I-VECTOR BASED LANGUAGE MODELING FOR SPOKEN DOCUMENT RETRIEVAL

Kuan-Yu Chen[†#], *Hung-Shin Lee*[†#], *Hsin-Min Wang*[†], *Berlin Chen*[*], *Hsin-Hsi Chen*[#]

[†]Institute of Information Science, Academia Sinica, Taipei, Taiwan
[*]National Taiwan Normal University, Taipei, Taiwan
[#]National Taiwan University, Taipei, Taiwan
E-mail: [†]{kychen, hslee, whm}@iis.sinica.edu.tw, [*]berlin@ntnu.edu.tw, [#]hhchen@ntu.edu.tw

ABSTRACT

Since more and more multimedia data associated with spoken documents have been made available to the public, spoken document retrieval (SDR) has become an important research subject in the past two decades. The i-vector based framework has been proposed and introduced to language identification (LID) and speaker recognition (SR) tasks recently. The major contribution of the i-vector framework is to reduce a series of acoustic feature vectors of a speech utterance to a low-dimensional vector representation, and then numbers of well-developed post-processing techniques (such as probabilistic linear discriminative analysis, PLDA) can be readily and effectively used. However, to our best knowledge, there is no research up to date on applying the i-vector framework for SDR or information retrieval (IR). In this paper, we make a step forward to formulate an i-vector based language modeling (IVLM) framework for SDR. Furthermore, we evaluate the proposed IVLM framework with both inductive and transductive learning strategies. We also exploit multi-levels of index features, including word- and subword-level units, in concert with the proposed framework. The results of SDR experiments conducted on the TDT-2 (Topic Detection and Tracking) collection demonstrate the performance merits of our proposed framework when compared to several existing approaches.

Index Terms— Spoken document retrieval, i-vector, language modeling, inductive, transductive

1. INTRODUCTION

Over the past two decades, spoken document retrieval (SDR) [1, 2] has become an interesting research subject in the speech processing community due to large volumes of multimedia data associated with spoken documents being made available to the public. A significant amount of research effort has been devoted towards developing robust indexing (or representation) techniques [3-6] so as to extract probable spoken terms or phrases inherent in a spoken document that could match the query words or phrases literally. More recently, SDR research has also revolved around the notion of relevance of a spoken document in response to a query. It is generally agreed that a document is relevant to a query if it can address the stated information need of the query, but not because it happens to contain all the words in the query [7].

In the past, the vector space model (VSM) [7, 8], the Okapi BM25 model [7, 9], and the unigram language model (ULM) [10, 11] are well-representative ones for many information retrieval (IR) applications, including SDR. Their efficient and effective abilities have been proved by many researchers and practitioners for a wide variety of IR-related tasks. Yet, the later effort for further extending these methods to capture context dependence based on *n*-grams of various orders or some grammar structures mostly lead to mild gains or even spoiled results [10, 11]. The reasons are two-fold. On one hand, this is due to the fact that these methods might suffer from the problem of word usage diversity, which sometimes degrades the retrieval performance severely as a given query and its relevant documents use quite different sets of words (e.g. synonyms). On the other hand, lots of polysemy words have different meanings in different contexts. As such, merely matching words occurring in the original query and a document may not capture the semantic intent of the query.

To mitigate the above problems, topic models [6, 12-16] attempt to discover a set of latent topics, for which the relevance between a query and a document is not computed directly based on the co-occurrence frequencies of the query words and the document words. Latent semantic analysis (LSA) [16, 17], probabilistic latent semantic analysis (PLSA) [12, 15], and latent Dirichlet allocation (LDA) [13-15] are best representatives of methods which introduce a set of latent topic variables to describe the "*word-document*" co-occurrence characteristics. LSA assumes that the latent topics are orthogonal and can be constructed by decomposing a pre-defined "*word-by-document*" matrix of a training document collection with singular value decomposition (SVD). Each document (and query) is subsequently characterized by a vector of weights indicating the strength with respect to each concept. The relevance degree between a query and a document can be estimated by the cosine similarity measure between the query and the document representations (vectors) [7]. PLSA and LDA derive the latent topics by using maximum likelihood training, and the relevance between a query and a document is computed based on the frequencies of the query words in the latent topics as well as the likelihood that the document generates the respective topics [14].

Recently, the i-vector based framework has become one of the state-of-the-art approaches for language identification (LID) [18-21] and speaker recognition (SR) [22-24]. One challenge of these tasks is the need to process and analyze a high-dimensional vector, which is constructed from the variable-length series of acoustic feature vectors of each input utterance based on some reference models. The i-vector framework proposed an elegant way to reduce such rough input utterance to a corresponding low-dimensional vector representation while retaining the most representative (e.g., language-specific for LID or speaker-specific for SR) information embedded in the original input utterance. Since a document is composed by a series of words, our idea is to apply the i-vector framework to represent a document by a low-dimensional vector, which retains the most representative information of the document. To our best knowledge, there is no research that investigates the i-vector framework for SDR or IR. In this paper, we make a step forward to formulate an i-vector based

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

language modeling (IVLM) framework for SDR. We also evaluate the proposed IVLM framework with both inductive and transductive learning strategies [25], and with both word- and subword-level index features.

2. RELATED WORK

The wide spectrum of IR models that have been developed so far may roughly fall into two main categories: non-probabilistic approaches and probabilistic approaches.

2.1. Non-Probabilistic Approaches

The vector space model (VSM) [5-8] is the basis for most of the IR researches until now. The major advantage of VSM is that it is simple and intuitive, but efficient and effective. In VSM, each document (and query) is represented by a high-dimensional vector, where each dimension specifies the occurrence statistics associated with an index term (e.g., word, subword, or their *n*-grams) in the document (and query). To eliminate the noisy words (e.g., the function words) and promote the discriminative words (e.g., the content words), the statistics is usually the term frequency (TF) that is weighted by the inverse document frequency (IDF) [7]. The relevance degree between a pair of query and document is estimated by the cosine measure of their vector presentations. The flaws of VSM are two-fold. On one hand, VSM might suffer from the word usage diversity, which sometimes degrades the retrieval performance severely as a given query and its relevant documents may use different sets of words (e.g., synonyms). On the other hand, lots of polysemy words have different meanings in different contexts. Hence, merely matching the query words with the words in a document may not capture the semantic intent of the query.

To complement the above drawbacks of VSM, LSA [6, 7] assumes that there is an implicit semantic structure between words and documents, which can be explored by performing SVD on a pre-defined word-by-document matrix:

$$\mathbf{A}_{M\times N} \approx \mathbf{U}_{M\times K}\, \Sigma_{K\times K}\, \mathbf{V}^{T}_{K\times N} = \widetilde{\mathbf{A}}_{M\times N}, \tag{1}$$

where **A** is the word-by-document matrix consisting of the statistics of the *M* distinct words occurring in *N* documents, and *K* is a desired number of most significant eigenvalues. Each element $\mathbf{A}_{wd}$ of **A** is the weighted statistics of word *w* in document *d*. After SVD, each word is uniquely associated with a row vector of matrix **U**, while each document is uniquely associated with a column vector of matrix $\mathbf{V}^T$. In the retrieval phase, a query is viewed as a new document, and its *K*-dimensional vector representation is computed by a "fold-in" process. The relevance degree between a pair of query and document is estimated by the cosine measure of their *K*-dimensional vector representations.

Semantic context inference (SCI) [3] is another specially designed approach for concept mapping and context expansion of spoken documents in SDR. The major difference between SCI and LSA is that SCI takes the word-word associations into account, while LSA considers the word-document co-occurrence relationships. In addition, some LSA-based language models [4, 5, 26] attempt to construct a matrix to render the word-ordering information. Regularized latent semantic indexing (RLSI) [27] formulates topic models as a problem of minimizing a quadratic loss function regularized by different norms, and the problem can be decomposed into multiple sub-optimization problems that can be solved in parallel. Weighted matrix factorization (WMF) [28] proposes a systematic way to modulate the impact of the occurring and non-occurring words on the semantic analysis.

2.2. Probabilistic Approaches

A recent trend in building SDR systems is to use the language modeling (LM) approach [2, 7, 15, 28]. This is due to the fact that the LM approach has sound theoretical underpinnings and excellent empirical performance. The fundamental formulation of the LM approach to SDR is to compute the conditional probability $P(Q|d)$, i.e., the likelihood of a query *Q* generated by a spoken document *d* (the so-called query-likelihood measure). A spoken document *d* is deemed to be relevant to the query *Q* if the corresponding document model is more likely to generate the query. If the query *Q* is treated as a sequence of words, $Q = q_1, q_2, \cdots, q_L$, where the query words are assumed to be conditionally independent given the document *d* and their order is also assumed to be of no importance (i.e., the so-called "*bag-of-words*" assumption), the similarity measure $P(Q|d)$ can be further decomposed as a product of the probabilities of the query words generated by the document [2, 15]:

$$P(Q \mid d) = \Pi_{l=1}^{L} P(q_l|d), \tag{2}$$

where $P(q_l|d)$ is the likelihood of generating q_l by document *d*, which is estimated based on the occurrence frequency of q_l in *d* by the maximum-likelihood estimator. To model the general properties of a language as well as to avoid the problem of zero probability, $P(w|d)$ is usually smoothed by a background unigram model $P(w|BG)$ [2, 15, 28].

Similarly, probabilistic topic models [12-15, 29, 30] (e.g., PLSA and LDA) have been proposed to complement the LM approach. For probabilistic topic models, each document *d* is taken as a document topic model M_d, consisting of a set of *K* shared latent topics $\{T_1,\ldots,T_k,\ldots,T_K\}$ associated with the document-specific weights $P(T_k|\mathrm{M}_d)$, where each topic T_k in turn offers a unigram distribution $P(w_i|T_k)$ for observing an arbitrary word of the language. For example, in the PLSA model [12, 14], the probability of a word w_i generated by a document *d* is expressed by:

$$P_{\mathrm{PLSA}}(w_i|\mathrm{M}_d) = \Sigma_{k=1}^{K} P(w_i|T_k) P(T_k|\mathrm{M}_d). \tag{3}$$

A document is believed to be more relevant to the query if the query words appear frequently in the topics on which the document has higher weights.

On the other hand, LDA [13-15, 30], having a formula analogous to PLSA for document modeling, is thought of as a natural extension to PLSA, and has enjoyed much empirical success for various IR tasks. LDA differs from PLSA mainly in the inference of model parameters: PLSA assumes that the model parameters are fixed and unknown; while LDA places additional a priori constraints on the model parameters, i.e., thinking of them as random variables that follow some Dirichlet distributions. Since LDA has a more complex form for model optimization, which is hard to solve by exact inference, several approximate inference algorithms, such as the variational Bayes approximation [13, 14] and the Gibbs sampling algorithm [30], have been proposed to facilitate the estimation of the parameters of LDA according to different training strategies.

3. I-VECTOR BASED LANGUAGE MODELING FOR SPOKEN DOCUMENT RETRIEVAL

3.1. I-Vector based Language Modeling (IVLM)

The i-vector framework [18-24] is a simplified variant of the joint factor analysis (JFA) approach [31, 32], and both are well-known approaches for LID and SR. Their major contribution is to provide an elegant way to convert the cepstral coefficient vector sequence of a variable-length utterance into a low-dimensional vector representation. To do so, first, a Gaussian mixture model is used to collect the Baum-Welch statistics from the utterance. Then, the first-order statistics from each mixture component are concatenated to form a high-dimensional "supervector" S, which is assumed to obey an affine linear model [21, 31, 32]:

$$S = \mathbf{m} + \mathbf{T} \cdot \varphi_S, \tag{4}$$

where $\mathbf{T}$ is a total variability matrix, φ_S is an utterance specific latent variable, and $\mathbf{m}$ denotes a global statistics vector. In detail, the column vectors of $\mathbf{T}$ form a set of bases spanning a subspace covering the important variability, e.g., the language-specific evidences for LID or the speaker-specific evidences for SR, and the utterance specific variable φ_S indicates the combination of the variability of the utterance. In this way, a variable-length utterance is represented by a low-dimensional vector φ. Finally, the low-dimensional vector is applied to some well-developed post-processing techniques, such as PLDA, for LID and SR. Since the i-vector framework can be trained in an unsupervised manner while JFA must be trained along with manual annotation information, the former has become one of the state-of-the-art approaches for LID and SR recently. In this paper, we investigate the same idea in the context of spoken document retrieval.

Specifically speaking, each document d is first represented by a high-dimensional feature vector $v_d \in \mathbb{R}^{\beta}$. All of the representative (e.g., lexical-, semantic-, and structure-specific) statistics are encoded in the β-dimensional vector, which obeys an affine linear model:

$$v_d = \mathbf{m} + \mathbf{T} \cdot \varphi_d, \tag{5}$$

where $\mathbf{T} \in \mathbb{R}^{\beta \times \gamma}$ is a total variability matrix, γ is a desired value ($\gamma \ll \beta$), and $\mathbf{m} \in \mathbb{R}^{\beta}$ denotes a global statistics vector. Similarly, the column vectors of $\mathbf{T}$ span a subspace covering the important characteristics for documents. Moreover, each document has a document specific variable $\varphi_d \in \mathbb{R}^{\gamma}$, which indicates the combination of the variability of the document. Based on the methodology, a disengaged version is to characterize the representative information of a document only by words. Consequently, each element of the β-dimensional vector is corresponding to a distinct word, and the probability of a word w occurring in a document d can be defined as a log-linear function:

$$P(w \mid d, \mathbf{T}, \mathbf{m}, \varphi_d) = \frac{\exp(\mathbf{T}_w \varphi_d + \mathbf{m}_w)}{\sum\limits_{w' \in V} \exp(\mathbf{T}_{w'} \varphi_d + \mathbf{m}_{w'})}, \tag{6}$$

where $\mathbf{T}_w$ denotes the row vector of $\mathbf{T}$ corresponding to word w, $\mathbf{m}_w$ denotes the statistics value of $\mathbf{m}$ corresponding to word w, and V denotes the vocabulary inventory in the language. We name this model as the i-vector based language model (IVLM). Based on Eqs. (5) and (6), the model parameters (i.e., $\mathbf{T}$, φ_d and $\mathbf{m}$) of the proposed IVLM can be estimated by maximizing the total likelihood over all training documents:

$$L = \prod_{d} \prod_{w \in d} \left(\frac{\exp(\mathbf{T}_w \varphi_d + \mathbf{m}_w)}{\sum\limits_{w' \in V} \exp(\mathbf{T}_{w'} \varphi_d + \mathbf{m}_{w'})} \right)^{c(w,d)}, \tag{7}$$

where $c(w,d)$ denotes the number of times the word w occurs in document d. Since estimating all the parameters jointly is intractable, we estimate them through an iterative process, i.e., we estimate $\mathbf{T}$ and $\mathbf{m}$ with fixed φ_d, and then estimate φ_d with fixed $\mathbf{T}$ and $\mathbf{m}$:

$$\begin{aligned} \varphi_d^{\tau+1} &= \varphi_d^{\tau} + \lambda \cdot \frac{\partial L}{\partial \varphi_d} \\ &= \varphi_d^{\tau} + \lambda \cdot \left\{ \sum_{w} \left[c(w,d) - |d| \cdot \frac{\exp(\mathbf{T}_w^{\tau} \varphi_d^{\tau} + \mathbf{m}_w^{\tau})}{\Sigma_{w'} \exp(\mathbf{T}_{w'}^{\tau} \varphi_d^{\tau} + \mathbf{m}_{w'}^{\tau})} \right] \mathbf{T}_w^{\tau} \right\}, \end{aligned} \tag{8}$$

$$\begin{aligned} \mathbf{T}_w^{\tau+1} &= \mathbf{T}_w^{\tau} + \lambda \cdot \frac{\partial L}{\partial \mathbf{T}_w} \\ &= \mathbf{T}_w^{\tau} + \lambda \cdot \left\{ \sum_{d} \left[c(w,d) - |d| \cdot \frac{\exp(\mathbf{T}_w^{\tau} \varphi_d^{\tau} + \mathbf{m}_w^{\tau})}{\Sigma_{w'} \exp(\mathbf{T}_{w'}^{\tau} \varphi_d^{\tau} + \mathbf{m}_{w'}^{\tau})} \right] \varphi_d^{\tau} \right\}, \end{aligned} \tag{9}$$

$$\begin{aligned} \mathbf{m}_w^{\tau+1} &= \mathbf{m}_w^{\tau} + \lambda \cdot \frac{\partial L}{\partial \mathbf{m}_w} \\ &= \mathbf{m}_w^{\tau} + \lambda \cdot \sum_{d} \left[c(w,d) - |d| \cdot \frac{\exp(\mathbf{T}_w^{\tau} \varphi_d^{\tau} + \mathbf{m}_w^{\tau})}{\Sigma_{w'} \exp(\mathbf{T}_{w'}^{\tau} \varphi_d^{\tau} + \mathbf{m}_{w'}^{\tau})} \right], \end{aligned} \tag{10}$$

where λ is the step size, $|d|$ is the length of document d, and τ is the iterative index. The Frobenius norm can be used to govern $\mathbf{T}$ and φ_d in the training process and the step size can be set empirically or by calculating the Hessian matrix [21, 33].

In the retrieval phase, each document d has its own IVLM, including the document specific variable φ_d and common $\mathbf{T}$ and $\mathbf{m}$. As such, the probability of word w occurring in document d computed by IVLM in Eq. (6) can be linearly combined with or used to replace $P(q_l|d)$ in the query-likelihood measure (*c.f.* Eq. (2)) to distinguish relevant documents from irrelevant ones.

The concept of the proposed IVLM is similar to that of LSA, RLSI, and PLSA, but differences do exist among them. First, IVLM and PLSA are probabilistic models while LSA and RLSI are not. Second, IVLM not only has a different formulation to PLSA, but it does not assume that the total variability is governed by some distribution. Since the parameters of IVLM are real numbers rather than positive real numbers in PLSA, IVLM is more flexible and general than PLSA. Moreover, the parameters of IVLM can be solved in parallel while the parameters of PLSA have to be estimated in a batch mode. It is worth noting that IVLM is a special (disengaged) case of the proposed i-vector based language modeling framework for SDR. We will try to discover and couple with more representative information in the future work.

3.2. Using Subword-level Index Units

In this paper, we also integrate subword-level information into various approaches for SDR. To do this, syllable pairs are taken as the basic units for indexing in addition to words. The recognition transcript of each spoken document, in form of a word stream, was automatically converted into a stream of overlapping syllable pairs. Then, all the distinct syllable pairs occurring in the spoken document collection were then identified to form a vocabulary of syllable pairs for indexing. We can simply use syllable pairs, in replace of words, to represent the spoken documents and construct the associated language models accordingly.

3.3. Inductive & Transductive Learning Strategies

In this paper, we will compare the use of inductive and transductive learning strategies [25] in IVLM. Inductive learning means that the models are trained from an external document collection. After training, $\mathbf{T}$ and $\mathbf{m}$ are used to fold-in each document d in the document collection to be retrieved to get the corresponding document specific variable φ_d. Transductive learning uses the document collection to be retrieved to train the models. After training, φ_d for each document d is used in the retrieval phase.

4. EXPERIMENTAL SETUP

We used the Topic Detection and Tracking collection (TDT-2) [34] in the experiments. The Mandarin news stories from Voice of America news broadcasts were used as the spoken documents. All news stories were exhaustively tagged with event-based topic labels, which served as the relevance judgments for performance evaluation. The average word error rate obtained for the spoken documents is about 35%. The Chinese news stories from Xinhua News Agency were used as our test queries. In the experiments, we will either use a whole news story as a "long query," or merely extract the tittle field from a news story as a "short query." Table 1 shows some basic statistics of the TDT-2 collection. It is known that the way to systemically determine the optimal number of latent variables is still an open issue and needs further investigation. In this paper, the number of latent variables is set to 8. The retrieval results are expressed in terms of non-interpolated mean average precision (MAP) following the TREC evaluation [35].

5. EXPERIMENTAL RESULTS

First, Table 2 reports the retrieval results of the proposed IVLM approach for both short and long queries with respect to two learning strategies using word- or subword-level index features. We use a set of Chinese news stories from Xinhua News Agency as a contemporaneous external document set for inductive learning. It is generally believed that transductive learning should be better than inductive learning. However, as can be seen from Table 2, inductive learning achieves slightly better performance than transductive learning in most cases, except when using word-level index features with short queries for SDR. Since the document collection to be retrieved (2,265 documents in total) is much smaller than the external collection (18,461 documents in total), transductive learning may suffer from the data sparseness problem while inductive learning can obtain more robust model parameters from a larger set of contemporaneous documents.

Next, the proposed IVLM approach is compared with several well-known non-probabilistic and probabilistic approaches, namely VSM, LSA, SCI, and ULM, and topic models such as PLSA and LDA. To bypass the impact of the data sparseness problem, all the approaches are trained by inductive learning. The results when using word- and subword-level index features are shown in Table 3. From the table, at first glance, it can be seen that the proposed IVLM framework outperforms all the non-probabilistic approaches (*c.f.* VSM, LSA, and SCI) and the probabilistic approach (*c.f.* ULM, PLSA, and LDA) in most cases. The reason why it does not perform as well with subword-level index features for short queries is not clear, and is worthy of further studying. The results indicate that the proposed IVLM approach is a novel and alternative way for SDR. In addition, it can also be seen that most IR approaches seem to benefit more from the use of subword-level index features than word-level index features, probably because the subword-level index units can shadow the impact of imperfect speech recognition results.

Table 1. Statistics of the TDT-2 collection.

	TDT-2 (Development Set) 1998, 02~06			
# Spoken documents	2,265 stories, 46.03 hours of audio			
# Distinct test queries	16 Xinhua text stories (Topics 20001~20096)			
	Min.	Max.	Med.	Mean
Doc. length (in characters)	23	4,841	153	287.1
Short query length (in characters)	8	27	13	14
Long query length (in characters)	183	2,623	329	532.9
# Relevant documents per query	2	95	13	29.3

Table 2. Retrieval results (in MAP) of IVLM with word- and subword-level index features for short and long queries using inductive and transductive learning strategies.

IVLM	**Inductive**		**Transductive**	
	Word	**Subword**	**Word**	**Subword**
short	0.336	0.360	0.382	0.350
long	0.582	0.584	0.563	0.574

Table 3. Retrieval results (in MAP) of different approaches with word- and subword-level index features for short and long queries.

	Word		Subword	
	short	**long**	**short**	**long**
VSM	0.273	0.484	0.257	0.499
LSA	0.296	0.364	**0.384**	0.527
SCI	0.270	0.413	0.270	0.349
ULM	0.321	0.563	0.329	0.570
PLSA	0.328	0.567	0.376	0.584
LDA	0.328	0.566	0.377	0.584
IVLM	**0.336**	**0.582**	0.360	**0.584**

Moreover, two general observations can be made from the results. First, probabilistic approaches in general outperform non-probabilistic approaches. The results indicate that probabilistic approaches are a school of simple but powerful methods for SDR, and there are still potential research areas for non-probabilistic approaches. It should also be noticed that, the frequency count of a word is weighted by using the standard IDF method for non-probabilistic approaches while probabilistic approaches (including IVLM) only take the frequency count of a word into account. Second, a topic modeling approach outperforms its non-topic modeling counterpart (e.g., LSA vs. VSM, IVLM vs. ULM). The results indicate that the relevance between a pair of query and document should not be estimated only based on "literal term matching," concept information is useful and should be considered in SDR.

6. CONCLUSIONS & FUTURE WORK

This paper has proposed an i-vector based language modeling approach for spoken document retrieval, which suggests an alternative way to improve SDR performance. The utility of the proposed framework has been validated by extensive comparisons with several existing information retrieval approaches. Our future work includes the development of supervised training, incorporation of various representative information or knowledge, and applying the proposed IVLM approach to speech recognition and document summarization.

REFERENCES

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

[2] L. S. Lee and B. Chen, "Spoken document understanding and organization," *IEEE Signal Processing Magazine*, 22(5), pp. 42-60, 2005.

[3] C. L. Huang, B. Ma, H. Li, and C. H. Wu, "Speech indexing using semantic context inference," in *Proc. INTERSPEECH*, pp. 717-720, 2011.

[4] W. Naptali, M. Tsuchiya, and S. Nakagawa, "Context dependent class language model based on word co-occurrence matrix in LSA framework for speech recognition," in *Proc. ACS*, pp. 275-280, 2008.

[5] W. Naptali, M. Tsuchiya, and S. Nakagawa, "Word co-occurrence matrix and context dependent class in LSA based language model for speech recognition", *International Journal of Computers*, pp. 85-95, 2009.

[6] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by latent semantic analysis", *Journal of the American Society of Information Science*, 41(6), pp. 391-407, 1990.

[7] C. D. Manning, P. Raghavan and H. Schtze, *Introduction to Information Retrieval*, New York: Cambridge University Press, 2008.

[8] G. Salton , A. Wong , and C. S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, 18(11), pp. 613-620, Nov. 1975

[9] K. S. Jones, S. Walker, and S. E. Robertson. "A probabilistic model of information retrieval: development and comparative experiments (parts 1 and 2)," *Information Processing and Management*, 36(6), pp. 779-840, 2000.

[10] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in *Proc. SIGIR*, pp. 275-281, 1998.

[11] W. B. Croft and J. Lafferty (eds.), "Language modeling for information retrieval," Kluwer International Series on Information Retrieval, Volume 13, Kluwer Academic Publishers, 2003.

[12] T. Hoffmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning,* 42, pp. 177-196, 2001.

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, pp. 993-1022, 2003.

[14] D. M. Blei and J. Lafferty, "Topic models," in A. Srivastava and M. Sahami, (eds.), *Text Mining: Theory and Applications*, Taylor and Francis, 2009.

[15] K. Y. Chen, H. M. Wang, and B. Chen, "Spoken document retrieval leveraging unsupervised and supervised topic modeling techniques," *IEICE Transactions on Information and Systems*, pp. 1195-1205, 2012.

[16] M. Berry, S. Dumais, and G. O'Brien, "Using linear algebra for intelligent information retrieval," *SIAM Review*, pp. 573-595, 1995.

[17] W. Guo and M. Diab, "Modeling sentences in the latent space," in *Proc. ACL*, pp. 864-872, 2012.

[18] D. Martinez, O. Plchot, L. Burget, O. Glembek, and P. Matejka, "Language recognition in ivector space," in *Proc. INTERSPEECH*, pp. 861-864, 2011.

[19] M. Soufifar, S. Cumani, L. Burget, and J. Cernocky, "Discriminative classifiers for phonotactic language recognition with ivectors," in *Proc. ICASSP*, pp. 4853-4856, 2012.

[20] L. F. D'Haro, O. Glembek, O. Plchot, P. Matejka, M. Soufifar, R. Cordoba, and J. Cernocky, "Phonotactic language recognition using i-vectors and phoneme posteriogram counts," in *Proc. INTERSPEECH*, pp. 42-45, 2012.

[21] M. Soufifar, M. Kockmann, L. Burget, and O. Plchot, O. Glembek, and T. Svendsen, "Ivector approach to phonotactic language recognition," in *Proc. INTERSPEECH*, pp. 2913-2916, 2011.

[22] O. Glembek, L. Burget, P. Matejka, M. Karafiat, and P. Kenny, "Simplification and optimization of i-vector extraction," in *Proc. ICASSP*, pp. 4516-4519, 2011.

[23] D. Garcia-Romero, and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. INTERSPEECH*, pp. 249-252, 2011.

[24] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, "I-vector based speaker recognition on short utterances," in *Proc. INTERSPEECH*, pp. 2341-2344, 2011.

[25] S. Pang, and N. Kasabov, "Inductive vs transductive inference, global vs local models: SVM, TSVM, and SVMT for gene expression classification problems," in *Proc. IJCNN*, pp. 1197-1202, 2004.

[26] J. T. Chien, M. S. Wu and H. J. Peng, "Latent semantic language modeling and smoothing," *International Journal of Computational Linguistics and Chinese Language Processing*, 9(2), pp. 29-44, 2004.

[27] Q. Wang, J. Xu, H. Li, and N. Craswell, "Regularized latent semantic indexing," in *Proc. SIGIR*, pp. 685-694, 2011.

[28] K. Y. Chen, H. M. Wang, B. Chen, and H. H. Chen, "Weighted matrix factorization for spoken document retrieval," in *Proc. ICASSP*, pp. 8530-8534, 2013.

[29] X. Wei and W. B. Croft, "LDA-based document models for ad-hoc retrieval," in *Proc. SIGIR*, pp. 178-185, 2006.

[30] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. of the National Academy of Sciences*, pp. 5228-5235, 2004.

[31] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4), pp. 1435-1447, 2007.

[32] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in GMM-based speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4), pp. 1448-1460, 2007.

[33] A. L. Maas, and A. Y. Ng, "A probabilistic model for semantic word vectors," in *Proc. NIPS Workshop*, 2010.

[34] LDC, "Project topic detection and tracking," *Linguistic Data Consortium,* 2000.

[35] J. Garofolo, G. Auzanne, and E. Voorhees, "The TREC spoken document retrieval track: A success story," in *Proc. TREC*, pp. 107-129, 2000.