AUTOMATIC TOPIC DETECTION STRATEGY FOR INFORMATION RETRIEVAL IN SPOKEN DOCUMENT

Shan Jin[1], *Hemant Misra*[2*], *Thomas Sikora*[1], *Joemon Jose*[2]

[1]Department of Telecommunication Systems
Technical University of Berlin, Germany
[2]Department of Computing Science
University of Glasgow, United Kingdom

ABSTRACT

This paper suggests an alternative solution for the task of spoken document retrieval (SDR). The proposed system runs retrieval on multi-level transcriptions (word and phone) produced by word and phone recognizers respectively, and their outputs are combined. We propose to use latent Dirichlet allocation (LDA) model for capturing the semantic information on word transcription. The LDA model is employed for estimating topic distribution in queries and word transcribed spoken documents, and the matching is performed at the topic level. Acoustic matching between query words and phonetically transcribed spoken documents is performed using phone-based matching algorithm. The results of acoustic and topic level matching methods are compared and shown to be complementary.

1. INTRODUCTION

The amount of accessible online audio-visual material is growing rapidly. Audio streams of multimedia documents often contain spoken parts which include a lot of semantic information. Therefore development of efficient and effective methods for spoken information retrieval has become a key requirement for retrieving multimedia document.

The traditional spoken document retrieval (SDR) strategy is to run text retrieving methods on transcription of spoken documents produced by a large vocabulary automatic speech recognition (ASR) system. Though such strategies are able to achieve a reasonable performance, the size of the recognizable vocabulary restricts the number of queries. In [1], the authors reported that approximately 13% of user queries contain out-of-vocabulary (OOV) words. Moreover, OOV words pose a serious problem in a word based SDR system, particularly in domains where new words appear frequently over a short period of time.

A phone-based matching algorithm could address the issue of OOV words encountered in a word-based matching algorithm. However, its performance depends heavily on the accuracy of the phonetic transcription. A phone recognizer is very sensitive to background noise. Typically, a phone recognizer can achieve an accuracy of 50% only as compared to 80% accuracy of a domain dependent word recognizer.

We propose a solution that combines (unsupervised) topic matching and acoustic matching algorithms for OOV-robust spoken information retrieval. The proposed system runs retrieval on multi-level transcriptions (word and phone) produced by word and phone recognizers, respectively. Latent Dirichlet allocation (LDA) [2] model is used to estimate the topic distribution, and capture the semantic information present in word transcription of spoken documents and text queries. One of the aims of LDA and similar topic modeling methods, including probabilistic latent semantic analysis (PLSA) [3] is to produce low dimensionality representations of texts in a "semantic space" while preserving their inherent statistical characteristics. A reduction in dimensionality facilitates storage as well as faster retrieval. In this paper, the results of topic based matching are compared with those of acoustic matching performed at phone-level, and are found to be complementary.

The rest of the paper is organized as follows: in Section 2.1, the system used in this paper for spoken document retrieval is explained. In Section 2.2, we describe the LDA model, the method used for its training and process of computing topic distribution for unseen text document. The experimental setup, database and results are discussed in Section 3, and the conclusions of this study are drawn in Section 4.

2. MATCHING AND SCORING

2.1. System Overview

Figure 1 gives an overview of the proposed system. It can be divided into three main modules: acoustic matching, semantic matching and fusion. First of all, spoken documents in archive are transcribed using word and phone recognizers. Query-words that have no semantic bearing (also called stop-words or function words) are removed with Content word detection module. Once the list of content words (in a query) is obtained, it is sent to the acoustic and semantic matching modules to perform information retrieval at acoustic and semantic levels respectively. The fusion module enables the exploitation of the complementary characteristics of acoustic and semantic information to improve the robustness of the system.

2.2. LDA-based Semantic Matching

In [2], the LDA model was proposed for unsupervised topic detection, and the authors investigated its use for the task of text modeling, text classification and collaborative filtering. In [4], the authors examined LDA for identifying "hot topics" by observing temporal dynamics of topics over a period of time and illustrated the role that words play in the semantic content of a document. Recently LDA has also been used for applications such as unsupervised language model adaptation in ASR [5], fraud detection in telecommunications [6] and detecting coherence of documents [7] .

*This work was primarily performed when the author was working with Telecom-ParisTech, Paris

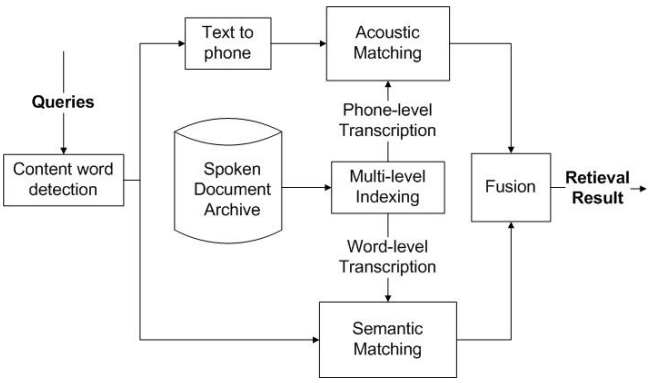

Fig. 1. System Architecture

Like most of the IR methodologies, the LDA model considers documents as bag-of-words (the ordering of the words in a document is unimportant). Two basic assumptions in the LDA model are: 1) the documents are made up of topics (every document is represented by a topic distribution), and 2) each topic has an underlying word distribution.

LDA is a generative model and defines a probabilistic method for generating a new document. Assuming a fixed and known number of topics, T, for each topic t, a distribution ϕ_t is drawn from a Dirichlet distribution of order W, where W is the vocabulary size. The first step for generating a document d is choosing a topic distribution, $\theta_{dt}, t = 1...T$, for that document from a Dirichlet distribution of order T. Next, for each word in the document, a topic, z_i, is chosen from this distribution and a word is selected from this topic. Given the topic distribution, each word is thus drawn independently from every other word using a *document specific* mixture model.

Thus, the probability of i^{th} word token w_i in document d is:

$$P(w_i|\theta_d, \phi) = \sum_{t=1}^{T} P(z_i = t|\theta_d) P(w_i|z_i = t, \phi) \quad (1)$$

$$= \sum_{t=1}^{T} \theta_{dt} \phi_{tw_i} \quad (2)$$

where $P(z_i = t|\theta_d)$ is the probability that the t^{th} topic was chosen for the i^{th} word token and $P(w_i|z_i = t, \phi)$ is the probability of word w_i given topic t.

The likelihood of document d is a product of terms such as (2), and can be written as:

$$P(C_d|\theta_d) = \prod_{w=1}^{W} [\sum_{t=1}^{T} (\theta_{dt} \phi_{tw})]^{C_{dw}} \quad (3)$$

where C_{dw} is the count of word w in d.

2.2.1. LDA: Training

In LDA model, the training step consists of estimating the following two parameters from a set of training documents: the topic distribution in each document d ($\theta_{dt}, t = 1...T, d = 1...D$) and word distribution in each topic ($\phi_{tw}, t = 1...T, w = 1...W$). Both θ and ϕ are assumed to be a multinomial distribution and represent which topics are important for a particular document and which words are important for a particular topic respectively.

The task of estimating parameters can be accomplished using statistical techniques such as variational Bayes [2] and Gibbs sampling [4]. In this paper, we have used Gibbs sampling method to estimate these two distributions. In Gibbs sampling, two hyperparameters α and β are considered which define the non-informative Dirichlet priors on θ and ϕ respectively.

The estimation procedure for LDA model using Gibbs sampling has been explained in [4]. For each word token in the training data, the probability of assigning the current word token to each topic is conditioned on the topic assigned to all other word tokens except the current word token. A topic is sampled from this conditional distribution and assigned to the current word token. In every pass of Gibbs sampling, this process of assigning a topic for all the word tokens in the training data constitutes one Gibbs sample. The initial Gibbs samples are discarded as they are not a reliable estimate of the posterior. For a particular Gibbs sample, the estimates for θ and ϕ are given by

$$\phi_{tw} = \frac{J_{tw} + \beta}{\sum_{k=1}^{W} J_{tk} + W\beta} \quad (4)$$

$$\theta_{dt} = \frac{K_{dt} + \alpha}{\sum_{k=1}^{T} K_{dk} + T\alpha} \quad (5)$$

where J_{tw} is the number of times word w is assigned to topic t and K_{dt} is the number of times topic t is assigned to some word token in document d.

2.2.2. LDA: Testing

Training LDA on a text collection provides insights regarding the thematic structure of the collection. This has been the primary application of LDA in [2, 4]. Even better, LDA being a generative model can also be used to make prediction regarding novel documents. In a typical IR setting, where the main focus is on computing the similarity between a document d and a query d', a natural similarity measure is given by $P(C_{d'}|\theta_d, \phi)$, computed according to (3) [8].

An alternative would be to compute the KL divergence between the topic distribution in d and d'. However, this requires to infer $\theta_{d'}$. As the topic distribution of a (new) document gives its representation along the latent semantic dimensions, computing this distribution is helpful for many applications, including text segmentation or text classification.

In this paper, we use the approach suggested in [5, 7] for estimating topic distribution. The approach essentially implements an iterative procedure based on the following update rule:

$$\theta_{dt} \leftarrow \frac{1}{l_d} \sum_{w=1}^{W} \frac{C_{dw} \theta_{dt} \phi_{tw}}{\sum_{t'=1}^{T} \theta_{dt'} \phi_{t'w}} \quad (6)$$

where l_d is the length of the document in terms of number of content words. Although no justification was given in [5], it has been shown in [7] that this update rule converges towards a local optimum of the likelihood.

2.2.3. Matching Component

Topic distribution, θ, was estimated for word transcription of spoken documents as well as text queries using (6). The similarity between queries and word transcription of spoken documents was measured by the KL-divergence between their respective θs. In the topic space, the most similar document to a query is the one which gives the least

KL-divergence between their respective θs. The results with KL-divergence are denoted by **KL** in the following sections.

The parameter ϕ (distribution of words in each topic) of the LDA model learned on a training corpus, and topic distribution (θ) of each word transcribed spoken document can be used to estimate the conditional probability of a query given a document (3). The most relevant documents are the ones which maximize the conditional probability of a query given the candidate document [8]. The results obtained with conditional probability based matching are denoted by **LL** in the following sections.

2.3. Acoustic Matching

The goal of Acoustic-Matching Module is to find portions of a spoken document that are acoustically similar to the query words. The probabilistic string matching method described in [9] is selected for this task. This method is based on one-best phone transcription and consists of search term location, search term weighting and scoring stages. After possible occurrences (slots) of query phoneme sequence are identified with slot-detection component, slot-probability estimation component assigns probabilities to each of these detected slots. Finally, the similarity score between query and document is computed.

2.3.1. Slot-Detection Component

The task of slot-detection component is to find all possible slots in each document which may contain the keyword sequence. It is assumed that most of the errors produced during phoneme recognition are substitution errors. The substitution-tolerant slot detection method estimates the slots that have sufficient conformity with the query phoneme sequence. This conformity is measured as the number of common phoneme (the same phoneme occurring at the same position within the query phoneme sequence and slots). A slot is verified when its number of "common" phonemes is greater than a pre-defined threshold.

2.3.2. Slot-Probability Estimation Component

This component assigns a probability to each slot detected in the previous stage corresponding to a spoken occurrence of the query phoneme sequence. Probability estimation using confusion information allows to model the error-production behavior of the underlying phoneme recognizer. Statistical information about the inter-phoneme recognition errors, also called confusion information, is captured and subsequently used for slot-probability estimation with string similarity function based on dynamic programming. Slot-probability could be considered as a measure of certainty with which a slot corresponds to an occurrence of the query phone sequence.

2.3.3. Similarity Score Computation Component

Acoustic similarity score, S_{ac}, between query and document is computed as

$$Prob(w_i) = \max_{C} [slot\_prob(C_j)] \quad (7)$$

$$S_{ac} = \left[\sum_{i=1}^{n} Prob(w_i) \right] \quad (8)$$

where C is all candidates detected for word i in query and C_j is the j^{th} candidate; the probability of word i in query is expressed by $Prob(w_i)$. n indicates the number of content words in query. Finally, documents from archive are presented to a user, sorted by decreasing similarity score.

2.4. Combining Acoustic and Semantic Scores

Since acoustic matching and semantic matching retrieve documents from archive with different information sources, we investigate if a combination of the two matching algorithms will remove some randomly distributed errors. We linearly combine the acoustic and semantic scores to form a weighted score as follows:

$$S_{co} = S_{am} * \lambda + S_{sm} * (1 - \lambda). \quad (9)$$

where S_{co} is the combined score, S_{sm} is the similarity score using semantic matching and λ ($0 < \lambda < 1$) is an interpolation weight.

3. EXPERIMENTAL SETUP

The data subset *si-dt-s2* from Wall Street Journal (WSJ) corpus is selected to evaluate the retrieval effectiveness of the proposed system. The subset *si-dt-s2* consists of a set of single-sentence documents covering ten different domains. There are total 207 sentences spoken by 10 persons (5 females and 5 males). 79 queries are defined to evaluate the retrieval performance of each individual system and the combined system.

Word Recognizer: We model 8,000 tied states using Gaussian mixture models (16 Gaussians per state). The acoustic models are initialized with TIMIT train set and trained on WSJ training data (full set). A bi-gram language model (LM) with a vocabulary size of 20k words is used. The LM is trained using NoV-92 LM training data. This speaker-independent ASR system produces a word error rate of 40% on the TIMIT test data set and has an OOV rate of approximately 7%.

Phone Recognizer: Left-to-right HMMs with 128 Gaussians per state are used to model the 39 phonemes. TIMIT training data set is used to initialize the HMM parameters. 64 phones of TIMIT train set are grouped into 39 phonemes to make phoneme recognizer less sensitive to background noise. WSJ's *si-tr-s* and *si-tr-l* data subsets are selected for further training. The constructed phoneme recognizer was evaluated on TIMIT test data set and yields a phoneme error rate of approximately 49%.

LDA training data: It is a subset of English news text feeds from Reuters from the years 1996-1997 [10]. The database is normalized, followed by removal of function words and finally converting it into an appropriate format to run the LDA analysis.

Evaluation Measures and Results: We use the measures proposed by Choi [11] to evaluate retrieval effectiveness. These measures are:

- $E1$ is the Number of queries for which the relevant document is at first place.
- $E2$ is the Number of queries for which the answer document is within the top 10 documents.
- $E3$ is the Mean answer rank.
- $E4$ is the Mean answer rank after removing the outliers.
- Mean reciprocal rank is represented by $E5$.

We perform retrieval using 79 queries with semantic matching methods (KL & LL) and phoneme-based acoustic matching algorithm. λ value is fixed empirically during testing to obtain the best results. The results are shown in Table 1. It is observed from Table 1 that the semantic matching algorithms ($E1 = 42, 34$) outperform the acoustic matching algorithm ($E1 = 22$). The poor performance of

Evaluation Measures	Matching		
	KL	LL	Acoustic
$E1$	42	34	22
$E2$	70	65	39
$E3$	8.12	12.35	40.4
$E4$	6.29	9.15	37.03
$E5$	0.62	0.55	0.32

Table 1. Retrieval performance by different systems

acoustic matching could be because the phoneme transcription contains much more errors than word transcription. Further, in semantic matching, KL ($E1 = 42$) performs better than LL ($E1 = 34$).

In Figure 2, we plot the performance ($E1$ evaluation measure) of the acoustic and semantic matching algorithms with respect to the query length (number of content words in a query). The average query length in our case is 15. Figure 2 shows that the phone-based acoustic matching outperforms KL-based semantic matching for shorter queries (less than 8 content words). In this case, 50% of the queries find their answer at rank 1 for acoustic matching whereas KL-based semantic matching achieves only 37% answers at rank 1. In contrast, the acoustic matching based retrieval performs poorly for longer queries. Interestingly, it was shown in [7] that topic estimation by LDA is poor for short documents. Though the longer queries are better for the semantic matching, they create more confusion in the phone-based acoustic matching.

It is also observed from Figure 2 that a simple linear combination of semantic and acoustic scores improves the system performance, especially for shorter queries. It suggests that the acoustic and the semantic matching approaches are complementary, and a combination system similar to the one proposed in this paper can be readily exploited to improve over the performance of the individual systems.

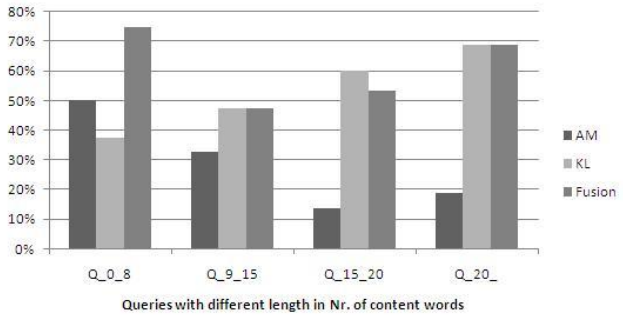

Fig. 2. Performance Comparison in $E1$ (%)

4. CONCLUSION

This paper proposed that LDA-based semantic matching and phone-based acoustic matching have complementary performances as they use different information sources, and a combination of these two systems could improve the performance of an SDR system. Despite fairly high word error rates in word-level transcriptions, the LDA-based semantic matching could find answer for about 53% of the queries at rank 1. It was also found that the phone-based acoustic matching algorithm performs well for shorter queries whereas KL-based semantic matching provides reliable retrieval performance for longer queries. A simple linear combination of the scores obtained by the two systems was able to achieve an improvement over the performance of the individual systems.

Future work will focus on exploring different fusion algorithms and evaluating the variants of the system on a large spoken document archive. Further, a noise robust ASR system will also be incorporated into the system to produce more reliable multi-level transcriptions.

Acknowledgment

This research was supported by the European Commission under the contracts *FP6-027026-K-Space*, *IST-1-038398-VISNET-II* and *FP6-027122-Salero*. The views expressed in this paper are those of the authors and do not necessarily represent the views of the commission.

5. REFERENCES

[1] B. Logan, P. Moreno, JM. van thong, and E. Whittaker, "An experimental study of an audio indexing system for the Web," in *ICSLP*, Beijing, China, Oct. 2000.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent Dirichlet allocation," in *Advances in Neural Information Processing Systems (NIPS)*, Thomas G. Dietterich, Suzanna Becker, and Zoubin Ghahramani, Eds., Cambridge, MA, 2002, vol. 14, pp. 601–608, MIT Press.

[3] Thomas Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning Journal*, vol. 42, no. 1, pp. 177–196, 2001.

[4] Thomas L. Griffiths and Mark Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101 (supl 1), pp. 5228–5235, 2004.

[5] Aaron Heidel, Hung an Chang, and Lin shan Lee, "Language model adaptation using latent Dirichlet allocation and an efficient topic inference algorithm," in *EuroSpeech*, Antwerp, Belgium, 2007.

[6] Dongshan Xing and Mark Girolami, "Employing latent Dirichlet allocation for fraud detection in telecommunications," *Pattern Recognition Letters*, vol. 28, no. 13, pp. 1727–1734, Oct. 2007.

[7] Hemant Misra, Olivier Cappè, and Françio Yvon, "Using LDA to detect semantically incoherent documents," in *CoNLL*, Manchester, U.K., Aug. 2008.

[8] Wray Buntine, Jaakko Löfström, Jukka Perkiö, Sami Perttu, Vladimir Poroshin, Tomi Silander, Henry Tirri, Antti Tuominen, and Ville Tuulos, "A scalable topic-based open source search engine," in *IEEE/WIC/ACM International Conference on Web Intelligence*, Beijing, China, 2004, pp. 228–234.

[9] M. Wechsler, *Spoken Document Retrieval based on Phoneme Recognition*, Ph.D. thesis, Swiss Federal Institute of Technology (ETH), 1998.

[10] D.D. Lewis, Y. Yang, T. Rose, and F. Li, "Rcv1: A new benchmark collection for text categorization research," *Machine Learning Research*, vol. 5, pp. 361–397, 2004.

[11] John Choi, Don Hindle, Julia Hirschberg, Ivan Magrin-chagnolleau, Christine Nakatani, O Pereira, Amit Singhal, and Stev Whittaker, "An overview of the AT&T spoken document retrieval," in *DARPA Broadcast News transcription and Understanding Workshop*, Feb. 1998.