# CORRELATED LATENT SEMANTIC MODEL FOR UNSUPERVISED LM ADAPTATION

*Yik-Cheung Tam and Tanja Schultz*

InterACT, Language Technologies Institute,
Carnegie Mellon University,
Pittsburgh, PA 15213
{yct,tanja}@cs.cmu.edu

**ABSTRACT**

We propose a Latent Dirichlet-Tree Allocation (LDTA) model - a correlated latent semantic model - for unsupervised language model adaptation. The LDTA model extends the Latent Dirichlet Allocation (LDA) model by replacing a Dirichlet prior with a Dirichlet-Tree prior over the topic proportions. Latent topics under the same subtree are expected to be more correlated than topics under different subtrees. The LDTA model falls back to the LDA model using a depth-one Dirichlet-Tree, and the model fits to the variational Bayes inference framework employed in the LDA model. Empirical results show that the LDTA model has a faster training convergence than the LDA model with the same initial flat model. Experimental results show that LDTA-adapted LM performed better than LDA-adapted LM on the Mandarin RT04-eval set when the models were trained using a small text corpus, while both models had the same recognition performance when the models were trained using a big text corpus. We observed 0.4% absolute CER reduction after LM adaptation using LSA marginals.

***Index Terms***— correlated topics, Dirichlet-Tree, LSA, unsupervised LM adaptation

## 1. INTRODUCTION

Latent Dirichlet Allocation (LDA) [1] has been proposed to model the latent topics of a text corpus. The LDA model has been found useful in unsupervised LM adaptation on large-vocabulary automatic speech recognition systems [2, 3, 4]. The LDA model can be viewed as a Bayesian extension of the unigram mixture model by putting a Dirichlet prior over the topic mixture weights (or topic proportions). One assumption made in the Dirichlet prior is that apart from the constraint that the topic proportions sum to unity, they are basically independent. That means that knowing the proportion of one topic does not provide any information about the proportion of another topic. In reality, the assumption may not be true since topics may be correlated. For instance, news articles in a newspaper website are usually organized into main-topic and sub-topic hierarchy. From a human point of view, it would be advantageous to model the correlation among topics. We are interested in using machine learning technique to discover them in an unsupervised fashion. In this paper, we propose an extension of the LDA model - the Latent Dirichlet-Tree Allocation (LDTA) - as a correlated latent semantic model. The idea is to employ a Dirichlet-Tree prior [5, 6] over the topic proportions instead of using a single Dirichlet prior. Each node in the tree is represented by a Dirichlet distribution over the branches to its child nodes. Each latent topic is attached to the leaf node of the tree as illustrated in Figure 1. Apart from using different prior, the LDTA and LDA models are essentially the same in terms of their generative nature. To sample a vector of topic proportions from the Dirichlet-Tree prior, we first sample the branching probabilities from a Dirichlet distribution in each node independently. We compute the topic proportion as the product of branching probabilities realized as walking through a path from the root node to the leaf node which corresponds to a topic index. Correlation among topic proportions can be modeled. Topics under a common subtree are more correlated than topics under different subtrees.

Related work includes the Correlated Topic Model (CTM) [7] and the Pachinko Allocation Model (PAM) [8]. Essentially the CTM model is also an extension of the LDA model by replacing the Dirichlet prior with a logistic-normal prior. Correlation among topic proportions are modeled by first sampling a vector from a multivariate Gaussian distribution. Then the vector is mapped to a vector of topic proportions through the logistic normal distribution. Topic correlations are thus modeled through the covariance matrix of the Gaussian distribution. Despite its flexibility of modeling pairwise topic correlation, the non-conjugate logistic-normal prior poses complication on model training and inference. On the other hand, an advantage of the proposed LDTA model is that it enjoys the simplicity and similarity in training and inference as a LDA model. We can view the LDA model as a special case of the LDTA model with a depth-one Dirichlet-Tree. PAM [8] uses a direct-acyclic graph (DAG) to model the correlation among topics. Each node in the DAG is represented by a Dirichlet distribution over the child links which can be viewed as a Dirichlet-DAG prior. PAM can be viewed as a generalization of the LDTA model, and Gibbs sampling technique was employed for training and inference in their work. In this paper, we present a variational Bayes inference framework for efficient LDTA training and inference.

The paper is organized as follows: In Section 2, we describe the LDTA model training and inference, and review a LM adaptation approach to integrate latent semantic analysis (LSA) into a background N-gram LM. In Section 3, we analyze and compare the LDTA and LDA models with recognition experiments, followed by conclusions and future work in Section 4.

## 2. LATENT DIRICHLET-TREE ALLOCATION

In the LDA model, we sample the topic proportions from a Dirichlet prior which implicitly assumes that the latent topics are independent.

This work is partly supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-2-0001. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

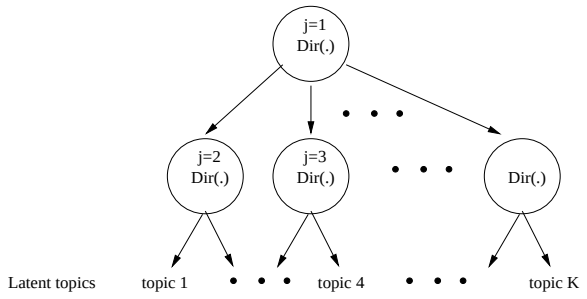


**Fig. 1**. Dirichlet-Tree prior of depth two: Each internal node is represented by a Dirichlet distribution over the branches.

Our target is to relax the independence assumption by employing a Dirichlet-Tree prior in which the topic correlations are modeled through the tree structure. Each internal node in the tree represents a Dirichlet distribution over the branches to its child nodes. Figure 1 illustrates a depth-two Dirichlet-Tree. We can see that a tree of depth one is simply a single Dirichlet distribution corresponding to the LDA model. The LDTA model is a generative model and enjoys the simplicity and similarity as a LDA model. Given a Dirichlet-Tree T of a fixed structure parameterized by a set of Dirichlet parameters $\{\alpha_j\}$ (i.e. the "pseudo-counts" of the branches), we generate a document $w_1^n$ as follows:

1. Sample a vector of branch probabilities $b_j \sim Dir(\alpha_j)$ for each node j=1...J.
2. Compute the topic multinomial (proportions) as:

$$\theta_k = \prod_{jc} b_{jc}^{\delta_{jc}(k)} \quad (1)$$

   where $\delta_{jc}(k)$ is an indicator function which sets to unity when the c-th branch of the j-th node leads to the leaf node of topic k and zero otherwise. The k-th topic proportion $\theta_k$ is computed as the product of branch probabilities from the root node to the leaf node of topic k.
3. Generate a document using the topic multinomial for each word $w_i$:

$$\begin{aligned} z_i &\sim Mult(\theta) \\ w_i &\sim Mult(\beta_{.z_i}) \end{aligned}$$

   where $\beta_{.z_i}$ denotes the topic-dependent unigram LM indexed by $z_i$.

The latent variables in the LDTA model are the topic sequence $z_1^n$ and a set of branch probabilities $\{b_j\}$ in the Dirichlet-Tree. The joint distribution [1] of the latent variables and an observed document is as follows:

$$p(w_1^n, z_1^n, b_1^J) = p(b_1^J|\{\alpha_j\}, T) \prod_i^n \beta_{w_i z_i} \cdot \theta_{z_i}$$

$$\begin{aligned} \text{where} \quad p(b_1^J|\{\alpha_j\}, T) &= \prod_j^J Dir(b_j; \alpha_j) \\ &\propto \prod_{jc} b_{jc}^{\alpha_{jc}-1} \end{aligned}$$

[1] We assume that $\theta$ is not a latent variable for computational convenience.

Similar to LDA training, we apply variational Bayes approach by optimizing the lower bound of the marginalized likelihood of a document $w_1^n$:

$$\begin{aligned} L &= E_q[\log \frac{p(w_1^n, z_1^n, b_1^J; \Lambda)}{q(z_1^n, b_1^J; \Gamma)}] \\ &= E_q[\log p(w_1^n|z_1^n)] + E_q[\log \frac{p(z_1^n|b_1^J)}{q(z_1^n)}] \\ &\quad + E_q[\log \frac{p(b_1^J; \{\alpha_j\})}{q(b_1^J; \{\gamma_j\})}] \end{aligned}$$

where $q(z_1^n, b_1^J; \Gamma) = \prod_i^n q(z_i) \cdot \prod_j^J q(b_j)$ is a factorizable variational posterior distribution over the latent variables parameterized by $\Gamma$ which are determined in the E-step. $\Lambda$ is the actual model parameters for a Dirichlet-Tree $\{\alpha_j\}$ and the topic-dependent unigram LM $\{\beta_{vk}\}$.

Notice that the Dirichlet-Tree has a conjugate counterpart like a Dirichlet distribution. That means that the posterior Dirichlet-Tree has the same form as the Dirichlet-Tree prior given the topic sequence $z_1^n$:

$$\begin{aligned} p(b_1^J|z_1^n) &\propto p(z_1^n|b_1^J) \cdot p(b_1^J; \{\alpha_j\}) \\ &\propto \prod_i^n \prod_{jc} b_{jc}^{\delta_{jc}(z_i)} \cdot \prod_{jc} b_{jc}^{\alpha_{jc}-1} \\ &= \prod_{jc} b_{jc}^{(\alpha_{jc}+\sum_i^n \delta_{jc}(z_i))-1} \end{aligned}$$

The conjugateness suggests an E-step similar to the LDA model [1]:

E-Steps:

$$\gamma_{jc} = \alpha_{jc} + \sum_i^n \sum_k^K q(z_i = k) \cdot \delta_{jc}(k) \quad (2)$$

$$q(z_i = k) \propto \beta_{w_i k} e^{E_q[\log \theta_k]} \quad (3)$$

$$\begin{aligned} \text{where } E_q[\log \theta_k] &= \sum_{jc} \delta_{jc}(k) E_q[\log b_{jc}] \\ &= \sum_{jc} \delta_{jc}(k) \left( \Psi(\gamma_{jc}) - \Psi(\sum_c \gamma_{jc}) \right) \end{aligned}$$

Equation 2–3 are executed iteratively until convergence is reached. Intuitively, Equation 2 can be implemented as the propagation of fractional topic posterior counts from the leaf nodes to the internal nodes in a bottom-up fashion.

M-Step:

$$\beta_{vk} \propto \sum_i^n q(z_i = k) \cdot \delta(w_i, v)$$

where $\delta(w_i, v)$ is a Kronecker Delta function. Similar to LDA training, the alpha parameters can be estimated with iterative methods such as Newton-Raphson or simple gradient ascent procedure.

### 2.1. LM adaptation approach

We followed our previous work [3] on LM adaptation by minimizing the Kullback-Leibler divergence between the adapted LM and

the background LM. The approach has two steps. Firstly, we estimated the in-domain $Pr(w)$ using the LSA marginals. With the LDTA model, we applied variational Bayes inference (Equation 2–3) to obtain the branch posterior counts $\gamma_{jc}$. We computed the relative frequencies of the counts and computed the topic posterior probabilities as follows:

$$Pr_{ldta}(w) = \sum_{k=1}^{K} \beta_{wk} \cdot \hat{\theta}_k \quad (4)$$

$$\text{where } \hat{\theta}_k \propto \prod_{jc} \left( \frac{\gamma_{jc}}{\sum_{c'} \gamma_{jc'}} \right)^{\delta_{jc}(k)} \quad (k = 1...K) \quad (5)$$

Secondly, we integrated the in-domain LSA marginals into the background LM using the following equation [9]:

$$Pr_a(w|h) \propto \left( \frac{Pr_{ldta}(w)}{Pr_{bg}(w)} \right)^{\beta} \cdot Pr_{bg}(w|h) \quad (6)$$

where $\beta$ is set to 0.5 in all reported experiments.

## 3. EXPERIMENTAL SETUP

We evaluated the LM adaptation approach on the ISL-RT04 Mandarin Broadcast News evaluation system [10] using the JANUS speech recognition toolkit. The system employed context-dependent Initial-Final acoustic model. We trained the acoustic models using 27 hours of the Mandarin HUB4 1997 training set and 69 hours of the TDT4 Mandarin data. We used the 42-dimensional features after Linear Discriminant Analysis projected from a window of MFCC and energy features for the front-end processing. The system employed a two-pass decoding strategy using speaker-independent and speaker-adaptive acoustic models for the first-pass and the second-pass decoding respectively. In the second-pass decoding, we applied the state-of-the-art acoustic adaptations [10]: Vocal Tract Length Normalization (VTLN), Feature Space Adaptation (FSA), and Maximum Likelihood Linear Regression (MLLR). The vocabulary size is 108K words. Performance metrics are the word perplexity and the character error rates (CER) evaluated on the RT04-eval set containing three shows: CCTV, RFA and NTDTV. We trained the background 4-gram LM using the modified Kneser-Ney smoothing scheme using the SRI LM toolkit. We trained the LDA and the LDTA models with 200 topics. We first performed first-pass decoding on the test audios to obtain the automatic transcription for each show. Treating the automatic transcription of a show as a single "document", we applied variational Bayes inference to estimate the LSA marginals for each show. We adapted the background 4-gram LM using the LM adaptation technique described in Section 2.1 and used the LSA-adapted LM for second-pass decoding.

### 3.1. Analysis of the LDTA model

We first investigated the likelihood convergence behavior of the LDTA and LDA models when the number of training iterations increases. Both models shared the same initial flat model with $\{\beta_{vk}\}$ initialized with uniform distributions, while the Dirichlet and the Dirichlet-Tree priors were initialized randomly. We employed a balanced binary tree of depth $log_2(K)$ with $K = 200$. Figure 2 shows that LDTA training converges faster than LDA training. The fast convergence is attributed to the Dirichlet-Tree prior, which models topic correlations. In other words, an observed topic would somehow trigger its correlated topics. On the contrary, the Dirichlet prior in the LDA model assumes topic independence which may explain the slow convergence. In particular, LDA training was trapped in the likelihood plateau in the early training stage and the model initialization sounds an important issue in LDA training, and it is even severe when the number of latent topics increases. On the other hand, model initialization appears not an issue for LDTA training empirically, even when the number of topics increases. We believe that the tree structure helps constrain the model parameter space compared to a flat Dirichlet prior and thus helps speed up the training convergence.

| Latent topic index | Top words (translated from Chinese) |
|---|---|
| "topic-61" | education, student, school, teacher, learning |
| "topic-62" | university, expert, high-level, education, training |
| "topic-63" | employment, expert, labor, work, career |
| "topic-64" | research, china, science, technology, scientist |
| "topic-65" | gene, human, clone, research, biology |
| "topic-66" | research, discover, cell, gene, treatment |
| "topic-67" | transplant, surgery, patient, liver, hospital |
| "topic-68" | information, network, service, web, client |
| "topic-69" | system, computer, technology, computer, chip, software |

**Table 1**. Sample contiguous fragment of latent topics extracted from the LDTA model.

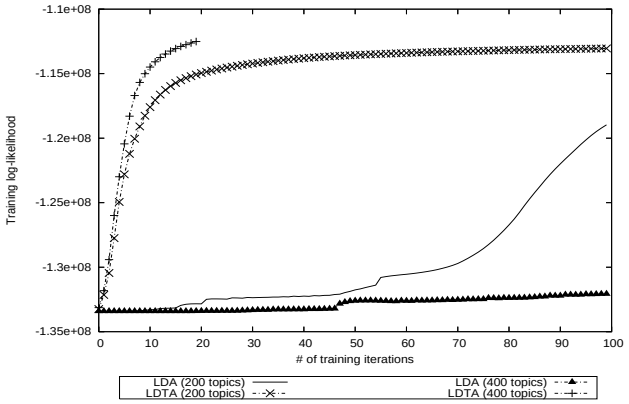


**Fig. 2**. Log-likelihood comparison of the LDA and LDTA models on the Xinhua News 2002 training corpus.

Next, we explore how the LDTA latent topics look like. By inspecting the top-N words of the topic-dependent unigram LM, we observed contiguous fragments of correlated topics shown in Table 1 with the topic indices assigned in the left-to-right fashion in the tree illustrated in Figure 1. For instance, topics 61–63 are closely related to a general topic "education" and topics 68–69 are closely related to a general topic "information technology". The results agree with our intuition that the tree structure enforces proximity constraint over the topics. From Table 1, we conjecture that the LDTA model is able to extract fine-grained correlated topics and potentially supports more topics.

### 3.2. LSA training using small corpora

We compared the LDA and LDTA models on unsupervised LM adaptation using LSA marginals explored in our previous work [3]. We trained the LDA and LDTA models using the Xinhua News 2002 corpus containing 13M words and 64k documents. We performed first-pass decoding to obtain the word hypotheses. For each test show, we concatenated the decoded hypotheses to form a single "document", and then estimated the in-domain LDA/LDTA marginals. The marginals were used to adapt the background 4-gram

| LM (13M) | CCTV | RFA | NTDTV |
|---|---|---|---|
| BG LM | 748 | 3655 | 1718 |
| +LDA | 695 | 3451 | 1669 |
| +LDTA | 629 | 3015 | 1547 |

**Table 2**. Perplexity (PPL) on the RT04 test set with LM trained on a small corpus compared to the unadapted background LM (BG).

| LM (13M) | CCTV | RFA | NTDTV | Overall |
|---|---|---|---|---|
| BG LM | 15.8% | 40.1 | 22.0 | 25.3 |
| +LDA | 15.1 | 39.6 | 21.6 | 24.8 |
| +LDTA | 14.8 | 39.0 | 21.5 | 24.5 |

**Table 3**. Character Error Rates (%) on the RT04 test set after the 2nd-pass decoding with LM trained on a small corpus.

LM separately. We used the adapted LM for second-pass decoding. Table 2 and Table 3 shows the word perplexity and the character error rates (CER) results. Results showed that the LDTA model yields lower word perplexity and CER than the LDA model. In particular, the LDTA model reduces the absolute overall CER by 0.3% compared to the LDA model. The LDTA model achieved 10%–17.5% relative word perplexity reduction and 0.8% absolute CER reduction compared to the unadapted 4-gram LM.

### 3.3. LSA training using large corpora

We evaluated the LDA and LDTA models on the CMU-InterACT Mandarin transcription system for the GALE 2006 evaluation. Similar to the ISL-RT04 system [10], the CMU-InterACT system employs multi-pass (3-pass) decoding strategies. We trained the CMU-InterACT system with over 500 hours of quickly transcribed speech data released by the GALE program and the baseline 4-gram LM with over 800M-word corpus including the Mandarin Gigaword corpora V2, broadcast news and broadcast conversation training transcripts and the webdata. The final corpus has over 3M documents for training the LSA and the background 4-gram LM. During testing, we performed show-dependent LSA adaptation using the second-pass word hypotheses as a "document". We used the LSA-adapted LM for third-pass decoding. We trained the LDA and LDTA models with 50 and 20 iterations respectively starting with the same initial flat model. Table 4 and Table 5 shows the word perplexity and the CER results respectively. We found that the LDTA model performs better than the LDA model on word perplexity, and achieved the same CER reduction. Both LSA models reduced perplexity and CER on a large-scale evaluation system. In particular, the LDTA-adapted LM reduces the perplexity relatively in the range of 8.9%–14.5%, and achieved 0.4% absolute CER reduction compared to the unadapted baseline LM.

## 4. CONCLUSIONS AND FUTURE WORKS

We proposed the Latent Dirichlet-Tree Allocation - a correlated latent semantic model. LDTA uses a Dirichlet-Tree prior which generalizes LDA when the depth of the tree is one. We showed that LDTA can be trained within the variational Bayes framework similar to LDA training. Empirically, LDTA training converges much faster than LDA training initialized with the same flat model. The fast convergence holds when the number of latent topics increases. We compared LDA and LDTA and found that LDTA performs better than LDA when they were trained using a small corpus. Both achieved the same recognition performance on a large-scale LM adaptation. Future work includes the investigation of extracting more correlated topics using the proposed technique on a big text corpus and automatic determination of the optimal number of topics.

| LM (800M) | CCTV | RFA | NTDTV |
|---|---|---|---|
| BG LM | 359 | 778 | 868 |
| +LDA (50 iter) | 332 | 703 | 834 |
| +LDTA (20 iter) | 313 | 665 | 791 |

**Table 4**. Perplexity (PPL) on the RT04 test set with LM trained on a big corpus compared to the unadapted background LM (BG).

| LM (800M) | CCTV | RFA | NTDTV | Overall |
|---|---|---|---|---|
| BG LM | 8.3% | 26.3 | 14.4 | 15.9 |
| +LDA (50 iter) | 8.1 | 25.6 | 14.0 | 15.5 |
| +LDTA (20 iter) | 8.3 | 25.3 | 14.2 | 15.5 |

**Table 5**. Character Error Rates (%) on the RT04 test set using the CMU-InterACT Mandarin transcription system for the GALE 2006 evaluation.

## 5. ACKNOWLEDGEMENT

We would like to thank Bing Zhao for helpful discussion on the LDTA model.

## 6. REFERENCES

[1] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation," in *Journal of Machine Learning Research*, 2003, pp. 1107–1135.

[2] Y. C. Tam and T. Schultz, "Language model adaptation using variational bayes inference," in *Proc. of Interspeech*, 2005.

[3] Y. C. Tam and T. Schultz, "Unsupervised Language Model Adaptation using Latent Semantic Marginals," in *Proceedings of the Interspeech*, 2006.

[4] D. Mrva and P. C. Woodland, "Unsupervised language model adaptation for mandarin broadcast conversation transcription," in *Proc. of Interspeech*, 2006.

[5] R. J. Connor and J. E. Mosimann, "Concepts of independence for proportions with a generalization of the dirichlet distribution," *Journal of the American Statistical Association*, vol. 64, pp. 194–206, 1969.

[6] T. Minka, "The dirichlet-tree distribution," in *http://research.microsoft.com/∼minka/papers/dirichlet/minka-dirtree.pdf*, 1999.

[7] D. Blei and J. Lafferty, "Correlated topic models," in *Advances in Neural Information Processing Systems*, 2005.

[8] W. Li and A. McCallum, "Pachinko allocation: DAG-structured mixture models of topic correlations," in *International Conference on Machine Learning*, 2006.

[9] R. Kneser, J. Peters, and D. Klakow, "Language model adaptation using dynamic marginals," in *Proc. of Eurospeech*, 1997, pp. 1971–1974.

[10] H. Yu, Y. C. Tam, T. Schaaf, S. Stüker, Q. Jin, M. Noamany, and T. Schultz, "The ISL RT04 Mandarin Broadcast News Evaluation System," in *EARS Rich Transcription Workshop*, 2004.