LATENT SEMANTIC RETRIEVAL OF SPOKEN DOCUMENTS OVER POSITION SPECIFIC POSTERIOR LATTICES

Hung-lin Chang, Yi-cheng Pan and Lin-shan Lee

Graduate Institute of Computer Science and Information Engineering, National Taiwan University

`komkon@speech.ee.ntu.edu.tw`, `thomas02@iis.sinica.edu.tw`, `lslee@gate.sinica.edu.tw`

ABSTRACT

This paper presents a new approach of latent semantic retrieval of spoken documents over Position Specific Posterior Lattices (PSPL). This approach performs concept matching instead of literal term matching during retrieval based on the Probabilistic Latent Semantic Analysis (PLSA), so as to solve the problem of term mismatch between the query and the desired spoken documents. This approach is performed over PSPL to consider the multiple hypotheses generated by ASR process, as well as the position information for these hypotheses, so as to alleviate the problem of relatively poor ASR accuracy. We establish a framework to evaluate semantic relevance between terms and the relevance score between a query and a PSPL, both based on the latent topic information from PLSA. Preliminary experiments on Chinese broadcast news segments showed significant improvements can be obtained with the proposed approach.

Index Terms— Spoken Document Retrieval, Semantics

1. INTRODUCTION

In recent years, the traditional text-form information has clearly become insufficient for people's information need. With the ever increasing Internet bandwidth and fast falling storage costs, multimedia data such as broadcast programs, lecture and meeting records, as well as many other video/audio materials are now the most attractive network content. But compared to texts, multimedia/audio content are quite difficult to retrieve and browse, while the included speech information in such content very often indicate the subjects or topic areas of them. As a result, efficient technologies for retrieving spoken documents to provide users with easy access to the desired multimedia/spoken documents out of the huge quantities of Internet content is becoming more and more important.

Very good results for spoken document retrieval have been obtained in the Text REtrieval Conference (TREC) Spoken Document Retrieval (SDR) track considering relatively long queries, relatively long spoken documents with relatively low error rates, even using only one-best ASR results [1]. In recent years, the research interests were shifted to more realistic situations – short queries, short audio segments and relatively poor recognition accuracies [2, 3, 4, 5]. In such cases, with very limited available information for retrieval, the recognition errors may seriously degrade the retrieval performance. As a result, it is important to utilize ASR alternatives so as to retain multiple hypotheses during recognition process [2, 3, 4, 5]. In addition, even if all correct terms in the spoken segments are completely preserved as possible hypotheses, it is still very likely that the query and the highly relevant spoken segments may not have any common terms since the same concept may be expressed by many completely different terms. Obviously, the literal term matching approach which directly matches the possible terms in the spoken segments and the

query terms is no longer adequate in such situation, but instead the concept matching approach well developed for text-based information retrieval considering the latent semantic concepts carried by the spoken segments and the query will be highly desired, although such approach taking into account multiple hypotheses in ASR has not been well discussed in the literatures yet.

In this paper, we propose a new approach to retrieve short spoken segments with concept matching, i.e., based on the latent semantic concepts carried by the spoken segments and the query terms using Probabilistic Latent Semantic Analysis (PLSA) [6] over the Position Specific Posterior Lattices (PSPL) [2] to consider the multiple hypotheses generated in the ASR process.

2. PROBABILISTIC LATENT SEMANTIC ANALYSIS AND POSITION SPECIFIC POSTERIOR LATTICES

Probabilistic Latent Semantic Analysis (PLSA) [6] was proposed to analyze the latent topical information carried by terms and documents using a probabilistic framework. By introducing a set of latent topics $\{z_k, k = 1, 2, \dots, K\}$, PLSA tries to establish the relationship between a term w_i and a document d_j by

$$P(w_i|d_j) = \sum_{k=1}^K P(w_i|z_k) \cdot P(z_k|d_j). \quad (1)$$

Assuming the query terms q_l in the query Q are independent, the relevance score between Q and the document d_j can then be expressed by [7]

$$P(Q|d_j) = \prod_{q_l} \left[\sum_{k=1}^K P(q_l|z_k) P(z_k|d_j) \right], \quad (2)$$

which can be used to retrieve documents which are semantically relevant to the query Q , but not necessarily include the query terms q_l .

The basic idea of Position Specific Posterior Lattices (PSPL) is to calculate the position specific posterior probability $P(w_i, n|L_j)$ of a word w_i at a specific position n in a lattice L_j for a spoken segment d_j . Such information is actually hidden in the lattice L_j of d_j since in each path of L_j we clearly know each word's position. Because it is very likely that more than one path includes the same word in the same position, we need to aggregate the probabilities for a given word at a given position over all possible paths in a lattice [2]. In this way, many possible hypotheses in the ASR output for a spoken segment, obtained in the form of a lattice, can be expressed in a structure efficient for retrieval [2]. An example word-based PSPL (W-PSPL) is in Fig. 1, which can be further extended to Subword-based PSPL (S-PSPL) for better indexing of OOV words or rare words. S-PSPL is very similar to W-PSPL, except all words w_i in Fig. 1 is replaced by subword units and so on [8].

Fig. 1. (a) An ASR lattice, (b) all paths in (a), (c) the constructed PSPL structure, where $w_1, w_2 \dots$ are word hypotheses in the lattice.

3. PROPOSED APPROACH

The approach proposed in this paper is presented in this section.

3.1. Overall System

The overall system diagram for the proposed approach is shown in Fig. 2. In the upper middle of the figure, the PLSA model is constructed based on a large scale text corpus. Reliable probability distributions can thus be estimated. Next, in the upper left of the figure, PSPLs are constructed for all spoken segments in the spoken archive, from which the possible terms w_i in each segment can be easily obtained. The core of the proposed approach, Latent Semantic Retrieval over PSPLs, is then in the lower part of the figure. We first use the probability distributions obtained from PLSA to derive a more precise semantic relevance measure between terms, as will be presented in Sec. 3.2. Meanwhile, topically representative or semantically important terms are also extracted as key terms using the probability distributions from the PLSA model, as will be presented in Sec. 3.3. Finally, the relevance score between the query Q and each spoken document d_j can then be computed based on the semantic relevance measure $R(w_i, q_l)$ between all possible terms w_i in each spoken segment and all query terms q_l in the query Q , the posterior probabilities $P(w_i | L_j)$ of all possible terms from W-PSPL/S-PSPL, and the key term weight λ_{il} , from which the relevant spoken segments are then retrieved. This will be discussed in Sec. 3.4.

3.2. Semantic Relevance between Terms

With reliable estimation of probabilities $P(w_i | z_k)$ for all words w_i and all topics z_k from a well-trained PLSA model, by Bayes' Theorem, the probability $P(z_k | w_i)$ can be obtained as follows:

$$P(z_k | w_i) = \frac{P(w_i | z_k) P(z_k)}{\sum_{k=1}^K P(w_i | z_k) P(z_k)}. \quad (3)$$

In this way, for each term w_i a latent topic vector can be constructed with dimension K , whose components are simply the K probabilities $P(z_k | w_i)$ in Equ. (3) for the k latent topics $\{z_1, z_2, \dots, z_K\}$.

With the latent topic vectors constructed for all terms w_i , the semantic relevance $R(w_i, w_{i'})$ between two terms w_i and $w_{i'}$ can be easily derived from the distance between their corresponding latent

Fig. 2. Overall System Diagram for the Proposed Approach.

topic vectors or probability distributions. In this research, a total of 5 different distance measures were used, including the cosine similarity, the correlation coefficient, Euclidean distance, Kullback-Leibler distance and Bhattacharyya distance.

3.3. Key Term Extraction

The purpose here is to de-emphasize semantically less meaningful terms, and select topically more clear and representative terms and weight them more in retrieval. In addition to such features as occurrence counts in training corpus, document frequencies and so on, we preliminarily use two entropy measures, the segment entropy and the latent topic entropy. The segment entropy $E_{seg}(w_i)$ for a term w_i is defined as

$$E_{seg}(w_i) = - \sum_{j=1}^N (c_{ij} / t_i) \log(c_{ij} / t_i), \quad (4)$$

where c_{ij} is the count of term w_i appearing in spoken document d_j , N is the total number of segments in the spoken document archive, and $t_i = \sum_{j=1}^N c_{ij}$ is the total count of term w_i in the spoken document archive. Although this entropy measure is useful to identify the key terms, another entropy measure turns out to be even more useful, the latent topic entropy defined by

$$E_{top}(w_i) = - \sum_{k=1}^K P(z_k | w_i) \log P(z_k | w_i), \quad (5)$$

where $P(z_k | w_i)$ is in Equ. (3). This latent topic entropy was shown to be an outstanding feature for selecting key terms for SDR [7]. Higher latent topic entropy implies the term is more uniformly used in many different latent topics, or less specific semantically. In contrast, lower latent topic entropy indicates the term is concentrated on fewer topics or carrying more topical information, and thus are possible key terms for these topics. Key term extraction can then be performed very well using all these parameters mentioned above.

3.4. Spoken Segment Retrieval

This is the core of the bottom part of Fig. 2. We first extract all possible terms w_i from the PSPL constructed for each spoken segment d_j , as shown in the upper left part of Fig. 2. Extracting possible terms from a word-based PSPL (W-PSPL) is quite straightforward, but for subword-based PSPL (S-PSPL), we need to select lexical terms composed by subword units appearing in consecutive clusters in the S-PSPL structure, and then compute the posterior probability for the lexical terms from the posterior probabilities of the component subword units.

The relevance score for a spoken segment d_j and a query Q is then defined as follows:

$$S(d_j, Q) = \frac{\sum_{i=1}^{|d_j|} \sum_{l=1}^{|Q|} \lambda_{il} \cdot R(w_i, q_l)^\alpha \cdot P(w_i|L_j)}{\sum_{i=1}^{|d_j|} P(w_i|L_j)}, \quad (6)$$

where $|d_j|$ is the number of all possible terms in d_j , $|Q|$ the number of query terms in Q , λ_{il} is a weight considering either w_i and q_l are key terms or not, $R(w_i, q_l)$ is the semantic relevance between terms w_i and q_l as discussed in Sec. 3.2. $P(w_i|L_j)$ is simply $\sum_{n \in \mathcal{N}} P(w_i, n|L_j)$ for W-PSPL case while for S-PSPL $P(w_i|L_j) = \sum_{n \in \mathcal{N}} \prod_{m=0}^{|w_i|-1} P(s_n, m+n|L_j)$, in which s_n is the n^{th} subword unit in w_i and $|w_i|$ is the number of subword units in w_i . In other word, $S(d_j, Q)$ above considers the semantic relevance between all possible terms in the spoken segment and all query terms in the query. Therefore even the spoken segment and the query do not have any terms in common, the spoken segment can still be retrieved as long as they have some terms which are highly relevant semantically. Also, with the posterior probability $P(w_i|L_j)$, the occurrence of the term w_i is no longer binary, but with a probability. So many possible hypothesis terms in the spoken segment with reasonable probabilities can also be considered. Note that in PSPL sequences of words in addition to single words can be matched between the target segments and the query. This can also be done here, by replacing the terms w_i and q_l in Equ. (6) with sequences of terms and evaluate the Semantic Relevance $R(w_i, q_l)$ by a PLSA model trained for sequences [9]. But this is not done here in the preliminary experiments.

4. PRELIMINARY EXPERIMENTAL RESULTS

Preliminary results for experiments performed for the approach proposed are reported in this section.

4.1. Results of Semantic Relevance between Terms

We used a corpus of 44590 Mandarin text news stories (454 Chinese characters each in average) collected in 2001 to train a PLSA model. We manually selected 20 single words scattered over different topics to evaluate the 5 different distance measures mentioned in Sec. 3.2 in their capabilities to estimate the semantic relevance between terms. These 20 single words were used to select and rank the words from the lexicon most relevant to them based on the 5 different distance measures.

By manually checking whether the terms ranked in the top N are semantically relevant or not, the results for precision at top N ($P@N$) are listed in Table 1. All the 5 distance measures mentioned in Sec. 3.2 were tested, including the cosine similarity (Cos), Correlation Coefficient (Corr), Euclidean distance (EU), Kullback-Leibler distance (KL) and Bhattacharyya distance (BC). From Table 1, We see Kullback-Leibler (KL) and Bhattacharyya (BC) distances turned

out to be slightly better, probably because they are defined for probability distributions, and our vector components are in fact probability distributions.

	Cos	Corr	EU	KL	BC
P@5	0.81	0.81	0.78	0.91	0.84
P@10	0.68	0.68	0.68	0.83	0.71
P@15	0.57	0.55	0.62	0.74	0.63
P@20	0.47	0.46	0.52	0.62	0.47

Table 1. Precision at top N ($P@N$) for 5 different distance measures: cosine Similarity (Cos), Correlation Coefficient (Corr), Euclidean distance (EU), Kullback-Leibler distance (KL) and Bhattacharyya distance (BC).

4.2. Retrieval Experiment Setup

The spoken document archive to be retrieved in the experiments are Mandarin broadcast news stories collected daily from local radio stations in Taiwan in 2001. We manually segmented these stories into 5034 segments, each with one to three utterances. We used TTK decoder developed in National Taiwan University to generate the bigram lattices for these segments, from which the corresponding W-PSPL/S-PSPL structures were obtained. The subword units used in S-PSPL were Chinese characters.

A trigram language model estimated from a 40M news corpus collected in 1999 was used in estimating the posterior probabilities in W-PSPL/S-PSPL and in obtaining the one-best results. The lexicon used in the decoder consisted of 62K words. The acoustic models used here included 151 intra-syllable right-context-dependent Initial-Final (I-F) models, trained using 8 hrs of broadcast news stories collected in 2000. The one-best recognition character accuracy obtained for the 5034 segments was 75.27% (under trigram one-pass decoding).

We used the same PLSA model mentioned in Sec. 4.1 and 6181 words were selected from the lexicon as key terms. 30 single word queries were manually selected for the retrieval experiments. The corresponding relevant spoken segments (171 for each query in average) were also manually generated. The literal term matching approach using conventional W-PSPL/S-PSPL [8] was used as the baseline for comparison. In addition to precision at top N , the evaluation metric Mean Average Precision (MAP) and R-Precision in standard *trec_eval* package used by the TREC evaluations [10] were also used.

4.3. Parameter Selection for Latent Semantic Retrieval

When performing latent semantic retrieval using W-PSPL/S-PSPL, several parameters in Equ. 6 need to be selected first. For the semantic relevance $R(w_i, q_l)$, it was found that properly combining the 5 different distance measures listed in Table 1 using relatively simple weights gave better results than any single distance measure. Similarly, the parameter α and λ_{il} in Equ. (6) can also be carefully chosen to achieve better results. Here only two values of λ_{il} were used, one when both w_i and q_l are key terms, and the other when either one or both of them are not key terms. The results for W-PSPL as an example for selection of α and λ_{il} are listed in Table 2. Obviously properly selecting these parameters can offer better results. In our experiments, a single set of parameters was used for both W-PSPL and S-PSPL.

4.4. Results for Latent Semantic Retrieval

The retrieval results are listed in Table 3. The results for baseline (BL) approach of literal term matching including using conventional

	MAP	R-Prec	P@5	P@10	P@30	P@100
Initial	0.4918	0.5047	0.8200	0.8000	0.7489	0.5783
α selection	0.5063	0.5097	0.8667	0.8367	0.7778	0.5887
λ_{il} selection	0.5294	0.5346	0.8467	0.8233	0.7922	0.6053

Table 2. An example of selection of parameters α and λ_{il} in Equ. (6) for latent semantic retrieval using W-PSPL.

W-PSPL and S-PSPL are in the upper half of the table. First, it is obvious that S-PSPL significantly outperformed W-PSPL, and W-PSPL significantly outperformed one-best. This is consistent to all previous results [2, 8, 11]. Also, it is clear that the spoken segments including the query terms were precisely retrieved, this is why very high precision rates were obtained at P@5 and P@10. However, after those spoken segments including the query terms were retrieved, other segments which are semantically relevant to the query but do not include the query terms could never be found by the literal term matching. As a result, the precision of P@30 and P@100 dropped dramatically and seriously affected the overall performance of MAP and R-Precision.

The results for the proposed latent semantic (LS) retrieval based on concept matching are in the lower left of Table 3, where the second last row of Table 3 is exactly the last row of Table 2 for W-PSPL. We can see that the precision of P@5 and P@10 were at a very high level, either higher or compatible to those obtained with the baseline (BL) approach of literal term matching. The main difference became clear at P@30 and P@100, where the proposed approach (LS) could still achieve relatively high precision, because even if the spoken segment didn't include the query terms, the semantic relationship with the queries could still be established by the latent semantics. As a result, very remarkable improvements can be observed in MAP and R-precision. Note that the improvements here (LS vs. BL) are much more significant than those obtained by parameter selection as shown in Table 2. Therefore the parameter selection discussed in Sec. 4.3 is not necessarily critical, though it certainly provided better performance. The recall-precision plots for the six cases listed in Table 3 are shown in Fig. 3. Clearly we can see the proposed latent semantic (LS) retrieval significantly outperformed the baseline (BL) approach of literal term matching.

		MAP	R-Prec	P@5	P@10	P@30	P@100
BL	one-best	0.1097	0.1265	0.7667	0.6633	0.3900	0.1817
	W-PSPL	0.1497	0.1709	0.8267	0.7467	0.4933	0.2290
	S-PSPL	0.2611	0.3127	0.9000	0.8300	0.6111	0.3907
LS	one-best	0.4811	0.5169	0.8267	0.8300	0.7656	0.5763
	W-PSPL	0.5294	0.5346	0.8467	0.8233	0.7922	0.6053
	S-PSPL	0.5354	0.5388	0.8733	0.8667	0.7933	0.6077

Table 3. An example of selection of parameters α and λ_{il} in Equ. (6) for latent semantic retrieval using W-PSPL.

There is still one more parameter which can be selected, the threshold of posterior probabilities $P(w_i|L_j)$ for pruning of possible term hypotheses in W-PSPL/S-PSPL. So all possible terms in the W-PSPL/S-PSPL with posterior probabilities below the threshold can be pruned. This can not only remove many noisy terms, but reduce the necessary index size. Fig. 4 is the results for four different thresholds: 0, 0.01, 0.1, 0.5, respectively for the four points on each curve in the figure. The right most points correspond to the threshold being 0, or no pruning. Clearly a threshold of 0.01 significantly reduced the index size without degrading the retrieval accuracy, but a threshold of 0.1 degraded the performance. The results here are those using a single Cosine similarity for the semantic relevance $R(w_i, q_i)$ without any parameter selection, therefore the numbers are different from those in Table 3.

Fig. 3. Recall-Precision plots respectively for the six cases in Table 3, one-best, W-PSPL and S-PSPL for baseline (BS) approach and latent semantic (LS) retrieval.

Fig. 4. Pruning of possible terms by thresholding the posterior probabilities and the tradeoff between the retrieval accuracy and the index size.

5. CONCLUSION

In this paper we propose a new approach of latent semantic retrieval of spoken documents over Position Specific Posterior Lattices (PSPL) based on latent topic information derived from the Probabilistic Latent Semantic Analysis (PLSA). This offers a new framework for Spoken Document Retrieval based on concept matching to cope with the problem of term mismatch between the query and the target segments, while properly consider the multiple hypotheses generated in ASR process.

6. REFERENCES

- [1] J. Garofolo, G. Auzanne, and E. Voorhees, "The trec spoken document retrieval track: A success story," in *Recherched Informations Assisté par Ordinateur: ContentBased Multimedia Information Access Conference*, 2000.
- [2] C. Chelba, J. Silva, and A. Acero, "Soft indexing of speech content for search in spoken documents," *Computer Speech and Language*, vol. 21, no. 3, pp. 458–478, July 2007.
- [3] J. Mamou, D. Carmel, and R. Hoory, "Spoken document retrieval from call-center conversations," in *SIGIR*, 2006, pp. 51–58.
- [4] Z.-Y. Zhou, P. Yu, C. Chelba, and F. Seide, "Towards spoken-document retrieval for the internet: Lattice indexing for large-scale web-search architectures," in *HLT*, 2006, pp. 415–422.
- [5] T. Hori, I.L. Hetherington, T.J. Hazen, and J.R. Glass, "Open-vocabulary spoken utterance retrieval using confusion networks," in *ICASSP*, 2007, pp. 73–76.
- [6] Thomas Hofmann, "Probabilistic latent semantic analysis," in *Proc. of Uncertainty in Artificial Intelligence, UAI'99*, Stockholm, 1999.
- [7] Y.-C. Hsieh, Y.-T. Huang, C.-C. Wang, and L.-S. Lee, "Improved spoken document retrieval with dynamic key term lexicon and probabilistic latent semantic analysis (PLSA)," in *ICASSP*, 2006.
- [8] Y.-C. Pan, H.-L. Chang, and L.-S. Lee, "Subword-based position specific posterior lattices (S-PSPL) for indexing speech information," in *Interspeech*, 2007.
- [9] J. Z. Nie, R. X. Li, D. S. Luo, and X. H. Wu, "Refine bigram plsa model by assigning latent topics unevenly," in *ASRU*, 2007, pp. 401–406.
- [10] <http://trec.nist.gov/>.
- [11] Y.-C. Pan, H.-L. Chang, and L.-S. Lee, "Analytical comparison between position specific posterior lattices and confusion networks based on words and subword units for spoken document indexing," in *ASRU*, 2007.