag{9}$$

This probability can be further interpolated with the probability $P(w|\theta'_Q)$ obtained by maximizing (7).

$$P(w|\theta'_Q + \theta^T_Q) = \lambda P(w|\theta'_Q) + (1-\lambda) P(w|\theta^T_Q). \tag{10}$$

4. CONTEXT-BASED DISCRIMINATIVE TERM MODELING

We see from the above two sections that besides the query model, the document model θ_d is another component here. It is used in both the relevance scores in (1) and the query expansion in (5). For text information retrieval, the estimation of the document model θ_d is trivial from the text directly. For lattice-based speech information retrieval, the document model θ_d is based on the normalized term frequency $P(w|X)$ from (2) to (4) instead. Note the lattices include recognition errors and noisy term hypotheses. As a result, the normalized term frequency $P(w|X)$ in (2) may not reflect the true existence of a term in a lattice. On the other hand, occurrences of a given term are usually characterized by similar context; widely-varying contexts typically imply different terms. Hence the context consistency has been proposed to verify the presence of some spoken terms [6]. Here with similar concept we propose to use the context consistency to train discriminative term models to enhance the normalized term frequencies $P(w|X)$ to be used.

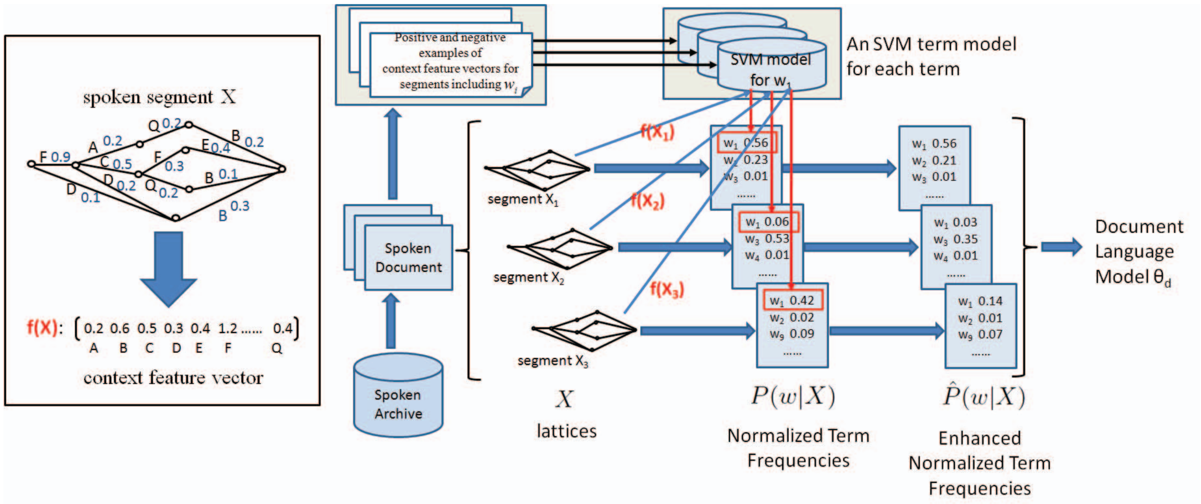

Fig. 1: Context-based discriminative term modeling using SVM

Fig. 1 shows the framework of the proposed context-based discriminative term modeling approach. Given all the spoken documents in the archive, all segments in all documents are first transcribed into lattices. Each segment X is then represented by a context feature vector $f(X)$. The left hand side of Fig. 1 shows an example of X and $f(X)$, where the word hypotheses (e.g., A, B, etc.) and posterior probabilities are shown beside the lattice arcs. Each dimension of $f(X)$ corresponds to a lexical word, so the dimensionality of $f(X)$ is the number of words in the lexicon. The value of each component of $f(X)$ is the posterior probability summed over all word arcs in the lattice having the same word hypotheses. Thus $f(X)$ contains the context information for all terms appearing in the lattice based on the entire segment.

For each term w to be indexed for retrieval, the N segments having the highest expected term frequencies for w are considered as most likely to contain the term w, thus taken as positive examples. The N segments having the lowest expected term frequencies for w are considered as least likely to contain the term w and taken as negative examples. Note that here positive and negative examples of each term w are selected automatically in an unsupervised way, similar to the scenarios of pseudo-relevance feedback. The positive examples do not necessarily contain the term w, although very possibly they do. The context feature vector $f(X)$ of these positive and negative examples is then used to train an SVM model for the term w.

With the SVM models for all words, now for all the segments in the archive having the term w in their lattices, their context feature vector $f(X)$ is classified by the SVM for term w, giving a score $d_w(X)$ derived from the distance from the context feature vector $f(X)$ to the SVM hyperplane: $d_w(X)$ is positive when $f(X)$ is close to the positive examples and thus X is classified as very possibly containing the term w, and negative when $f(X)$ is close to the negative examples and thus very possibly the term w not contained. So the normalized term frequency $P(w|X)$ in (2) can be replaced by

$$\hat{P}(w|X) = \begin{cases} P(w|X)(\frac{1}{1+\exp^{-d_w(X)}})^{\alpha} & d_w(X) < 0 \\ P(w|X) & \text{otherwise}, \end{cases} \tag{11}$$

or the value is reduced by a factor related to $d_w(X)$ if the lattice is classified by the SVM model as not containing the term w, while unchanged otherwise. θ_d is then modified accordingly. The complete process is shown in Fig. 1.

5. EXPERIMENTS

5.1. Experimental Setup

In the experiments, we used a broadcast news corpus in Mandarin Chinese as the spoken document archive to be retrieved from. 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 96 hours. The story length ranged from 68 to 2934 characters, with an average of 411 characters per story. 163 queries and their relevant spoken documents were provided by 22 graduate students. The number of desired 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. For recognition we used a 60K-word lexicon, a tri-gram language model trained on 39M words of Yahoo news, a set of acoustic models with 64 Gaussian mixtures per state trained on a corpus of 24.5 hours of broadcast news different from the archive tested here, with cepstral mean and variance normalization (CMVN) applied. The character accuracy for the archive was 54.43%. λ in (10) was 0.9. 100 positive and 100 negative examples were used in SVM training, and α in (11) was 10. The number of latent topics K in (9) was 64, with latent topics obtained from PLSA. We used mean average precision (MAP) as the evaluation measure for the following experiments.

5.2. Experimental Results

The MAP results are in Table 1. The four columns are for four query models: baseline query model θ_Q estimated by MLE without expansion (MLE), word-based query expansion with query model θ'_Q estimated in (7) (QE_{word}), topic-based semantic query expansion with query model θ^T_Q in (9) (QE_{topic}), and the combination of both in (10) ($QE_{word+topic}$). Here $\mathcal{D}$ and $\mathcal{D}'$ in (7) (8) both included 10 documents. On the other hand, the four rows are for different document models generated from different transcriptions: manual transcriptions **(Manual)**, 1-best transcriptions **(1-Best)**, lattice using

(4) **(Lattice)**, and lattice with normalized term frequencies enhanced by (11) **(Enhanced Lattice)**. All language models were interpolated with background models. For retrieving manual transcriptions, all the manual transcriptions in the corpus were collected to train a background language model; while for other cases, 1-best transcriptions of all documents in the corpus were used to train the background model. The PLSA model was trained from manual transcriptions when manual transcriptions were retrieved, and trained from 1-best transcriptions otherwise.

First compare the results of different query models in different columns. We found that the word-based query-regularized mixture model outperformed the baseline without query expansion (QE_{word} vs MLE). Although the performance of using topic-based semantic query expansion (QE_{topic}) alone was poor, the combination of the word-based and topic-based query expansion ($QE_{word+topic}$) surpassed the individuals. This is consistent with recent studies [9] that latent semantic approaches alone fail for TREC collections, but they can help other retrieval methods via combination. A possible reason may be that there were only 64 topics here, so the topic-based model couldn't be very precise, but could certainly complement the relatively precise word-based model by considering the semantics. Next compare the results for different document models in different rows. Obviously manual transcriptions were by far the best in all cases **(Manual)**, verifying that the recognition accuracy plays a very important role here. It is also clear that the **Lattice** was slightly better than **1-Best** **(Lattice** vs **1-Best)** except for topic-based query expansion alone. The reason for the latter might be that the lattices included more information but also more noise, and it was not easy for the relatively less precise topic-based model with only 64 topics to differentiate noise from information. However, the proposed SVM term models offered about 1 to 2 percent improvements compared to the original lattices for all query models **(Enhanced Lattice** vs **Lattice)**.

In the next experiment, $QE_{word+topic}$ was tested but with number of documents used in the query expansion, M for set $\mathcal{D}$ in (7), ranged from 0 to 100, while M' for set $\mathcal{D}'$ in (8) was fixed to 10. The results are shown in Fig. 2. M=0 means without query expansion or the results of MLE in Table 1. From Fig. 2 it is clear that the proposed **Enhanced Lattice** was always better than **1-Best** or **Lattice** for all choices of M. The best result achieved was 47.72 using **Enhanced Lattice** at M=70, compared to 46.49 for a conventional lattice at M=50. This verified again that the proposed SVM term models are useful.

6. CONCLUSION

In this paper, we propose a new model for semantic query expansion considering the latent topics. We also propose a new method for enhancing the lattices with SVM term models considering the context consistency of terms in the segment. Improved performance was observed in tests with different query models and different document models over a corpus of broadcast news in Mandarin Chinese.

Table 1: MAP results under different query and document models. The four columns are for four query models: baseline without expansion (MLE), word-based query expansion (QE_{word}), topic-based semantic query expansion (QE_{topic}), and their combination ($QE_{word+topic}$). The four rows are for different document models: obtained from manual transcriptions (**Manual**), 1-best ASR transcriptions (**1-Best**), lattices (**Lattice**), and lattices enhanced by the SVM term models (**Enhanced Lattice**). The number of documents M and M' for $\mathcal{D}$ and $\mathcal{D}'$ in (7)(8) were both 10.

MAP	MLE	QE_{word}	QE_{topic}	$QE_{word+topic}$
Manual	53.99	57.38	17.37	61.93
1-Best	37.07	40.17	13.49	40.53
Lattice	38.09	40.48	13.21	40.79
Enhanced Lattice	40.05	41.73	13.73	41.90

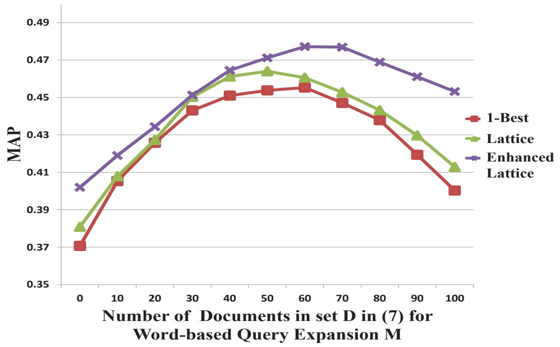

Fig. 2: The MAP results for the integration of word-based and topic-based query expansion techniques ($QE_{word+topic}$) for the number of documents M for set $\mathcal{D}$ in (7) ranged from 0 to 100, while M' for set $\mathcal{D}'$ in (8) was fixed to 10.

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

[3] 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.

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

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

[6] Hung yi Lee, Tsung wei Tu, Chia ping Chen, Chao yu Huang, and Lin shan Lee, "Improved spoken term detection using support vector machines based on lattice context consistency," in *ICASSP*, 2011.

[7] 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.

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

[9] Avinash Atreya and Charles Elkan, "Latent semantic indexing (LSI) fails for TREC collections," *SIGKDD Explor. Newsl.*, vol. 12, pp. 5–10, 2011.