EXPLOITING SEMANTIC ASSOCIATIVE INFORMATION IN TOPIC MODELING

Meng-Sung Wu, Hung-Shin Lee, and Hsin-Min Wang

Institute of Information Science, Academia Sinica, Taipei, Taiwan
{wums, hslee, whm}@iis.sinica.edu.tw

ABSTRACT

Topic modeling has been widely applied in a variety of text modeling tasks as well as in speech recognition systems for effectively capturing the semantic and statistic information in documents or speech utterances. Most topic models rely on the bag-of-words assumption that results in learned latent topics composed of lists of individual words. Unfortunately, these words may convey topical information but lack accurate semantic knowledge of the text. In this paper, we present the *semantic associative topic model*, where the concept of the semantic association terms is extended to topic modeling, which provides guidance on modeling the semantic associations that occur among single words by expressing a document as an association of multiple words. Further, the pointwise KL-divergence metric is used to measure the significance of the association. We also integrate original PLSA and SATM models, which have mixed feature representations. Experimental results on WSJ and AP datasets show that the proposed approaches achieved higher performance compared to other methods.

Index Terms— topic model, semantic association, language model, information retrieval

1. INTRODUCTION

In recent years, topic-based language models for speech recognition and information retrieval have increased in popularity. Topic models are unsupervised probabilistic models for document collection and are generally used for extracting course-grained semantic information from a collection of text documents. In a topic model, each document in the collection of D documents is modeled as a multinomial distribution over t topics, where each topic is a multinomial distribution over W words. Of the different topic models, probabilistic latent semantic analysis (PLSA) [11] and latent Dirichlet allocation (LDA) [2] are the two most frequently investigated topic models. These two models were originally developed for document representation and recently improved for statistical language modeling in document modeling [7], spoken documents retrieval [4], and robust speech recognition [8]. In PLSA, each document is considered as a mixture model containing latent semantic mixtures. Parameters of mixture probabilities are then estimated by the maximum likelihood (ML) principle. In LDA, each document is seen as a mixture distribution over latent topics.

Unfortunately, one of the important restrictions in most existing topic models may lie in that individual terms are usually too general and treated independently without considering the relationships among words in the documents. In some cases of retrieval, several individual words are not enough to represent the accurate semantic information of the text. For example, the word "Apple" often occurs together with "mobile", "iPhone", "computer" or "HON-HAI" in financial or economics news. A query requesting information about Apple might be satisfied by a document about the iPhone or HON-HAI. If this document is represented by these individual words, many unrelated documents will be assumed to satisfy this query. For this reason, we propose using word association to add new words to the document representations that are related to the original words. Adding word associations to represent the document information increases the model's complexity, but it avoids the ambiguity mentioned above. Generally, any set of words appearing in the contexts which have a strong semantic association can be collected as the *associated words*. For example, a meaningful combination of words such as <clairvoyant, Paul, Octopus> can be used to refer to a World Cup football games, although they do not necessarily occur adjacently in a document.

Many attempts have been made to incorporate word order and co-occurrence knowledge into topic modeling [5][12][14]. One such example is bigram topic models [12][14], which inherit the merits of both the traditional n-gram model and the topic model to improve statistical language modeling on document retrieval. In this case, the relationship between words is limited to two adjacent words. Chen et al. [5] presented a bag-of-word pairs (BoWP)-based LSA method, where a document is expressed by a group of word pairs. Although the word order and co-occurrence were compensated for, it was not feasible to incorporate the associations of more distant words. Chien and Chen [6] showed that using term associations could improve the effectiveness of language modeling. However, to our best knowledge, no one has yet tried to incorporate semantic association within the topic language modeling framework. We ignore the order in which the words occur, and instead focus on the words and their statistical distribution in documents. The extracted associations identify the relations between features in the document collection. In this work, we propose an approach called semantic associative topic models (SATM) that is built on the notion of associated words and the well-known PLSA paradigm. This approach extends the existing topic modeling approach by relaxing the independence assumption. We concentrate on exploring the relationship of multiple words, so as to effectively solve the problem of insufficient semantic information in topic modeling. We apply the association mining method [1] and pointwise KL divergence approach [13] as a metric to discover sets of associated words in the documents. We also present the hybrid topic model, which combines the PLSA with our proposed model. The proposed models are evaluated by the metric of perplexity, and the model's performance is validated by applying it to an information retrieval task.

The rest of the paper is organized as follows. Section 2 reviews previous work on probabilistic topic modeling and the

association rule. Section 3 presents our semantic association topic language model. In section 4, we present the hybrid topic model framework. Experimental results are presented in section 5. Section 6 provides a conclusion and discussion of future work.

2. SURVEY OF RELATED WORK

In this section, we give an overview of two topic models: probabilistic latent semantic analysis (PLSA) and bag-of-word pairs (BoWP).

2.1. Probabilistic latent semantic analysis

In topic modeling, PLSA [11] is a general machine learning technique, which adopts the aspect model to represent the co-occurrence data associated with a topic or hidden variable. As a generative model for word/document co-occurrences, PLSA can be described by the following procedure:

1. choose a document d_j with probability $p(d_j)$,
2. select a latent topic z_k with probability $p(z_k \mid d_j)$,
3. generate a word w with probability $p(w_i \mid z_k)$.

The conditional probability of a document d_j generating a word w_i can be represented by

$$p(w_i \mid d_j) = \sum_{k=1}^{K} p(w_i \mid z_k)\, p(z_k \mid d_j), \tag{1}$$

assuming that d_j and w_i are conditionally independent on the mixture of associated topic z_k. We can accumulate the log likelihood of the overall training data $\{w_i, d_j\}$ as follows

$$L = \sum_{j=1}^{M} \sum_{i=1}^{N} n(w_i, d_j) \log \sum_{k=1}^{K} p(w_i \mid z_k)\, p(z_k \mid d_j), \tag{2}$$

where $n(w_i, d_j)$ is the count of word w_j occurring in document d_i, M is number of documents in the training data set, N is the number of words in the vocabulary and K is the number of topics or hidden variables in the model. Parameters of mixture probabilities were estimated by the maximum-likelihood (ML) principle and the expectation-maximization (EM) algorithm is used to resolve missing data or the latent variable problem in parameter estimation [11]. A drawback of PLSA is that the number of parameters grows linearly with the size of the document collection.

2.2. Bag-of-word pairs

In natural language, if a word w_i is significantly associated with a future word w_n, the word pair $w_i \to w_n$ is produced. The bag-of-word pairs (BoWP) [5] algorithm is based on a latent semantic analysis (LSA) framework, which expresses the document as a group of word pairs. LSA is a conceptual-indexing method, which uses the singular value decomposition (SVD) [2] to find the latent semantic structure of the word-to-document association. BoWP employs $n(w_i, w_n, d_j)$, which take into account the normalized frequency of the word pair (w_i, w_n) in document d_j, which can be represented as a collection of matrix decompositions as shown in Figure 1. Given the word-pair-by-document co-occurrence matrix $\mathbf{A}$, an SVD operation is carried out to generate the LSA feature space of BoWP. The SVD decomposes matrix $\mathbf{A}$ into three sub-matrices such that $\mathbf{A} = \mathbf{USV}^T$, where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices ($\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}_R$), and $\mathbf{S}$ is a diagonal matrix. In word pair selection, a simple way to measure the significance of the association is to measure the *distance in the LSA space* between words w_i and w_n [5]

$$\mathbf{A} = \begin{bmatrix} n(w_1,w_1,d_1) & \cdots & n(w_1,w_1,d_j) & \cdots & n(w_1,w_1,d_M) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ n(w_i,w_n,d_1) & \cdots & n(w_i,w_n,d_j) & \cdots & n(w_i,w_n,d_M) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ n(w_N,w_N,d_1) & \cdots & n(w_N,w_N,d_j) & \cdots & n(w_N,w_N,d_M) \end{bmatrix}$$

Figure 1. Matrix decomposition of BoWP

$$WP(w_i, w_n) = \Big\{ w_i, w_n \Big| idf(w_i) > \alpha, idf(w_n) > \alpha, \frac{\mathbf{u}(w_i)\mathbf{S}^2\mathbf{u}(w_n)}{\|\mathbf{u}(w_i)\mathbf{S}\| \cdot \|\mathbf{u}(w_n)\mathbf{S}\|} > \beta \Big\}, \tag{3}$$

where α and β are empirical thresholds, and *idf* means inverse document frequency.

3. SEMANTIC ASSOCIATION TOPIC MODEL

To enhance the topic language modeling via the use of more complex and detailed semantic information, we extend the PLSA model, which assumes words are generated independently from each other, to a new associative topic model, which explores the semantic associations of more than two words for topic modeling.

3.1. Selection of associated words

Association is a powerful data analysis concept that appears in data mining task. An association rule is an implication of the form $A \Rightarrow B$ where A and B are disjoint sets of related words. To construct the associated terms and merge semantic information into topic models, we first generate the patterns of frequently associated words from the training sets. We apply a slightly modified association mining method [1] to discover sets of associated words in documents. The discovery of useful associated words is a two-step process: (1) find all frequent word sets and (2) use the identified frequent word sets to generate strongly associated words.

Assume a set of words $W = \{w_1, \cdots, w_N\}$ and a collection of documents $D = \{d_1, \cdots, d_M\}$. The frequency of each word is counted in the training corpus. The frequent one-word subset, denoted as $L_1 = \{w_i\}$, has no associated words. The frequent *aw*-word subset L_{aw} is discovered from the (*aw*-1)-word subset L_{aw-1}. Analogous to [6], we denote W_{aw}^i as an associated term in the frequent *aw*-word subset $L_{aw} = \{W_{aw}^i\} = \{W_{aw-1}^i \Rightarrow w_i\}$. To find L_{aw}, a two-step process is followed, consisting of joining and pruning actions, which are iteratively performed until no more frequent *aw*-word sets are found.

Joining step: To find L_{aw}, a set of candidate *aw*-word sets is generated by joining frequent word sets of L_{aw-1} with itself. This candidate sets is denoted as C_{aw}. Let W_{aw-1}^a and W_{aw-1}^b be word sets in L_{aw-1}, the candidate *aw*-word sets C_{aw} is generated by

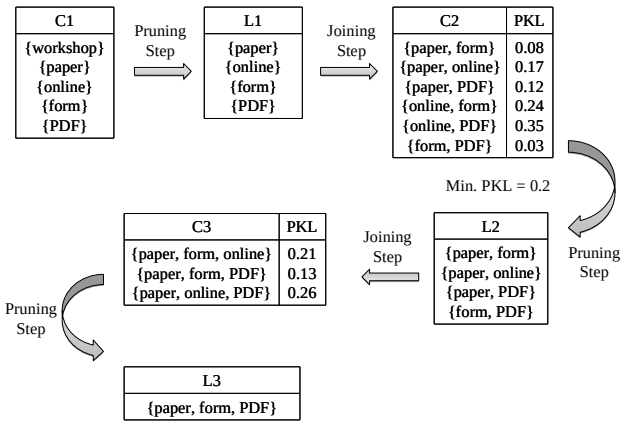

Figure 2. Example of generating process for associated words

merging W^a_{aw-1} and W^b_{aw-1}, where the preceding (*aw*-2)-words are identical. Therefore, the last word w_b of W^b_{aw-1} is appended to W^a_{aw-1} to generate the candidate word sets $C_{aw} = \{W^a_{aw-1} \cup w_b\}$.

Pruning step: Any (*aw*-1)-word sets that do not frequently appear cannot be a subset of a frequent *aw*-word sets. If any (*aw*-1)-word subset of a candidate *aw*-word sets is not in L_{aw-1}. The candidate word sets that are not frequent and so can removed from C_{aw}. Figure shows an example of generating process for associated words. To ensure the selection quality, we choose pointwise KL-divergence (PKL) [13] as the measure of word association used in a language modeling framework, which can be computed as follows

$$\mathrm{PKL}(W^i_{aw-1} \Rightarrow w_i) = p(W^i_{aw-1}, w_i) \log \frac{p(W^i_{aw-1}, w_i)}{p(W^i_{aw-1})p(w_i)}. \quad (4)$$

Note that if two words are statistically dependent, then their PKL value is smaller. Finally, the associated words W^i_{aw} that have a measure score smaller than or equal to the minimum PKL are selected to form the frequent *aw*-word subset L_{aw}.

3.2. Topic models with term association

Similar to the PLSA framework, SATM defines the joint probability of associated words $(W^i_{aw-1} \Rightarrow w_i)$ in a document d_j as $p(W^i_{aw-1} \Rightarrow w_i, d_j)$. The joint probability of an observed tuple $(W^i_{aw-1} \Rightarrow w_i, d_j)$ can be generated in asymmetric parameterization form

$$P(W^i_{aw-1} \Rightarrow w_i, d_j) = P(d_j) \sum_{k=1}^{K} P(W^i_{aw-1} \Rightarrow w_i | z_k) P(z_k | d_j).$$

Figure 3 shows our proposed model represented as a collection of matrix decompositions. The parameters can be estimated by maximizing the log-likelihood function

$$\begin{aligned} L &= \log \sum_{k=1}^{K} n(W^i_{aw}, d_j)\, p(W^i_{aw-1} \Rightarrow w_i | z_k)\, p(z_k | d_j) \\ &= \log \sum_{k=1}^{K} n(W^i_{aw}, d_j)\, p(W^i_{aw} | z_k)\, p(z_k | d_j) \end{aligned} \quad (5)$$

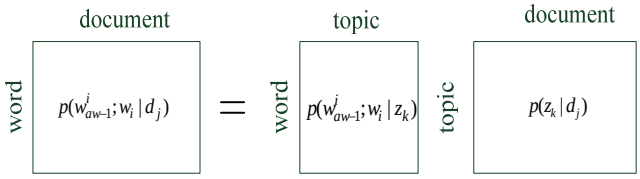

Figure 3. Matrix decomposition of SATM

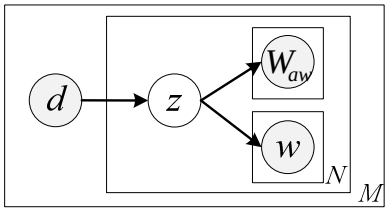

Figure 4. Graphical model representations of hybrid topic models

where the shorthand $p(W^i_{aw}) = p(W^i_{aw-1} \Rightarrow w_i)$ is used to represent the probability of observing w_i when W^i_{aw-1} is given, and $n(W^i_{aw}, d_j)$ denotes the number of times the associated word (W^i_{aw-1}, w_i) occurred in document d_i.

3.3. Parameter estimation

The training procedure of SATM is similar to PLSA, where the log likelihood function is updated as seen in equation (5). The goal of the model is to estimate the parameters $p(z_k)$, $p(z_k | d_j)$ and $p(W^i_{aw} | z_k)$ given a set of observations (W^i_{aw}, d_j). Due to the latent variable appearing in SATM, we should apply the EM algorithm to solve the ML parameter estimation. In the E-step, we compute

$$p(z_k | W^i_{aw}, d_j) = \frac{p(W^i_{aw} | z_k) p(z_k | d_j)}{\sum_{l=1}^{K} p(W^i_{aw} | z_l) p(z_l | d_j)}. \quad (6)$$

In the M-step, we aim to maximize the expectation of the complete data likelihood

$$p(W^i_{aw} | z_k) = \frac{\sum_{j=1}^{M} n(W^i_{aw}, d_j)\, p(z_k | W^i_{aw}, d_j)}{\sum_{W^{i'}_{aw}} \sum_{j=1}^{M} n(W^{i'}_{aw}, d_j)\, p(z_k | W^{i'}_{aw}, d_j)}. \quad (7)$$

$$p(z_k | d_j) = \frac{\sum_{W^i_{aw}} n(W^i_{aw}, d_j)\, p(z_k | W^i_{aw}, d_j)}{n(d_j)}. \quad (8)$$

With an initial random guess of $\{p(z_k | d_j), p(W^i_{aw} | z_k)\}$, SATM alternately applies the E-step equation (6) and M-step equation (7,8) until a termination condition is met.

4. HYBRID TOPIC MODELS

To achieve the best results, language models that model different aspects of language have been used together. The proposed method uses a hybrid topic model combining individual and associated word units. Based on the graphical model representation in Figure

4, we derive the log likelihood function with relative weight λ, as follows

$$L = \sum_{d_j} \Bigg[\lambda \sum_{i=1}^{N} \log \sum_{k=1}^{K} p(w_i \mid z_k) p(z_k \mid d_j) + (1-\lambda) \sum_{W_{aw}^i} \log \sum_{k=1}^{K} p(W_{aw}^i \mid z_k) p(z_k \mid d_j) \Bigg]. \quad (9)$$

In extreme case when $\lambda = 1$, the log likelihood function ignores all the biases from the semantic association information, and degenerates to the traditional PLSA model. Now, the objective is to maximize the log likelihood in Equation (9). Following the EM approach it is straightforward to derive a set of re-estimation equations. For the E-step, the posterior probabilities of the latent variables with each observation $p(z_k \mid w_i, d_j)$ and $p(z_k \mid W_{aw}^i, d_j)$, can be computer by

$$p(z_k \mid w_i, d_j) = \frac{p(w_i \mid z_k) p(z_k \mid d_j)}{\sum_{l=1}^{K} p(w_i \mid z_l) p(z_l \mid d_j)}, \quad (10)$$

$$p(z_k \mid W_{aw}^i, d_j) = \frac{p(W_{aw}^i \mid z_k) p(z_k \mid d_j)}{\sum_{l=1}^{K} p(W_{aw}^i \mid z_l) p(z_l \mid d_j)}. \quad (11)$$

The conditional distributions are recomputed in the M-step according to

$$p(w_i \mid z_k) = \frac{\sum_{j=1}^{M} n(w_i, d_j) p(z_k \mid d_j, w_i)}{\sum_{l=1}^{N} \sum_{j=1}^{M} n(w_l, d_j) p(z_k \mid d_j, w_l)}, \quad (12)$$

$$p(W_{aw}^i \mid z_k) = \frac{\sum_{j=1}^{M} n(W_{aw}^i, d_j) p(z_k \mid W_{aw}^i, d_j)}{\sum_{W_{aw}^{i'}} \sum_{j=1}^{M} n(W_{aw}^{i'}, d_j) p(z_k \mid W_{aw}^{i'}, d_j)}, \quad (13)$$

along with the mixture portion of posterior probability of latent variables

$$p(z_k \mid d_j) \propto \lambda \sum_{i=1}^{N} p(z_k \mid w_i, d_j) + (1-\lambda) \sum_{W_{aw}^i} p(z_k \mid W_{aw}^i, d_j). \quad (14)$$

5. EXPERIMENTAL RESULTS

5.1. Experimental setting

We evaluated the model described in the previous sections using two different TREC collections [15]. One is the Wall Street Journal 1987 (WSJ) dataset consisting of 46,488 documents; the other is the Associated Press Newswire 1988 (AP) dataset consisting of 79,919 documents. The baseline of our experiment is the PLSA model and unigram model smoothed by interpolated Witten-Bell algorithm. The effectiveness of IR is measured by the standard mean average precision (mAP). We also calculated the perplexity for document modeling. We performed preprocessing stages of stemming and stop word removal for all documents. In the experiments, the window size of associated words is set to be sentence length.

5.2. Experimental results of model perplexity

The WSJ corpus was used to evaluate the proposed method for document modeling. 10% of the dataset was reserved for testing purposes. The models were trained by the remaining 90%. Perplexity was used to measure the average word branching factor of the document models. In this set of experiments, the number of latent topics k was fixed to be 32. The perplexity of PLSA and unigram model was 795.94 and 970.48, respectively. Table 1 displays the perplexity results using a bigram model, association pattern language model (APLM) [6], bag-of-word pairs (BoWP) [5] model, and our SATM methods (with the number of association words set to 2). We can see that the bigram has a perplexity of 262.73. The topic-based models perform better than language model without merging topic information. SATM obtains lower perplexity compared to the other models.

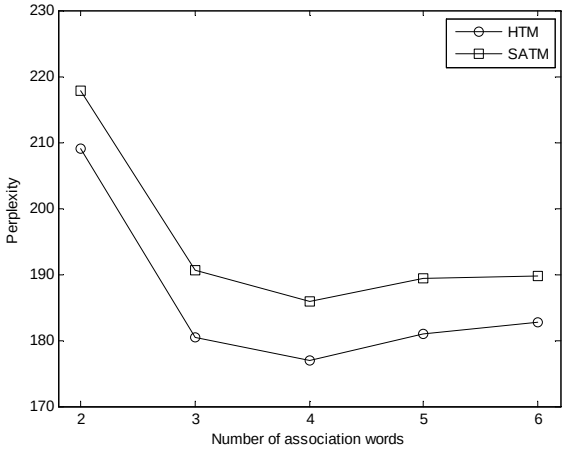

Figure 5. Performance of SATM and HTM in perplexity measure

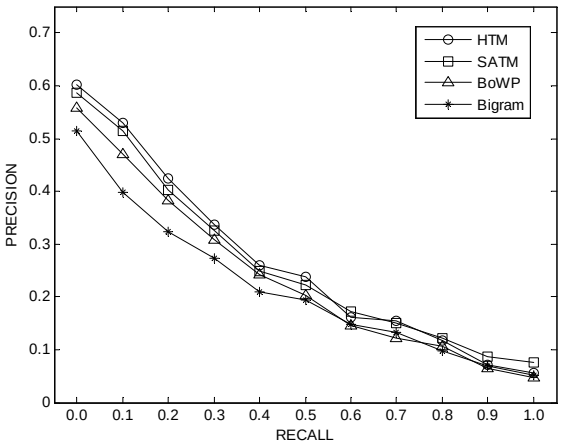

Figure 6. Precision-recall curves for different methods on AP dataset.

	bigram	APLM	BoWP	SATM
Perplexity	262.73	248.6	232.49	217.8

Table 1. Comparison of perplexity using bigram, APLM, BoWP and SATM in case of aw = 2

To examine the effect on different numbers of associated words (aw), we calculated perplexities for cases of aw = 2 to 6. The experimental results of the performance comparison our proposed SATM and hybrid topic model (HTM) under different maximal association steps is shown in Figure 5. From the Figure 5, it is

clear that HTM outperforms the SATM consistently, and that more than 3% relative perplexity reduction is achieved for all trials. The lowest perplexity of 176.92 was attained for the case of aw = 4, which is better than the 185.93 achieved using SATM.

5.3. Experimental results on information retrieval

In the next set of experiments, we evaluated the performance of our proposed models on the TREC ad-hoc information retrieval tasks. We used the bigram language model as the baseline system. The precision-recall curve for different methods was compared. Here, we only report the associated words with a two association step. Figure 6 displays the comparison of precision-recall curves for the different methods using the AP dataset. Compared to the bigram model, APLM, and BoWP, the proposed models obtain the highest performance. We also calculated the mean average precision (mAP), which denotes the mean of average precision over a set of queries. In each query test, we retrieved 1000 documents and calculated the average precision over all topics. The mAP of HTM was 0.2384, which is better than those obtained by the bigram model (0.2082), BoWP (0.2296), and SATM (0.2328). These promising results illustrate the advantage of using a bag-of-associated words model for information retrieval.

6. CONCLUSION AND FUTURE WORK

This work relaxed the assumptions of the bag-of-words in topic modeling paradigm, and considered the useful information of semantic associated words in the latent topic. In this paper, we presented a novel semantic associative topic model in which the word associations of the frequent word sets consisted of more than two non-contiguous words were merged in topic model. We used information-theoretic criteria and measurements to judge whether the selected word sets are frequent and enables us to extract more semantic information. We also present the hybrid topic model, which combines the PLSA with our proposed model. From the experimental results, the proposed methods achieved better performance on perplexity metrics and mean average precision compared to baseline *n*-gram model and other topic-based language model methods. In the future, we will apply our method to other topic models such as LDA and their variations. We will also investigate the effect of the feature selection method of associated word mining. We are applying the proposed model to the tasks of spoken document classification and retrieval.

7. REFERENCES

[1] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules”, in *Proc. of the International Conference on Very Large Data Bases*, pp. 48-499, 1994.

[2] M. W. Berry, S. T. Dumais and G. W. O’Brien, “Using Linear algebra for Intelligent Information Retrieval”, *Society for Industrial and Applied Mathematics: Review*, vol. 37, no. 4, pp. 573-595, 1995.

[3] D. M. Blei, A. Y. Ng and M. I. Jordan, “Latent Dirichlet allocation”, *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.

[4] H. L. Chang, Y. C. Pan and L. S. Lee, “Latent Semantic Retrieval of Spoken Documents over Position Specific Posterior Lattices”, *IEEE Workshop on Spoken Language Technology*, pp.285-288, 2008.

[5] L. Chen, KK Chin and K. Knill, “Improved language modelling using bag of word pairs”, in *Proc. of Interspeech*, pp.2671-2674, 2009.

[6] J. T. Chien and H. Y. Chen, “Mining of association patterns for language modeling”, in *Proc. of International Conference on Spoken Language Processing*, pp. 1369-1372, 2004.

[7] J. T. Chien and M. S. Wu, “Adaptive Bayesian latent semantic analysis”, *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 198-207, 2008.

[8] C. H. Chueh and J. T. Chien, “Nonstationary latent Dirichlet allocation for speech recognition”, in *Proc. of Interspeech*, pp. 372-375, 2009.

[9] K. W. Church and P. Hanks, “Word association norms, mutual information, and lexicography”, *Computational Linguistics*, vol. 16, pp. 22-29, 1990.

[10] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, “Indexing by Latent Semantic Analysis”, *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391-407, 1990.

[11] T. Hofmann, “Probabilistic latent semantic analysis”, *in Proc. of Uncertainty in Artificial Intelligence*, pp.289-296, 1999.

[12] J. Nie, R. Li, D. Luo and X. Wu, “Refine bigram PLSA model by assigning latent topics unevenly”, in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, pp.401-406, 2007.

[13] T. Tomokiyo and M. Hurst, “A language model approach to keyphrase extraction”, in *Proc. of the ACL Workshop on Multiword Expressions: Analysis, Acquisition & Treatment*, pp. 33-40, 2003.

[14] H. M. Wallach, “Topic modeling: beyond bag-of-words”, in *Proc. of the 23rd International Conference on Machine Learning*, pp.977-984, 2006.

[15] M. S. Wu and J. T. Chien, "Minimum rank error training for language model", in *European Conference on Speech Communication and Technology*, pp. 614-617, 2007.