Dynamic Language Model Adaptation using Variational Bayes Inference

Yik-Cheung Tam and Tanja Schultz

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

Abstract

We propose an unsupervised dynamic language model (LM) adaptation framework using long-distance latent topic mixtures. The framework employs the Latent Dirichlet Allocation model (LDA) which models the latent topics of a document collection in an unsupervised and Bayesian fashion. In the LDA model, each word is modeled as a mixture of latent topics. Varying topics within a context can be modeled by re-sampling the mixture weights of the latent topics from a prior Dirichlet distribution. The model can be trained using the variational Bayes Expectation Maximization algorithm. During decoding, mixture weights of the latent topics are adapted dynamically using the hypotheses of previously decoded utterances. In our work, the LDA model is combined with the trigram language model using linear interpolation. We evaluated the approach on the CCTV episode of the RT04 Mandarin Broadcast News test set. Results show that the proposed approach reduces the perplexity by up to 15.4% relative and the character error rate by 4.9% relative depending on the size and setup of the training set.

1. Introduction

Statistical N-gram language models (LM) are widely used in automatic speech recognition. Despite its popularity, N-gram LM (typically N=3) can only capture local context reliably due to data limitations. Intuitively, word history beyond the N-gram unit should be useful for better word prediction because it somehow tells what the current context or topic is about. We propose a LM adaptation approach which utilizes the observed word context to adapt to the current latent (hidden) topics. Each word is modeled as a mixture of latent topic unigrams. The topic mixture weights are adapted dynamically using the word context which can be arbitrarily long. The adapted model is then interpolated with the standard trigram LM.

Using distant word history for LM adaptation has been investigated in earlier work: Cache-based LM [1, 2] adapts the current context by keeping track of the recently occurred words in an exponentially decaying N-gram cache. The cache LM is shown to be effective especially in dictation tasks where recently occurred words such as name entities tend to occur again in the nearby context. One difference between our approach and the cache-based LM is that we model the word probabilities via the latent topics, whereas the cache-based LM models the word probabilities directly using the word surface form. Another difference is that our approach adapts the mixture weights of the latent topic unigrams whereas cache-based LM adapts the word probabilities directly. Since the number of latent topics is usually much smaller than the vocabulary size, adapting the topic mixture weights would be reliable and can be performed with small amount of adaptation data.

LM adaptation using latent topic mixtures has been explored previously. [3] uses "hard" clustering schemes to cluster documents into topic clusters and each document is assumed coming from a single topic, whereas we employ variational Bayes Expectation-Maximization (VB-EM) algorithm for implicit document clustering. Each document can be associated with multiple topics in our model. Another difference is that [3] employs sentence-level topic mixtures which assumes that the topic mixture weights are fixed within a sentence, whereas the mixture weights in our model can vary within a sentence.

Our work shares similarity with the latent semantic indexing (LSI) approach for LM adaptation. [4] employed LSI to discover latent topics via the Singular-Value Decomposition (SVD) algorithm on the word-document co-occurrence matrix. Strictly speaking, LSI is a non-probabilistic model. It requires heuristics to compute the probability of an *unseen* test document, whereas we employ the Latent Dirichlet Allocation (LDA) [5] which provides solid probability foundation to compute the probability of an unseen document. Another approach known as the probabilistic LSI (pLSI) has been applied for language modeling [6, 7]. But due to the nature of the model, pLSI can suffer from overfitting problem [5], whereas LDA can be interpreted as a regularized version of pLSI.

The paper is organized as follows: In Section 2, the Latent Dirichlet Allocation model is introduced for document modeling. In Section 3, we describe our approach of employing the LDA model for LM adaptation. LM adaptation experiments are described in Section 4, followed by conclusions and future works in Section 5.

2. Latent Dirichlet Allocation

In the latent semantic analysis, one question is how to extract the latent topics from a text corpus which contains a set of documents. In broadcast news, a document usually refers to a piece of news story within which the latent topics are consistent. Documents are usually assumed to be independent. One popular choice in the information retrieval (IR) community is to apply SVD-based LSI, or pLSI using the EM algorithm. Recently, Latent Dirichlet Allocation [5] has been proposed which has shown to outperform LSI/pLSI in various IR tasks. One view of the LDA model is a Bayesian extension of a mixture of unigram models where the topic mixture weight θ is drawn from a prior Dirichlet distribution:

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

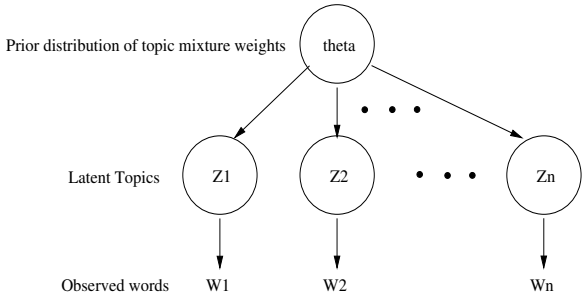

Figure 1: Graphical representation of Latent Dirichlet Allocation.

where $\alpha = \{\alpha_1, ..., \alpha_K\}$ represents the prior observation count of the K latent topics and $\alpha_k > 0$. LDA makes an assumption in which the words in a document $w_1^n = w_1 w_2 ... w_n$ are conditionally independent given the latent topic sequence $z_1^n = z_1 z_2 ... z_n$:

$$f(w_1^n|\theta) = \prod_{i=1}^{n}\sum_{k=1}^{K} f(w_i|z_i = k) \cdot f(z_i = k|\theta) \quad (2)$$

$$= \prod_{i=1}^{n}\sum_{k=1}^{K} \beta_{w_i k} \cdot \theta_k \quad (3)$$

where the model parameters $\Lambda = \{\{\alpha_k\}, \{\beta_{vk}\}\}$ are the Dirichlet prior $\{\alpha_k\}$ and the unigram probabilities of each latent topic $\{\beta_{vk}\}$ in which v denotes the vocabulary index. The model parameters are determined by maximizing the likelihood of a document:

$$f(w_1^n) = \int_\theta f(w_1^n|\theta) \cdot f(\theta; \alpha) d\theta \quad (4)$$

In other words, a document is generated by firstly sampling a mixture weight θ from its prior distribution. Then repeatedly sample a topic k from θ, and a word from the k-th latent unigram until all words in the document are generated. Figure 1 shows the document generation process using the graphical model representation. The above generative procedure is repeated to generate a set of documents of a text corpus. Optimizing the exact likelihood is computationally intractable. One alternative is to optimize the lower-bound of the log likelihood which can be derived using the Jensen's inequality: $\log \sum_i q_i \cdot \frac{f_i}{q_i} \geq \sum_i q_i \cdot \log \frac{f_i}{q_i} = E_q[\log \frac{f(.)}{q(.)}]$ where $\sum_i q_i = 1$. Therefore, the lower bound of the log likelihood has the following form:

$$Q(\Lambda, \Gamma) = E_q[\log \frac{f(\theta, w_1^n, z_1^n; \Lambda)}{q(\theta, z_1^n; \Gamma)}] \quad (5)$$

where $q(\theta, z_1^n)$ is constrained to be a fully factorized distribution over the latent variables parameterized by Γ:

$$q(\theta, z_1^n; \Gamma) = q(\theta) \cdot \prod_{i=1}^{n} q(z_i) \quad (6)$$

where $q(\theta)$ is a Dirichlet distribution and $\{q(z_i)\}$ are multinomial distributions. The method is known as the Variational Bayes approach [8] and q(.) is known as the variational distribution. The central problem in Bayesian inference is to perform the integration over the continuous variable like in Eqn 4 which can be interpreted as an infinite mixture model over all possible assignments of the topic mixture weight θ. The key of the variational approach is that by choosing the tractable variational distribution (e.g. distributions from the exponential family), the integration can be done in a tractable manner. We can interpret q(.) as an approximation to the posterior distribution over the latent variables $f(\theta, z_1^n|w_1^n)$. Optimizing the auxiliary function can be performed using the VB-EM algorithm. The E-step determines the hyper-parameters Γ of q(.) and the M-step uses the variational posteriors q(.) to weight the observations for model parameter estimation. Notice that the hyper-parameters are document-dependent which means that the procedure needs to be applied for each document. We only show the results of the parameter estimations for a single document. Complete derivations can be found in [5].

E-Step:

$$\gamma_k = \alpha_k + \sum_{i=1}^{n} q(z_i = k) \quad (7)$$

$$q(z_i = k) \propto \beta_{w_i k} \cdot e^{E_q[\log \theta_k]} \quad (8)$$

where $E_q[\log \theta_k] = digamma(\gamma_k) - digamma(\sum_{k=1}^{K} \gamma_k)$. Eqn 7 and Eqn 8 are applied iteratively until convergence.

M-Step:

$$\beta_{vk} \propto \sum_{i=1}^{n} q(z_i = k)\delta(w_i, v) \quad (9)$$

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.

3. Language Model Adaptation Approach

Given hypotheses h from a speech decoder, the adaptation procedure can be performed in two steps. In the first step, we apply VB-EM procedure on h to obtain the variational Dirichlet $q(\theta|h)$. In the second step, $q(\theta|h)$ is treated as the new Dirichlet prior, and perform another VB-EM procedure on every vocabulary w to obtain the lower-bound of $\log f(w|h)$. Then we could estimate $f(w|h)$ by normalizing the exponential of the lower-bounds. However, computing the normalization factor can be expensive since it requires to perform VB-EM procedure on every word in the vocabulary. To reduce computation, we compute the Maximum A-Posterior (MAP) likelihood instead as follows:

$$f(w|h) \approx \int_\theta \sum_{k=1}^{K} f(w|z = k) f(z = k|\theta) q(\theta|h) \quad (10)$$

$$\approx \sum_{k=1}^{K} \beta_{wk} \cdot \hat{\theta}_k \quad (11)$$

$$\text{and } \hat{\theta}_k = \frac{\gamma_k}{\sum_{k=1}^{K} \gamma_k} \quad (k = 1...K) \quad (12)$$

where $\hat{\theta}$ is the mode of the variational posterior $q(\theta|h)$ and $q(\hat{\theta}|h) \approx 1$. We implemented both approaches and found no significant performance difference between them despite the significant reduction of computational time for the MAP-based approach. This might be due to the fact that the variational likelihood is only a lower-bound of the exact likelihood. Therefore, we employed the MAP approach throughout all reported experiments. The adapted LDA model attempts to capture the long-term dependencies of words through the latent topics which are reflected by the variational Dirichlet posterior $q(\theta|h)$ providing the knowledge of what the likely latent topics will be given

Latent topics	Top words (translated from Chinese)
"economy"	development, economy, country, society, world, globe
"sport"	competition, candidate, rank, sport, result, champion
"health"	disease, therapy, AIDS, hospital, health, patient, people
"technology"	company, information, network, system, technology
"education"	hong kong, education, mainland, student, expert

Table 1: Examples of latent topics found in the LDA model.

the observed context. Another interpretation of the adaptation procedure is to dynamically boost the mixture weights of the likely topics given the observed context. As a result, *unobserved* words which are related to the likely topics have chances to be boosted while words which are out of the current context are de-emphasized. This is an important distinction from the cache-based approach [2] which boosts only the probabilities of the observed words. Since the LDA model ignores the word ordering information, it is necessary to combine it with the trigram LM. We interpolate the adapted LDA model with the trigram LM. In summary, the LM adaptation procedure begins by decoding the incoming speech utterance, and then stores the word hypotheses in a buffer. After the buffer contains more than M words,

1. perform the E-Step of VB-EM (Eqn 7, 8).
2. update the unigram probability (Eqn 11, 12).
3. update the Dirichlet prior as follows:

$$\alpha_k \quad \leftarrow \quad \lambda \cdot \alpha_k + \sum_{i=1}^{M} c_i \cdot q(z_i = k) \qquad (13)$$

   where $\lambda \in [0, 1]$ is a scaling factor of the history, and c_i denotes the confidence score of the i-th word.
4. clear the buffer.

Intuitively, the Dirichlet prior can be viewed as a dynamic cache table which caches the fractional counts of the *latent topics*. Discounting the prior counts with λ in Eqn 13 is necessary since document topics can change over time and information from past utterances can be irrelevant to the current topics. λ and M can be tuned using the heldout data. In our implementation, LM adaptation happens only after the speech decoder finishes decoding an input utterance. LM adaptation could be applied immediately during the Viterbi search of an input utterance but at the expense of more computation.

4. Experimental Setup

We evaluated the LM adaptation approach on the ISL-RT04 Mandarin Broadcast News evaluation system [9] using the JANUS speech recognition toolkit. The system employs two sets of context-dependent acoustic models: one using a syllable-based initial-final model, the other the phone-based model. The acoustic models were trained using 27 hours of the Mandarin HUB4 1997 training set and 69 hours of the TDT4 Mandarin data. 42-dimension features after Linear Discriminant Analysis were used for the front-end processing. The system employs a multi-pass decoding strategy in which cross adaptation among the syllable-based and the phone-based decoders were performed. The vocabulary size is 63K word.

We used subsets of the Mandarin Gigaword corpora to train the LDA model. We first organized the corpora into documents where each document is simply a piece of news story provided by the corpora. Then we applied the VB-EM procedure described in section 2. Motivated by [5], we set the number of

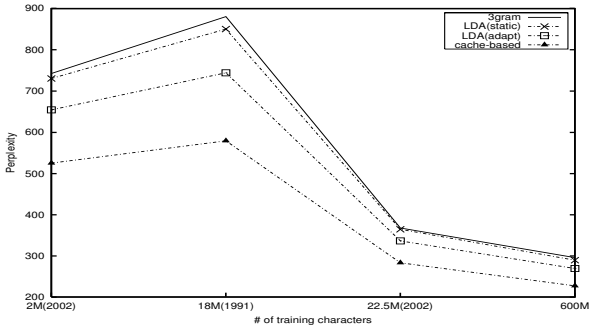

Figure 2: Perplexities over corpus type for different LM training scenarios.

	AM adapt	2M (2002)	18M (1991)	22.5M (2002)	600M
LDA(static)	no	1.3%	0.0%	3.0%	0.0%
LDA(adapt)	no	**2.5%**	**3.0%**	**3.5%**	**4.9%**
cache-based	no	0.8%	1.7%	2.0%	4.3%
LDA(static)	yes	-0.5%	2.1%	2.6%	-0.7%
LDA(adapt)	yes	**4.3%**	**4.2%**	**3.3%**	**3.5%**
cache-based	yes	0.0%	-2.1%	-2.0%	-1.4%

Table 2: Relative CER reduction compared with the trigram baseline over corpus type before/after acoustic adaptation.

latent topics in the LDA model to K=50. Table 1 shows examples of latent topics found in the LDA model. Unless specified in the following experiments, the word trigram LM and the LDA model were trained using the common training subsets. The trigram LM was estimated using the modified Kneser-Ney smoothing scheme. We used the official RT04 development set to tune the LM interpolation weight, the history scaling factor λ and the size of the buffer M for LM adaptation. The trigram LM interpolation weights are between 0.7-0.9. $\lambda = 0.4$ and M=20 words appear to be optimal in our settings which are kept fixed in all the reported experiments. Performance metrics are the word trigram perplexity and the character error rate (CER) evaluated on the CCTV episode of the RT04 test set. We report the best CER by rescoring with different values of the word insertion penalty and the language model weight so that we could see clearly the net effect from the LM adaptation. We are interested to see the impact of unsupervised acoustic adaptations on the LM adaptation performance because it is not uncommon that LM gains may disappear after the state-of-the-art acoustic adaptations (Vocal Tract Length Normalization (VTLN), Feature Space Adaptation (FSA), and Maximum Likelihood Linear Regression (MLLR) [9]) are applied. Therefore, we report results before and after acoustic adaptation. We compare the LDA-based and the cache-based LM adaptation under different training scenarios. The cache-based LM is a unigram model which dynamically adapts to the past decoded hypotheses using the decaying word counts, and is then interpolated with the trigram LM. We also evaluate the static interpolation of the trigram LM with the static LDA model (no adaptation) to compare with the adaptive one. Results are reported on the test set only.

4.1. LM training with large corpora

We utilized a large multiple corpora comprising 600M characters to train the trigram LM. We trained the LDA model using only one-year text (2002) containing 22.5M characters (48k documents) for fast turnaround time for evaluation experiments. The LDA training on the 22.5M corpora took 4 hours on a Pentium 4 machine with the whole corpora loaded into memory.

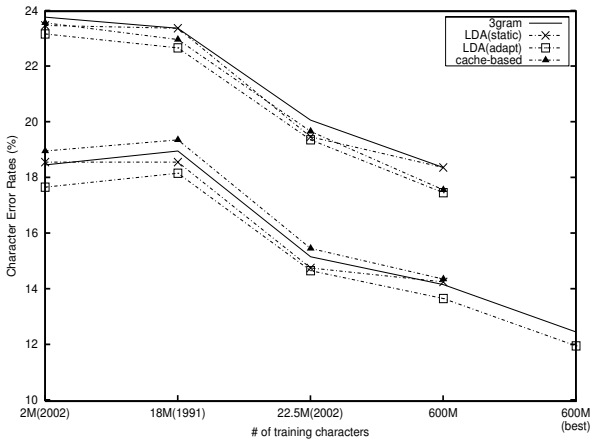

Figure 3: Character Error Rates over corpus type for different LM training scenarios before (upper line group) and after (lower line group) acoustic adaptation.

Figure 2 and Table 2 show the perplexity and the character error rate respectively. LDA-based LM adaptation helps reduce both the perplexity and CER after optimal tuning on the word insertion penalty, LM weight and standard acoustic adaptation. We observed similar results on the perplexity of the development set which implies that the development set is suitable for parameter tuning. When cross-adaptation passes [9] are applied among our syllable-based and the phone-based decoders, our final best CER with LDA-based LM adaptation is 12% which corresponds to 4% relative reduction compared with the counterpart of the trigram baseline shown in Figure 3. Table 2 also shows that static interpolation of the trigram LM and the LDA model does not help, implying the importance of dynamic LM adaptation using the current word context. It is intuitively satisfying to dynamically adjust the mixture weights of the latent topics because topics of the current word context can change over time. On the other hand, although the cache-based approach outperforms the LDA approach on the perplexity, improvement cannot be translated successfully to a lower CER after acoustic adaptation shown in Table 2. [10] also reported that cache-based LM adaptation could lead to poor performance despite substantial reduction on the perplexity.

4.2. LM training with small corpora

Although it is not difficult to obtain vast amount of training text for broadcast news, it is usually not the case for conversational tasks such as meetings or lectures. We further investigate the LM adaptation approaches where only (relatively) limited text data are available. We chose the 1991, 2002, and a 1-month 2002 subset to train the trigram LM and the LDA model. The 1991 training set contains 42k documents with 18M Chinese characters. The 1-month 2002 subset contains 3.9k documents with 2M characters. Using the 2M training set is to simulate the scenario of having very limited amount of data, whereas using the 1991 training set is to simulate a stronger mismatch compared to the 2002 training set. Since the test set is extracted from 2004, presumably the mismatch would be stronger for the 1991 training set. The degree of mismatch can be reflected from the perplexities shown in Figure 2.

Figure 2 and 3 show the perplexities and the CER over varying sizes of the training sets (from 2M to 600M characters). Table 2 shows the relative CER reduction compared with the corresponding trigram baseline. Similarly, LDA-based LM adaptation helps across different sizes of the training sets, and preserves performance gain after acoustic adaptation. Choosing a matching training set (2002) closer to the test condition is crucial. The 1-month training subset (2M characters) from 2002 and the 1-year training set (18M characters) from 1991 lead to comparable results. We repeated Experiment 4.1 but using the LDA model trained with the 1991 training set. We observed 2.2% relative CER reduction before acoustic adaptation, but observed no gain after acoustic adaptation which further supports the importance of choosing a matching training set.

5. Conclusions and Future Works

We proposed an unsupervised dynamic language model adaptation framework using latent topics under the Latent Dirichlet Allocation method. Model parameters can be estimated using variational Bayes EM algorithm. Only the mixture weights of the latent topics are adapted dynamically which can be performed using only a small amount of data. Empirically, the adaptation procedure is fast. We evaluated our approach on different training scenarios. Results show that the proposed approach reduces both the perplexity and the character error rates on the CCTV episode of the RT04 Mandarin Broadcast News. We achieve relative reduction on perplexity and the character error rate up to 15.4% and 4.9% respectively depending on the setup of the training sets. Future directions include vocabulary selection and LM adaptation across different task domains, integration of human-motivated information, and relaxation of the conditional independence assumption in the LDA model.

6. References

[1] R. Kuhn and R. D. Mori, "A cache-based natural language model for speech reproduction," in *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 1990.

[2] P. R. Clarkson and A. J. Robinson, "Language Model Adaptation using Mixtures and an Exponentially Decaying Cache," in *Proc. of ICASSP*, vol. 2, 1997.

[3] R. Iyer and M. Ostendorf, "Modeling long distance dependence in language: Topic mixtures vs. dynamic cache models," in *Proc. of ICSLP*, vol. 1, 1996.

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

[5] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation," in *Journal of Machine Learning Research*, 2003.

[6] D. Gildea and T. Hofmann, "Topic-based language models using EM," in *Proc. of Eurospeech*, 1999.

[7] D. Mrva and P. C. Woodland, "A PLSA-based language model for conversational telephone speech," in *Proc. of ICSLP*, 2004.

[8] 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.

[9] 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.

[10] P. Clarkson and A. Robinson, "The applicability of adaptive language modelling for the broadcast news task," in *Proc. of ICSLP*, 1998.