Segmented Topic Model for Text Classification and Speech Recognition

Chuang-Hua Chueh
Department of Computer Science and Information Engineering
National Cheng Kung University, Taiwan
chchueh@chien.csie.ncku.edu.tw

Jen-Tzung Chien
Department of Computer Science and Information Engineering
National Cheng Kung University, Taiwan
jtchien@mail.ncku.edu.tw

Abstract

This paper presents a new segmented topic model (STM) to explore the topic regularities and simultaneously partition the text or spoken documents into coherent segments. The topic model based on the latent Dirichlet allocation (LDA) is adopted to extract the topics and is strengthened by incorporating a Markov chain to detect the segments in a document. STM is trained according to a variational Bayesian procedure where a Viterbi decoder is inherent in carrying out the document segmentation. Each segment is represented by a Markov state, and so the nonstationary stylistic and contextual information are captured. The word variations within a document are compensated. In the experiments, STM outperformed LDA for text classification using ICASSP dataset and for speech recognition using WSJ corpus.

1 Introduction

Latent Dirichlet allocation (LDA) [4] has been proposed to generalize new documents and adopted for document representation in text classification and summarization systems [5] as well as language model adaptation in speech recognition system [6]. LDA represents the documents based on bag-of-words scheme and ignores the position of words. However, the usage of words is varied in different segments of a document even if it involves the same topic. Such variations affect the correctness of document representation and word prediction. In this study, a Markov chain is merged in LDA to detect the stylistically-similar segments and estimate the time-varying word statistics of a document. Each segment indicates a specific writing or spoken style in composition of a text or spoken document. This segmented topic model (STM) exploits the topic information across documents and the word variations within a document. In STM parameter inference, a Viterbi variational inference algorithm is presented by running a Viterbi decoding stage in a variational Bayesian EM (VB-EM) procedure. The proposed STM is evaluated for text classification and speech recognition.

2 Segmented topic model

The word distributions in different paragraphs of a text or spoken document are varied due to the composition style and document structure. In addition to topic information, the temporal positions of words are embedded in natural language, e.g. scientific articles and broadcast news documents. Particularly, when a scientific article is related to a specific scope of research topics, the word distributions are varied in the segments of abstract, introduction and experiment. To compensate the temporal variations, a Markov chain is merged to characterize the dynamics of words in different segments as displayed by the graphical representation of STM in Figure 1. The K-dimensional topic mixture vector $\boldsymbol{\theta}$ is drawn from a Dirichlet distribution with parameter $\boldsymbol{\alpha}$. The topic sequence $\mathbf{z}$ is generated by a multinomial distribution with parameter $\boldsymbol{\theta}$. The state sequence $\mathbf{s}=\{s_1,\cdots,s_N\}$ is generated by a Markov chain with an initial state parameter $\boldsymbol{\pi}$ and a $S\times S$ state transition probability matrix $\mathbf{A}=\{a_{s_{n-1}s_n}\}$. Each word is associated with a topic and a segment. The marginal likelihood of a document over unseen variables $\boldsymbol{\theta}$, $\mathbf{z}$ and $\mathbf{s}$ is yielded by

$$p(\mathbf{w} \mid \boldsymbol{\alpha}, \mathbf{B}, \boldsymbol{\pi}, \mathbf{A}) = \int p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) \sum_{\mathbf{s}} \prod_{n=1}^{N} \sum_{z_n=1}^{K} p(w_n \mid z_n, s_n, \mathbf{B}) p(z_n \mid \boldsymbol{\theta}) p(s_n \mid s_{n-1}, \boldsymbol{\pi}, \mathbf{A}) d\boldsymbol{\theta} \,. \qquad (1)$$

Since STM adopts Markov states $\{s_n\}$, the dynamics of words in different positions are characterized. STM calculates the word probability associated with the topic **z** and state **s**. The multinomial parameter matrix $\mathbf{B} = \{b_{s_n z_n w_n}\}$ contains the observation probability of word w_n given latent topic z_n and state s_n. Decoding the best state sequence is comparable of calculating the document probability and finding the segment boundaries at the same time.

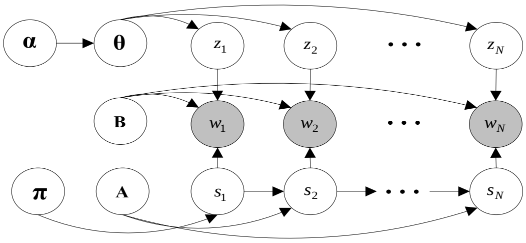

Figure 1: Graphical representation of STM.

3 Viterbi variational inference

3.1 Variational inference

STM parameters $\{\boldsymbol{\alpha}, \mathbf{B}, \boldsymbol{\pi}, \mathbf{A}\}$ are estimated by maximizing the logarithm of marginal likelihood in (1) which is accumulated from *M* documents $\{\mathbf{w}_d\}$. However, it is intractable to directly optimize the joint likelihood due to the coupling among $\{\boldsymbol{\theta}, \mathbf{B}, \boldsymbol{\pi}, \mathbf{A}\}$ in the summation over topics and states. Accordingly, the lower bound of the logarithm of marginal likelihood is maximized by adopting a factorized variational distribution $q_d(\boldsymbol{\theta}, \mathbf{z}, \mathbf{s}) = q_d(\boldsymbol{\theta}) q_d(\mathbf{z}) q_d(\mathbf{s})$ which approximates the true posterior $p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{s} \mid \mathbf{w}_d, \boldsymbol{\alpha}, \mathbf{B}, \boldsymbol{\pi}, \mathbf{A})$. An VB-EM procedure [4] is performed to estimate the variational distributions $\{q_d(\boldsymbol{\theta} \mid \boldsymbol{\gamma}), q_d(\mathbf{z} \mid \boldsymbol{\varphi}), q_d(\mathbf{s} \mid \boldsymbol{\zeta})\}$ and the model parameters $\{\boldsymbol{\alpha}, \mathbf{B}, \boldsymbol{\pi}, \mathbf{A}\}$. In VB-E step, the optimal variational parameters $\{\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\varphi}}, \hat{\boldsymbol{\zeta}}\}$ corresponding to latent variables $\{\boldsymbol{\theta}, \mathbf{z}, \mathbf{s}\}$ are derived to obtain the highest lower bound. In the VB-M step, the lower bound given the variational distribution $q_d(\boldsymbol{\theta}, \mathbf{z}, \mathbf{s} \mid \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\varphi}}, \hat{\boldsymbol{\zeta}})$ is further maximized to find the closed-form solutions to $\{\hat{\boldsymbol{\pi}}, \hat{\mathbf{A}}, \hat{\mathbf{B}}\}$ [6]. The Newton-Raphson algorithm [4] is employed to estimate the Dirichlet parameter $\hat{\boldsymbol{\alpha}}$.

3.2 Viterbi segmentation

However, such a procedure suffers from high computation due to the consideration of all possible state sequences. An efficient approach is to adopt the forward-backward algorithm, which is popular in hidden Markov model (HMM) training. For the purpose of document segmentation, we present a *Viterbi variational inference* by merging a Viterbi segmentation stage into the VB-EM procedure. The best state sequence $\hat{\mathbf{s}}_d$ of document $\mathbf{w}_d$ is obtained by maximizing the joint probability of $\mathbf{w}_d$ and $\mathbf{s}_d$

$$\hat{\mathbf{s}}_d = \arg\max_{\mathbf{s}} p(\mathbf{w}_d, \mathbf{s} \mid \boldsymbol{\alpha}, \boldsymbol{\pi}, \mathbf{A}, \mathbf{B}) = \arg\max_{\mathbf{s}} \int p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) \sum_{\mathbf{z}} p(\mathbf{z} \mid \boldsymbol{\theta}) p(\mathbf{w}_d \mid \mathbf{s}, \mathbf{z}, \mathbf{B}) p(\mathbf{s} \mid \boldsymbol{\pi}, \mathbf{A}) d\boldsymbol{\theta} \,. \qquad (2)$$

Considering the variational inference, the optimization problem turns out to maximize the lower bound of $p(\mathbf{w}_d, \mathbf{s} \mid \boldsymbol{\alpha}, \boldsymbol{\pi}, \mathbf{A}, \mathbf{B})$, i.e.

$$\begin{aligned} \hat{\mathbf{s}}_d &= \arg\max_{\mathbf{s}} < \log\{p(\mathbf{w}_d \mid \mathbf{z}, \mathbf{s}, \mathbf{B}) p(\mathbf{s} \mid \boldsymbol{\pi}, \mathbf{A})\} >_{q(\mathbf{z})} = \arg\max_{\mathbf{s}} \{ \log p(\mathbf{s} \mid \boldsymbol{\pi}, \mathbf{A}) \\ &+ < \log p(\mathbf{w}_d \mid \mathbf{z}, \mathbf{s}, \mathbf{B}) >_{q(\mathbf{z})} \} = \arg\max_{\mathbf{s}} \left[\log \pi_{s_1} + \sum_{n=2}^{N} \log a_{s_{n-1} s_n} + \sum_{n=1}^{N} \sum_{k=1}^{K} \phi_{dkn} \log b_{s_n k w_n} \right]. \end{aligned} \qquad (3)$$

where $< \log p(y) >_{q(y)}$ denotes an expectation function and ϕ_{dkn} is a variational parameter in $q_d(\mathbf{z} \mid \boldsymbol{\varphi})$ representing the approximate posterior probability $q_d(k \mid w_n, \mathbf{w}_d)$ of word w_n in document $\mathbf{w}_d$ belonging to topic *k*. Notably, (3) shows a kind of HMM calculation with new output probability calculated by $\exp(\sum_{k=1}^{K}\phi_{dkn}\log b_{s_n k w_n})$, which is derived from the expectation function over hidden topics with variational probabilities $\{\phi_{dkn}\}$. The best state sequence $\hat{\mathbf{s}}_d$ is searched to efficiently implement STM by

$$\hat{\pi}_i = \frac{\sum_{d=1}^{M}\delta(s_i, \hat{s}_{dn})}{\sum_{s=1}^{S}\sum_{d=1}^{M}\delta(s_s, \hat{s}_{dn})} \tag{4}$$

$$\hat{a}_{ij} = \frac{\sum_{d=1}^{M}\sum_{n=2}^{N_d}\delta(s_i, \hat{s}_{dn})\delta(s_j, \hat{s}_{d(n+1)})}{\sum_{s=1}^{S}\sum_{d=1}^{M}\sum_{n=2}^{N_d}\delta(s_i, \hat{s}_{dn})\delta(s_s, \hat{s}_{d(n+1)})} \tag{5}$$

$$\hat{b}_{jkv} = \frac{\sum_{d=1}^{M}[\sum_{n=1}^{N_d}\phi_{dkn}\delta(s_j, \hat{s}_n)\delta(w_v, w_n)]}{\sum_{m=1}^{V}\{\sum_{d=1}^{M}[\sum_{n=2}^{N_d}\phi_{dkn}\delta(s_j, \hat{s}_n)\delta(w_m, w_n)]\}} \tag{6}$$

where V is the vocabulary size and $\delta(w_v, w_n)$ is the Kronecker delta function that returns one when word w_v is identical to w_n and returns zero otherwise. In the initialization, the state alignment is presumed by uniform partition. The variational distributions are estimated according to the current state sequence $\hat{\mathbf{s}}_d$ in VB-E step. Next, we perform the Viterbi decoding and realign all training documents by accumulating the best score $G_n(s_n) = \max_{s_{n-1}} G_{n-1}(s_{n-1}) \cdot a_{s_{n-1}s_n} \cdot \exp(\sum_{k=1}^{K}\phi_{dkn}\log b_{s_n k w_n})$, and recording the most likely state $R_n(s_n) = \arg\max_{s_{n-1}} G_{n-1}(s_{n-1}) \cdot a_{s_{n-1}s_n} \cdot \exp(\sum_{k=1}^{K}\phi_{dkn}\log b_{s_n k w_n})$ for each word w_n. The score $G_n(s_n)$ approximates the lower bound of marginal likelihood. The optimal state label for each word is obtained by performing the backtracking stage as $\hat{s}_n = R_{n+1}(\hat{s}_{n+1})$. Next, the model parameters $\{\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\pi}}, \hat{\mathbf{A}}, \hat{\mathbf{B}}\}$ are estimated in VB-M step.

3.3 Comparison of different topic models

Some topic models were compared with the proposed STM. Blei *et al.* presented a hierarchical topic model [1] to build the hierarchy of topics and subtopics. Differently, hidden variables in STM indicate the topics and states which are used to compensate the variations of topic-based word distributions due to the positions within a document. The HMM-LDA [7][9] merged an HMM into LDA model to characterize the syntax regularities and focused on separately modeling the function words and the content words. An HMM was adopted to indicate the emission of a word either from a syntactic state or from a semantic topic and was estimated by Markov chain Monte Carlo method. A general Markov chain was employed. STM differs from HMM-LDA in aspects of state representation and word generation. STM jointly adopts the positional states and semantic topics in word generation for document segmentation. The temporal variations are compensated by left-to-right Markov chain without skip using few states. Also, the dynamic topic model [2] was proposed to capture the time evolution of topics in large document collections. A Gaussian noise model was adopted to chain the topic parameters in a state space model. In contrast, the proposed STM models the word variations within a document and detect stylistically-similar segments using a Markov chain. In addition, Gruber *et al.* [8] presented the hidden topic Markov model (HTMM) where the words were generated dependently by using an extended LDA by considering the topic transition. All words in the same sentence were associated with the same topic. The successive sentences were more likely to involve the same topic. The proposed STM emphasizes on extracting the nonstationary word distributions within a document. The word distributions in different positions were concerned rather than the topic variations of a document by using HTMM.

4 Experiments

In the experiments on document categorization, a total of 1211 ICASSP papers with four categories were collected. Here, 960 documents were used for training and the remaining 251 documents were used for testing. The most frequent 8598 words were extracted to build the lexicon. The case of 20 topics and three states was considered in implementation of STM. Table 1 shows the Kullback-Leibler (KL) divergence between different states in the first topic of a 3-state STM. The nonstationary property of word distributions in a document is obvious. The word distributions change at different states. The longer the KL divergence, the larger the difference between two states is measured. Table 2 displays the classification accuracies using the vector space model (VSM), LDA (or equivalently 1-state STM), 2-state and 3-state STMs.

2-state STM outperformed VSM and LDA. 3-state STM did not work well. This situation is caused by the overtraining problem. The issue of model complexity should be tackled.

Table 1: KL divergence between word distributions in various states

	State 1	State 2	State 3
State 1	0	6.16	16.87
State 2	6.3	0	18.1
State 3	13.8	10.91	0

Table 2: Classification accuracies of VSM, LDA and STM with various numbers of states

VSM	LDA (1-state STM)	2-state STM	3-state STM
82%	84%	89%	87%

The Wall Street Journal corpus (WSJ) [6] was adopted when evaluating STM for speech recognition. The probability of a word sequence W in a spoken document was calculated using n-gram model. We performed the Viterbi VB-EM procedure using W and obtained the marginal likelihood $p_{\text{STM}}(W)$ which was linearly interpolated with a background trigram model. The composite language model was employed in language model rescoring for speech recognition. The test data of WSJ consisted of individual utterances rather than the whole spoken documents. To employ STM in WSJ broadcast news transcription, the Viterbi VB-EM was modified in order to search the best state sequence $\hat{\mathbf{s}}_d$ for test utterances. Since the position of a test utterance in the document was unknown, the limitation of starting state and ending state was relaxed. The state initial probabilities were specified to be equal so that different states would be the starting state in calculation of sentence probability [6]. The word error rate (WER) using baseline trigram was 5.4%. The WERs of LDA and 3-state STM were reduced to 5.2% and 5.1%, respectively. The reduction of WER using STM was slight due to the limited length of test utterances. The improvement of STM in speech recognition should be significant when evaluating the overall performance of spoken document transcription.

5 Conclusions

We have presented a new segmented topic model to combine the text segmentation and topic-based model under a unified statistical model and a consistent objective function. This approach explored the topic regularities and partitioned the text documents into coherent segments. A Markov chain was embedded to detect similar segments in a document. Each segment was viewed as a composition unit of a document. The variational solutions were derived to maximize the lower bound of marginal likelihood of training data and develop the Viterbi VB-EM procedure for efficient implementation. STM was evaluated by the experiments of document classification and speech recognition. STM obtained better classification accuracy and word error rate than LDA. STM extracted the topic information and document structure. In the future, we will adopt the semi-Markov chain to analyze the duration of each segment in STM. We will automatically detect suitable state number to build a regularized document model. Also, the supervised version [3] of STM will be developed to improve the prediction of future unlabeled documents by jointly modeling the documents and the categories.

References

[1] Blei, D. M., Griffiths. T. L., Jordan, M. I. and Tenenbaum, J. B. (2003) Hierarchical topic models and the nested Chinese restaurant process. In *Advances in Neural Information Processing Systems* 16, pp. 17-24.

[2] Blei, D. M. and Lafferty, J. D. (2006) Dynamic topic model. In *Proceedings of the International Conference on Machine Learning* 148, pp. 113-120.

[3] Blei, D. M. and McAuliffe, J. D. (2008) Supervised topic models. In *Advances in Neural Information Processing Systems*.

[4] Blei, D. M., Ng, A.Y. and Jordan, M.I. (2003) Latent Dirichlet allocation. *Journal of Machine Learning Research* 3(5): 993-1022.

[5] Chang, Y.-L. and Chien, J.-T. (2009) Latent Dirichlet learning for document summarization. In *Proceedings of ICASSP*, pp. 1689-1692.

[6] Chueh, C.-H and Chien, J.-T. (2009) Nonstationary latent Dirichlet allocation for speech recognition. In *Proceedings of Interspeech*, pp. 372-375.

[7] Griffiths, T.L., Steyvers, M., Blei, D. M. and Tenenbaum, J. B. (2004) Integrating topics and syntax. In *Advances in Neural Information Processing Systems* 17, pp. 537-544.

[8] Gruber, A., Rosen-Zvi, M. and Weiss, Y. (2007) Hidden topic Markov models. In *Proceedings of the Conference on Artificial Intelligence and Statistics*.

[9] Hsu, B.-J. and Glass, J. (2006) Style & topic language model adaptation using HMM-LDA. In *Proceedings of Empirical Methods in Natural Language Processing*, pp. 373-381.