# Unsupervised Language Model Adaptation Using Latent Semantic Marginals

*Yik-Cheung Tam and Tanja Schultz*

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

## Abstract

We integrated the Latent Dirichlet Allocation (LDA) approach, a latent semantic analysis model, into unsupervised language model adaptation framework. We adapted a background language model by minimizing the Kullback-Leibler divergence between the adapted model and the background model subject to a constraint that the marginalized unigram probability distribution of the adapted model is equal to the corresponding distribution estimated by the LDA model – the latent semantic marginals. We evaluated our approach on the RT04 Mandarin Broadcast News test set and experimented with different LM training settings. Results showed that our approach reduces the perplexity and the character error rates using supervised and unsupervised adaptation.

**Index Terms**: unsupervised LM adaptation, LSA marginals, Latent Dirichlet Allocation, Mandarin Broadcast News

## 1. Introduction

In automatic speech recognition, unsupervised language model (LM) adaptation is an attractive research area since the automatic transcription from the decoder provides in-domain information which may be useful for adapting the background LM. One challenge is that the automatic transcription usually contains recognition errors. Minimizing their effect is important since it is undesirable to reinforce the errors back to the background LM after adaptation. Different LM adaptation techniques have been proposed in the literature. One technique proposed in [1] attempts to adapt the background LM by minimizing the Kullback-Leibler divergence between the adapted LM and the background LM subject to a constraint that the marginalized unigram distribution of the adapted LM is equal to some unigram distribution which is estimated using an in-domain text data. They called the latter as "dynamic marginals". Similar idea was also proposed earlier in [2]. The approach was shown to reduce the perplexity and the recognition errors successfully when in-domain supervised text data were available for LM adaptation. However, they [1] reported degradation of recognition performance when the background LM was adapted on automatic transcription. We postulate that this may be caused by the recognition errors that were not smoothed out properly in estimating the dynamic marginals based on relative word frequency.

In this paper, we revisit their approach but we propose using the Latent Dirichlet Allocation (LDA) model [3], a Bayesian latent semantic analysis approach, to estimate the dynamic marginals based on automatic transcription. As a latent semantic model, the LDA model contains a set of unigram LM each of which describes a word distribution of a latent topic. In our earlier work [4], we successfully applied the LDA model into unsupervised LM adaptation by interpolating the background LM with the dynamic unigram LM estimated by the LDA incrementally. In this paper, we propose using the LDA-adapted unigram as the dynamic marginal. One advantage is that we only need to estimate the topic mixture weights to compute the LDA-adapted unigram which can be done robustly on small amount of adaptation data compared to relative word frequency. We conjecture that the LDA model provides smoothing effect on recognition errors since the model is adapted by boosting the topic mixture weights instead of directly boosting the probability of misrecognized words in the automatic transcription. Similar approach has been explored using probabilistic Latent Semantic Analysis (pLSA) in [5] but on an *supervised* setting where short text descriptions of the test audios were utilized. We employed the LDA model which provides regularization over the pLSA model due to the Bayesian nature of the LDA model, and explored our approach on the unsupervised setting.

The paper is organized as follows: In Section 2, we provide an overview of the Latent Dirichlet Allocation model and the estimation of the LDA-adapted marginals using Variational Bayes inference. In Section 3, we describe the LM adaptation approach, followed by experiments in Section 4 and conclusions in Section 5.

## 2. Review of Latent Dirichlet Allocation

The goal of LSA is to extract the latent topics from a text corpus which contains a set of documents in an unsupervised fashion. In broadcast news, a document usually refers to a piece of news story within which the latent topics are consistent. Various LSA techniques has been proposed and applied across different research fields, such as SVD-based LSI [6] and its extension [7], pLSI using the EM algorithm [8], Latent Dirichlet Allocation [3] and its extension [9] to model the correlation among topics. The LDA model is a Bayesian extension of a mixture of unigram models where a vector of topic mixture weights $\theta$ is drawn from a prior Dirichlet distribution:

$$f(\theta;\alpha) \quad \propto \quad \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \tag{1}$$

where $\alpha = \{\alpha_1, ..., \alpha_K\}$ represents the prior observation count of the K latent topics and $\alpha_k > 0$. As a "bag-of-word" generative model, the LDA model assigns probability to a document $w_1^n =$

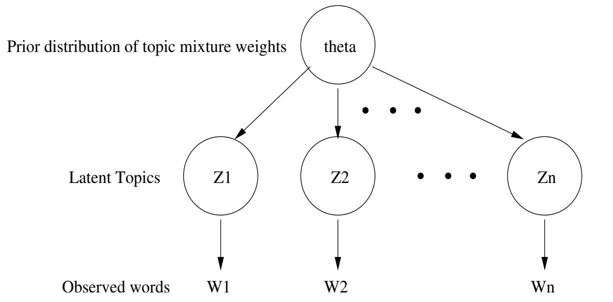


Figure 1: Graphical representation of Latent Dirichlet Allocation.

$w_1 w_2 ... w_n$ as follows:

$$Pr(w_1^n) \quad = \quad \int_\theta \left( \prod_{i=1}^{n} \sum_{k=1}^{K} \beta_{w_i k} \cdot \theta_k \right) f(\theta; \alpha) d\theta \qquad (2)$$

where $\beta_{w_i k}$ denotes the probability of a word $w_i$ given the k-th latent topic. Figure 1 shows the document generation process using the graphical model representation and the circles in the graph represent the latent variables in the model. Optimizing the exact likelihood is computationally intractable. One alternative is to optimize the lower-bound of the log likelihood. It turns out that the lower bound of the log likelihood has the following form:

$$Q(\Lambda, \Gamma) \quad = \quad E_q[\log \frac{f(\theta, w_1^n, z_1^n; \Lambda)}{q(\theta, z_1^n; \Gamma)}] \qquad (3)$$

where $q(\theta, z_1^n)$ is an approximate posterior distribution over all the latent variables given an observed document. In Variational Bayes inference [10], the distribution is factorizable and parameterized by $\Gamma$:

$$q(\theta, z_1^n; \Gamma) \quad = \quad q(\theta; \{\gamma_k\}) \cdot \prod_{i=1}^{n} q(z_i) \qquad (4)$$

where $q(\theta; \{\gamma_k\})$ is a Dirichlet distribution over topic mixture weights parameterized by the "pseudo" topic counts $\{\gamma_k\}$, and $\{q(z_i)\}$ is a set of multinomial distributions over topic indices. Optimizing the auxiliary function Q(.) can be performed using the VB-EM algorithm. The E-step determines the parameters $\Gamma$ of variational posteriors q(.) and the M-step uses q(.) to re-weight the observations to estimate the model parameters $\Lambda$. We only show the results of the parameter estimations of a single document. Complete derivations can be found in [3].

E-Step:

$$\gamma_k \quad = \quad \alpha_k + \sum_{i=1}^{n} q(z_i = k) \qquad (5)$$

$$q(z_i = k) \quad \propto \quad \beta_{w_i k} \cdot e^{E_q[\log \theta_k]} \qquad (6)$$

where $E_q[\log \theta_k] = digamma(\gamma_k) - digamma(\sum_{k=1}^{K} \gamma_k)$. Eqn 5 and Eqn 6 are applied iteratively until convergence.

M-Step:

$$\beta_{vk} \quad \propto \quad \sum_{i=1}^{n} q(z_i = k) \delta(w_i, v) \qquad (7)$$

where $\delta(.)$ is the Kronecker Delta function. Parameters of the Dirichlet prior $\{\alpha_k\}$ can be determined using the Newton-Raphson algorithm or gradient ascent procedure.

### 2.1. Estimating LDA-adapted marginals

We applied our idea proposed in [4], but in the context of estimating a "global" LDA-adapted marginal of the test domain. We first treated the automatic transcription as a single "document". Then we applied Variational Bayes inference (Eqn 5, 6) to estimate the variational Dirichlet posterior over the topic mixture weights. We computed the LDA-adapted marginal as follows:

$$Pr_{lda}(w) \quad = \quad \int_\theta \sum_{k=1}^{K} \beta_{wk} \cdot \theta_k \cdot q(\theta) d\theta \qquad (8)$$

$$= \quad \sum_{k=1}^{K} \beta_{wk} \cdot E_q[\theta_k] \qquad (9)$$

$$\text{where } \; E_q[\theta_k] \quad = \quad \frac{\gamma_k}{\sum_{k=1}^{K} \gamma_k} \quad (k = 1...K) \qquad (10)$$

## 3. LM adaptation approach

The goal of LM adaptation using dynamic marginals [1] is to find an adapted LM $Pr_a(w|h)$ such that the KL divergence between $Pr_a(w|h)$ and the background LM $Pr_{bg}(w|h)$ is minimized subject to the marginalization constraints for each word w in the vocabulary:

$$\sum_h Pr_a(w|h) \cdot Pr_a(h) \quad = \quad Pr_{lda}(w) \; \forall w \qquad (11)$$

The constraint optimization problem has close connection to the maximum entropy approach [11]. It turns out that the form of the adapted model is a rescaled version of the background LM:

$$Pr_a(w|h) \quad = \quad \frac{\alpha(w) \cdot Pr_{bg}(w|h)}{Z(h)} \qquad (12)$$

where Z(h) is a normalization term to guarantee that the probability sums to unity. $\alpha(w)$ is a scaling factor which is commonly approximated as follows:

$$\alpha(w) \quad \approx \quad \left( \frac{Pr_a(w)}{Pr_{bg}(w)} \right)^\beta \qquad (13)$$

where $\beta$ is a tuning factor between 0 and 1. In our reported experiments, we set $\beta$ equal to 0.5. We employed the same strategy proposed in [1] to compute the normalization factor Z(h) efficiently. The idea is to further impose a constraint that the total probability of the observed transition (h,w) in the background training corpus is conserved after LM adaptation:

$$\sum_{w:(h,w)} Pr_a(w|h) = \sum_{w:(h,w)} Pr_{bg}(w|h) = Mass(h)$$

where the summation is taken *only* on the observed history and word pair (h,w) in the training set. Given that our background LM has a standard backoff structure plus the above constraint, the adapted LM has the following recursive backoff formula:

$$Pr_a(w|h) \quad = \quad \begin{cases} \frac{\alpha(w)}{z_0(h)} \cdot Pr_{bg}(w|h) & \text{if (h,w) exists} \\ bo(h) \cdot Pr_a(w|\hat{h}) & \text{otherwise} \end{cases}$$

$$\text{where } z_0(h) \quad = \quad \frac{\sum_{w:(h,w)} \alpha(w) \cdot Pr_{bg}(w|h)}{Mass(h)}$$

$$\text{and } bo(h) \quad = \quad \frac{1 - Mass(h)}{1 - \sum_{w:(h,w)} Pr_a(w|\hat{h})}$$

$bo(h)$ denotes the backoff weight for context h to ensure that $Pr_a(w|h)$ sums to unity. $\hat{h}$ denotes the reduced word history of h. The intuition behind the factor $z_0(h)$ is to perform "normalization" similar to Eqn 12, but the summation is only over the observed alternative words with the same word history h in the LM. For incremental LM adaptation, we can modify the above formula by replacing the background model with the previously adapted model. For example, the alpha computation can be modified as follows:

$$\alpha^{(t)}(w) \approx \left( \frac{Pr_a^{(t)}(w)}{Pr_a^{(t-1)}(w)} \right)^{\beta} \tag{14}$$

where t denotes t-th online adaptation and $Pr_a^{(t=0)}(w)$ denotes the background unigram distribution. We did not evaluate the incremental LM adaptation approach on recognition experiments in this paper. Our preliminary results showed that it brought significant reduction in word perplexity. The reported results are only based on the *batch-mode* LM adaptation.

# 4. Experimental setup

We evaluated the LM adaptation approach on the ISL-RT04 Mandarin Broadcast News evaluation system [12] using the JANUS speech recognition toolkit. The system employs 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-dimension features after Linear Discriminant Analysis projected from a window of MFCC 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 (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 test set containing three episodes: CCTV, RFA and NTDTV. We trained the background 4-gram LM using the modified Kneser-Ney smoothing scheme using the SRI LM toolkit [13]. We trained the LDA model with 200 topics which was found optimal from our previous experience. The LM adaptation procedure is to first perform first-pass decoding on the test audios to obtain the automatic transcription. Treating the automatic transcription as a single "document", we applied the Variational Bayes inference described in Section 2.1 to estimate the LDA-adapted marginals for each test episode. We applied the LM adaptation technique described in Section 3 on the background LM and performed the second-pass decoding using the adapted LM of each episode. We compared adaptation performance using LDA-adapted marginals with word-adapted marginals which are estimated using smoothed relative word frequency with Good-Turing discounting scheme.

### 4.1. Out-Of-Vocabulary rate analysis

One way to measure the quality of the LDA-adapted marginals is to analyze the Out-Of-Vocabulary (OOV) rate with respect to different vocabulary sizes. Figure 2 shows the OOV rates of the CCTV test episode. Ideally, we want a sharp drop of OOV rates when the number of vocabularies is increased. We noticed that the LDA-adapted marginal gives a lower OOV rate compared to the

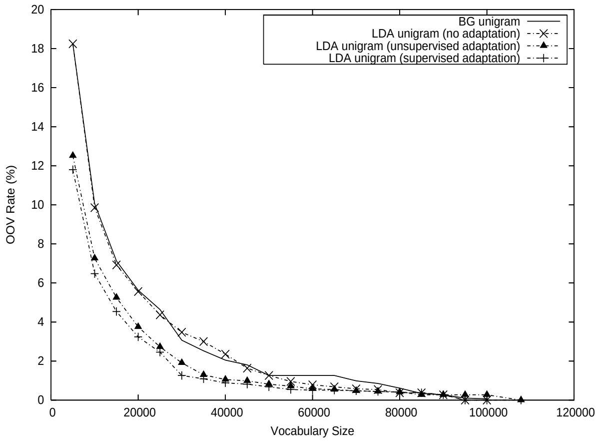


Figure 2: OOV rates using un-adapted and LDA-adapted marginals.

| LM (13M) | CCTV | RFA | NTDTV |
|---|---|---|---|
| BG LM | 748 | 3655 | 1718 |
| +word-based | 655 | 3718 | 1584 |
| +LDA | 673 | 3663 | 1638 |
| +LDA (supervised) | 613 | 2589 | 1518 |

Table 1: Perplexity (PPL) on the RT04 test set with LM trained on a small corpus.

un-adapted background unigram model, showing that the LDA-adapted marginal is more matched to the test data than the un-adapted one. We observed similar trends on the other two test episodes. The figure also shows the "oracle" OOV rate when we used the reference transcription to adapt the LSA model for comparison purpose.

### 4.2. LM training with a small corpus

We first evaluated the LM adaptation approach using a small training corpus which comprises 13M words of Xinhua News 2002 from the Mandarin Gigaword corpus to verify the correctness of our implementation and to simulate a scenario that only small amount of training data is available. We trained the LDA model using the same training corpus. Table 1 shows the word perplexity results. We achieved perplexity reduction with unsupervised LM adaptation using LDA-adapted marginals. One point worths noticing is that the perplexity number depends on the Chinese text segmentation which is determined by the vocabulary to segment the text. We used the 108k vocabulary for text segmentation for the reported experiment. We found that there is a substantial drop in perplexity when the text is segmented with a 63k vocabulary. Table 2 shows the second-pass recognition results. We found that unsupervised LM adaptation using LDA-adapted marginals gives 0.5% absolute reduction on the overall character error rates compared to the un-adapted background LM. We observed that the recognition results are comparable when we computed the exact normalization term Z(h) in Eqn 12 which is computationally expensive. The substantial reduction in computation makes the approach feasible in large scale application. The supervised LM adaptation reduces the absolute overall character error rates by 1.2% which

| LM (13M) | CCTV | RFA | NTDTV | Overall |
|---|---|---|---|---|
| BG LM | 15.8% | 40.1 | 22.0 | 25.3 |
| +word-based | 16.0 | 40.9 | 22.1 | 25.7 |
| +LDA | 15.1 | 39.7 | 21.5 | 24.8 |
| +LDA (supervised) | 14.7 | 38.8 | 20.7 | 24.1 |

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

| LM (600M) | CCTV | RFA | NTDTV |
|---|---|---|---|
| BG LM | 473 | 1159 | 839 |
| +LDA | 405 | 1086 | 778 |
| +LDA (supervised) | 385 | 875 | 738 |

Table 3: Perplexity (PPL) on the RT04 test set with LM trained on a large corpus.

serves as the upper-bound recognition performance. As we expected, there is a degradation in recognition performance for the word-adapted marginals compared to the un-adapted background LM. As mentioned earlier, the result may be explained by the hypothesis that the effect of recognition errors are reinforced after LM adaptation. On the other hand, the LSA-adapted marginals may provide smoothing effect on the recognition errors. The intuition is: each hypothesized word is first projected into the latent topic space where each word "votes" fractionally on how likely the topics are (Eqn 6). Then the "votes" of the words are smoothed by averaging the posterior counts of the topics (Eqn 10). Moreover, estimation in the low dimensional latent topic space is more robust than in the high dimensional vocabulary space due to data sparseness.

### 4.3. LM training with a large corpus

We evaluated the LM adaptation approach on the background LM trained on a large corpus with 600M characters. We trained the LDA model with the full Mandarin Gigaword corpus with over 1M documents. To reduce computation in the LDA training, we used the LDA model from the previous experiment in Section 4.2 as an initial model and applied few training iterations over the whole corpus. Table 3 shows that our approach on unsupervised LM adaptation achieves relative perplexity reduction between 6% – 14% depending on the test episodes. Table 4 shows that our approach yields 0.5% absolute reduction in character error rates compared to the un-adapted baseline. With the supervised LM adaptation, we observed an additional 0.4% absolute reduction in character error rates compared to the unsupervised setting.

| LM (600M) | CCTV | RFA | NTDTV | Overall |
|---|---|---|---|---|
| BG LM | 13.1% | 35.7 | 17.5 | 21.5 |
| +LDA | 12.7 | 34.4 | 17.7 | 21.0 |
| +LDA (supervised) | 12.3 | 34.1 | 17.1 | 20.6 |

Table 4: Character Error Rates (%) on the RT04 test set after the 2nd-pass decoding with LM trained on a large corpus.

## 5. Conclusions and Future Works

We proposed an unsupervised LM adaptation framework by integrating the LDA-adapted marginals into the background LM using Kullback-Leibler divergence criterion. We successfully reduced the word perplexity and character error rates on the RT04 Mandarin Broadcast News test set after applying unsupervised LM adaptation. The LDA-adapted marginals perform better than the word-adapted marginals estimated using relative word frequency. We also showed that incremental adaptation approach reduces the perplexity significantly which may have potential to improve the recognition performance. Future directions include exploring LM adaptation for statistical machine translation (SMT) on the text and automatic transcription, and incremental LM adaptation for ASR and SMT using the proposed approach.

## 6. References


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

[2] S. A. D. Pietra, V. J. D. Pietra, R. L. Mercer, and S. Roukos, "Adaptive language modeling using minimum discriminant estimation," in *Proc. of ICASSP*, 1992, pp. 1633–1636.

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

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

[5] M. Federico, "Language model adaptation through topic decomposition and mdi estimation," in *Proc. of ICASSP*, 2002.

[6] J. Bellegarda, "Exploiting latent semantic information in statistical language modeling," in *IEEE Trans. on ASSP*, vol. 88, no. 8, Aug 2000, pp. 63–75.

[7] J. Bellegarda, "Latent semantic mapping: Dimensionality reduction via globally optimal continuous parameter modeling," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, 2005, pp. 127–132.

[8] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. of UAI*, 1999.

[9] D. Blei and J. Lafferty, "Correlated topic models," in *Advances in Neural Information Processing Systems*, 2006.

[10] M. I. Jordan, Z. Ghahramani, T. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," in *Machine Learning*, vol. 37, no. 2, 1999, pp. 183–233.

[11] R. Rosenfeld, "Adaptive statistical language modeling: A maximum entropy approach," Ph.D. dissertation, School of Computer Science, Carnegie Mellon University, Pittsburgh, April 1994.

[12] 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.

[13] A. Stolcke, "Srilm - an extensible language modeling toolkit," in *Proc. of ICSLP*, 2002.